II-SDV

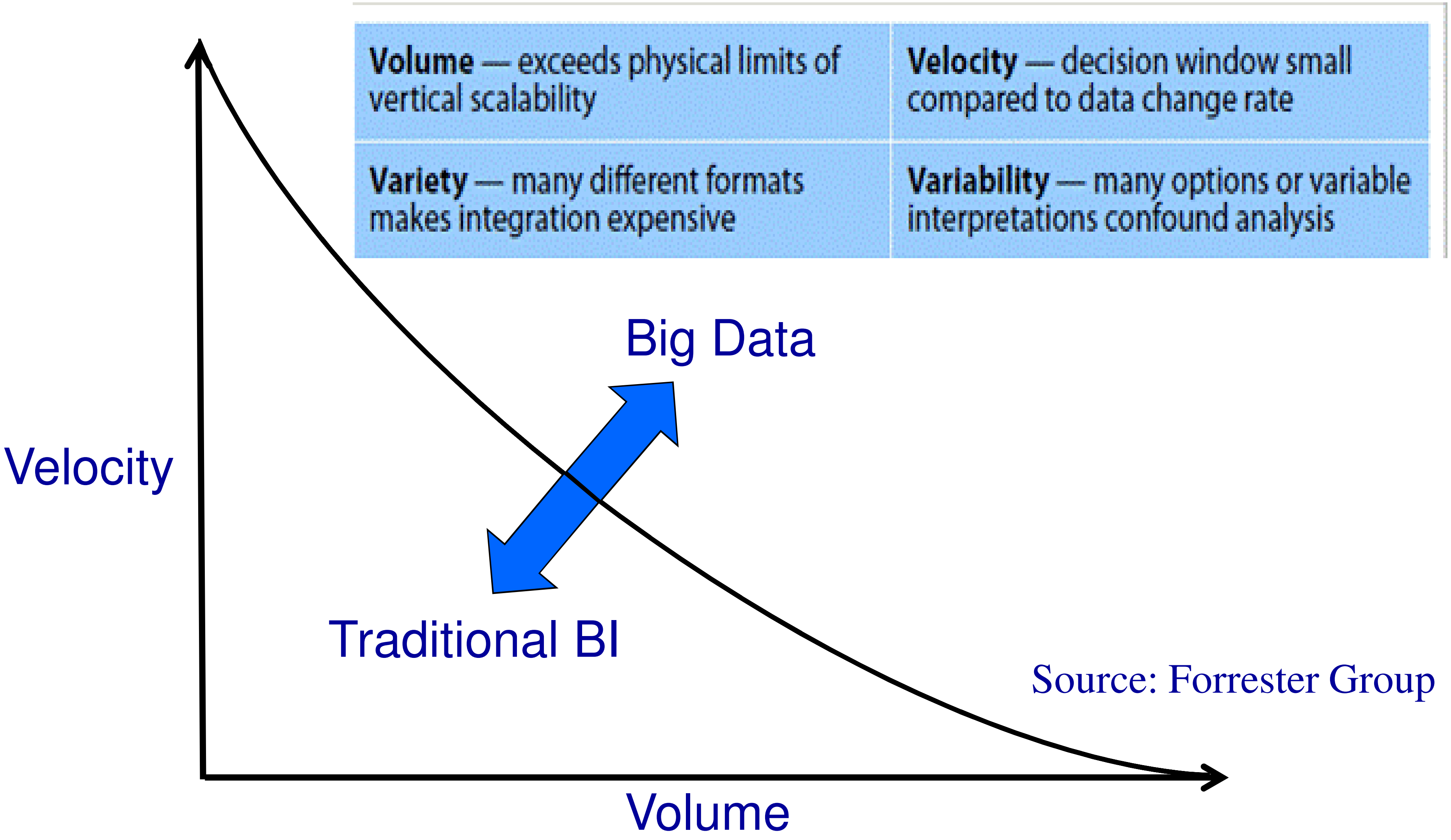
The Analytics Challenges Posed by Big Data

Roger Bradford

Agilex Technologies

15 April 2013

Standard Big Data View



Volume — exceeds physical limits of vertical scalability	**Velocity** — decision window small compared to data change rate
Variety — many different formats makes integration expensive	**Variability** — many options or variable interpretations confound analysis

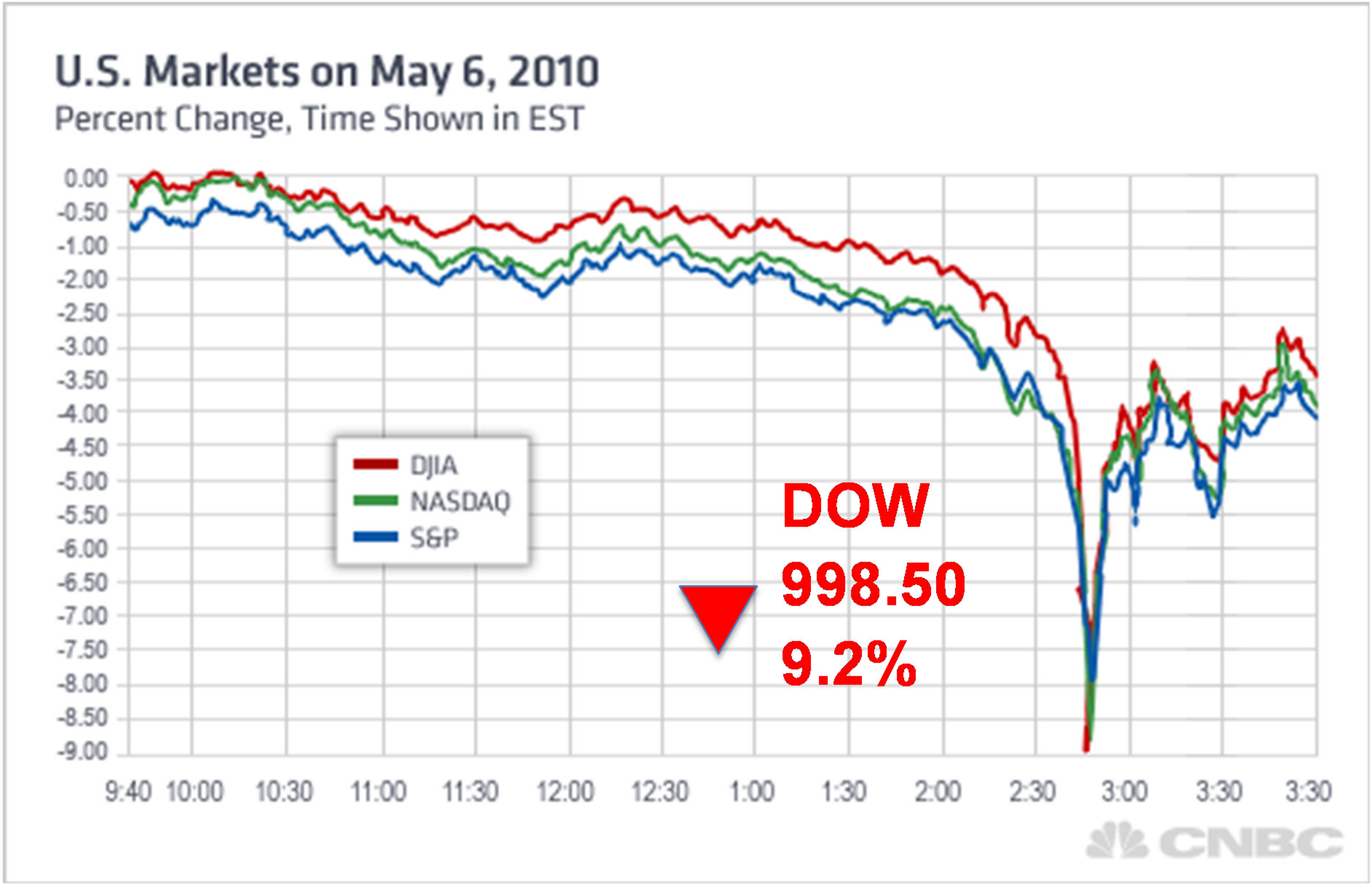
Source: Forrester Group

Big Data - Volume Examples

Activity	Rate
E-mail	>300 Billion*/Day
Text Messages	> 24 Billion/Day
Cell Phones	> 10 Billion Calls/Day
YouTube	> 1 Million New Videos/day
Twitter	> 500 Million Tweets/Day
Facebook	> 1 Billion Posts/Day

*Short Scale Billion = 1,000 Million = 10^9

Big Data - Velocity Example

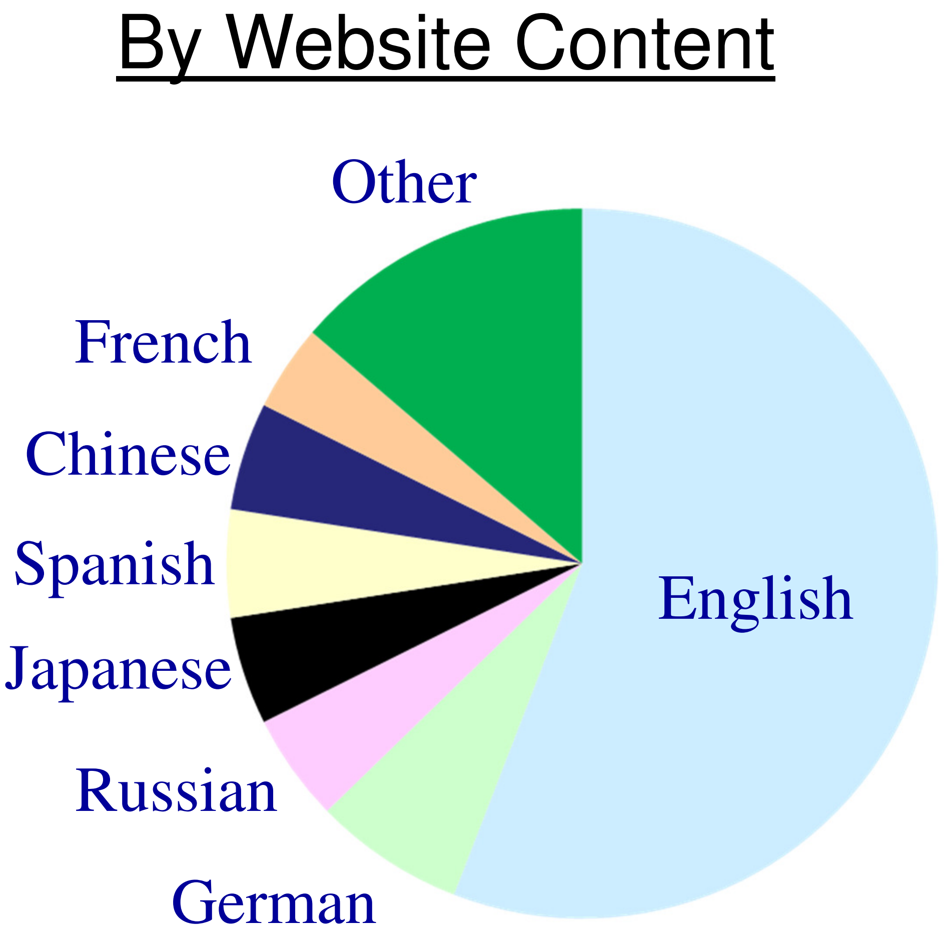


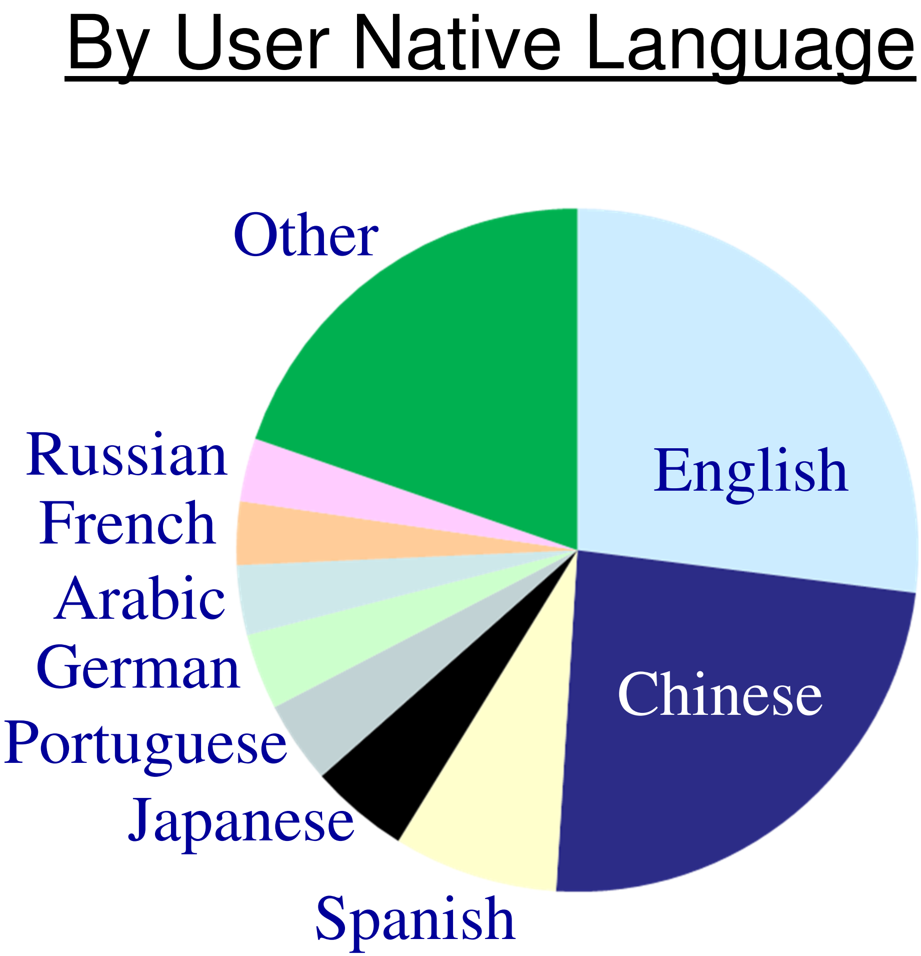
U.S. Markets on May 6, 2010

Percent Change, Time Shown in EST

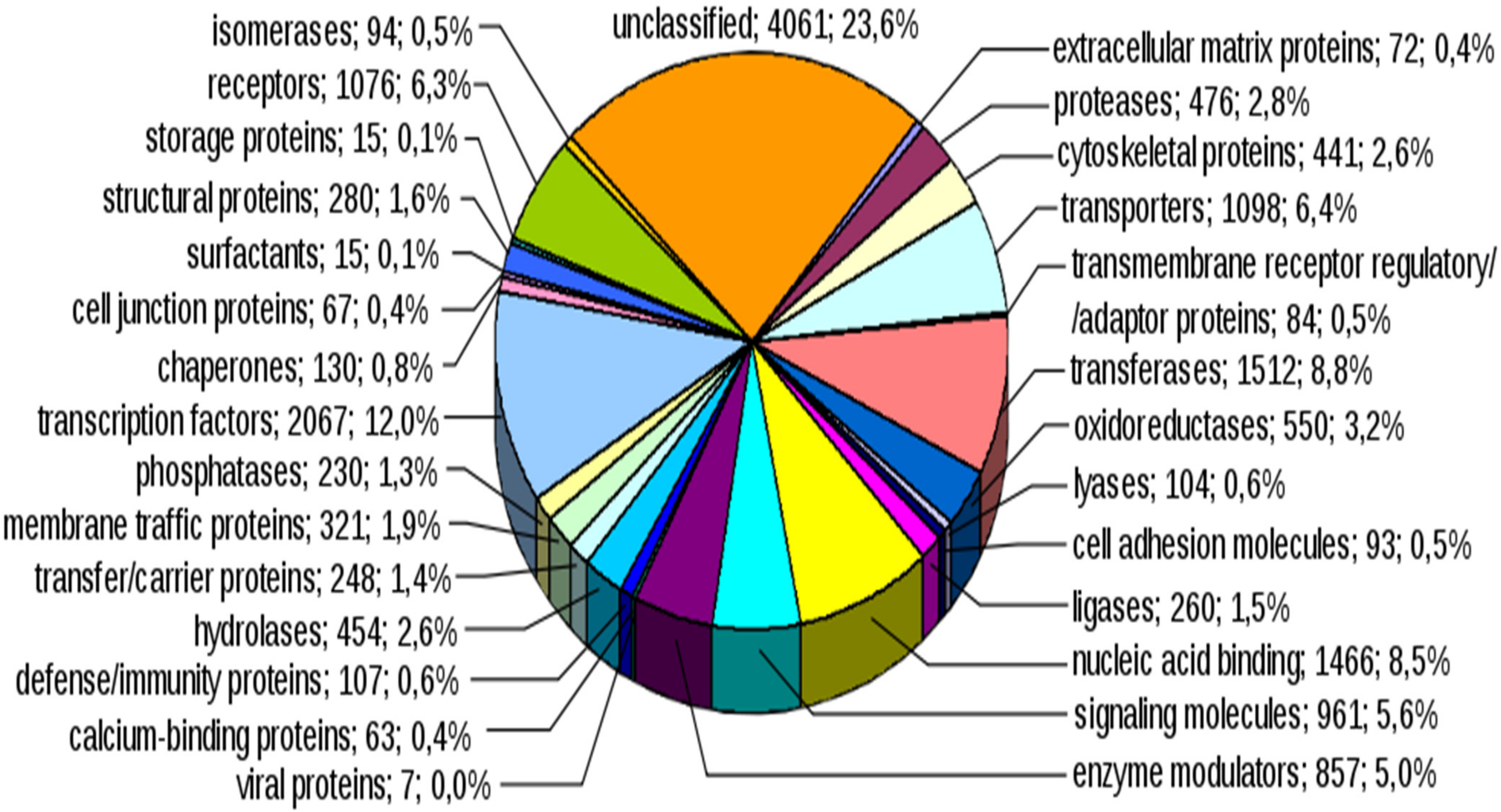
DOW
998.50
9.2%

Big Data Variety Example –
Internet Language Usage





Big Data - Variability Example



Functions of 17,209 Genes

Structured and Unstructured Data

Structured	Unstructured
Sales Data	E-mail
Financial Data	Instant messaging
Climate Data	Tweets
Census Data	Audio
Movie Ratings	Images
Sensor Measurements	Video

Unstructured Information Accounts for more than 80% of all Data in Organizations and is Growing 15X Faster than Structured Data

Challenges: Big Data vs. Hard Problems

Big Data
Volume
Velocity
Variety
Variability

Hard Problems
Ambiguity
Nth-order Relations
Cardinality
Non-locality

Ambiguity in Text

- Synonomy:
  - ➢Common English Nouns have 6-8 Close Synonyms
  - ➢Common English Verbs have 9-11
- Polysemy:
  - ➢The Word *Strike* has >30 Common Meanings
- Entity Ambiguity:
  - ➢ There are more than 45,000 People Named *John Smith* in the United States
  - ➢ There are more than 300,000 People Named *Zhang Wei* in China
- Entity Variability:
  - ➢Some Person Names in Collections of Interest Occur in over 100 Variants

Name Variant Example

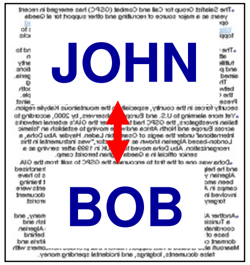
Vladimir Putin	Vladimir Poutine	Vladimir V. Putin
Vladmir Putin	Valdimir Putin	Vladimir Vladimirovich Putin
Vladamir Putin	Vladimr Putin	Vladimir Vladimirovitch Putin
Vlaidimir Putin	Vladimir Puttin	Vladimir Vladimirovic Putin
Vladimir Poutin	Putin, Vladimir	Putin, Vladimir Vladimirovitch
Vladimir Puttin	Vladamir Putin	Putin, Vladimir Vladimirovich
Vlademir Putin	Vladimier Putin	V.V. Putin

Nth-order Relationships

John ↔ Bob Relationship:

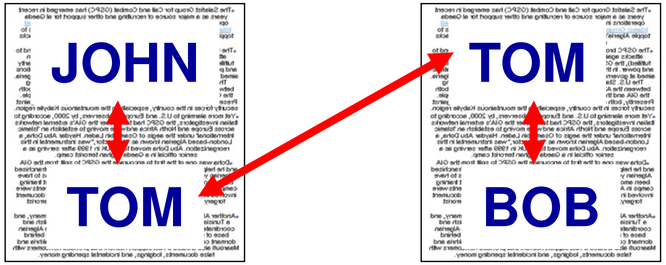
# of Relations in 5,998 Documents:

First Order:



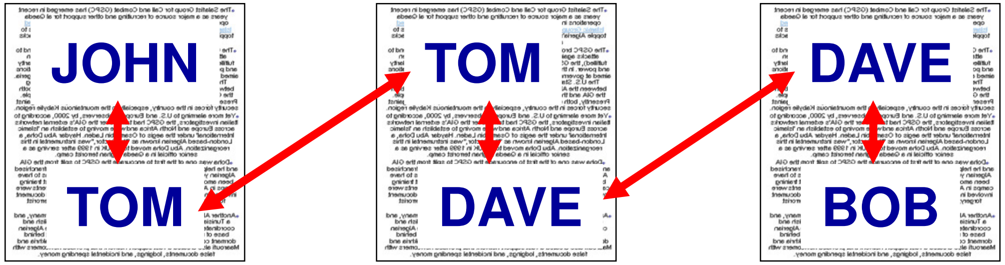
51,474

Second Order:



11,026,553

Third Order:



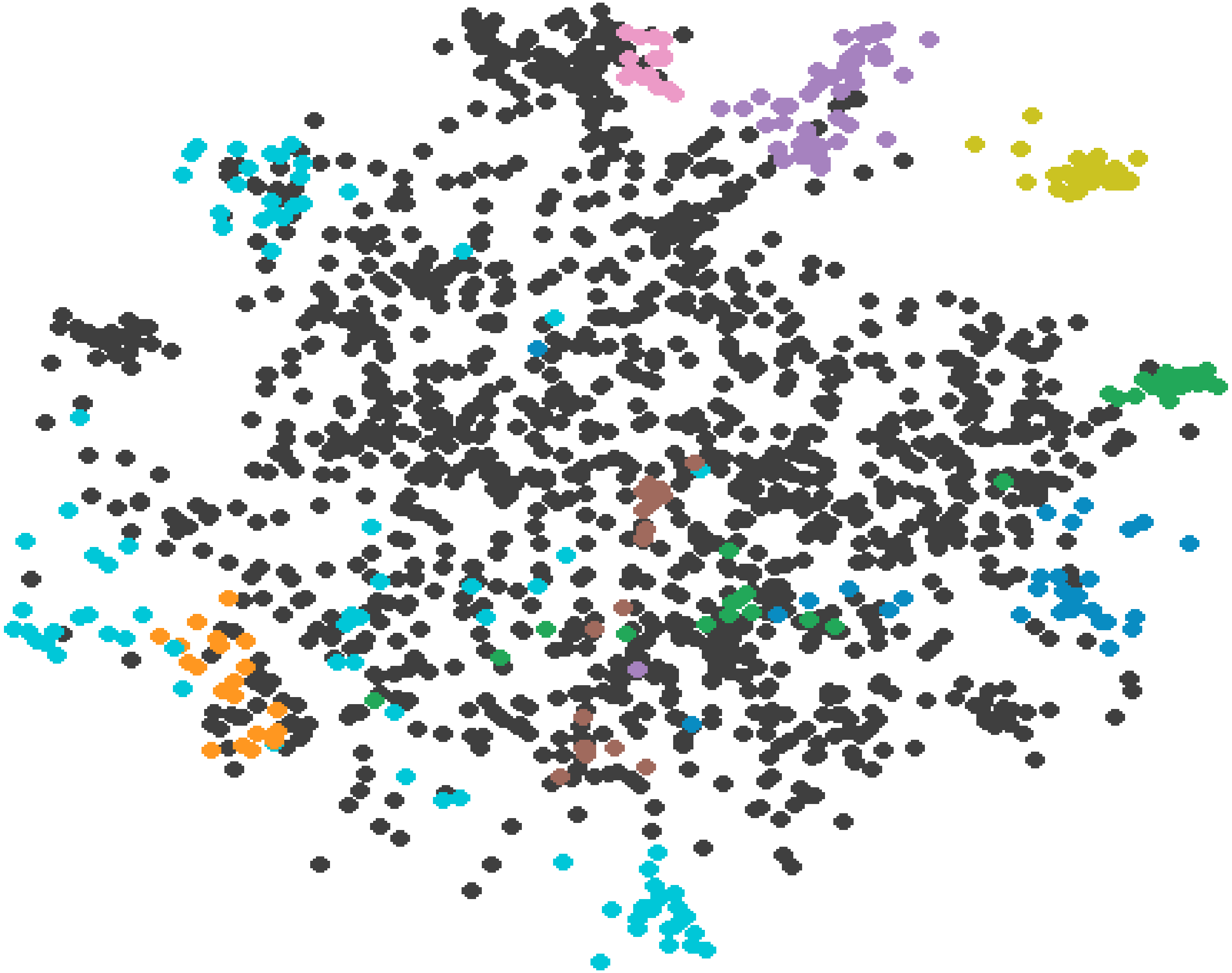
68,070,600

Cardinality Example – Alias Detection

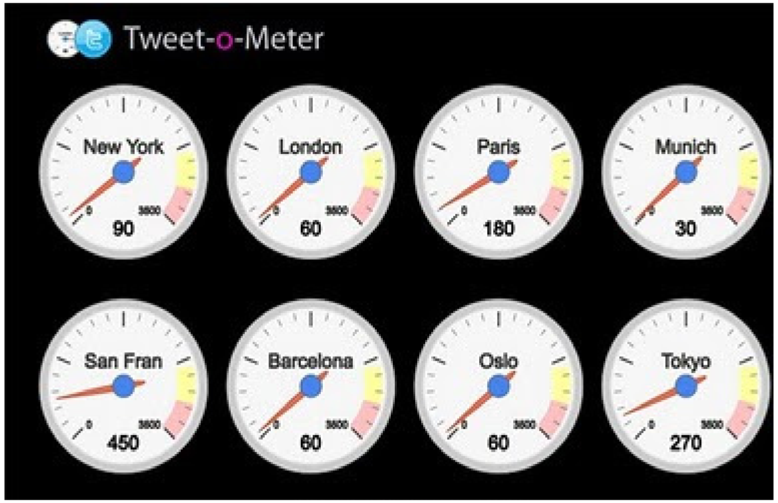
	Arthur Bishop	Raul Sanchez	Joel Rifkin	Jose Haddock	William Bonin
Arthur Bishop					
Raul Sanchez	.0366				
Joel Rifkin	-.0464	.0616			
Jose Haddock	.0366	.9675	.0616		
William Bonin	.1526	.0125	.0016	.0125	

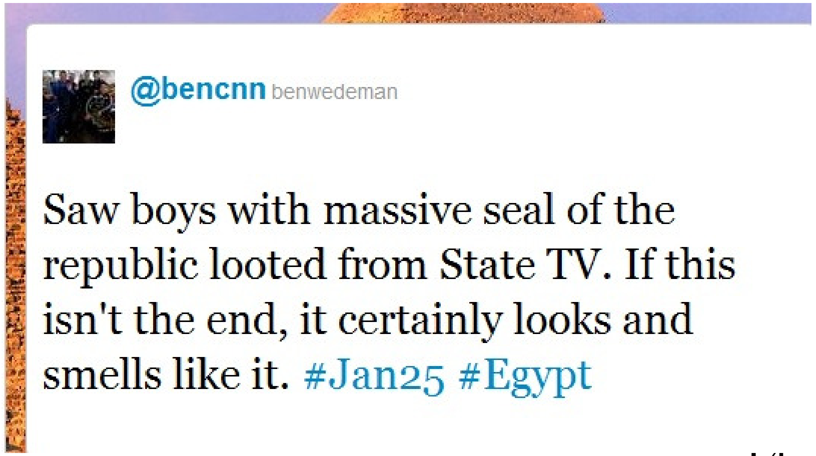
Challenge: Many by Many Comparisons- Processing 10 Million Names Requires 50 Trillion Comparisons

Non-locality Example– Clustering Documents



Twitter Example





@bencnn benwedeman

Saw boys with massive seal of the republic looted from State TV. If this isn't the end, it certainly looks and smells like it. #Jan25 #Egypt

The Tweet Analysis Problem

- Volume – 500 Million Tweets per Day Worldwide

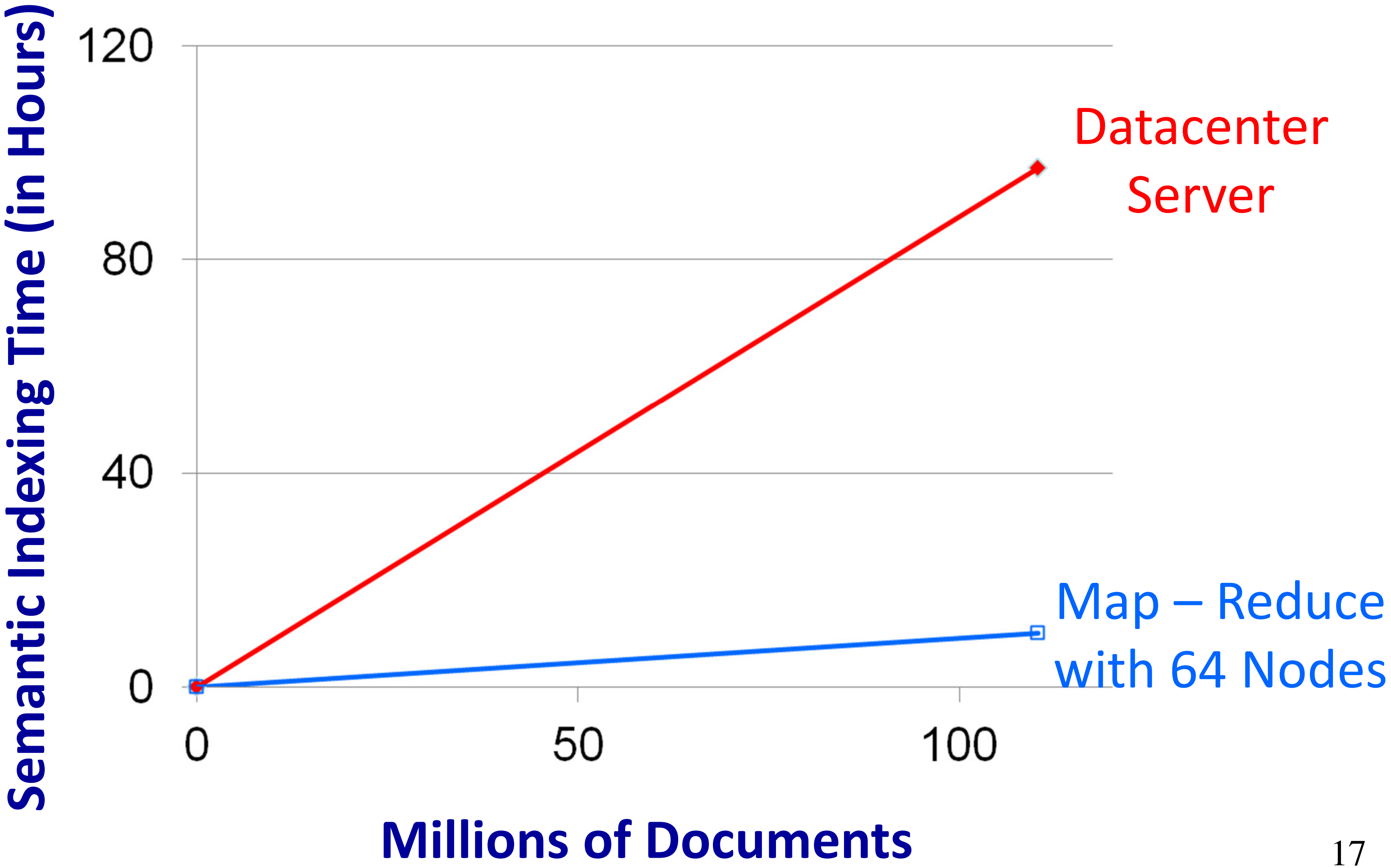
- Challenges:
  - ➢ Very Low Signal to Noise Ratio (31 Million People Follow Lady Gaga)
  - ➢ Implicit Context (“*Let’s all Meet at Bob’s House*”)
  - ➢ Incomplete, Conflicting, and Erroneous Information
  - ➢ Deliberate Deception (>50% of all Tweets are Machine-generated)

Applicable Analytic Techniques

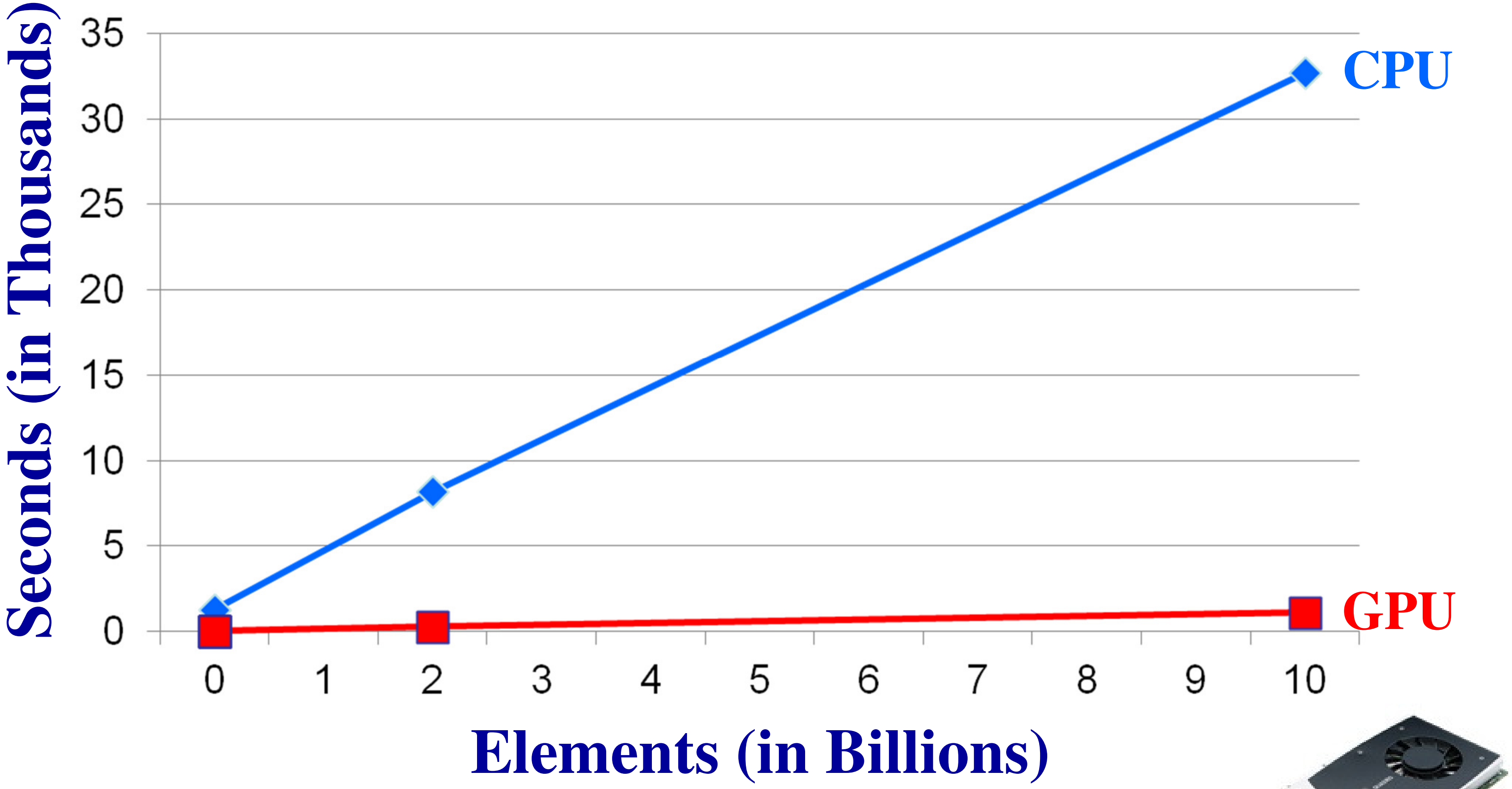
- Statistical Analysis
- Categorization
- Clustering
- NLP Techniques
- Semantic Analysis

In General, Application of such Techniques to Big Data Problems is Computationally Intensive

Cloud Enabling



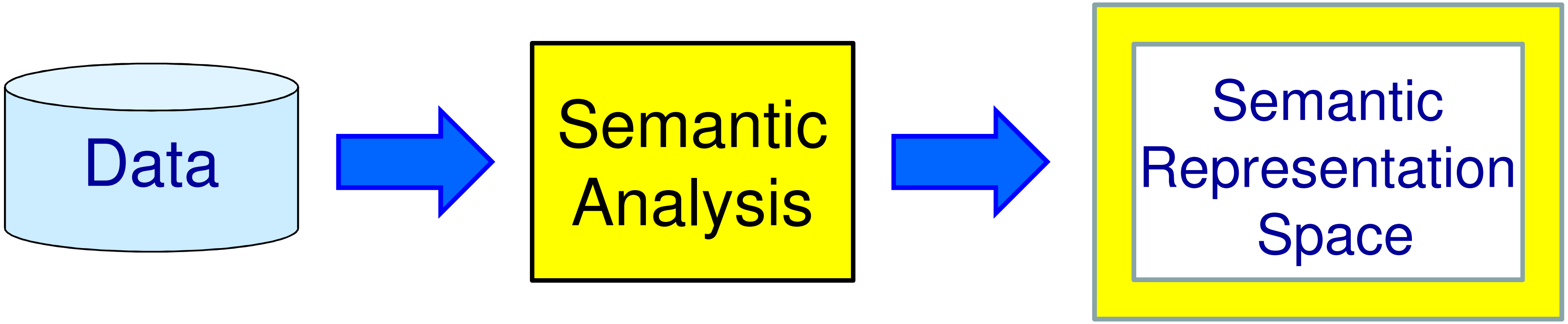
GPU Enabling



kNN Calculation

CPU: Intel Xeon X5660
GPU: Nvidia Quadro 2000

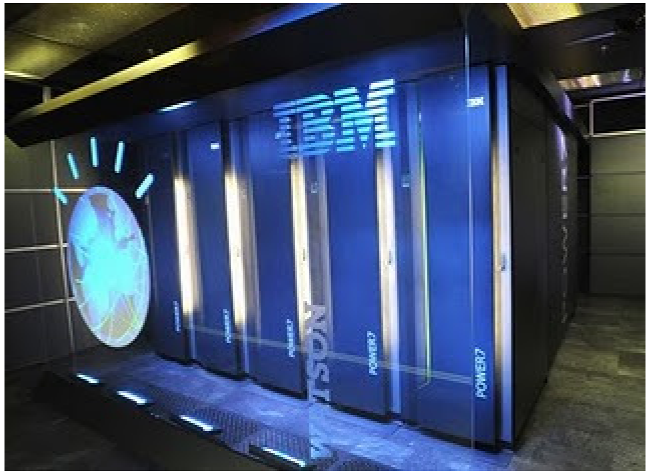
Semantic Enabling



- Accommodates Nth-order Relationships
- Automatically Coalesces Term Variants
- Supports Automated Entity Disambiguation
- Identifies Subtle Relationships
- Can Combine Structured and Unstructured Data

But Not as Well Understood as Structured Data Analysis Techniques

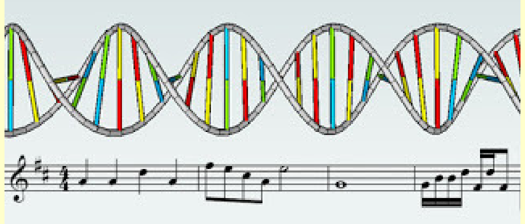
IBM WATSON Winning “Jeopardy”



- Volume: “Only” 1TB of Data (Mostly Text)
- Velocity: Meeting the 3-second Response Requirement of *Jeopardy* Required 80 Teraflops of Processing Power

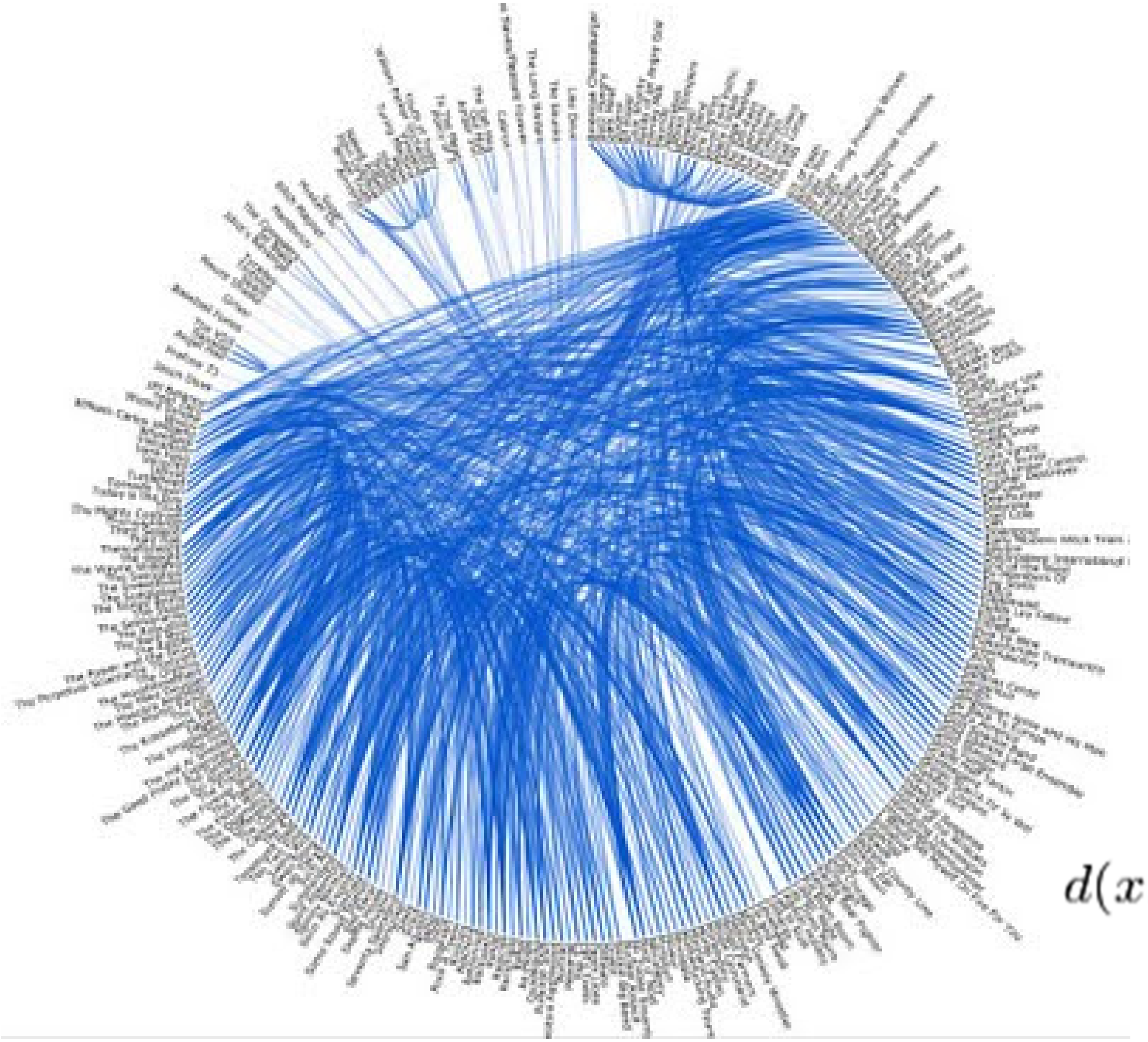
Challenge:

•Question Decomposition



Music Genome

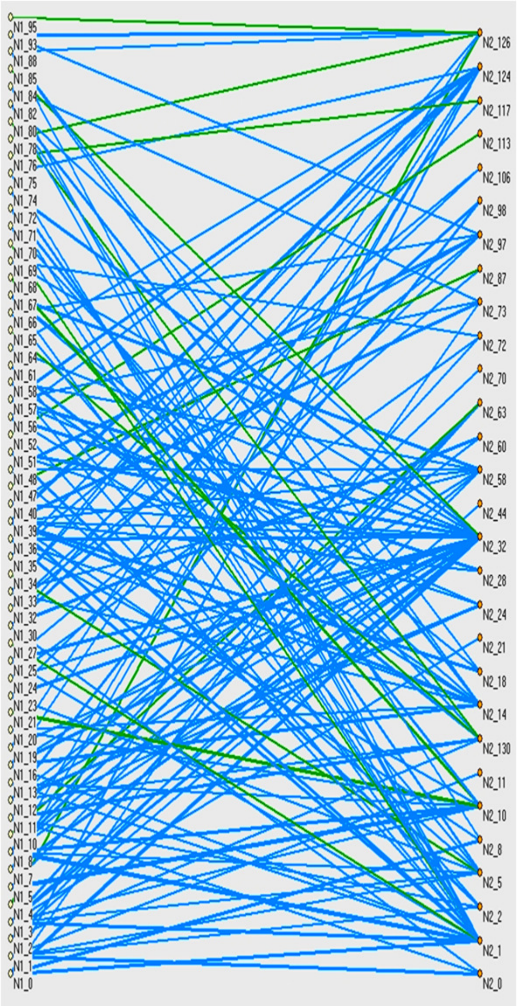
Objective: Match Liked Songs to Recommended Ones



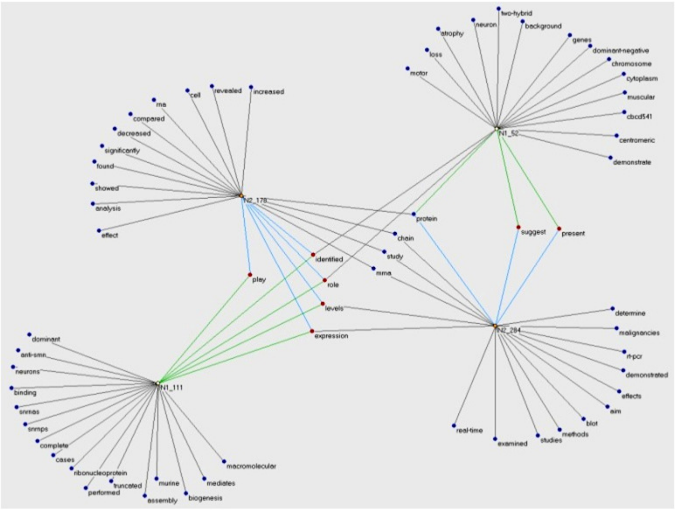
- ≈ 400 Attributes per Song
- 10 Million Songs
- Each Song Represented by a Vector of Elements
- 140 Trillion Elements
- Distance Function is Calculated between All Songs

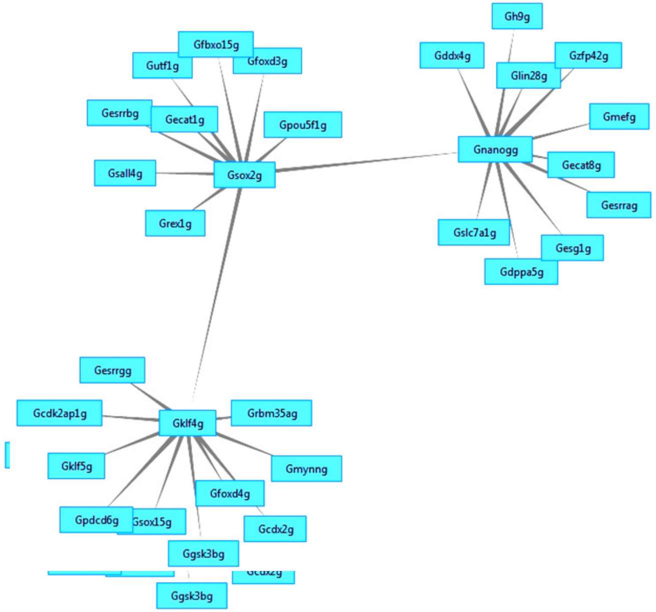
$$d(x,y) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{p_n(x-y)}{1+p_n(x-y)}$$

Literature-based Discovery



- PubMed Abstracts
- Gene – Function Relationships Derived Semantically
- 98,074,359 Potential Gene-function Associations.

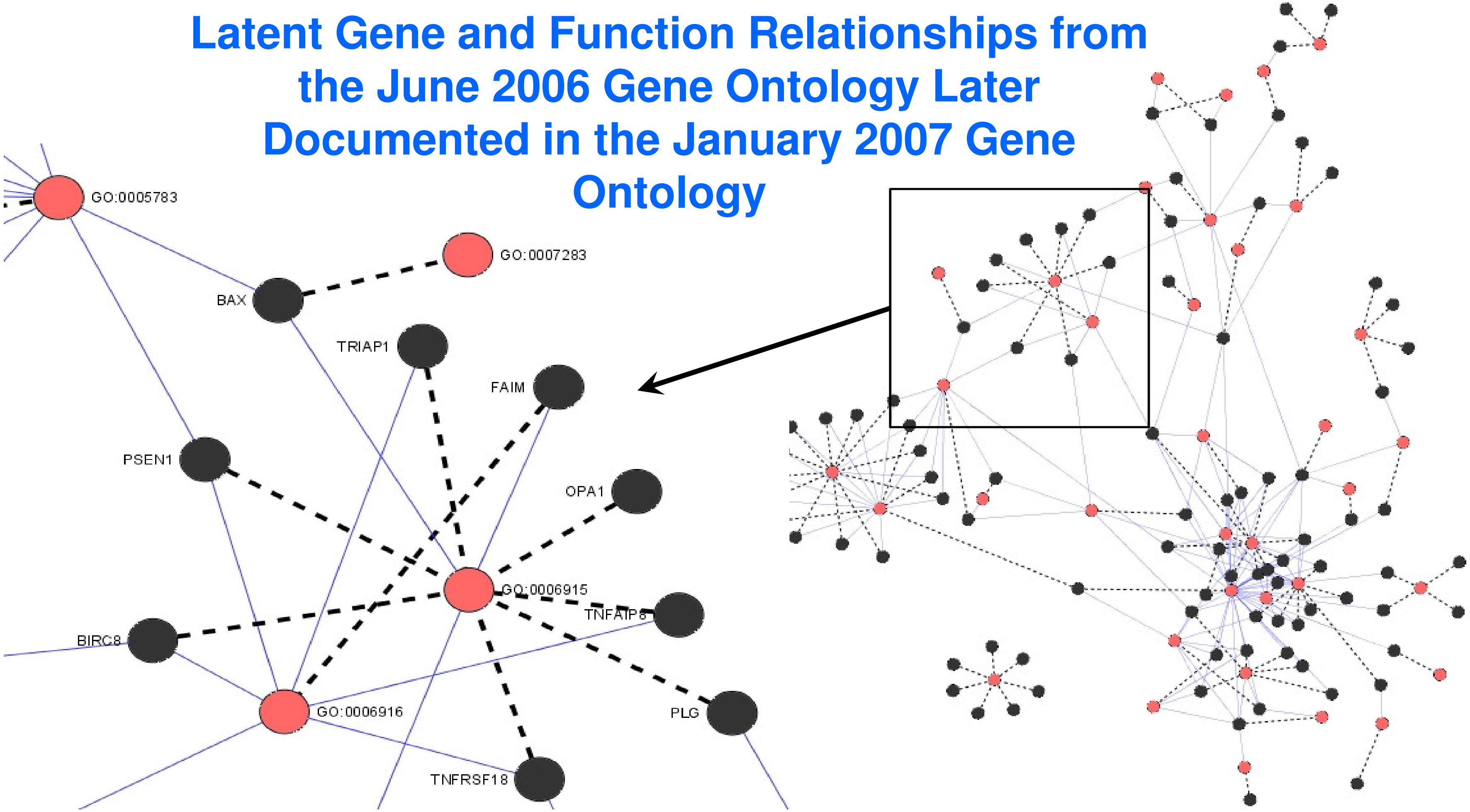




Zukas, A., GO-Driven Literature-Based Discovery using Semantic Analysis, MS Thesis, **George Mason University**, 2007.

Literature-based Discovery (Cont'd)

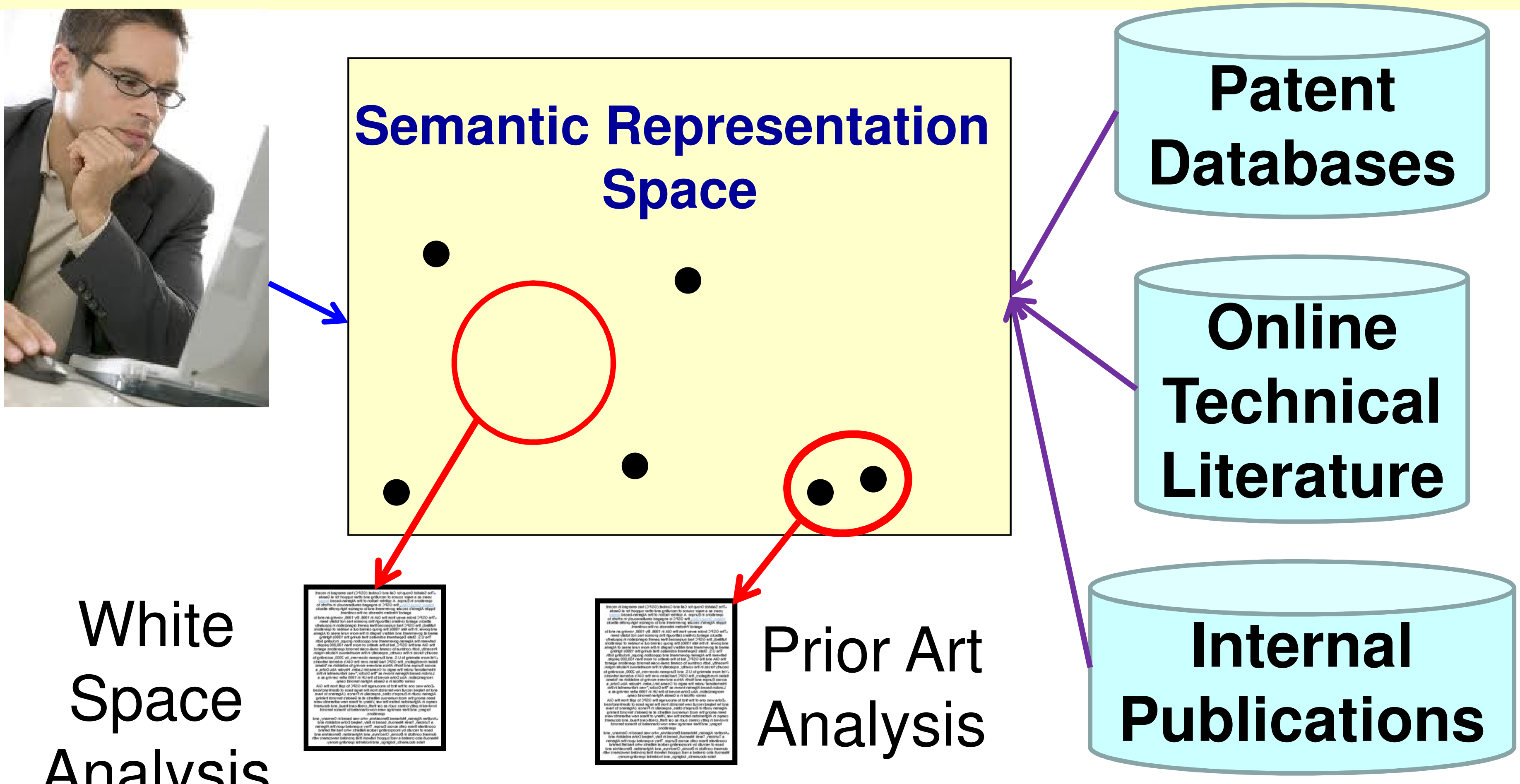
Latent Gene and Function Relationships from the June 2006 Gene Ontology Later Documented in the January 2007 Gene Ontology



Challenges:

- Nth-order Relationships
- Complexity of Relations

Patent Analysis



Semantic Representation Space

Patent Databases

Online Technical Literature

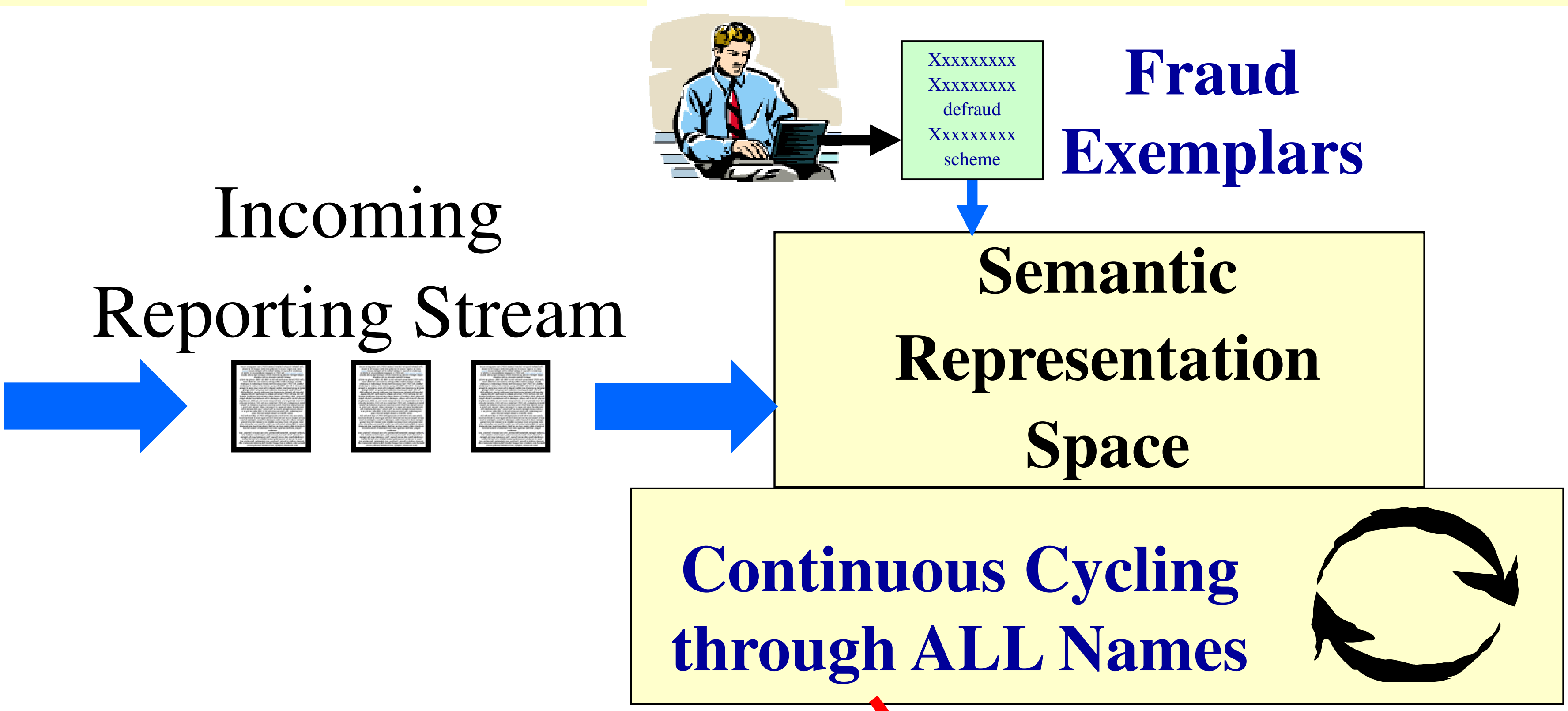
Internal Publications

White Space Analysis

Prior Art Analysis

Challenges:
- Need for Conceptual Comparisons
- Technical Terminology / Obfuscation
- Convoluted Structure (Claims)

Concept-driven Discovery



Xxxxxxxxx
Xxxxxxxxx
defraud
Xxxxxxxxx
scheme

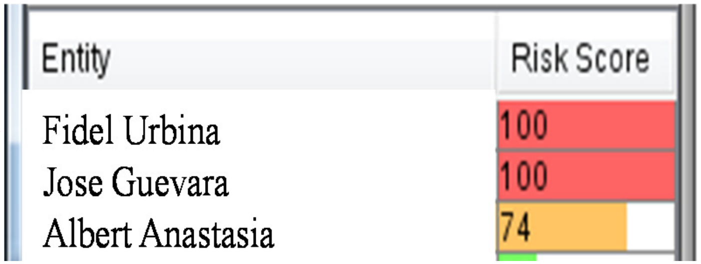
Fraud Exemplars

Incoming Reporting Stream

Semantic Representation Space

Continuous Cycling through ALL Names

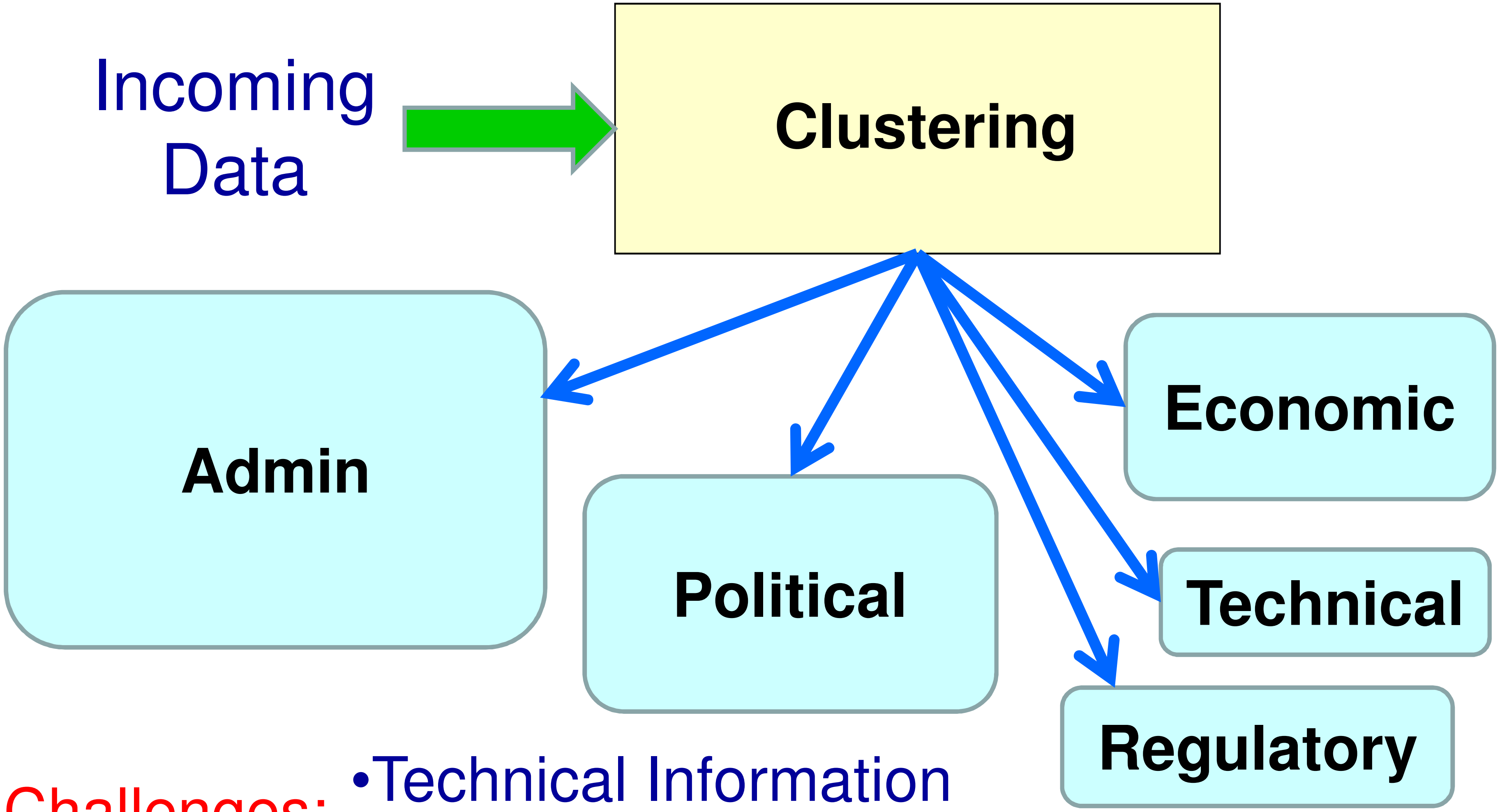
Generate Alerts



Entity	Risk Score
Fidel Urbina	100
Jose Guevara	100
Albert Anastasia	74

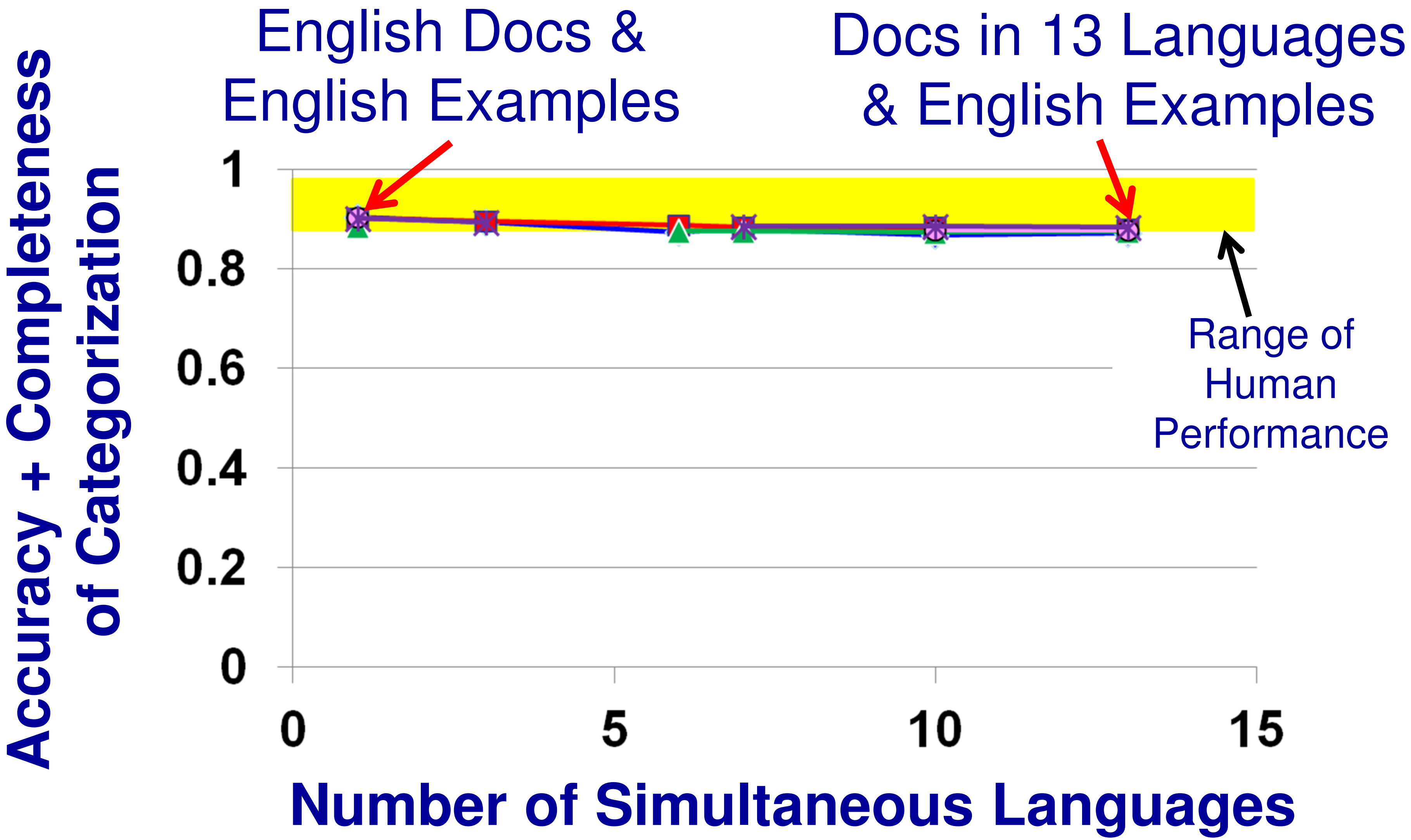
Issue: Name Disambiguation

Rapid Data Overview



Challenges:
- Technical Information
- Multilingual Data

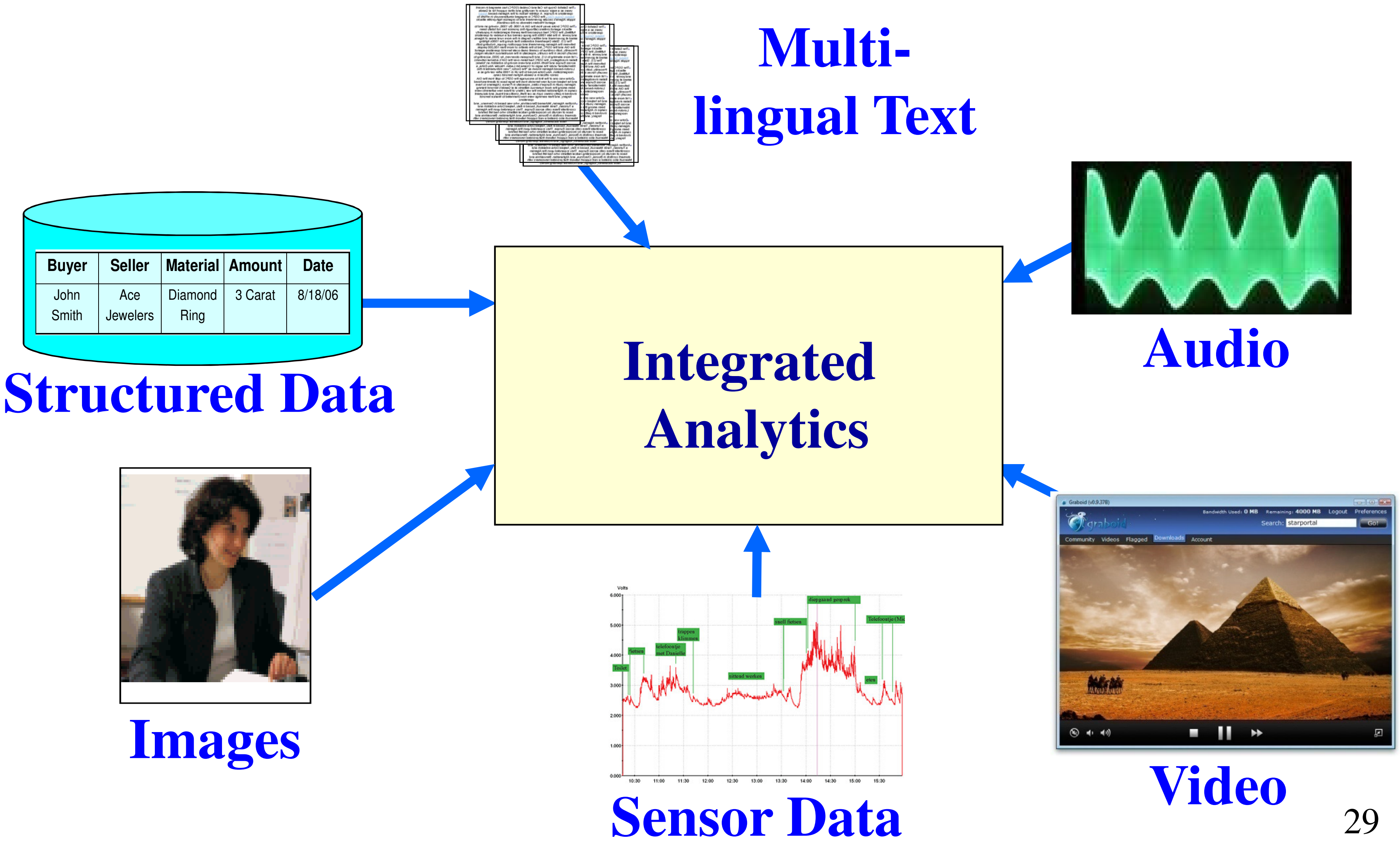
Crosslingual Document Categorization – Big Data Solution



Where is Big Data Analytics Going?

- Real-time Analysis
- Multimedia Collections
  - ➢ Text
  - ➢ Structured Data
  - ➢ Audio
  - ➢ Video
  - ➢ Sensor Data
- Temporal and Spatial Data Integration
- Interactive Visualization
- Continuous Retrospective Analysis
- Advanced Analytics (Especially Semantic Analysis)

Integration of Multimedia Data



Integrated Analytics

Multi-lingual Text

Buyer	Seller	Material	Amount	Date
John Smith	Ace Jewelers	Diamond Ring	3 Carat	8/18/06

Structured Data

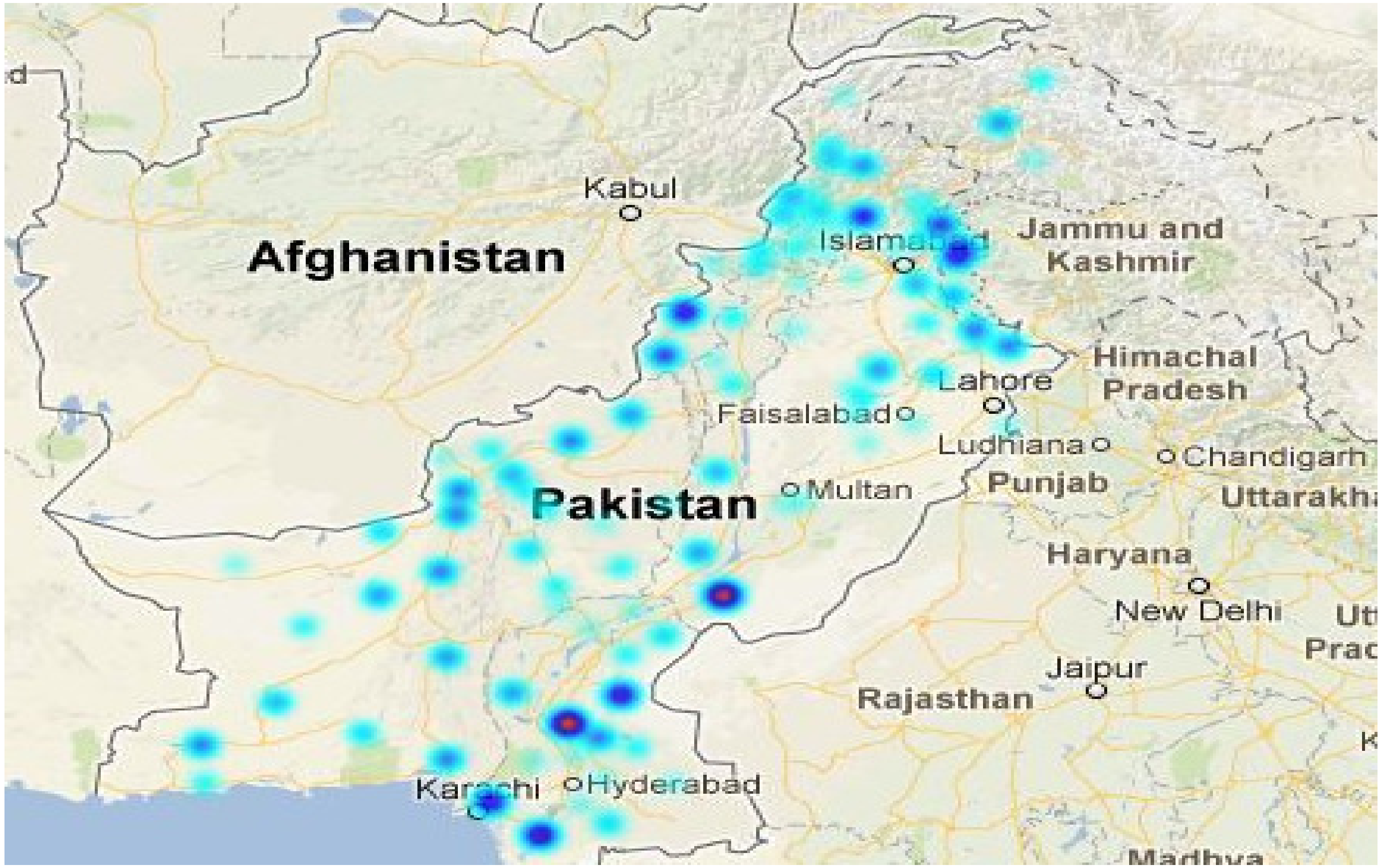
Audio

Images

Sensor Data

Video

Spatiotemporal Data Integration



Challenges:

- •Fully Automatic Integration of Spatial, Temporal, and Semantic Information
- •Location Disambiguation

Questions or Comments

Roger Bradford
Agilex Technologies Inc
1-703-889-3916
r.bradford@agilex.com