

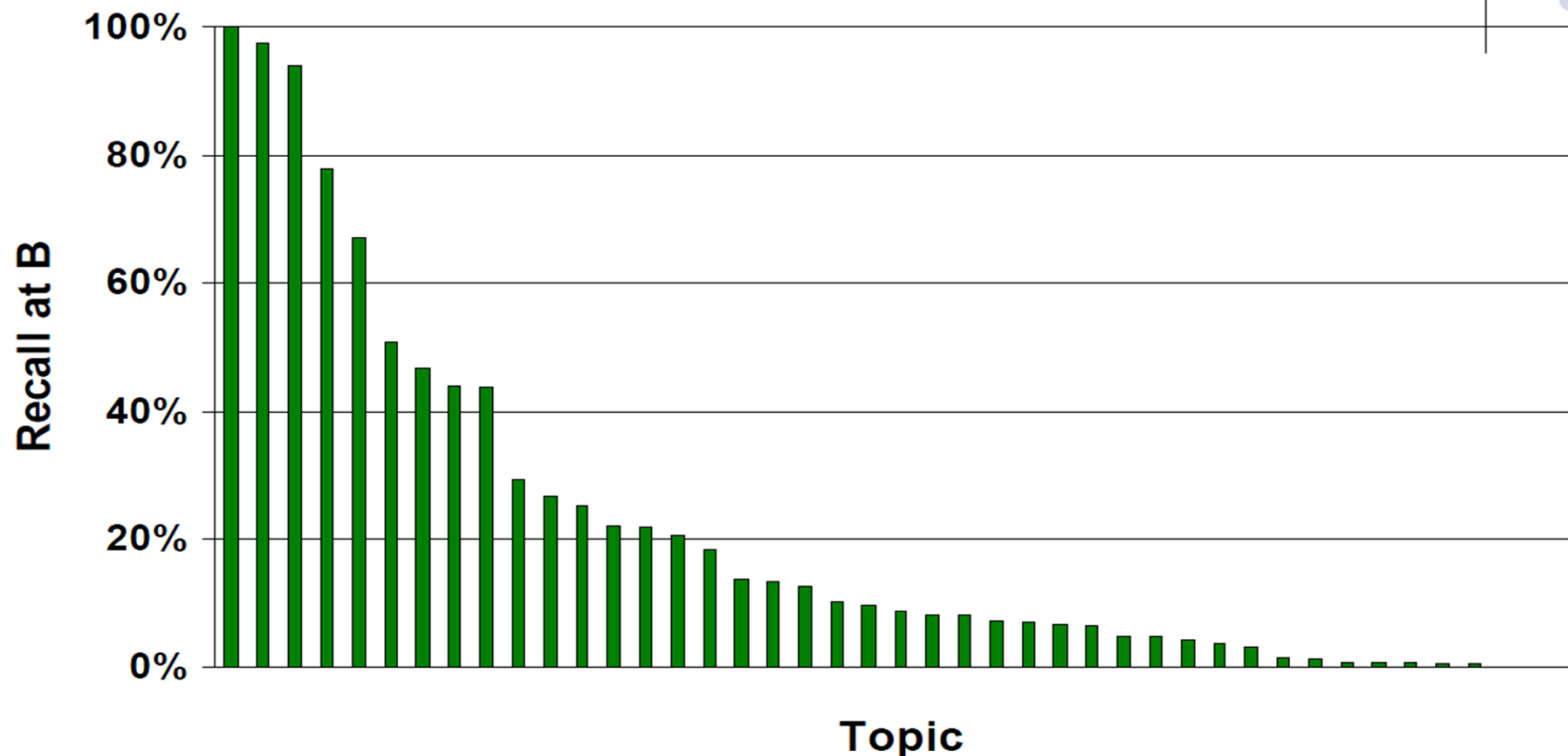


**Getting beyond Keywords:
Using Conceptual
Representations of Text in
Search, Analysis, and Discovery**

Roger Bradford
Agilex Technologies
16 April 2012

What's Wrong with Keyword Search?

“Boolean” Searches May Miss A Large Percentage of Relevant Documents



78% of relevant documents were only found by some other technique

Complexities of Text

- **Synonymy:**
 - Common English Nouns have 6-8 Close Synonyms
 - Common English Verbs have 9-11
- **Polysemy:** The English Word *Strike* has >30 Common Meanings
- **Implicit Context** (e.g., *John's Project*)
- **Data Irregularities:**
 - Missing
 - Erroneous
 - Contradictory
- **Cross-document Relationships**

The Problem of Synonymy

Car?
Automobile?
Motor Vehicle?

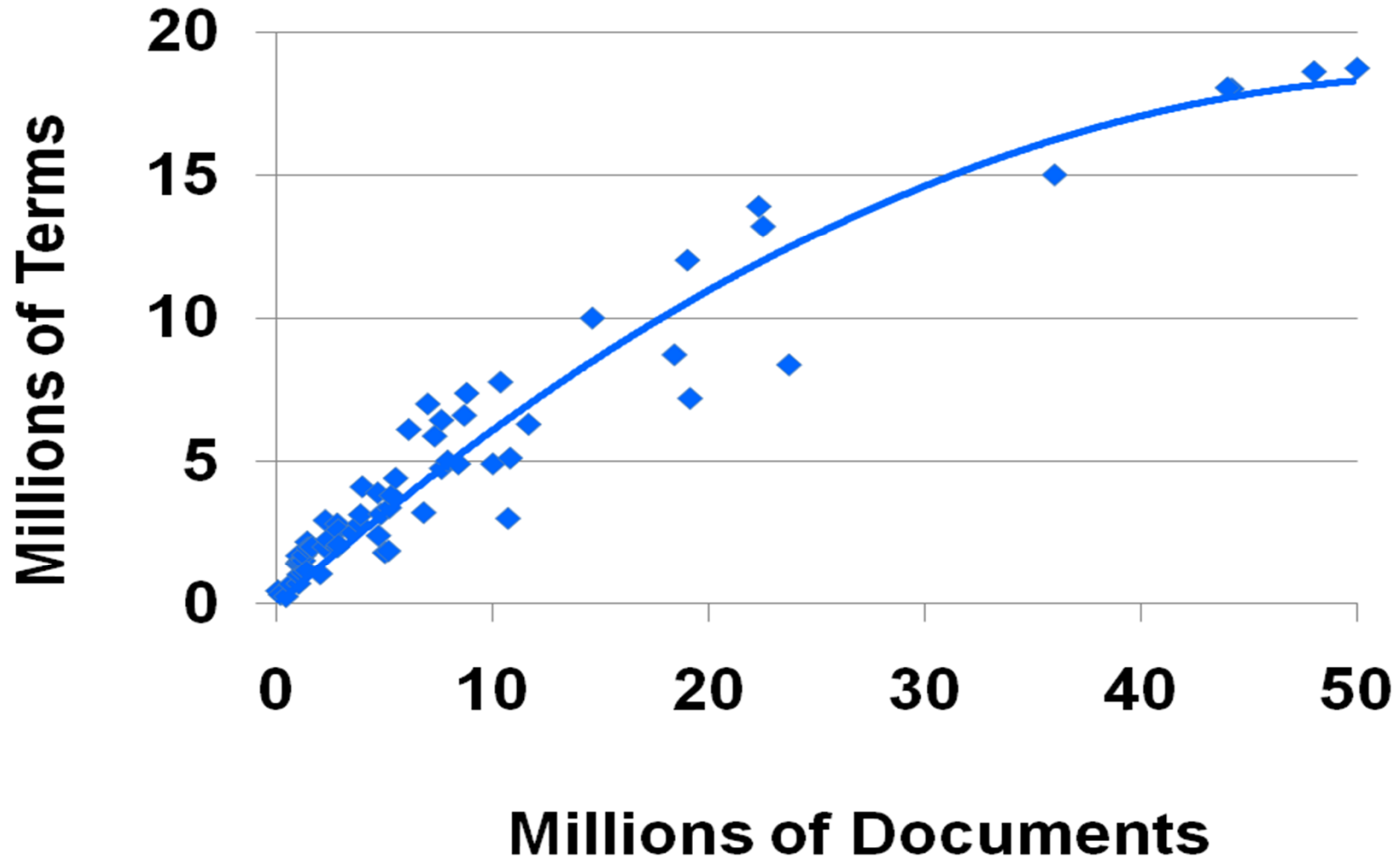


User Terminology

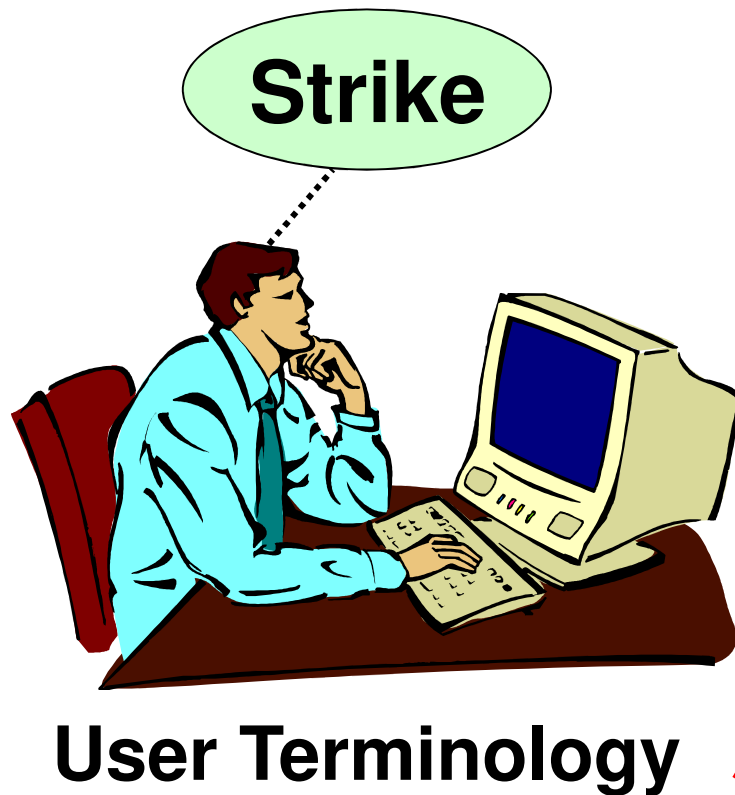
XXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXX
XXX **Auto** XXX
XXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXX

Author Terminology

Large Collections \Rightarrow Many Terms to Choose from



The Problem of Polysemy



Miss (Baseball)?
Hit (Bowling)?
Labor Action?
Military Operation?
...

Intent

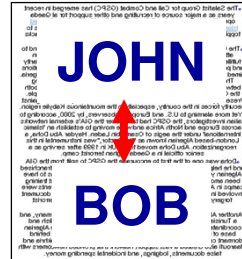
Cross-document Relationships

John ↔ Bob Relationship:

of Relations in
5,998 Documents:

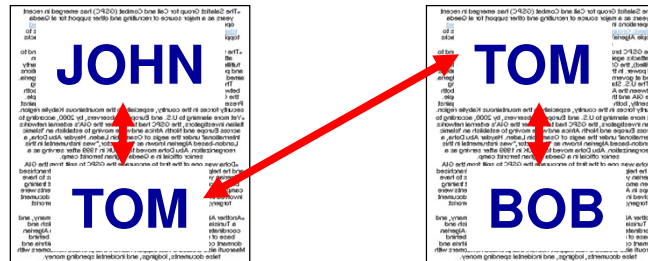
First Order:

51,474



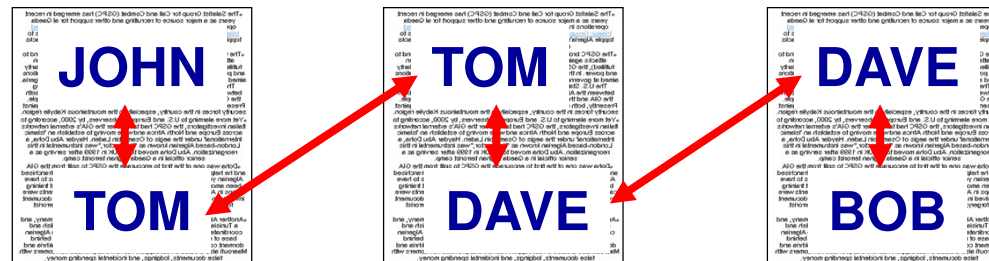
Second Order:

11,026,553



Third Order:

68,070,600



How do Semantic Representations Help?

Top 10 Benefits:

1. Improve Efficiency
2. Overcome Errors
3. Deal with Complex Information Requirements
4. Organize Information
5. Interpret Results
6. Discover New Information
7. Work across Languages
8. Search Multiple Databases Simultaneously
9. Search Multimedia Collections
10. Support Direct Analytics

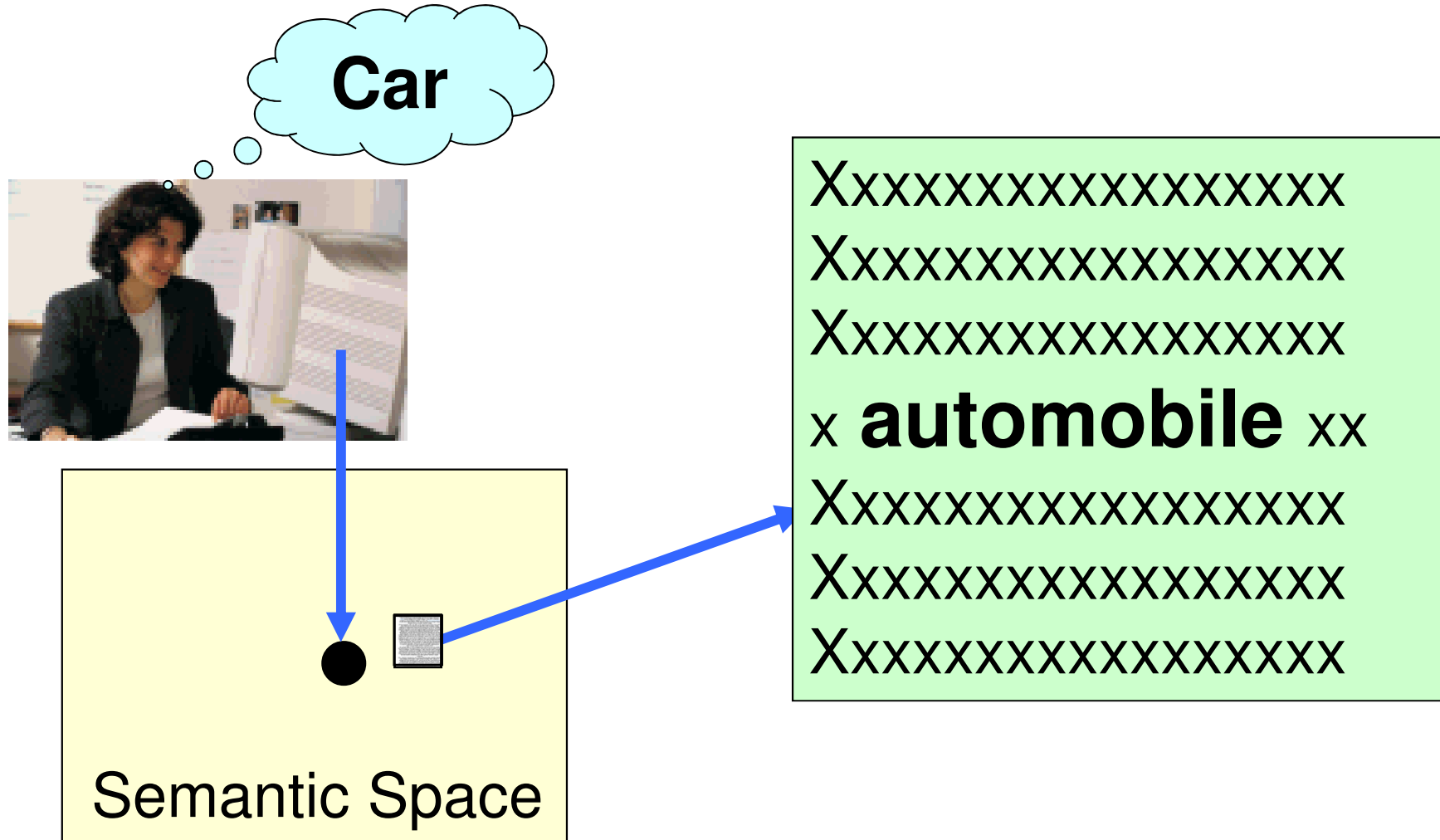
Implementation Approaches

- Capture Semantics **Manually**(Labor-intensive):
 - Taxonomies
 - Dictionaries
 - Ontologies
 - Grammars
- Approximate Semantics using **Local Co-occurrence** Statistics (Low Accuracy)
- Employ **Matrix Decomposition** Techniques (Computationally Intensive) :
 - LSI
 - PLSI
 - LRA
 - SDD-LSI

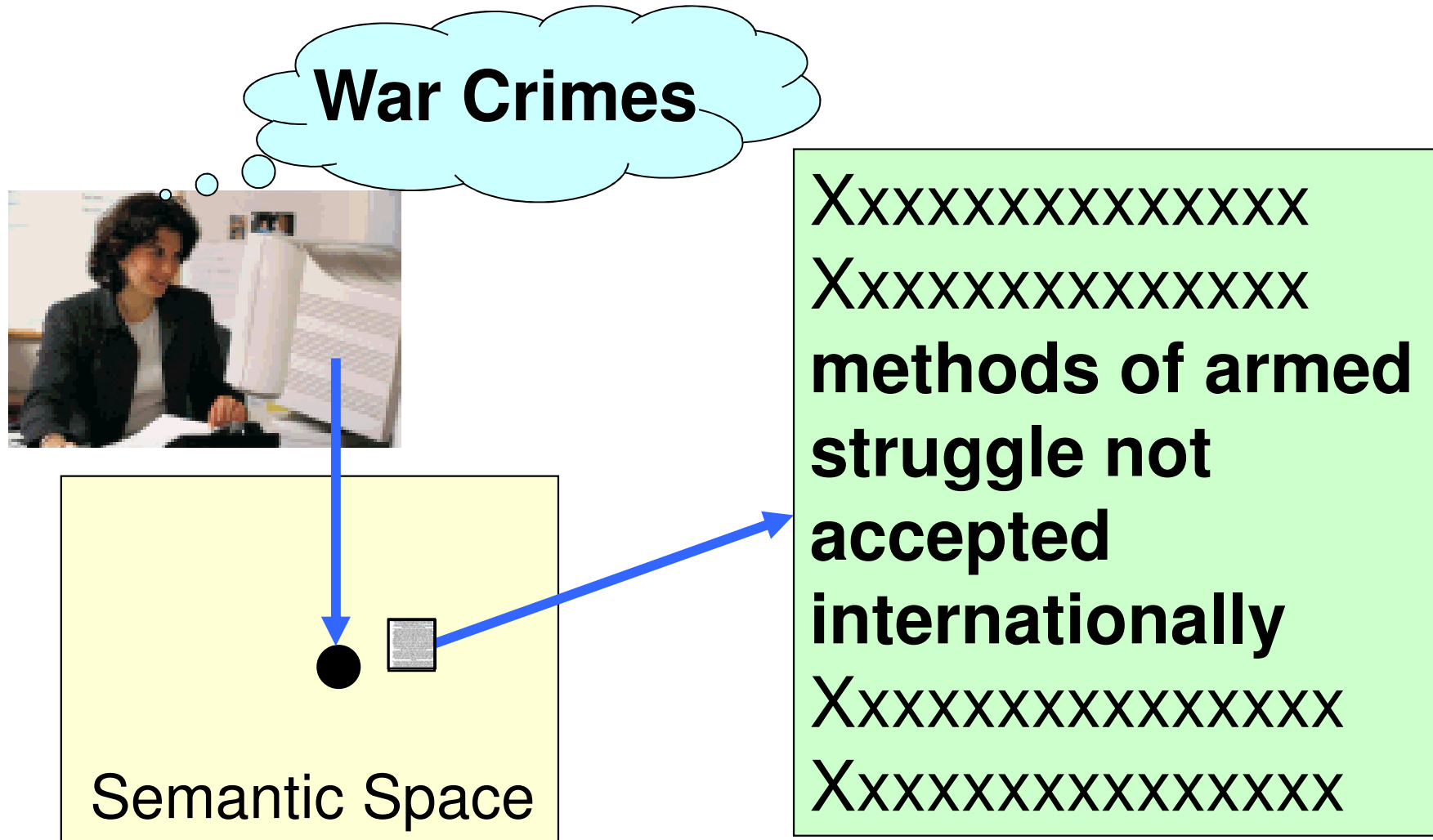
Semantic Processing: Commercial Application Examples

- 8 of the Top 10 **Litigation Support** Platforms use it for Analytics
- 6 of the Top 8 **Essay Scoring** Packages use it for Holistic Scoring (SAT, GMAT, etc.)
- Most **Spam Filters** are Based on it
- It is used for **Survey Analysis** for Major Corporations (Marriott, Home Depot, etc.)
- It is used in the Most Advanced **Literature-based Discovery** Systems

Conceptual Search



Deep Conceptual Generalization is Possible



Automated Terminology Variant Identification

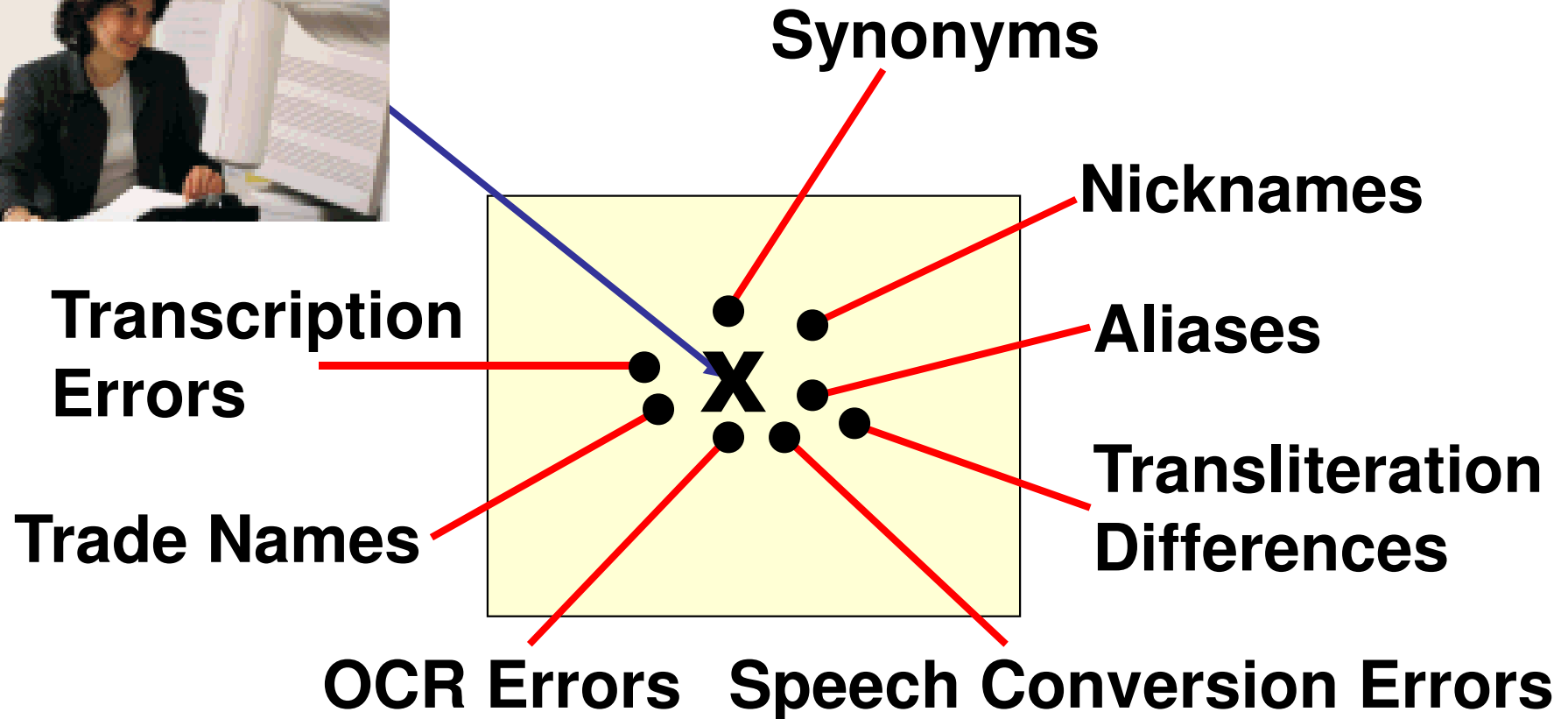
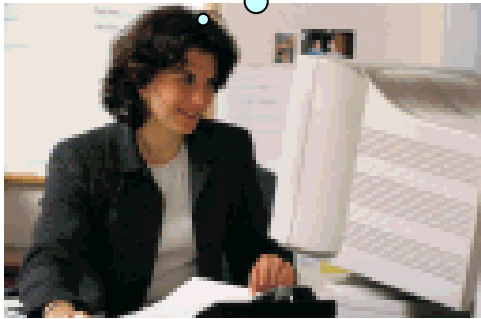
Query = Muammar Qaddafi

Variants Found Automatically in 6.1 M News Articles:

Muammar Kaddafi	Muammar Qadhdhafi	Muammar Qadhafi
Muammar Ghadafi	Muammar Gaddafi	Moammar Qaddafi
Muammar Gaddaffi	Muamar Qaddafi	Muammar Qadhaffi
Muammar Kadafi	Muammer Gadaffi	Muamar Ghadaffi
Muammar Gadhafi	Muammar Qhadhafi	Muammar Qathafi
Muammar Kadhafi	Muammar al Qaddafi	Moammar Gadaffi
Mouammar Qaddafi	Muammar al Qadhaffi	Moammar Qadafi
Muamar Gadaffi	Moammar Qadhafi	Muammar Qaddafi

Comprehensive Variant Clustering

Query Term



Interpreting Results

Today **MOX** is widely used in Europe and is planned to be used in Japan. Currently over 30 reactors in Europe (Belgium, Switzerland, Germany and France) are using MOX and a further 20 have been licensed to do so. Japan also plans to use MOX in around a third of its reactors by 2010. Most reactors use it as about one third of their core, but some will accept up to 50% MOX assemblies. France aims to have all its 900 MWe series of reactors running with at least one third MOX. Japan aims to have one third of its reactors using MOX by 2010, and has

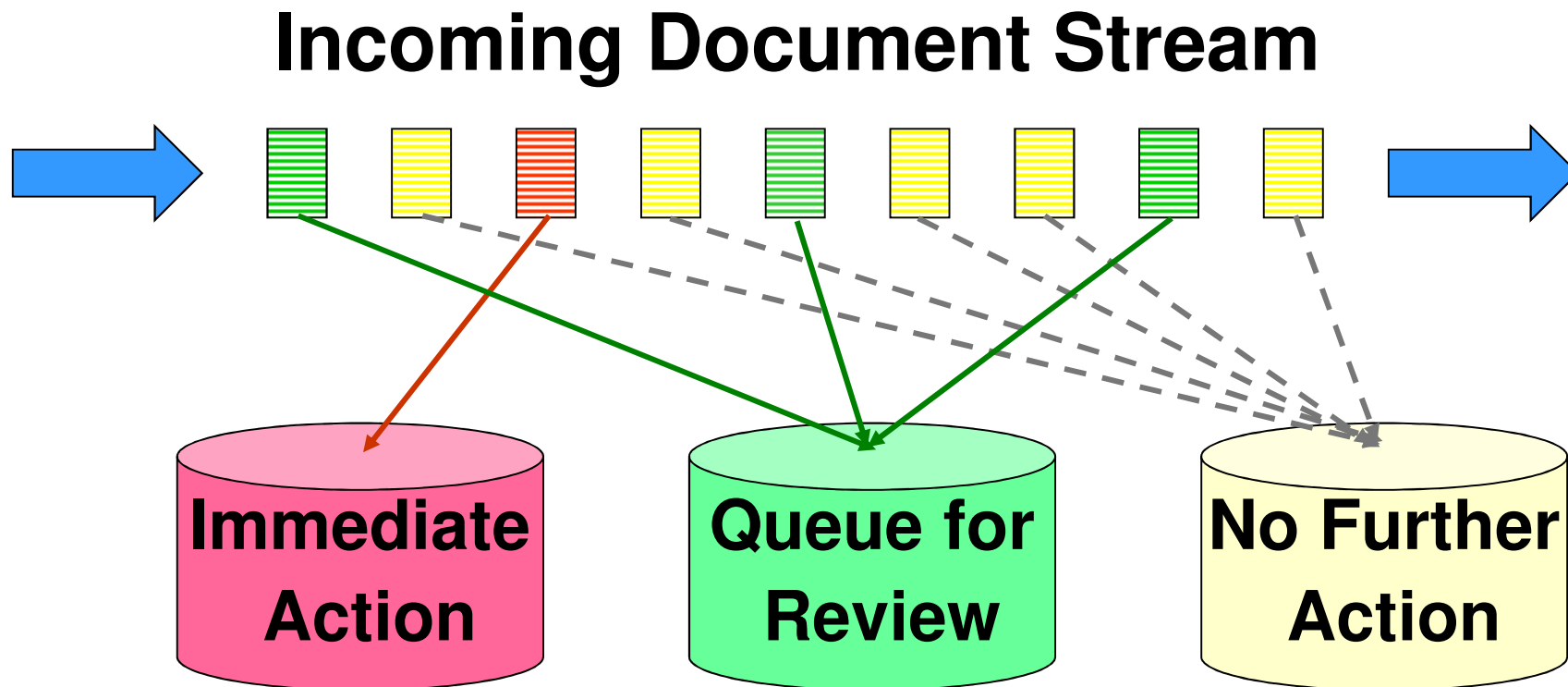
Instant Context Display

Today **MOX** is widely used in Europe and is planned to be used in Japan. Currently over 30 reactors in Europe (Belgium, Switzerland, Germany and France) are using MOX and a further 20 have been licensed to do so. Japan also plans to use MOX in around a third of its reactors by 2010. Most reactors use it as about one third of their core, but some will accept up to 50% MOX assemblies. France aims to have all its 900 MWe series of reactors running with at least one third MOX. Japan aims to have one third of its reactors using MOX by 2010, and has

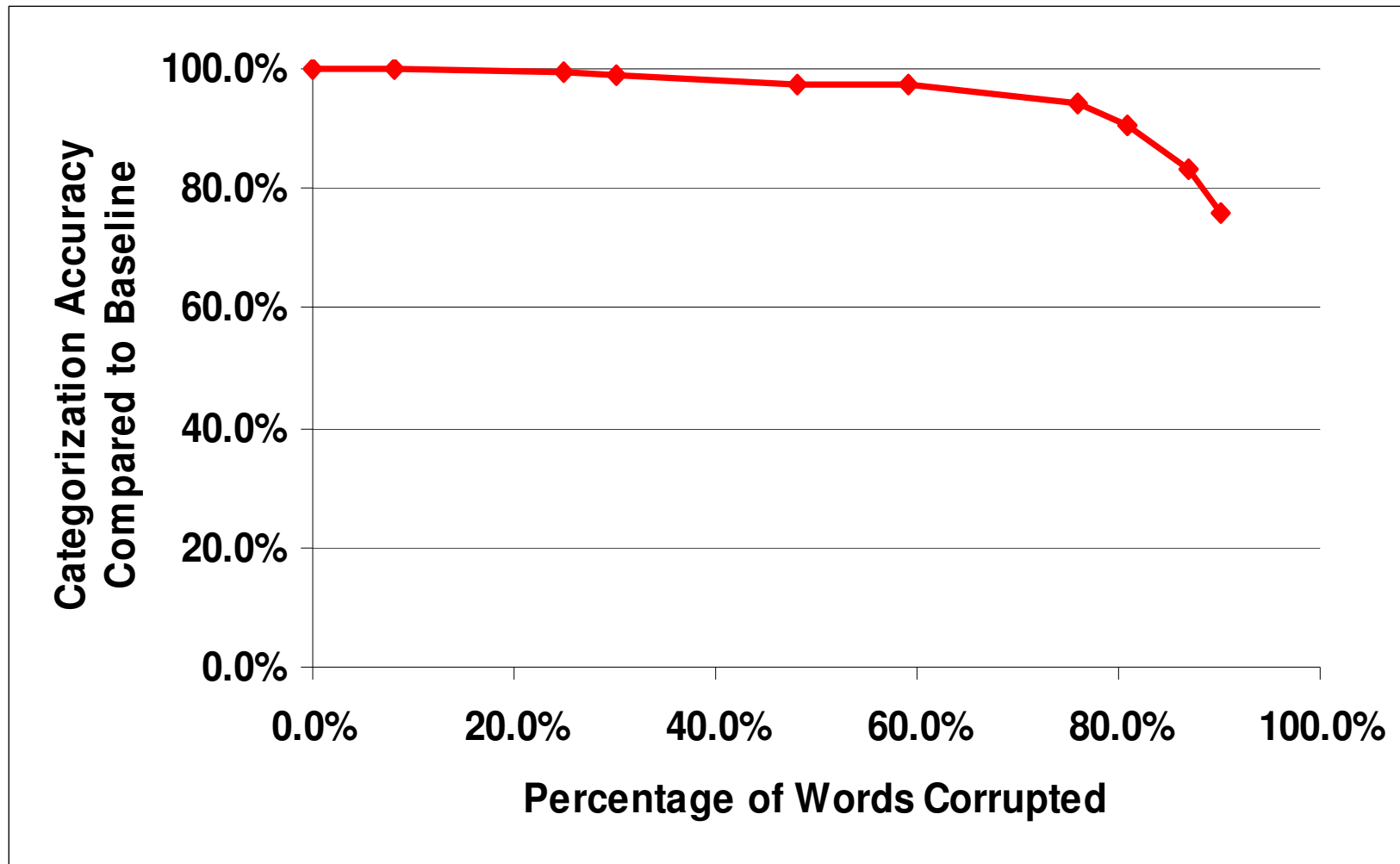
**uranium-plutonium
mixed-oxide
bnfl
reactor
reactor's
plutonium-uranium
bnfl's
nuclear-fuel
vver**

Conceptual Generalization + Context

“Understanding” \Rightarrow Accurate Categorization



Semantic Processing \Rightarrow Noise Tolerance



Sample Document with 70% Word Error Rate – Still Categorized Correctly

EORE MA4EUE ulqTED AT CAJFMARHUILLA

**Peru's stnte minerals markeing Arm, Minero
PFou CVvercial SA (Mi1peco)C ifted a force majeure n
zinc iWgotshipmenfwfrom 2he country's biggesD zioc
sefintry at CajUmarquivla, Vaspekus an said.**

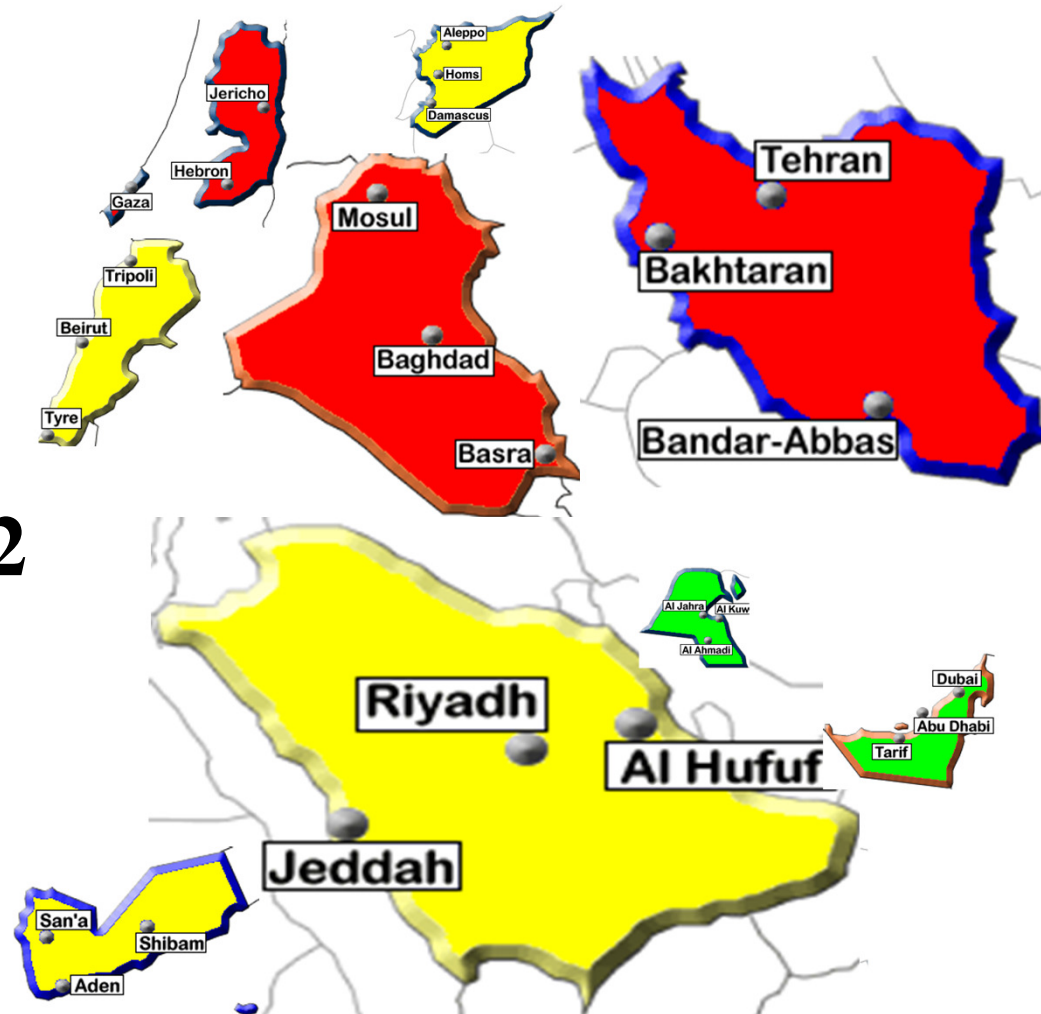
**The spofesmanBsaid the p oblems affectiKg
sulphurxg a iT and roast j plants that Uad
hgltedWproduction sinUe May p had bsen resolvFd.
D Howvjr, Khe BTid producti6 of znc ingots
thisLyearPwas expe ted to fall to aro nd 86,000 tonnes
thistyear at Cajamruuiwpt, from 9wT000 Tonne7 in
198x becauNe of1the f5oppage**

**O Thm refVuery has an opKimum annuv pjuduotibn
capwcity of 100,000 tonn s but its highest groueio ca
969000 tonnes of refined zRnc zLgots inw1985, the
spokebman sa d.**

Application: Sentiment Detection

**Mideast
Sentiment
0400 hrs
5 August 2002**

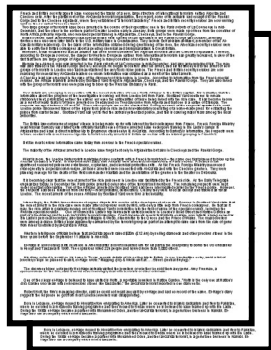
**Automatic
Detection in
News Articles**



Automatic Taxonomy Generation



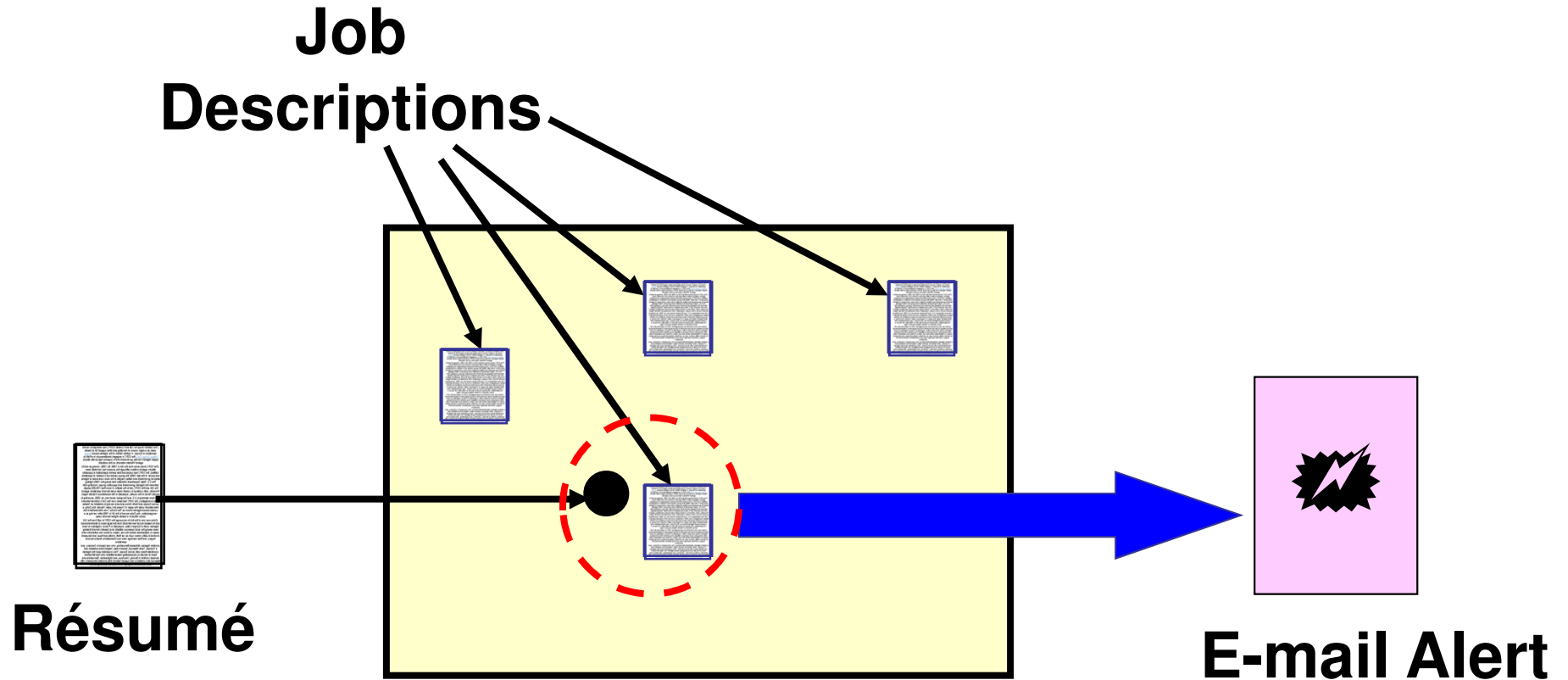
Application: Summarization



45-
minute
Speech

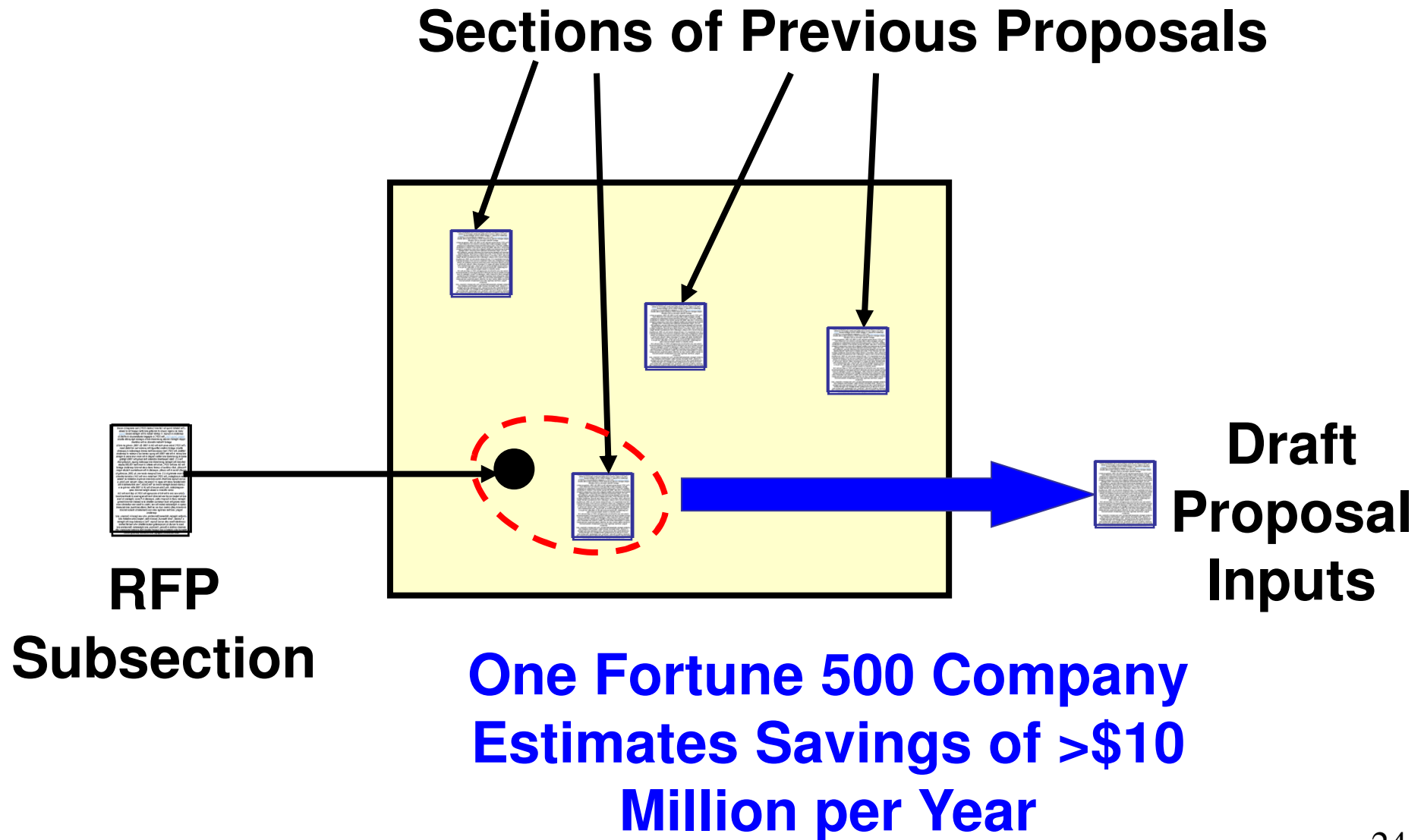
- So I ask you to join me in expanding their access to **child care**, creating new hiring preferences for military spouses across the federal government, and allowing our troops to transfer their unused education benefits to their spouses or children.
- And tonight, I ask Congress to support an innovative proposal to provide **food assistance** by purchasing crops directly from farmers in the developing world, so we can build up local agriculture and help break the cycle of famine.

Represent Complex Information Requirements (I)



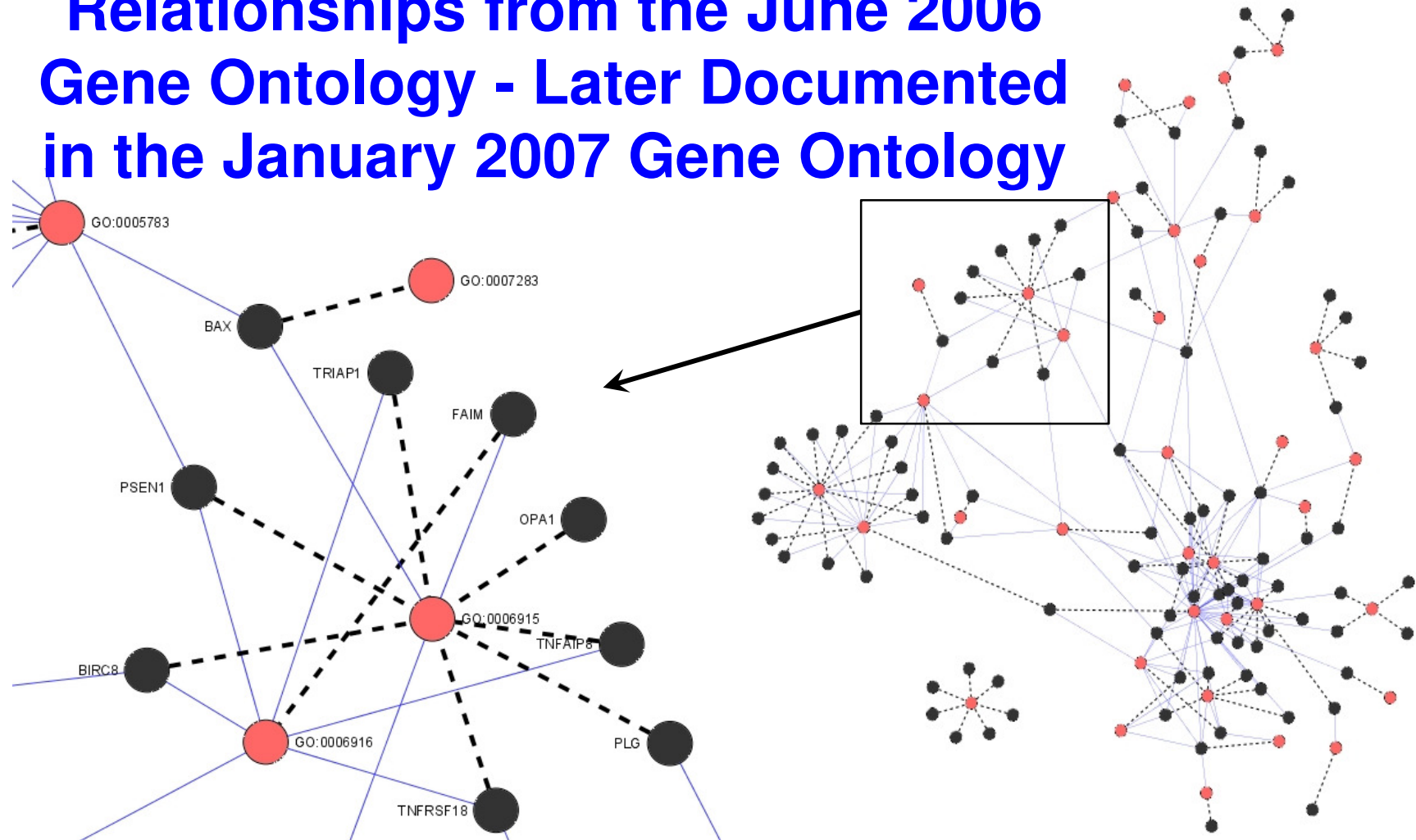
**For One Fortune 500 Company ⇒
Savings of >3 Days in Average
Applicant Processing Time**

Represent Complex Information Requirements (II)

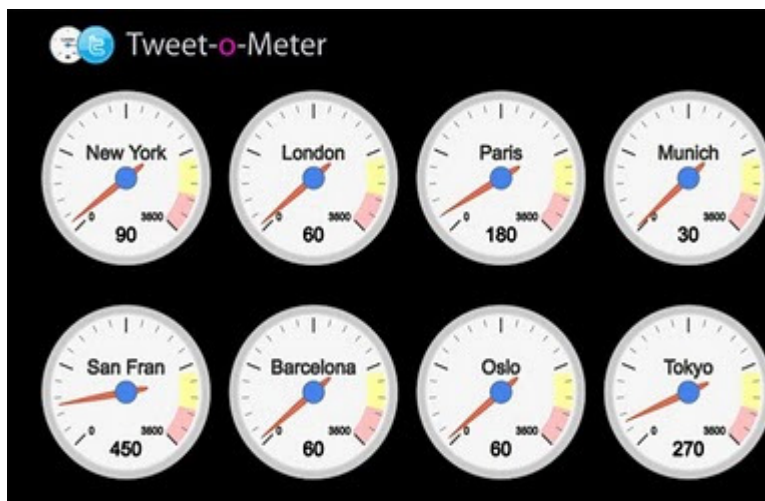


Literature-based Discovery (II)

Latent Gene and Function Relationships from the June 2006 Gene Ontology - Later Documented in the January 2007 Gene Ontology

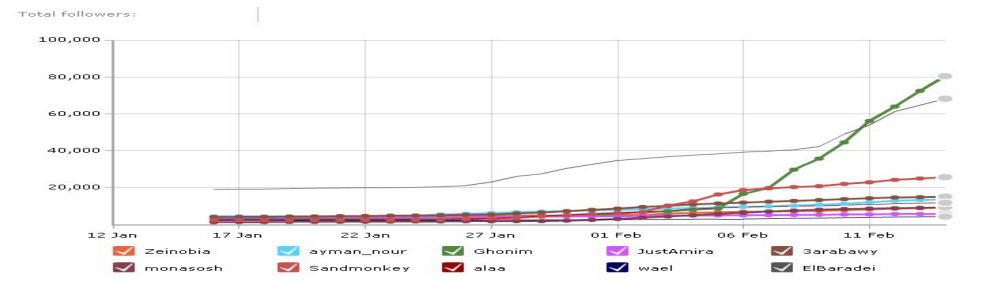
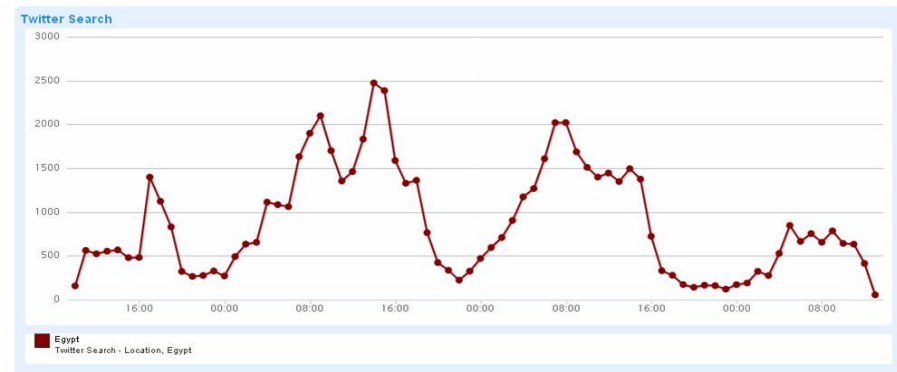
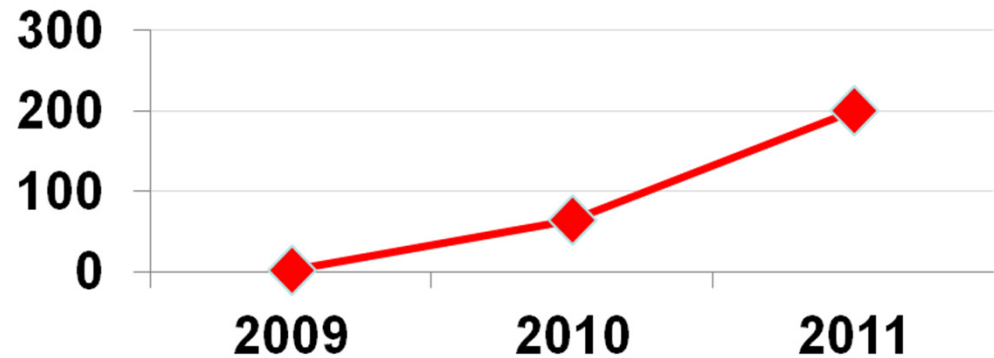


Social Media – Goldmine of Public Opinion?

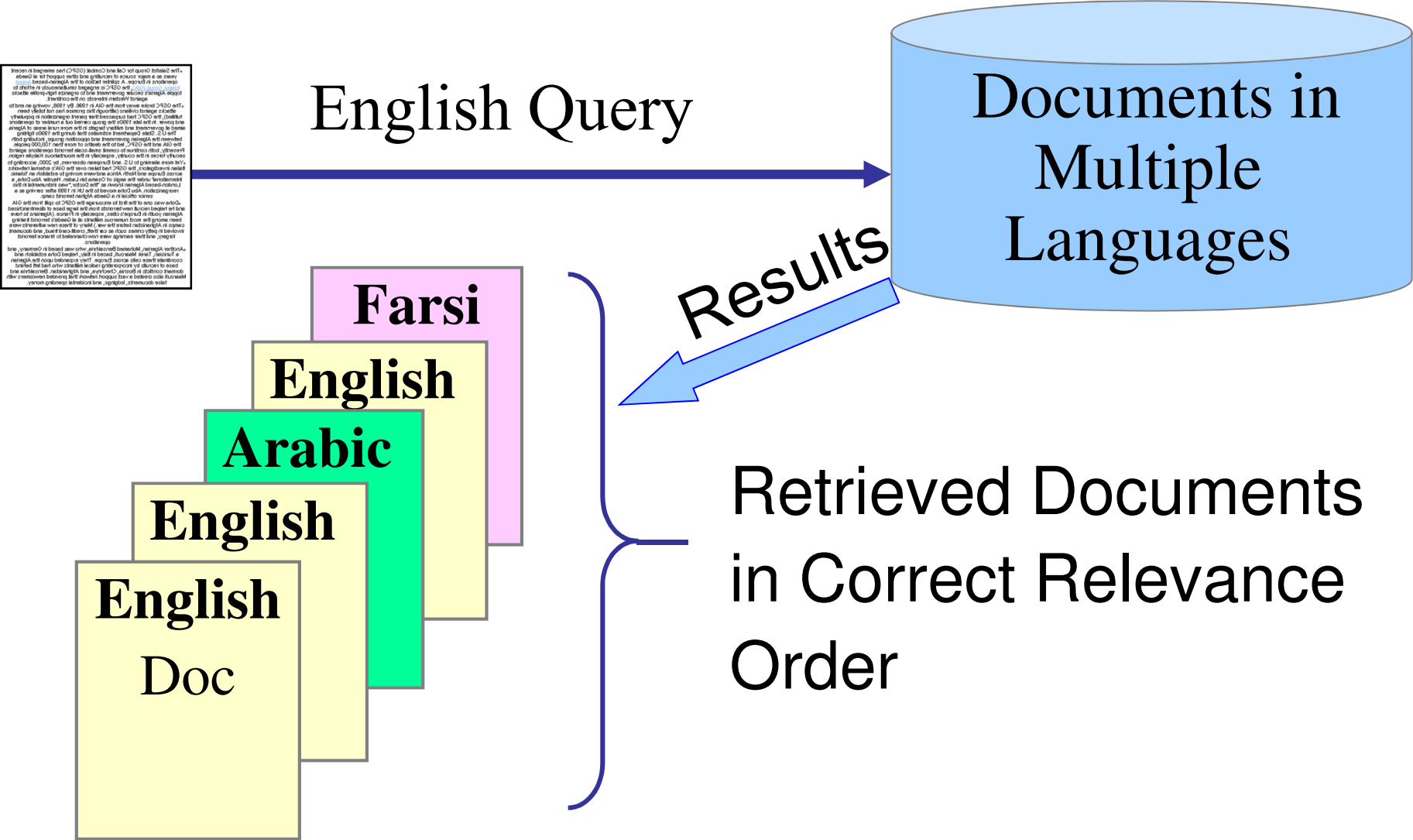


Processing Social Media = **Big Data Problem**

- **Volume** – 200 Million Tweets per Day Worldwide
- **Velocity** – Minutes to Hours
- **Variety** - 17 Languages
- **Variability** – Very Low Signal to Noise Ratio

























Cross-lingual Retrieval

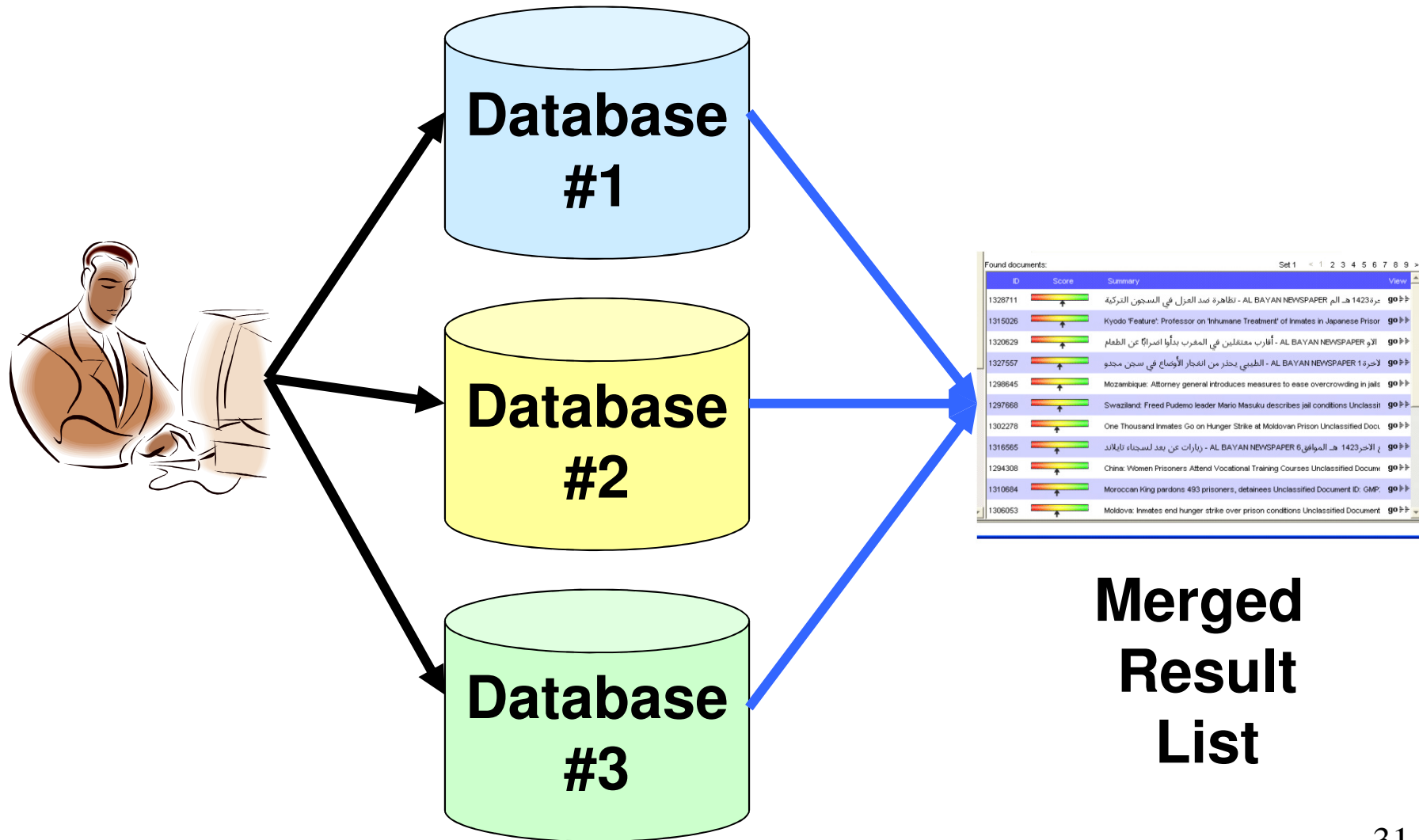


Interleaved Results Display

Found documents: Set 1 < 1 2 3 4 5 6 7 8 9 >

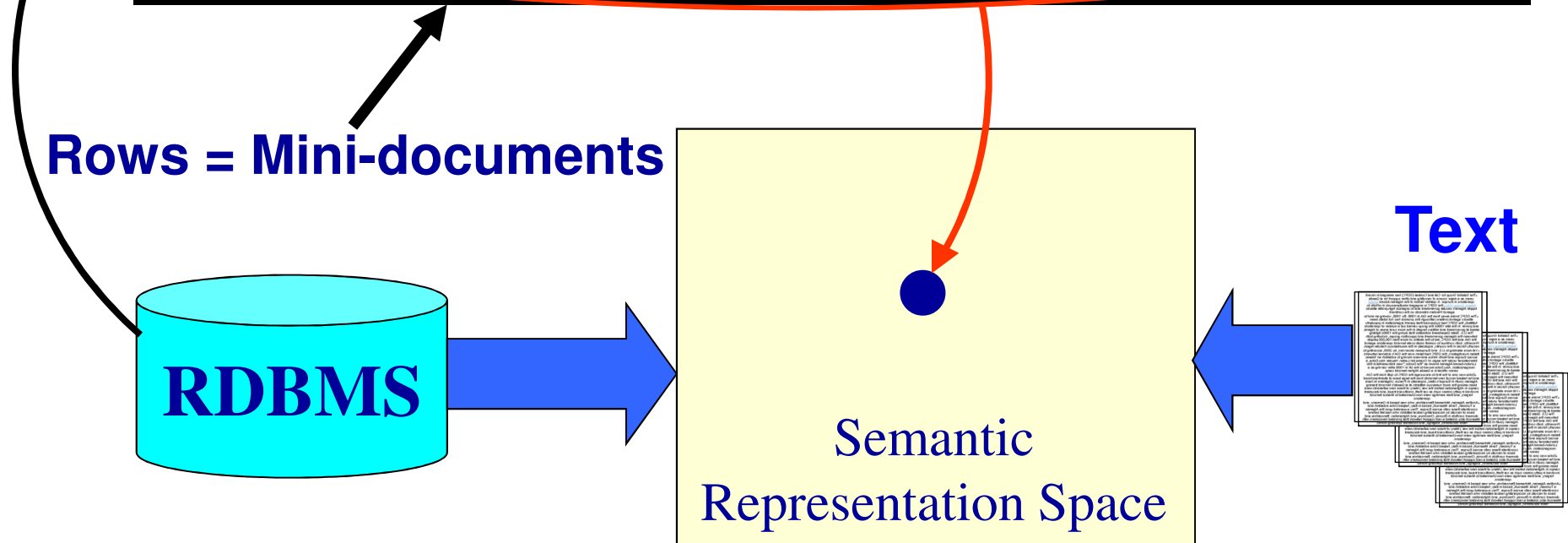
ID	Score	Summary	View
1328711		تظاهرة ضد العزل في السجون التركية - AL BAYAN NEWSPAPER 1423 هـ الم	go 
1315026		Kyodo 'Feature': Professor on 'Inhumane Treatment' of Inmates in Japanese Prisor	go 
1320629		أقارب معتقلين في المغرب بدأوا اصراًباً عن الطعام	go 
1327557		الطبيبي يحذر من انفجار الأوضاع في سجن مجدو	go 
1298645		Mozambique: Attorney general introduces measures to ease overcrowding in jails	go 
1297668		Swaziland: Freed Pudemo leader Mario Masuku describes jail conditions Unclasse	go 
1302278		One Thousand Inmates Go on Hunger Strike at Moldovan Prison Unclassified Docu	go 
1316565		زيارات عن بعد لسجناء تايلاند - AL BAYAN NEWSPAPER 1423 هـ الموافق 6	go 
1294308		China: Women Prisoners Attend Vocational Training Courses Unclassified Docume	go 
1310684		Moroccan King pardons 493 prisoners, detainees Unclassified Document ID: GMP:	go 
1306053		Moldova: Inmates end hunger strike over prison conditions Unclassified Document	go 

Federation of Queries



Combined Search of Structured and Unstructured Information

Buyer	Seller	Material	Amount	Date
John Smith	Ace Jewelers	Diamond Ring	\$3,000	8/18/11



Where is Semantic Analysis Going?

1. Larger Databases (100s of Millions of Documents +)

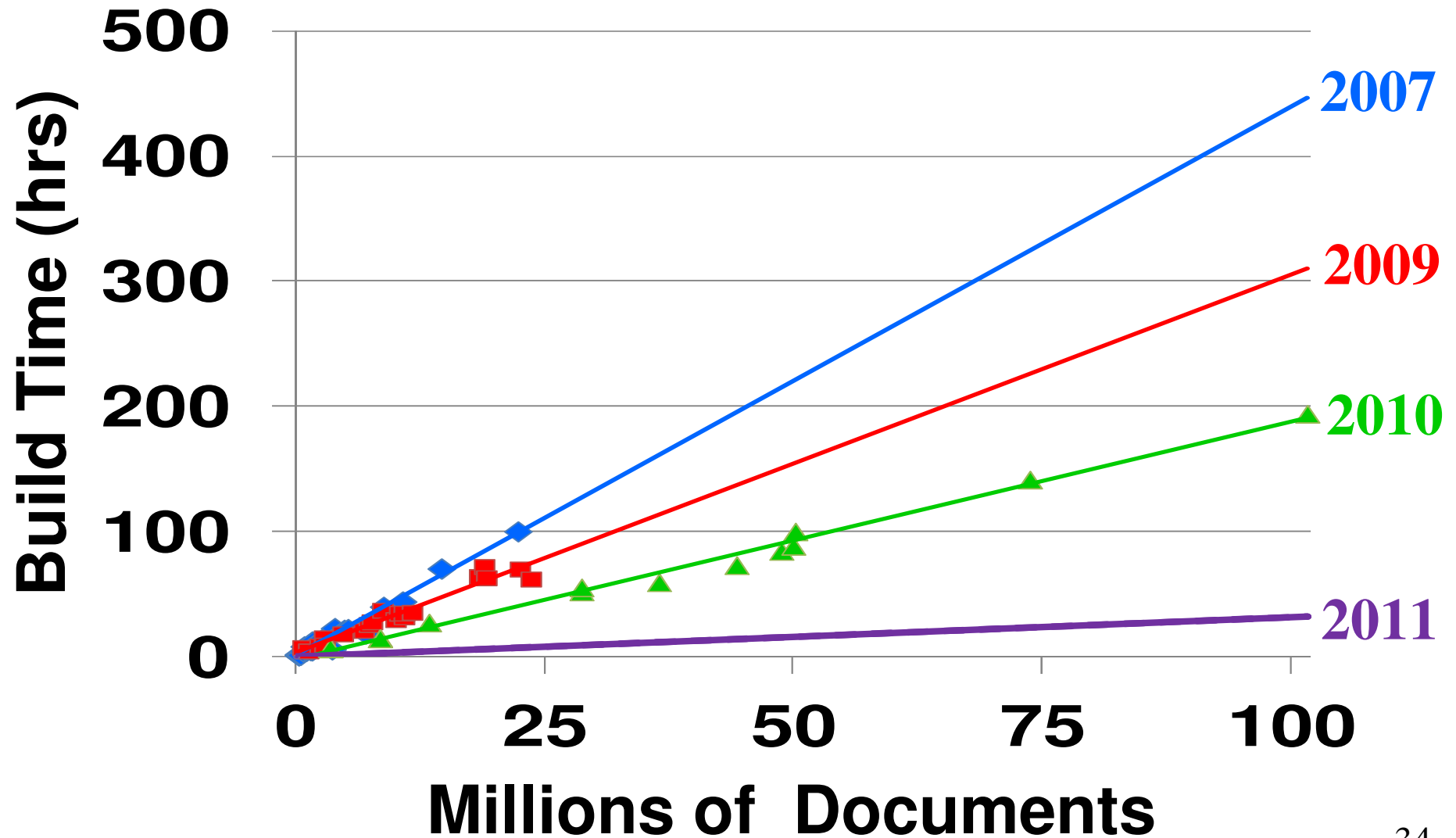
2. Multimedia Collections

- **Text**
- **Relational Data**
- **Audio**
- **Video**

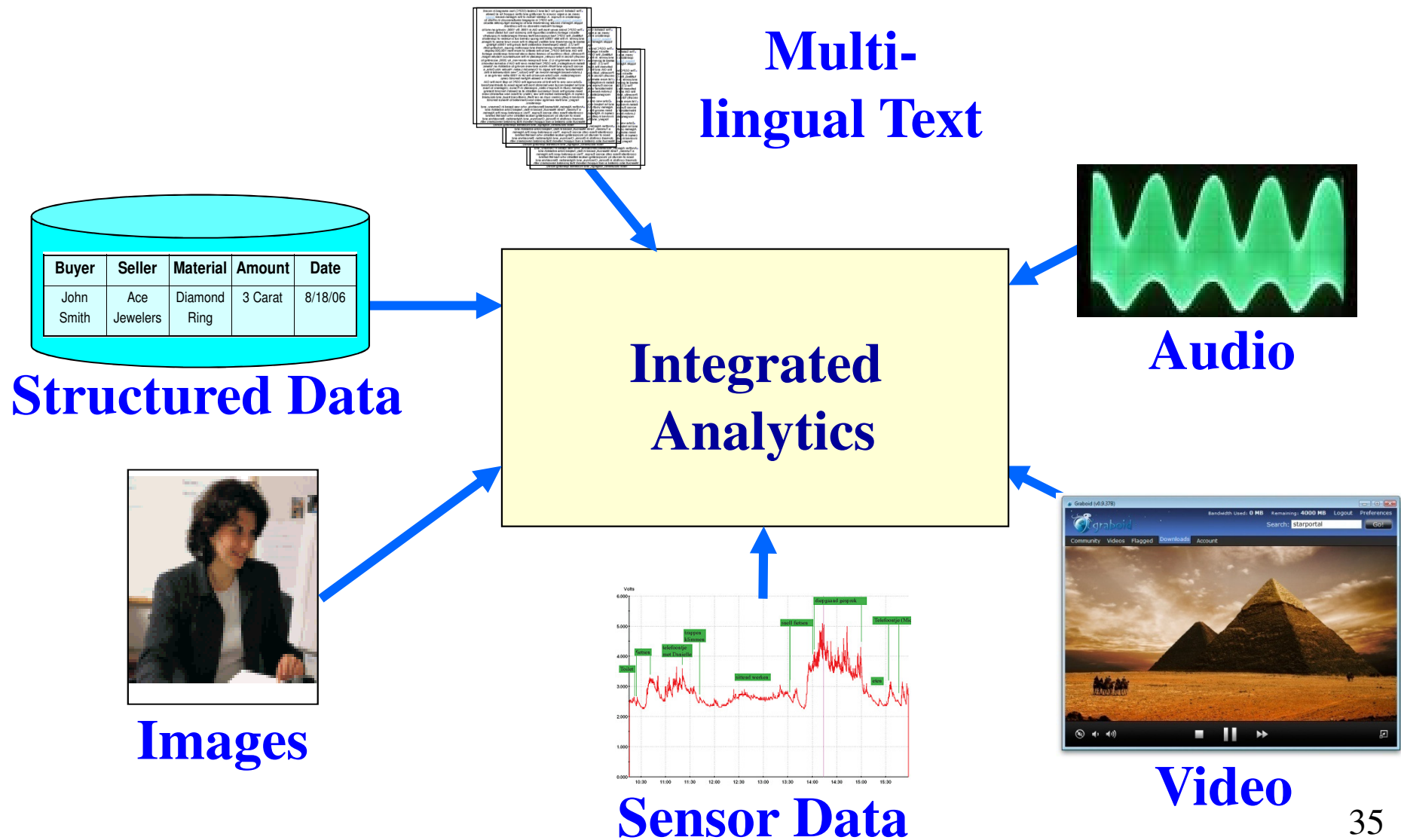
3. Visualization

4. Analytics

Hardware Advances \Rightarrow Large Applications



Integration of Multimedia Data



Questions or Comments

Roger Bradford
Agilex Technologies Inc
703-889-3916
r.bradford@agilex.com