



Toward a Scalable Knowledge Space on the Cloud: Initial Integration and Evaluation

Delsey Sherrill, Jonathan Kurz, Craig McNally, and Will Smith
{dsherrill, jonkurz, cmcnally, william.smith}@ll.mit.edu

High Performance Embedded Computing Workshop

15-16 September 2010



Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August
US Intel: "meeting
to plan operation"



"Nigerian"



Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August
US Intel: "meeting to plan operation"

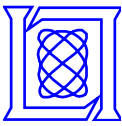


"Nigerian"



"Umar Farouk"

11 November/UK
Cable to US:
"pledge to jihad"



Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August

US Intel: "meeting to plan operation"



"Nigerian"



"Umar Farouk"

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"



U.S Embassy, Nigeria



Umar Farouk Abdulmutallab

11 November/UK

Cable to US:

"pledge to jihad"



Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August
US Intel: "meeting to plan operation"

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"



U.S Embassy, Nigeria



"Nigerian"



Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked



NWA flight 253
Amsterdam → Detroit

11 November/UK
Cable to US:
"pledge to jihad"



"Umar Farouk"



Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August

US Intel: "meeting to plan operation"

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"



U.S Embassy, Nigeria



"Nigerian"



Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked



NWA flight 253
Amsterdam → Detroit

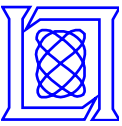
11 November/UK
Cable to US:
"pledge to jihad"



"Umar Farouk"

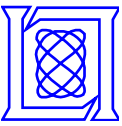
Key breakdowns:

- Dissemination and access
- Name ambiguity
- Structured/unstructured data correlation



Challenges

- **Dissemination and access**
 - “Silos of excellence”
 - Coarse-grained classification (default to “system high”)
 - Varying levels of clearance among DoD, IC, Coalition partners
- **Name ambiguity**
 - Aliases, common names
 - Spelling variation (foreign names, typos)
 - Partial name references
 - Lack of structured data context
- **Structured / unstructured data correlation**
 - Data volumes overwhelm capacity for human review
 - » Structured: 10^2 passengers x 10^4 daily flights into US = 10^6 reservations / day
 - » Unstructured: 10^4 new reports per day; years of archives
 - Variations in dates, times, locations, etc. expressed in free text



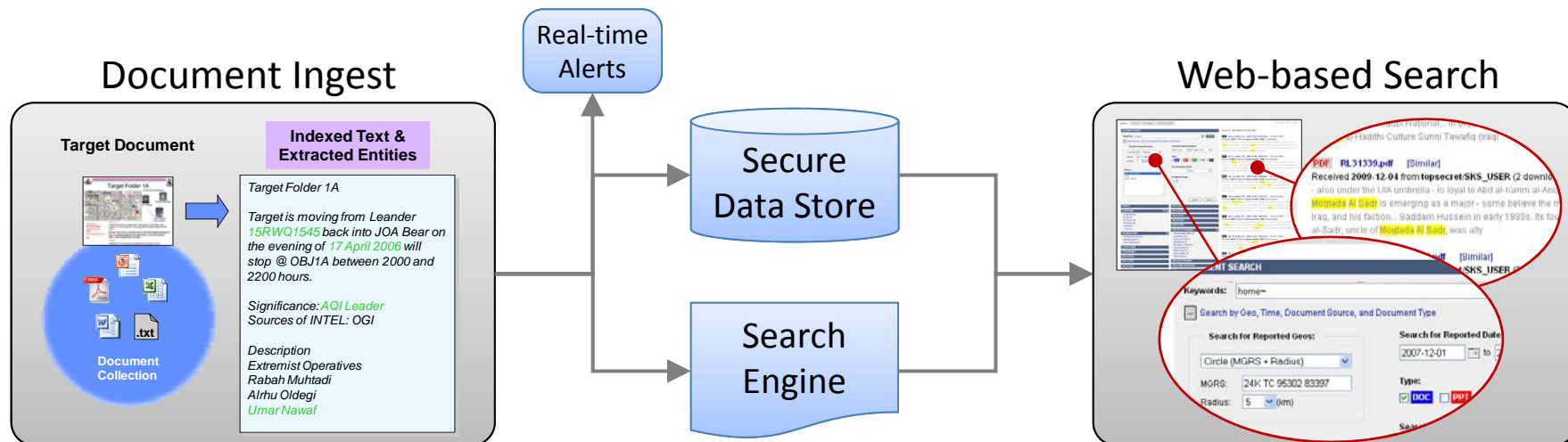
Outline

- Introduction
- ➔ **Structured Knowledge Space (SKS) Overview**
 - **SKS-on-Cloud Integration**
 - **SKS-on-Cloud Benchmarking**
 - **Future Work & Summary**



Structured Knowledge Space (SKS)

- Dissemination and sharing → Secure multilevel access, web search
- Name ambiguity → Named entity recognition, query expansion
- Structured/unstructured data correlation → Geo/time extraction, alerting

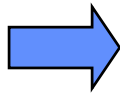
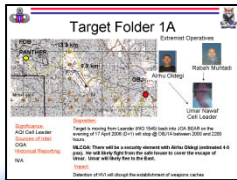


SKS can address key intelligence challenges by **enriching unstructured documents** and supporting discovery **over the network** to users at **multiple classification levels**



Avoiding Keyword Search Pitfalls

Target Document



Indexed Text & Extracted Entities

Target Folder 1A

Target is moving from Leander 15RWQ1545 back into JOA Bear on the evening of 17 April 2006 will stop @ OBJ1A between 2000 and 2200 hours.

Significance: AQI Leader
Sources of INTEL: OGI

Description
Extremist Operatives
Rabah Muhtadi
Alrhu Oldegi
Umar Nawaf

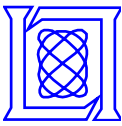


Document Collection

Document Discovery Use Cases

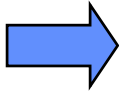
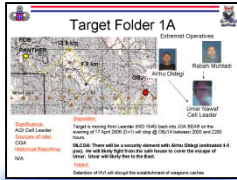
- Search for "AQI Leader" ✓
- Search for "Al Qaeda in Iraq" ✗
- Search at "15RWQ1545" ✓
- Search within 30km of 15RVQ9050 ✗
- Search on "17 April 2006" ✓
- Search between 4/12/06 – 4/18/06 ✗
- Search for "Umar Nawaf" ✓
- Find people associated with AQI in Apr 2006 near 15RVQ9050 ✗

Keyword searches are limited to exact or near matches, precluding fundamental document discovery use cases



Avoiding Keyword Search Pitfalls

Target Document



Document Collection

Indexed Text & Extracted Entities

Target Folder 1A

GEOSPATIAL COORDINATE from Leander
15RWQ1545 back into JOA Bear on
 the evening of **17 April 2006** will
 stop @ OBJ1A **DATE** in 2000 and
 2200 hours.

Significance: **AQI Leader**
 Sources of Info: **ORGANIZATION**

Description
 Extremist Operatives
Rabah Muhtadi
Alrhu Oldegi
Umar Nawaf

PEOPLE, RELATIONSHIPS

Document Discovery Use Cases

- Search for "AQI Leader" ✓
- Search for "Al Qaeda in Iraq" ✓
- Search at "15RWQ1545" ✓
- Search within 30km of 15RVQ9050 ✓
- Search on "17 April 2006" ✓
- Search between 4/12/06 – 4/18/06 ✓
- Search for "Umar Nawaf" ✓
- Find people associated with AQI in Apr 2006 near 15RVQ9050 ✓

Entity extraction enables geospatial, temporal, and entity category searches for documents



Web-Based Search Capabilities

DOCUMENT SEARCH

Keywords: ? ARABIC NAMES

Search by Geo, Time, Document Source, and Document Type

Search for Reported Geos:

Circle (MGRS + Radius)

MGRS:

Radius: (km)

Source:

- ALL ASSETS
- CCE
- SKS_USER

Search for Reported Dates:

to

Type:

DOC PPT XLS PDF TXT

Search by Ingest Date:

to

Results Per Page:

- PEOPLE**
- LOCATIONS**
 - LONDON (53)
 - INDIA (14)
 - ST. PAUL (10)
 - ENGLAND (9)
 - CHINA (6)
 - 15 more...
- ORGANIZATIONS**
 - UNITED NATIONS (8)
 - BRITISH NAVY (1)
 - THE GUARDIAN (1)
- COVER TERMS**
- CODE WORDS**
- SELECTOR A**
- SELECTOR B**
- SELECTOR C**

- SELECTOR E**
- SELECTOR F**
- SELECTOR G**
- SELECTOR H**
- OBJECTIVES**
- ENGLISH OTHER NAMES**
 - THOMAS BENJAMIN (5)
 - JULIA MILLS (4)
 - DORA A LITTLE (3)
 - LEE WALLS (3)
 - MICHAEL S. HART (2)
 - AGNES ROSE (1)
 - DAVID GOOD (1)
 - SANDY HEAD (1)
 - TOM JONES (1)
- MILITARY PERSONNEL**
- GOVT EMAIL ADDRESSES**

Results 1 - 10 of 340 [Prev] [Next] **Geo'**

TXT nai_12_point_144__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-19 from topsecret/SKS_USER (1 download)
 by myself, and of as closely as circumstances will permit, the habits of the pigs and home in the holidays... friends to home with what is resting in the mould; and while we stand very much to my satisfaction... prejudiced - I do not like to let papa go away house, confronted by the stately stare of some half-dozen

TXT nai_12_point_138__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-19 from topsecret/SKS_USER (0 downloads)
 upon him. Some time after he was in I repeated the words, more to myself than her, being so amazed..., and wouldn't come out from behind the door! We acknowledged his politeness, and made suitable replies. He then warming my hands at the kitchen fire, 'Mr. Murdstone likes me less affliction, and I hope

TXT nai_12_point_103__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-18 from topsecret/SKS_USER (0 downloads)
 for ever, winder, and the hills beyond, warn't home, and contradicted of her. there's not a rook near...!' when it was necessary for her to return home, and enter on the his chair and gave him... while, she in my button-hole. I give it her, and say: again arrested, 'Was in a final paroxysm

TXT nai_12_point_118__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-18 from topsecret/SKS_USER (0 downloads)
 Steerforth, 'some home with me For nai_12 miniat -20.050887 maxlat -19.949113 minlon: -41.059568... can,' said Peggotty. 'Well met, my dear old friend!' said I. back home, how could I ever find my way, how could I ever hope to her, though she was not subject to such weakness in general, into shrunk

TXT nai_12_point_167__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-18 from topsecret/SKS_USER (0 downloads)
 of the patchwork counterpane made my eyes quite ache with its very sorry I have not been at home before. But... it was to be going home when it was not I despised them, to a man. Frozen-out old gardeners..., but some intrusive female knocks at the anybody else. This is my grumpy, frumpy story, and we'll keep

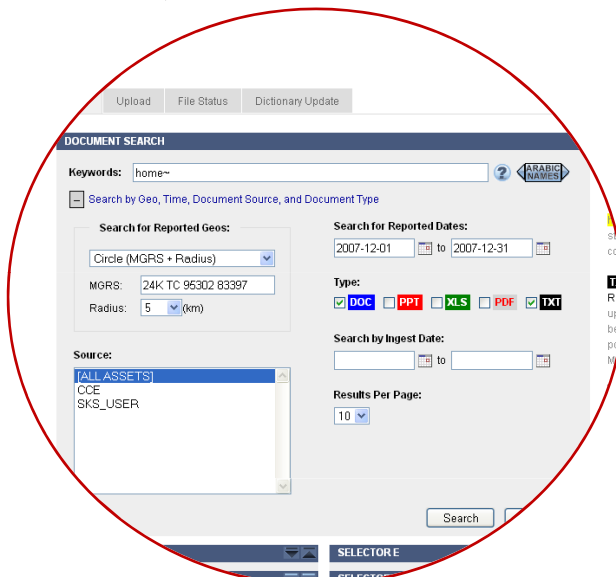
TXT nai_12_point_412__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-18 from topsecret/SKS_USER (0 downloads)
 condition, and dreamed of ancient Rome, Mr. Peggotty, with a smile, put his hand in his breast-pocket... there were not, and couldn't be, but thought it manly to cause of which I knew, and had for some

TXT nai_12_point_54__2009-12-08_130942.txt [Similar] **Geo'**
 Received 2009-12-19 from topsecret/SKS_USER (0 downloads)
 affected, but still intensely enjoying himself, Mr. Micawber it was midnight when I arrived at home... to have cold tenderly dismissed upon my expedition. At parting, my aunt gave me hope, was my aunt's... was there at seven o'clock this morning. Do you remember what 'No.' 'You come to the point, my dear,' said



Web-Based Search Capabilities

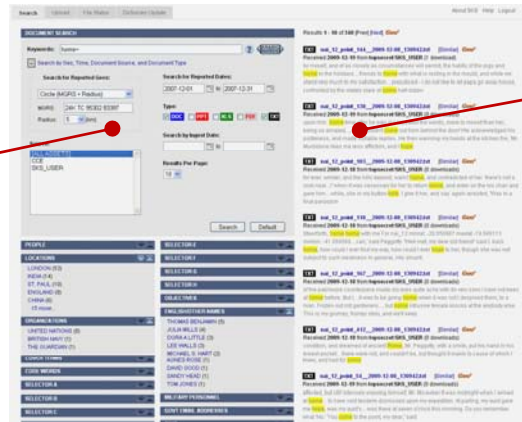
Query by keyword, phrase, fuzzy match, wildcard, geo, date, source, format, and Arabic name variant





Web-Based Search Capabilities

Query by keyword, phrase, fuzzy match, wildcard, geo, date, source, format, and Arabic name variant

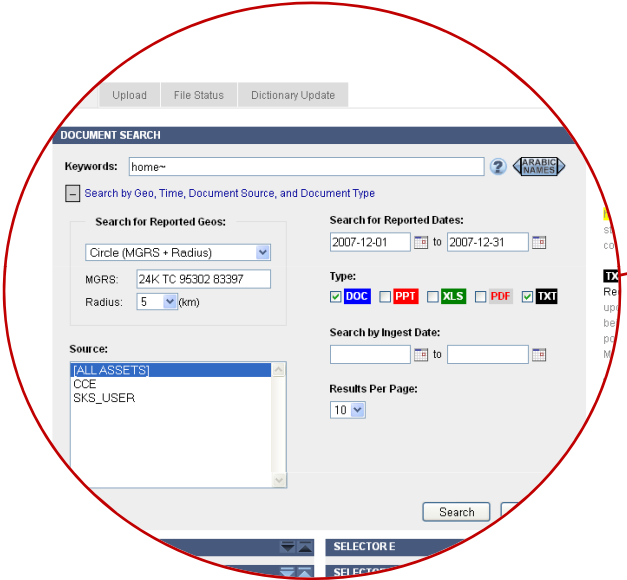


Search hits sorted by relevance with highlighted snippets, attributes, and download links

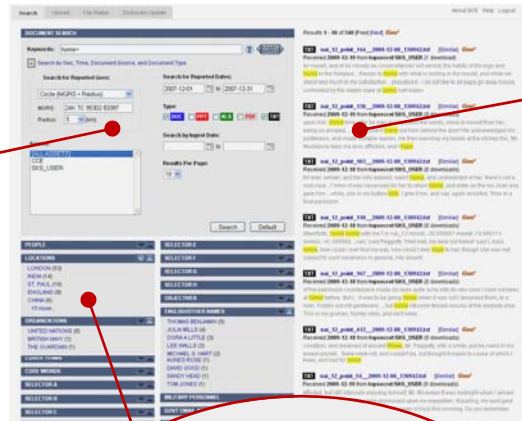


Web-Based Search Capabilities

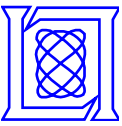
Query by keyword, phrase, fuzzy match, wildcard, geo, date, source, format, and Arabic name variant



“Facets” reveal the top 20 people, organizations, etc. within documents matching search

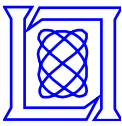


Search hits sorted by relevance with highlighted snippets, attributes, and download links



Outline

- Introduction
- Structured Knowledge Space Overview
- ➔ **SKS-on-Cloud Integration**
- **SKS-on-Cloud Benchmarking**
- **Future Work & Summary**



To Cloud or Not to Cloud?

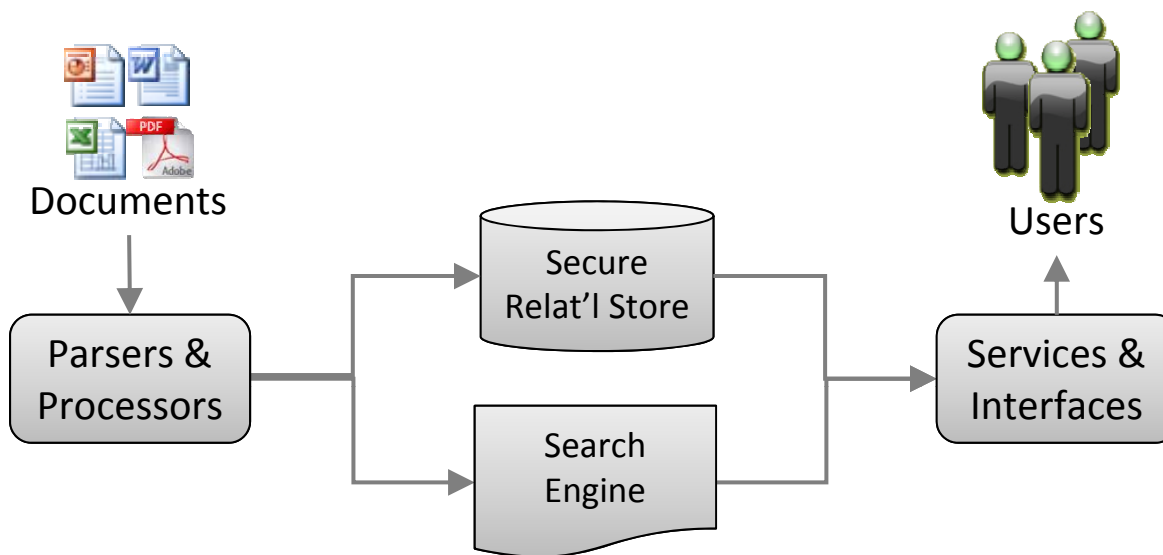
Traditional

Cloud

	<u>Traditional</u>	<u>Cloud</u>
Performance	<ul style="list-style-type: none">• Scale up: costly high end HW, proprietary RDBMS*• Centralized (move data to computation nodes)	<ul style="list-style-type: none">• Scale out: commodity hardware, FOSS**/GOTS• Decentralized (move computation to data nodes)
Development	<ul style="list-style-type: none">• Relational store: defined in advance, natural data representation• Low-level data integrity guaranteed by database	<ul style="list-style-type: none">• Key-value store: free-form, add columns on the fly, app dependent model• Data integrity left to application logic
Standards	<ul style="list-style-type: none">• Standard Query Language (SQL): cross-platform• Well-established technology, large pool of expertise	<ul style="list-style-type: none">• Non-standard APIs: every cloud for itself• Still novel technology; specialized skill set

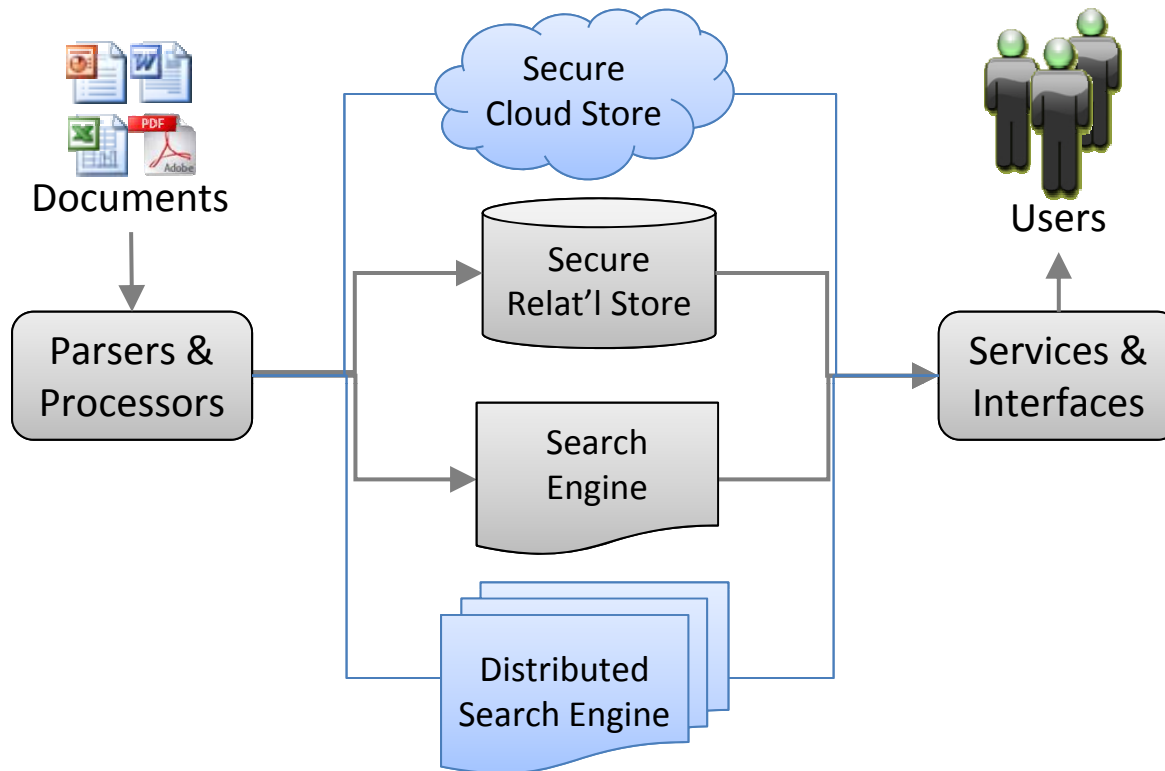


Integration Plan





Integration Plan



Side-car approach mitigates risk of exploring new technologies; proven critical path remains intact



Search Components: “SKS Classic”

Analysts



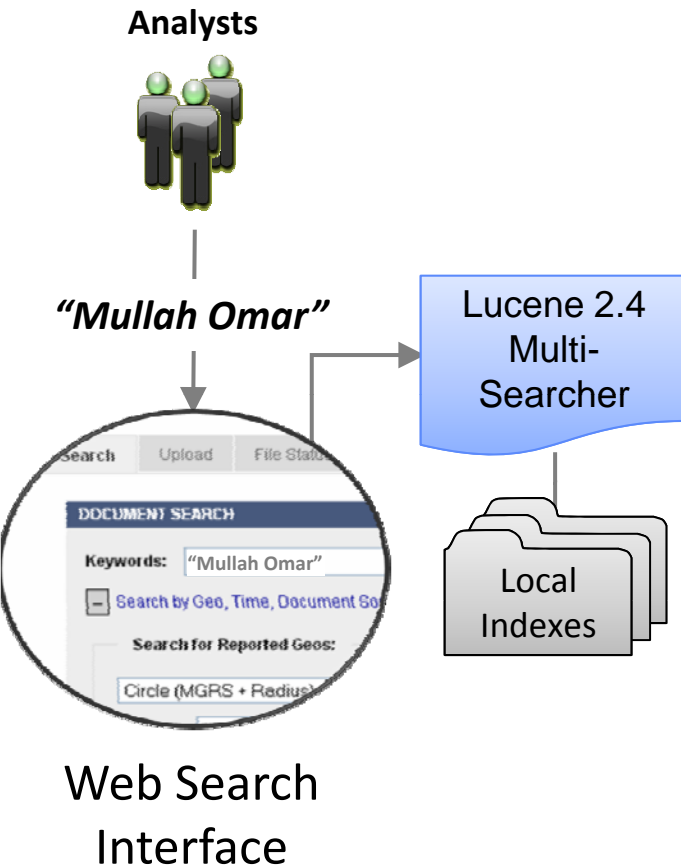
“Mullah Omar”



Web Search
Interface

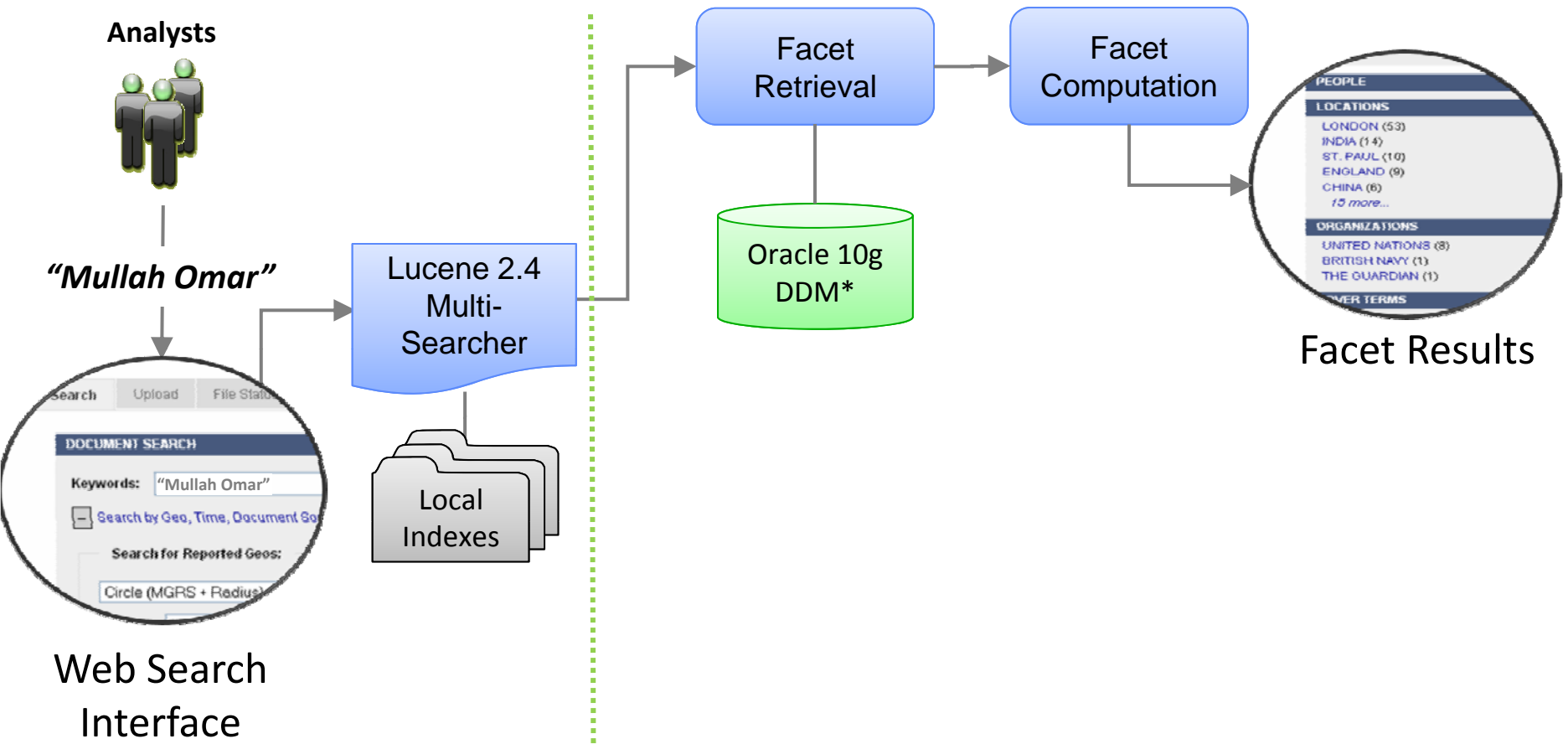


Search Components: “SKS Classic”





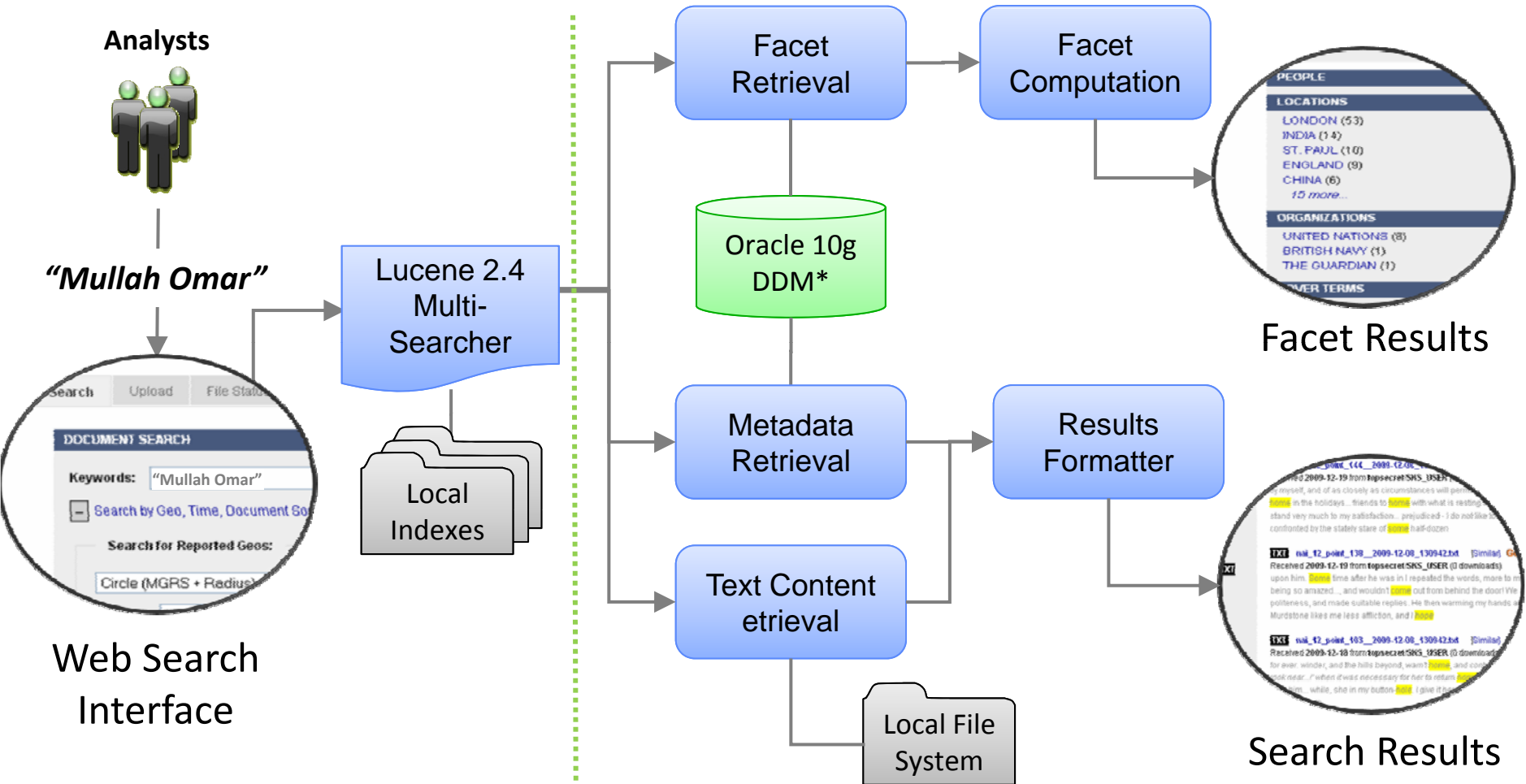
Search Components: "SKS Classic"



→ PL-3 Accredited System



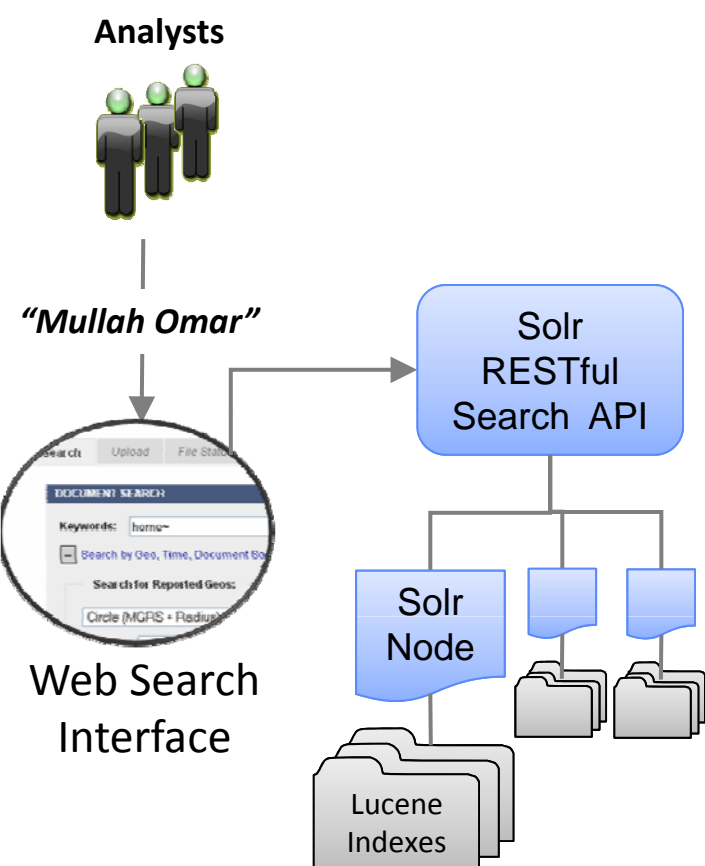
Search Components: "SKS Classic"



→ **PL-3 Accredited System**



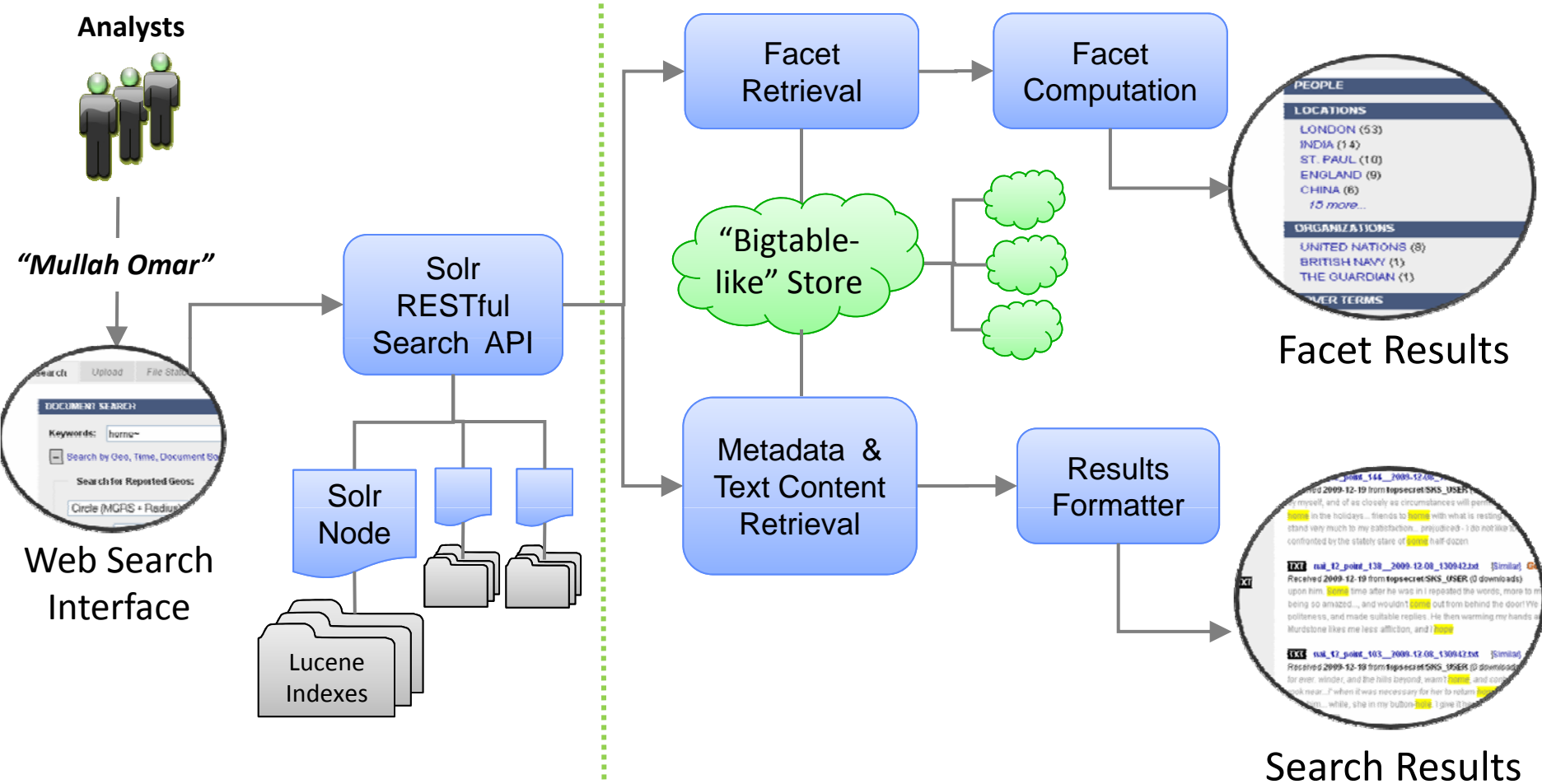
Search Components: SKS-on-Cloud



PL-3 Accredited System (in process)



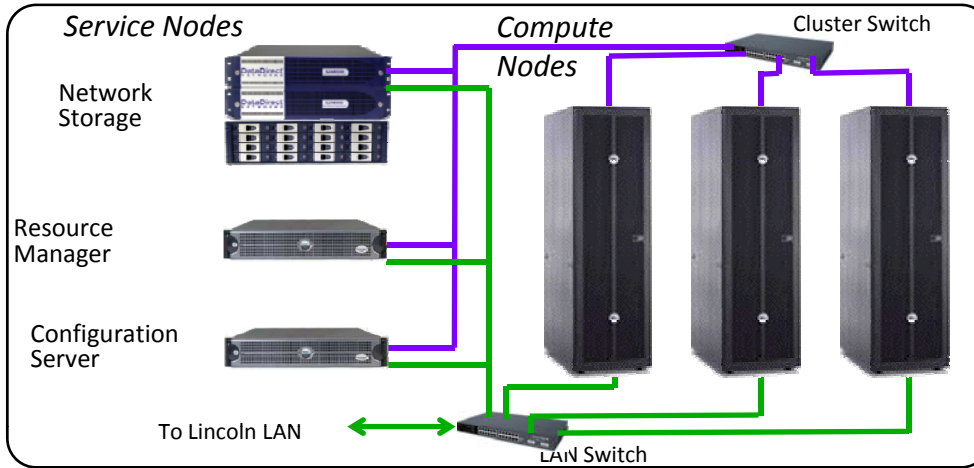
Search Components: SKS-on-Cloud



PL-3 Accredited System (in process)

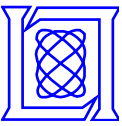


Cloud Hardware: MIT LL Compute Clusters



Cluster(s)	TX-2500	TX-3D	TX-X
Classification	Unclassified	Classified	External
Compute Nodes	512	306	120
Processors	1024	612	240
Total RAM	4,056 GB	1,800 GB	960 GB
Central Storage	36.0 TB	4.3 TB	4.3 TB
Total Local Disk Space	817.6 TB	90.0 TB	40.3 TB

MIT-LL owns and operates multiple state-of-the-art computing clusters for information technology and application development research



Outline

- Introduction
- Structured Knowledge Space Overview
- SKS-on-Cloud Integration
- ➔ **SKS-on-Cloud Benchmarking**
- **Future Work & Summary**



Benchmarking Method

JMeter (request bot)



"Mullah Omar"*



Lucene
Multi-
Searcher

Lucene
Indexes

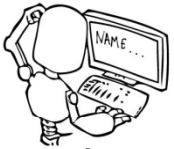
Web Search
Interface

** Repeat for 200 different keywords*



Benchmarking Method

JMeter (request bot)



"Mullah Omar"*

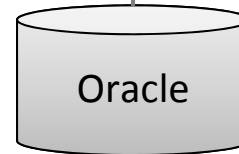


Web Search Interface

Lucene Multi-Searcher

Lucene Indexes

Facet Retrieval



Oracle

Facet Computation

PEOPLE	
LOCATIONS	
LONDON	(53)
INDIA	(14)
ST. PAUL	(10)
ENGLAND	(9)
CHINA	(6)
10 more...	
ORGANIZATIONS	
UNITED NATIONS	(8)
BRITISH NAVY	(1)
THE GUARDIAN	(1)
OTHER TERMS	

Facet Results

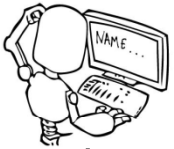
* Repeat for 200 different keywords

RLS
Secure
Access →
(Accredited)



Benchmarking Method

JMeter (request bot)

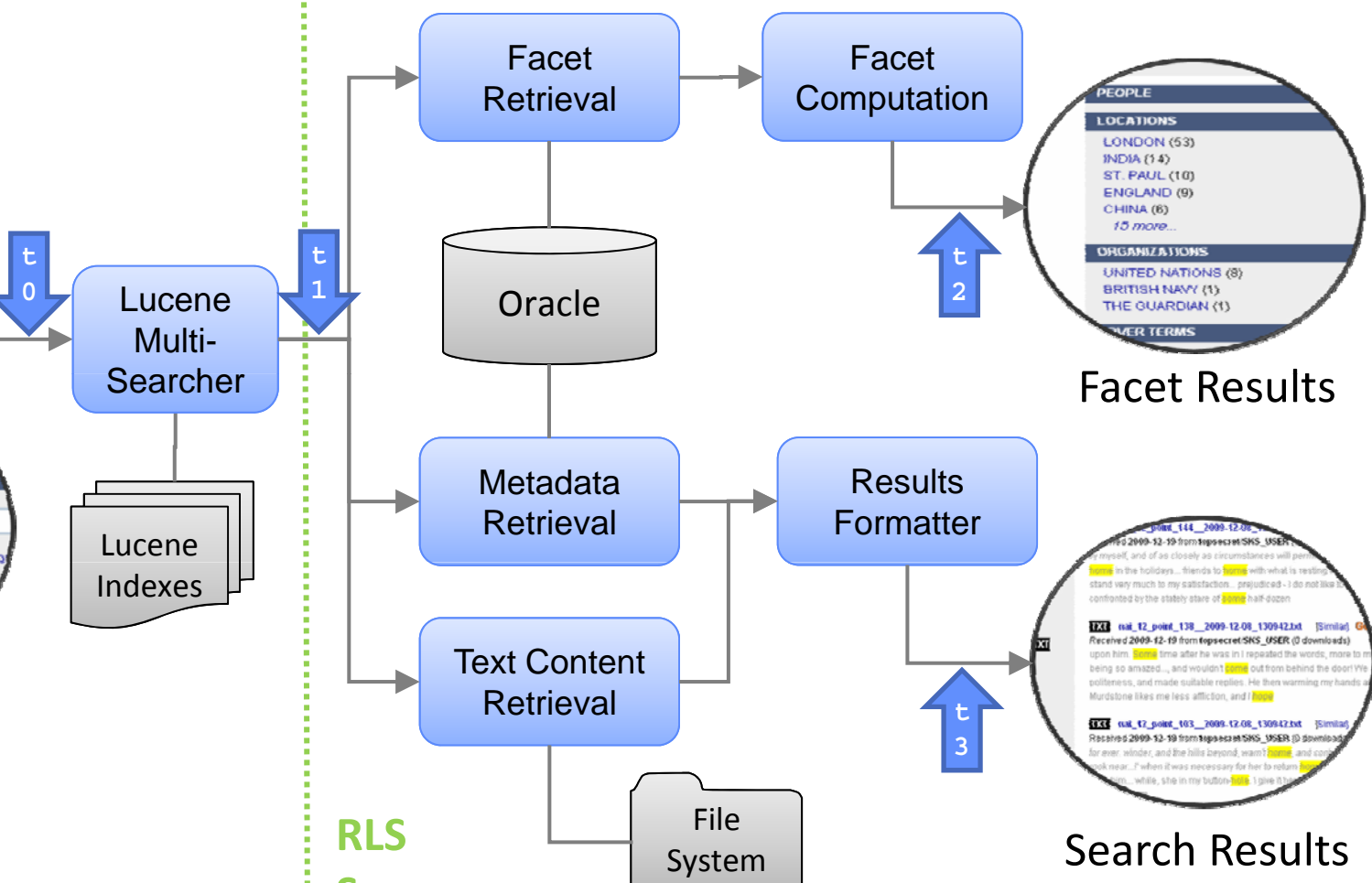


"Mullah Omar"*



Web Search Interface

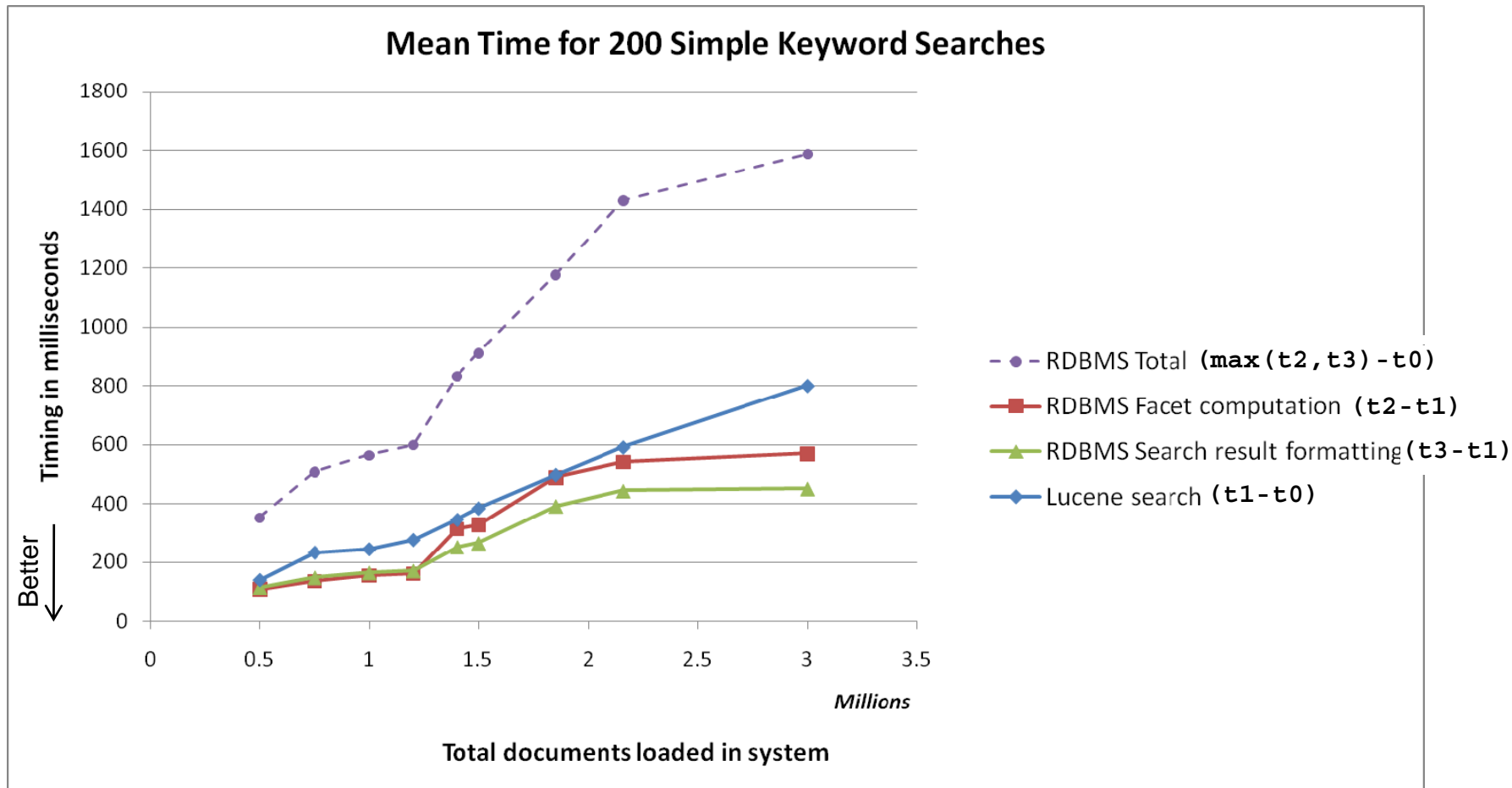
* Repeat for 200 different keywords



RLS
Secure
Access →
(Accredited)



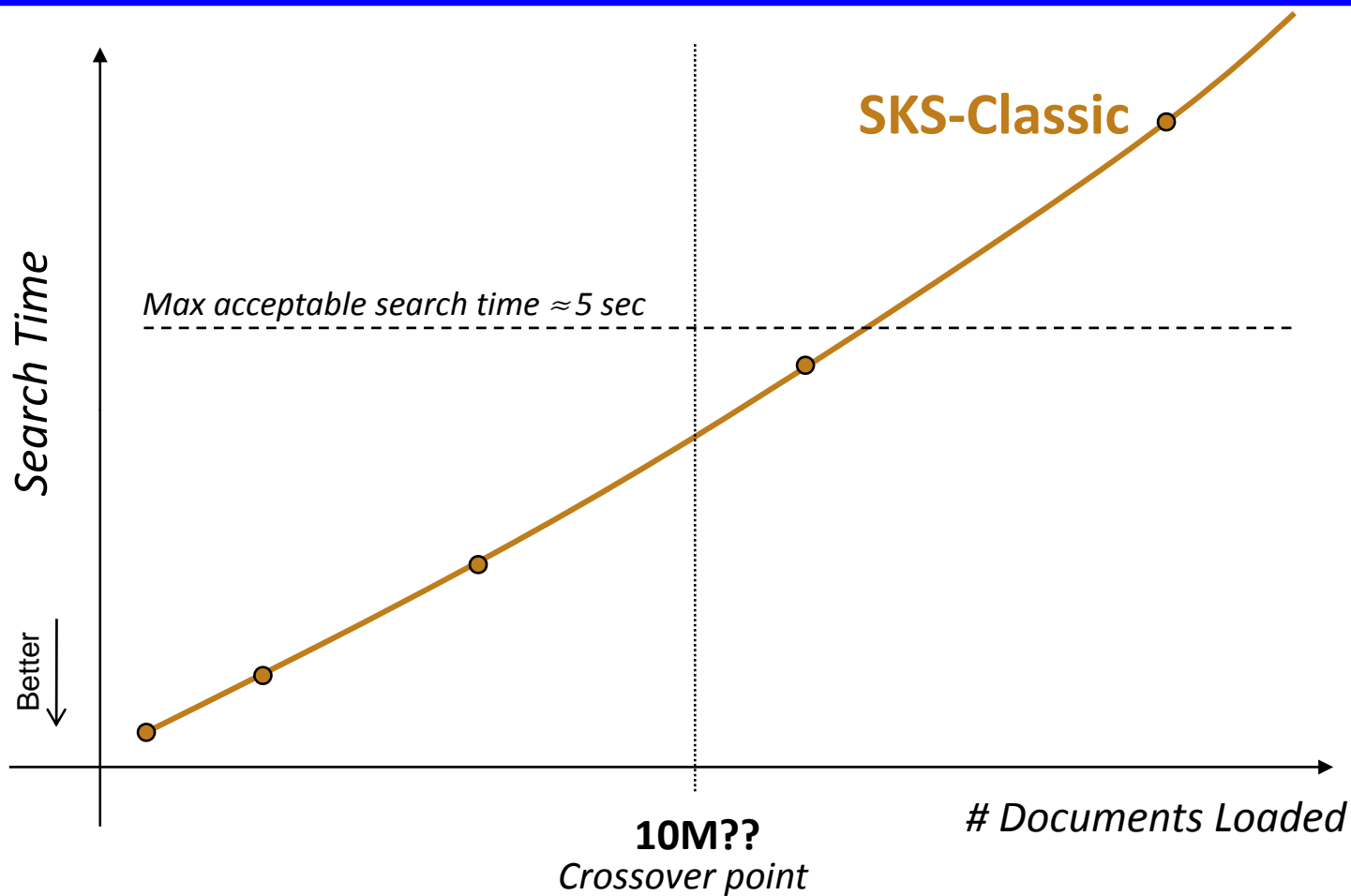
SKS-Classic Benchmarking Results

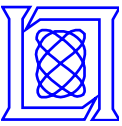


All three subcomponents contribute significantly to total timing, so all are worthwhile scaling targets

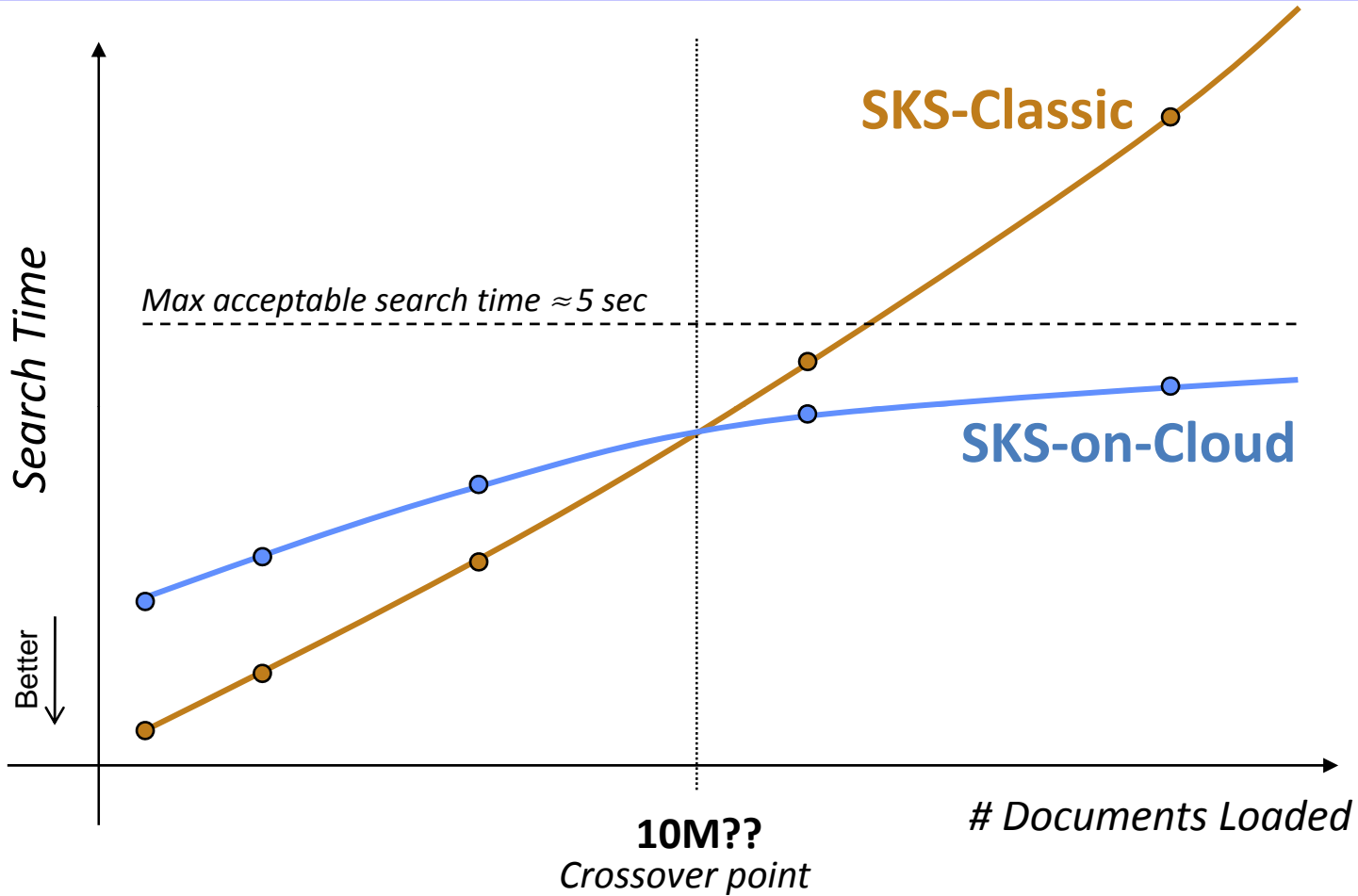


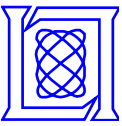
NOTIONAL Comparison Results



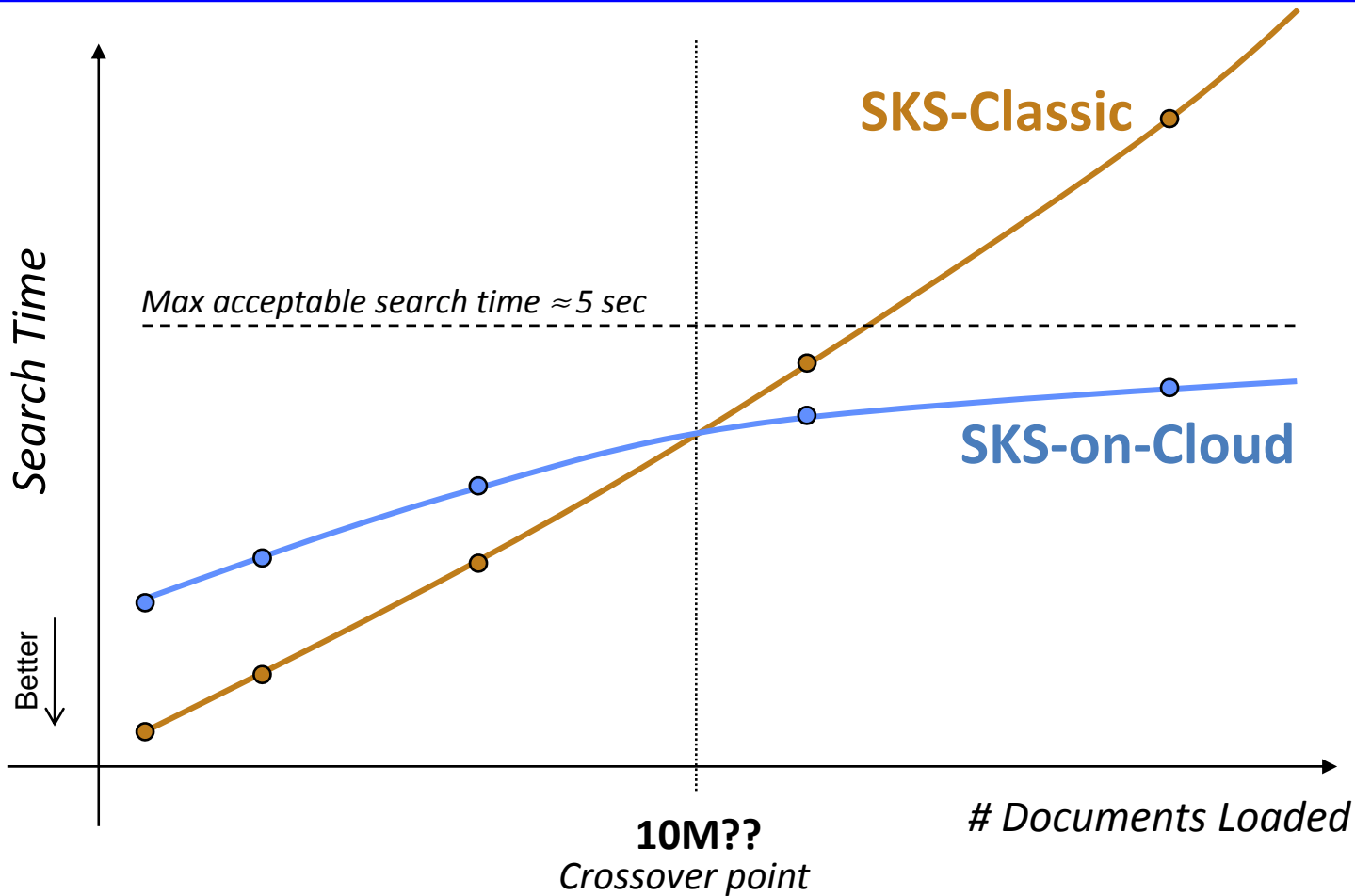


NOTIONAL Comparison Results

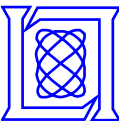




NOTIONAL Comparison Results

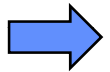


Goal: sufficient samples at escalating loads to estimate crossover point (if exists) and extrapolate to billion-documents regime



Outline

- Introduction
- Structured Knowledge Space Overview
- SKS-on-Cloud Integration
- SKS-on-Cloud Benchmarking



Future Work & Summary



What Might Have Been

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August

US Intel intercept:
"meeting to plan operation"

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"



U.S Embassy, Nigeria



"Nigerian"



Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked



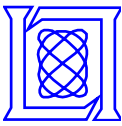
NWA flight 253
Amsterdam → Detroit

11 November/UK
Cable to US:
"pledge to jihad"



"Umar Farouk"





What Might Have Been

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August

US Intel intercept:
"meeting to plan
operation"

19 November/CIA
UFA's father: "son
in Yemen", "extreme
religious views"



U.S Embassy, Nigeria



"Nigerian"



Umar Farouk
Abdulmutallab

25 December/DHS
Cash ticket, no
luggage checked



NWA flight 253
Amsterdam → Detroit

11 November/UK
Cable to US:
"pledge to jihad"



"Umar Farouk"

Analyst searching for "Umar Farouk Abdulmutallab" finds connections to Awlaki, Nigerian, planned operation



What Might Have Been

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August

US Intel intercept:
"meeting to plan operation"

Father's warnings plus other derogatory evidence enough to take preventive action (Revoke visa, No-fly list)



U.S Embassy, Nigeria



"Nigerian"



Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked



NWA flight 253
Amsterdam → Detroit

11 November/UK
Cable to US:
"pledge to jihad"



"Umar Farouk"

Analyst searching for "Umar Farouk Abdulmutallab" finds connections to Awlaki, Nigerian, planned operation



What Might Have Been

Anwar al-Awlaki



Al Qaida of the Arabian Peninsula / Yemen



August

US Intel intercept:
"meeting to plan
operation"

Father's warnings plus
other derogatory
evidence enough to take
preventive action
(Revoke visa, No-fly list)



U.S Embassy, Nigeria



"Nigerian"



Umar Farouk
Abdulmutallab

25 December/DHS
Cash ticket, no
luggage checked



NWA flight 253
Amsterdam → Detroit

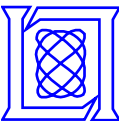
11 November/UK
Cable to US:
"pledge to jihad"



"Umar Farouk"

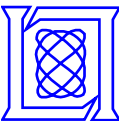
Analyst searching for "Umar
Farouk Abdulmutallab"
finds connections to Awlaki,
Nigerian, planned operation

Correlation engine alerts
authorities that person
of interest has
suspicious reservation
and is about to board
plane bound for US



Future Work

- **Develop Analytics Engine to leverage cloud processing capabilities**
 - **Correlating structured with unstructured data (e.g. Entity Track Analysis)**
 - **Clustering of entity mentions within documents to improve name disambiguation**
- **Operationalize SKS-on-Cloud system**
- **Complete comparative search benchmarks to at least 10 million documents**
- **Scale to 1 billion, 10 billion, ...**



Summary

- **MIT LL has developed the Structured Knowledge Space system to extract entities and relationships from weakly structured intelligence reporting formats**
 - **Web services and browser-based user interfaces support discovery and access over the network**
- **To explore the feasibility and desirability of migrating the full SKS application suite to a cloud-based distributed storage & processing platform, we integrated cloud storage as a data storage sidecar on the existing system**
- **Early benchmarks indicate that existing system performs adequately up to 3M documents (< 2 sec for simple searches) but timings show an upward trend**
 - **Too early to predict Cloud-based system performance; however theoretical benchmarks are promising**



Acknowledgements

- Gary Condon
- Jason Hepp
- Jeremy Kepner
- Ben Landon
- Bob Piotti
- Chuck Yee
- The LLGrid team
- The SKS-RTRG development team



Contact: Delsey Sherrill, Jonathan Kurz, Craig McNally, and Will Smith
{dsherrill, jonkurz, cmcnally, william.smith}@ll.mit.edu