Methods for Enabling Next Generation Knowledge Discovery

CRAY
THE SUPERCOMPUTER COMPANY

Shoaib Mufti

Director Knowledge Management

Outline

- Big data problem
- Semantic applications and databases to the rescue
- Issues with semantic applications and databases
- Some solutions

We have a BIG Data Problem!!

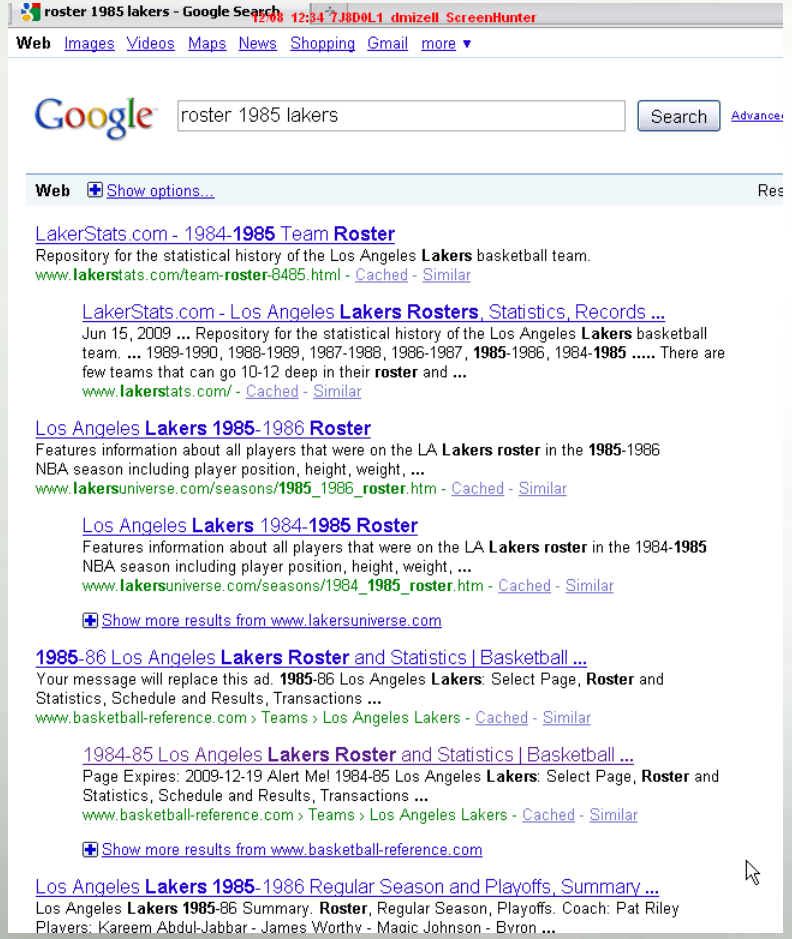
- **We are drowning in a sea of data!**
- **Web gave us the means to connect and produce data frequently and informally**
  - Web pages, emails, tweets, blogs, YouTube etc.
  - More than 2 billion new Web pages have been created since 1995, with an additional 200 million new pages being added every month (IDC)
- **Mobile is making data production 24x7**
- **Advances in computing is allowing production of massive amount of scientific data**
  - Life sciences, climate, space etc.
- **Homeland security data is becoming massive with the global war on terror**
- **Situation will get worse with the growth of population, access of web and technology by more people, and increased use of mobile technology**
- **Finding useful information and gathering knowledge from this massive amount of seemingly unrelated data will be the next big challenge for the computing world**

Most of the Data is Unstructured

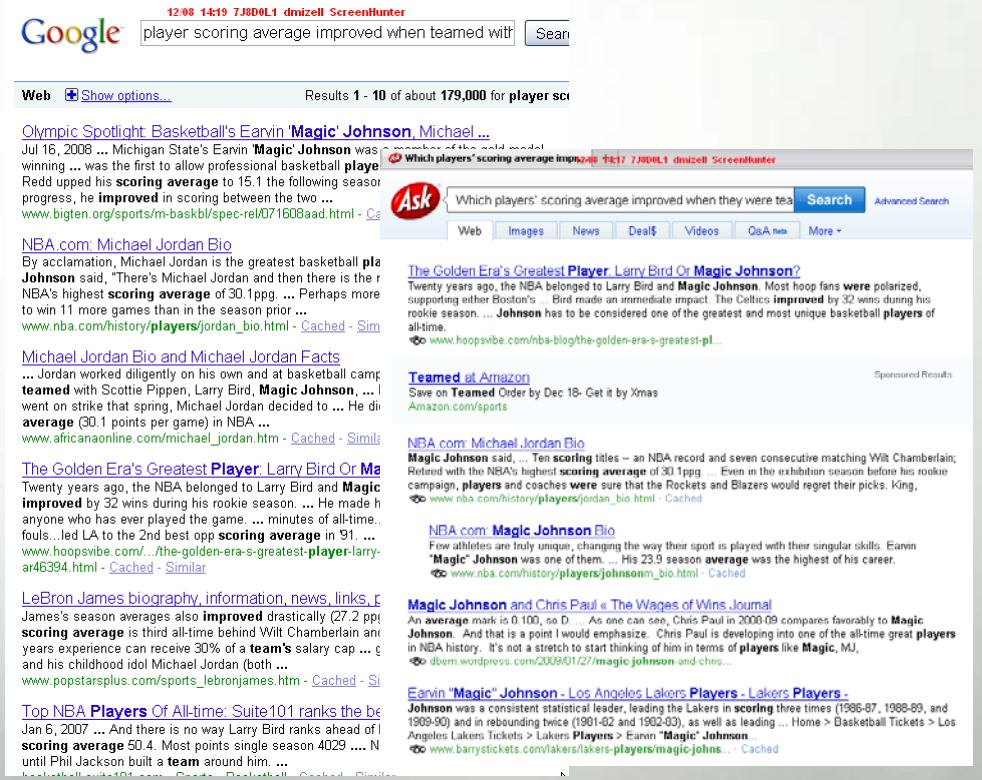
- **Most of the data is not in the databases and is unstructured**
  - Majority of data is residing in in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, letters, white papers, marketing material, research, presentations , Web pages, and simulation results.
  - 85 % of all business information exists as unstructured data (Merrill Lynch)
  - White-collar workers will spend from 30% to 40% of their time this year managing documents (Gartner)
- **Unstructured data analysis is a major unsolved problem**
  - **Tools and techniques** (e.g. RDBMS) successful in transforming structured data into knowledge simply don't work with unstructured data.
- **Keyword search solutions do exist but they fail with complex queries. Also, there are data representation and ontology issues when integrating data between different organizations and companies.**
- **Point and proprietary solutions are not based on standards and run into issues with the integration of multiple databases in disparate geographies and organizations.**
- **Current unstructured data solutions tend to create larger and larger databases as the available data increases. These larger data sets make the problem worse and make analysis slower.**

Complex Search Based on Keywords has Issues.....

- Who played for the 1985 Los Angeles Lakers?



- Which players' scoring averages improved when they were teamed with Magic Johnson?

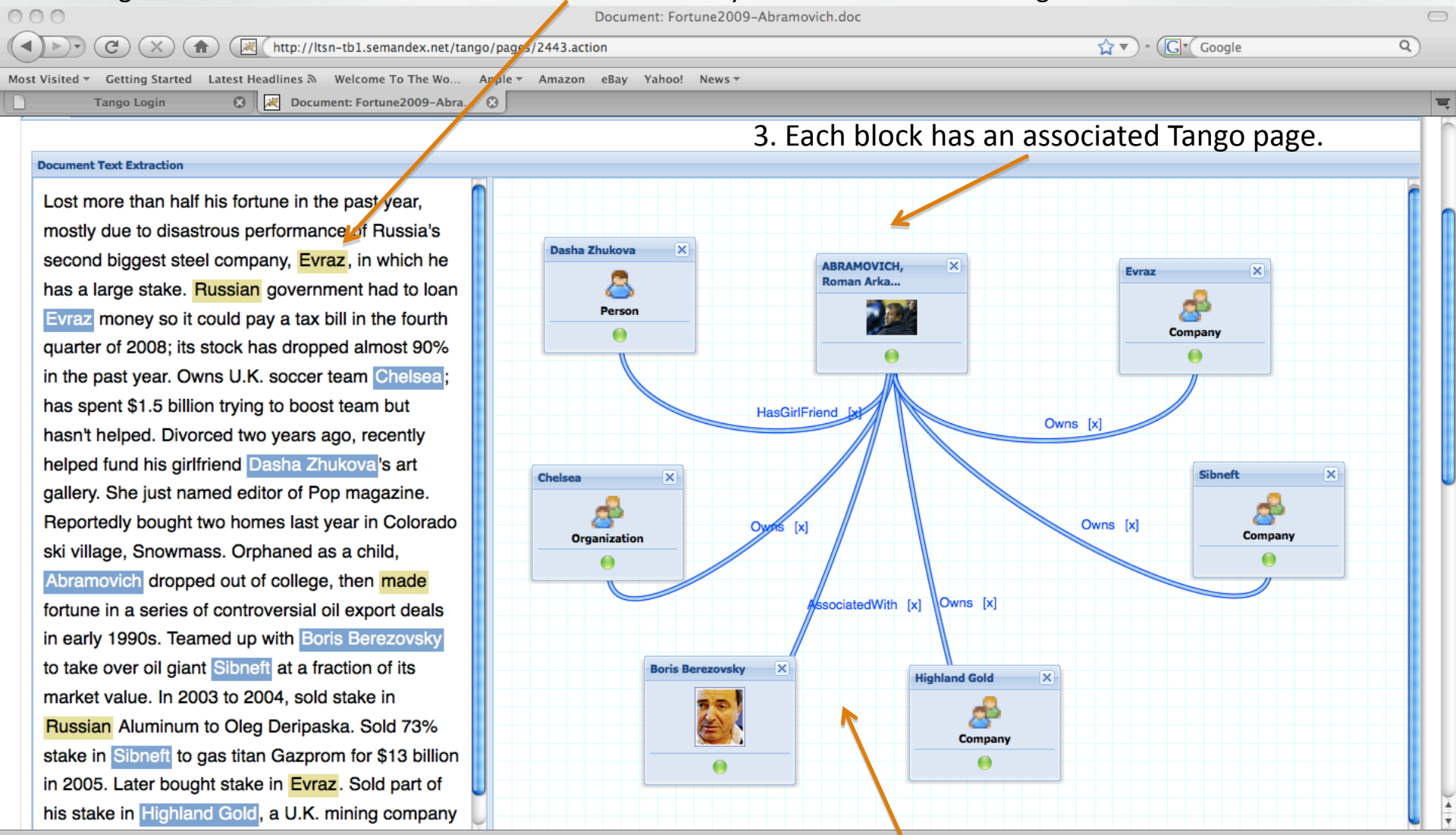


How do we solve this problem?

- **Semantic web applications and semantic databases are key technologies to facilitate unstructured data analysis problem.**
  - Methods to add structure/semantics to unstructured data are application-specific, but semantic network applications and semantic databases are key support technologies, which allow us to represent the results in useful ways.

An Example Application (Tango) that Creates Semantic Graphs from Documents and Other Sources

1. Tango extracts the text from the source and automatically annotates entities it recognizes



3. Each block has an associated Tango page.

2. Tango allows the user to rapidly build a structured representation (semantic graph) of the people and companies in the document

Semantic Web Databases

- **Motivation:**
  - Add “meaning” to web data
  - Support complex queries
  - Make it easier to merge multiple databases into one
- **The (W3C standard) SPARQL query language:**

Query:

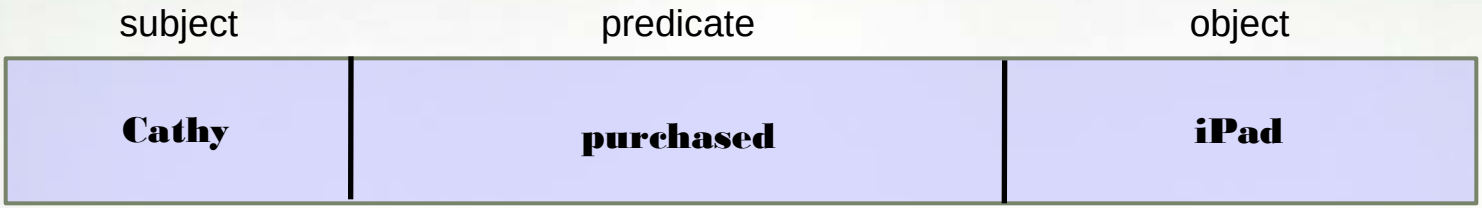
```
SELECT ?title
WHERE {    ?title    <http://www.recshop/example/cd#artist>  “Bob Dylan”    }
```

Results:

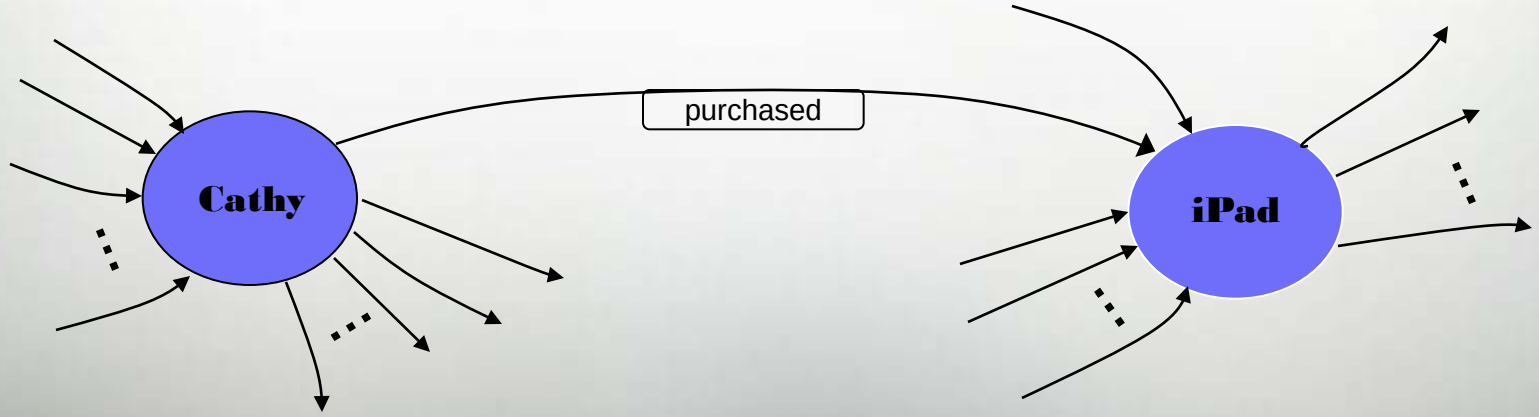
```
title
=========================================
http://www.recshop.example/cd/Empire Burlesque
http://www.recshop.example/cd/Blonde On Blonde
http://www.recshop.example/cd/Blood on the Tracks
http://www.recshop.example/cd/Nashville Skyline
http://www.recshop.example/cd/John Wesley Harding
http://www.recshop.example/cd/Highway 61 Revisited
…
```

Semantic databases look like graphs…

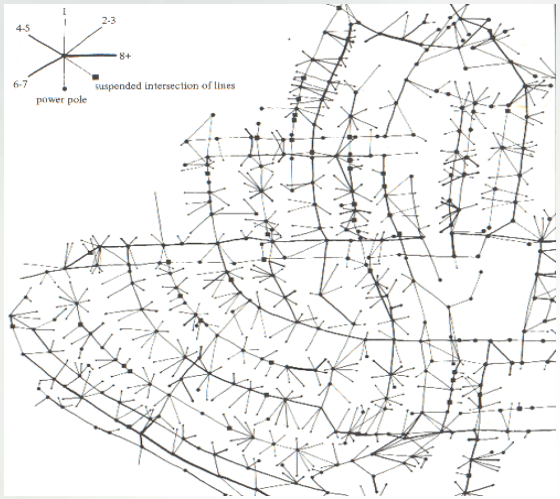
- Semantic network databases!



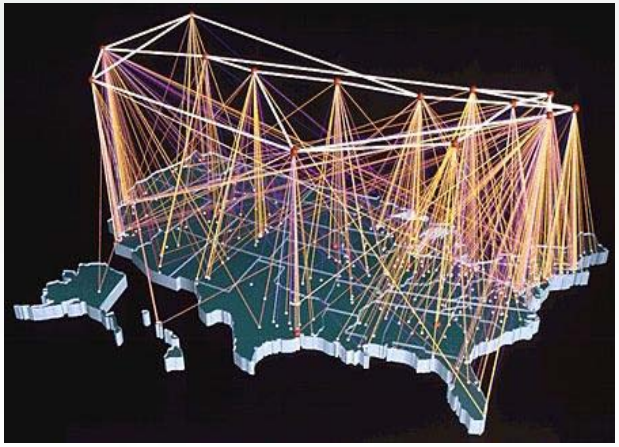
- Each RDF triple is a pair of vertices and an edge in a directed graph:



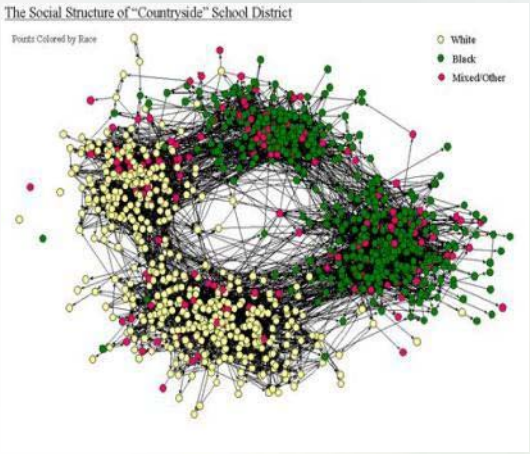
Power Distribution Networks



Internet backbone

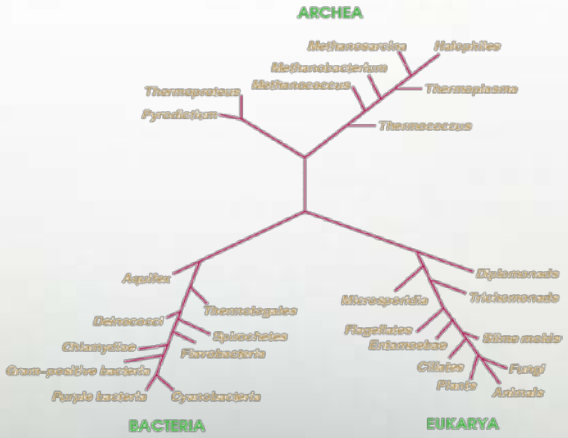


Social Networks

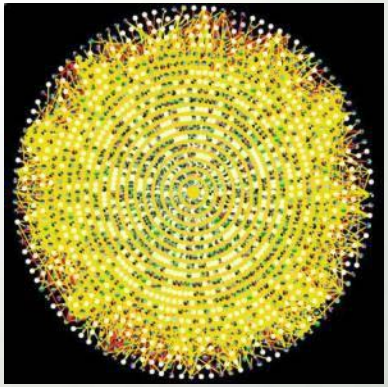


Graphs are everywhere!!!

Ground Transportation



Tree of Life



Protein-interaction networks

Relational Databases vs Semantic Network Databases

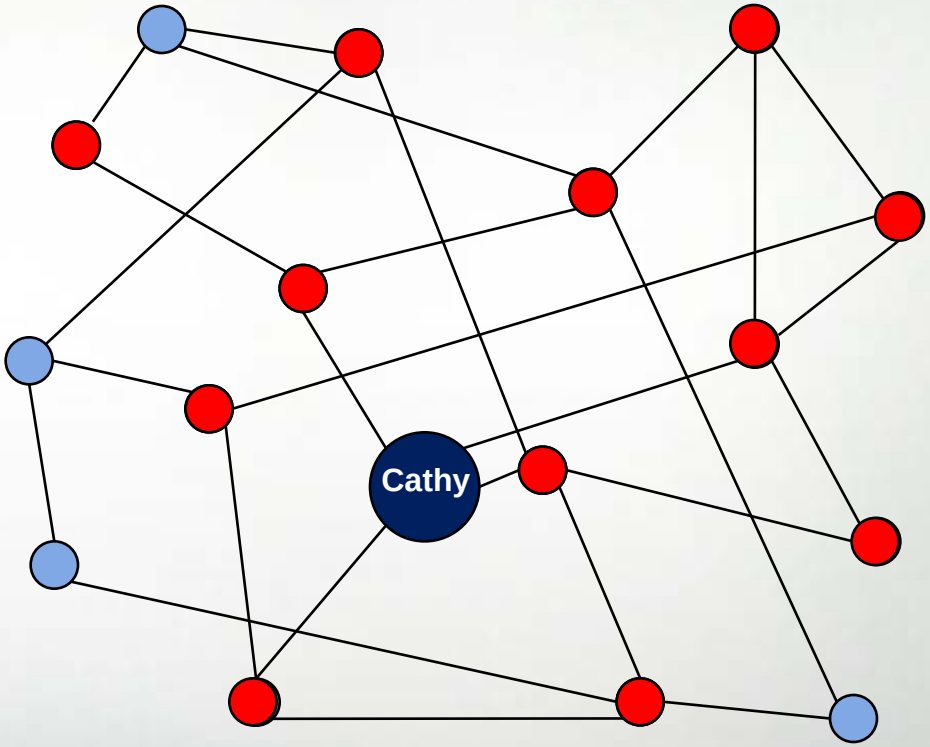
This type of query is easiest for RD:

"Show all company employees who are age 45 or older"

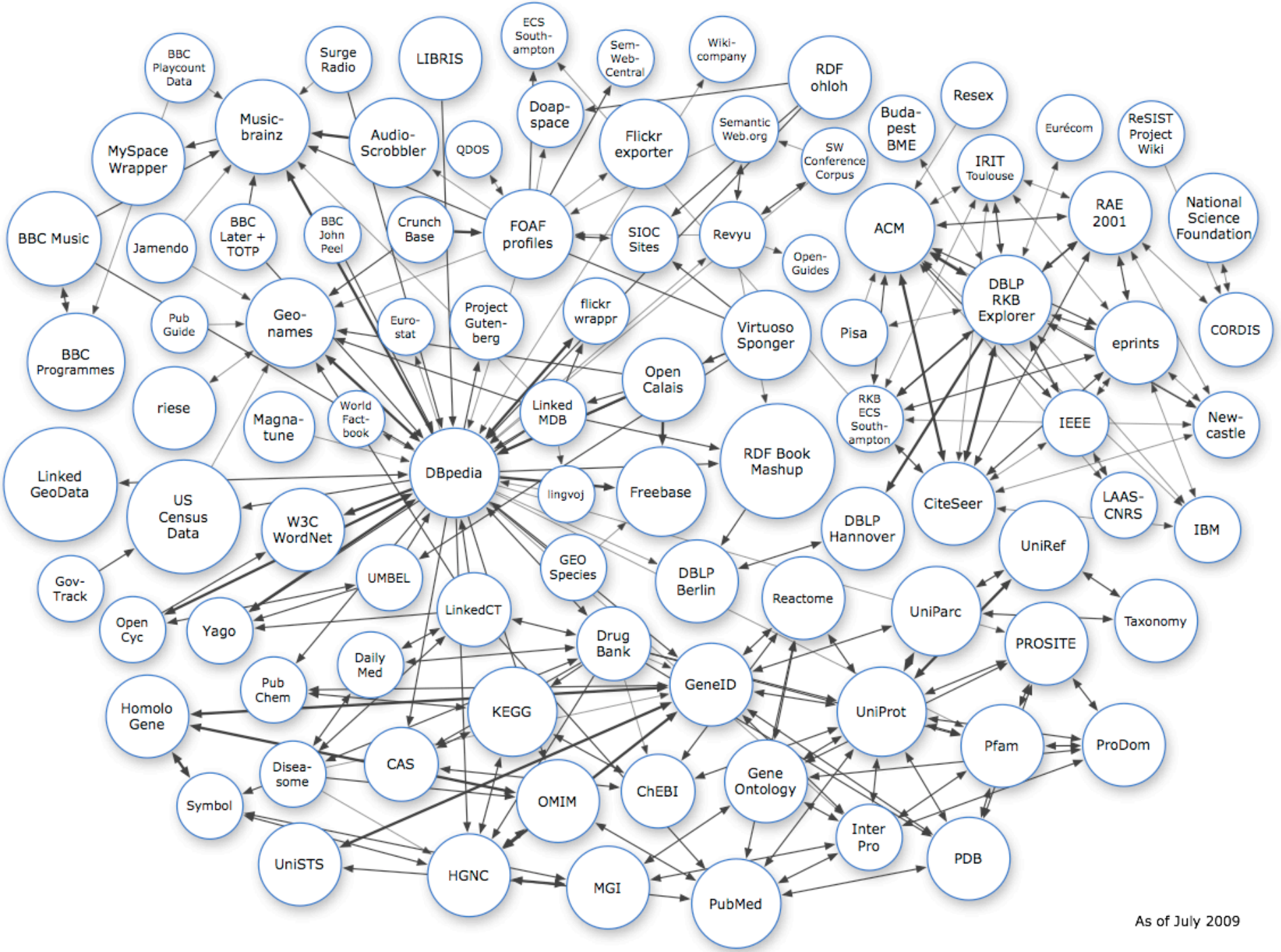
Smith	27
Jones	36
Johnson	29
Wilson	51
Peterson	48
Ordonez	34
Quigley	61
Roberts	53

This type of query is easiest for SND:

"Show everyone who has talked with Cathy or have talked with someone who talked with Cathy"



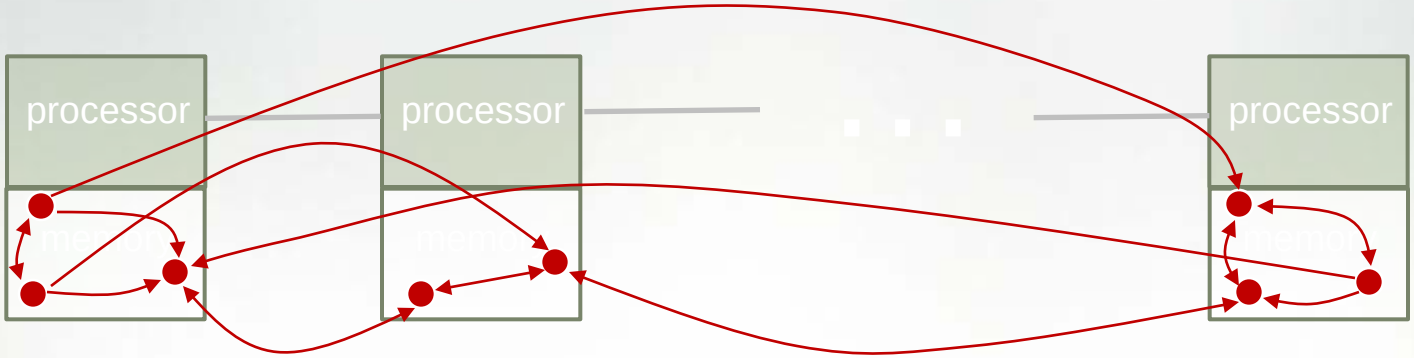
Semantic networks also support *reasoning*:
X attended meeting M &&
Y attended meeting M → X met with Y


BBC Playcount Data
Surge Radio
LIBRIS
ECS South-ampton
Sem-Web-Central
Wiki-company
RDF ohloh
Resex
Buda-pest BME
Eurécom
ReSIST Project Wiki
Music-brainz
Audio-Scrobbler
QDOS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
IRIT Toulouse
MySpace Wrapper
BBC Music
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
National Science Foundation
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
CORDIS
BBC Programmes
riese
Magna-tune
World Fact-book
Linked MDB
Open Calais
RKB ECS South-ampton
IEEE
eprints
New-castle
Linked GeoData
US Census Data
W3C WordNet
DBpedia
lingvoj
Freebase
RDF Book Mashup
CiteSeer
LAAS-CNRS
IBM
Gov-Track
UMBEL
GEO Species
DBLP Berlin
DBLP Hannover
UniRef
Open Cyc
Yago
LinkedCT
Drug Bank
Reactome
UniParc
Taxonomy
Daily Med
PROSITE
Pub Chem
Homolo Gene
KEGG
GeneID
UniProt
Pfam
ProDom
Disea-some
CAS
Symbol
OMIM
ChEBI
Gene Ontology
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB
As of July 2009

Issues with semantic databases and networks

- **Semantic databases are key technologies for next generation knowledge discovery but need to overcome several issues before wide adoption**
- **Ontologies (or relationship graphs) for a large number of entities do not exist today**
  - Situation is improving as more verticals are embracing semantic technologies
  - Tools and techniques to create ontologies are also improving
- **Scaling of semantic graphs and databases is a major issue**
- **Need to partition large graphs to run on distributed conventional clusters**
- **Partitioning large graph databases is a resource intensive task and consumes months of developers time**
- **Partitioning techniques fail (unacceptable performance) when doing complex sparse graph analysis**
  - Finding needle in haystack problems
  - Connecting the dots in massive amount of information

Problems with Conventional Parallelization Approaches



- **Large delays when one processor needs data from another's memory**
  - In graph analytics/semantic networks, this happens *almost all the time*
  - No way to "partition" the data so that references are usually local
- **Limited scaling – adding more boxes doesn't improve performance**
- **Architecture forces limitations on types of queries – no reasoning, spatial queries, approximate queries, etc. Graph-oriented queries about connectivity or relative closeness impossible or very slow – minutes to hours**

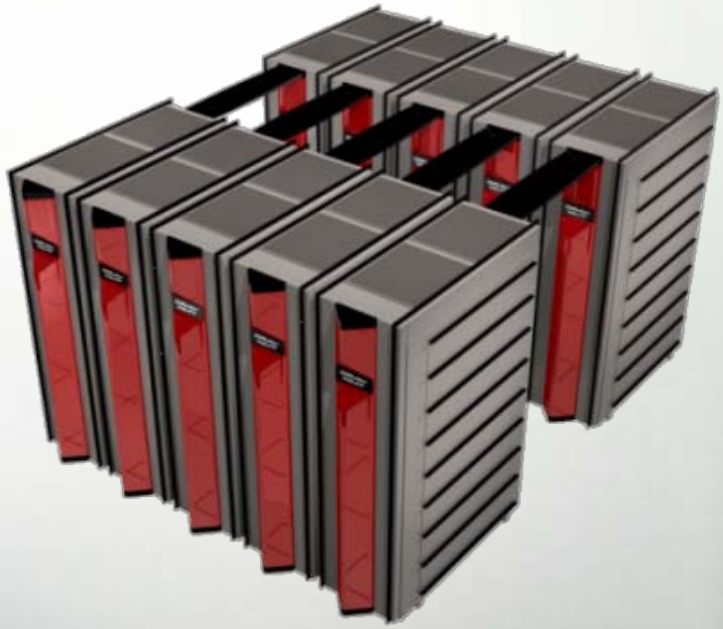
A Solution to Conventional Parallelization Issues

❖ Cray XMT

- An architecture like Cray XMT allows to get around some of the issues encountered by conventional systems for irregular data analysis problems

❖ Characteristics

- Very large shared memory
  - Architecture can support 128TB shared memory
- Multithreading
  - Architecture can support 8000 processors
  - 128 hardware threads per processor
  - Practically unlimited virtual threads
  - Hide memory latency
- Word level (fine grain) synchronization
  - Important for search problems
- Ease of parallel programming

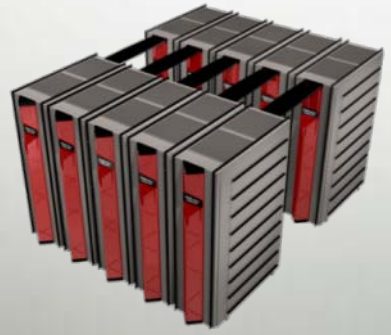


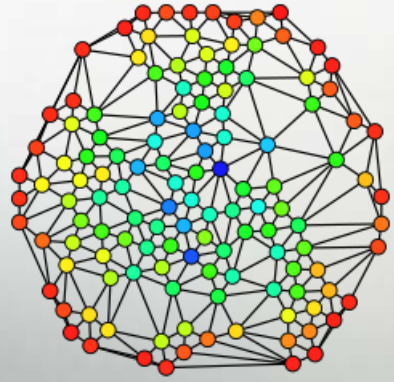
Advantages of the Shared Memory MT Architecture

- System design enables all memories to be shared by all processors; remote fetches of data get much higher throughput than in conventional parallel systems
- Outstanding performance on graph problems, other queries – often 100x-1000x faster than conventional parallel systems
- Almost linear scaling on graph problems
- Large shared memory obviates much file I/O → much faster response time

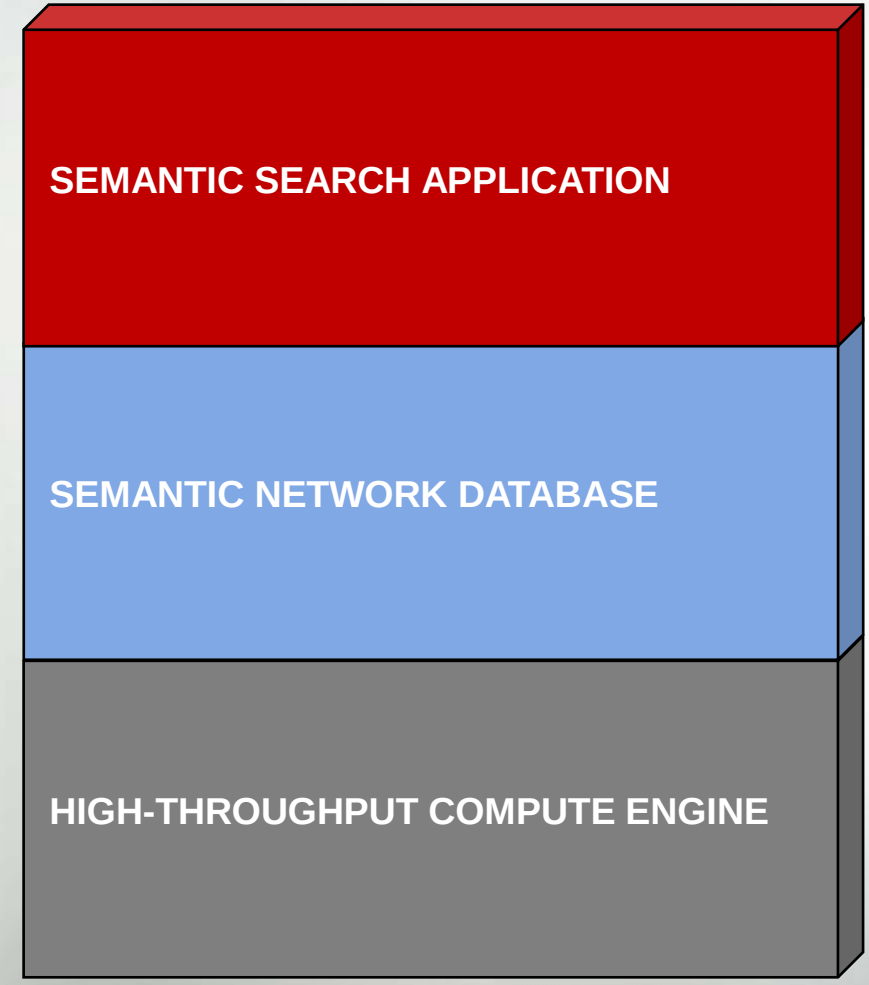
Performance on Graph Problems

- **Application Significance:**
  - ✓ Betweenness is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.
- **XMT vs. X86 Cluster:**
  - ✓ 64 processor Cray XMT vs. 64 processor X86 Cluster
- **XMT processed the graph in 10.6 seconds versus 65 minutes with the cluster**
  - ➢ ***Batch has become interactive***





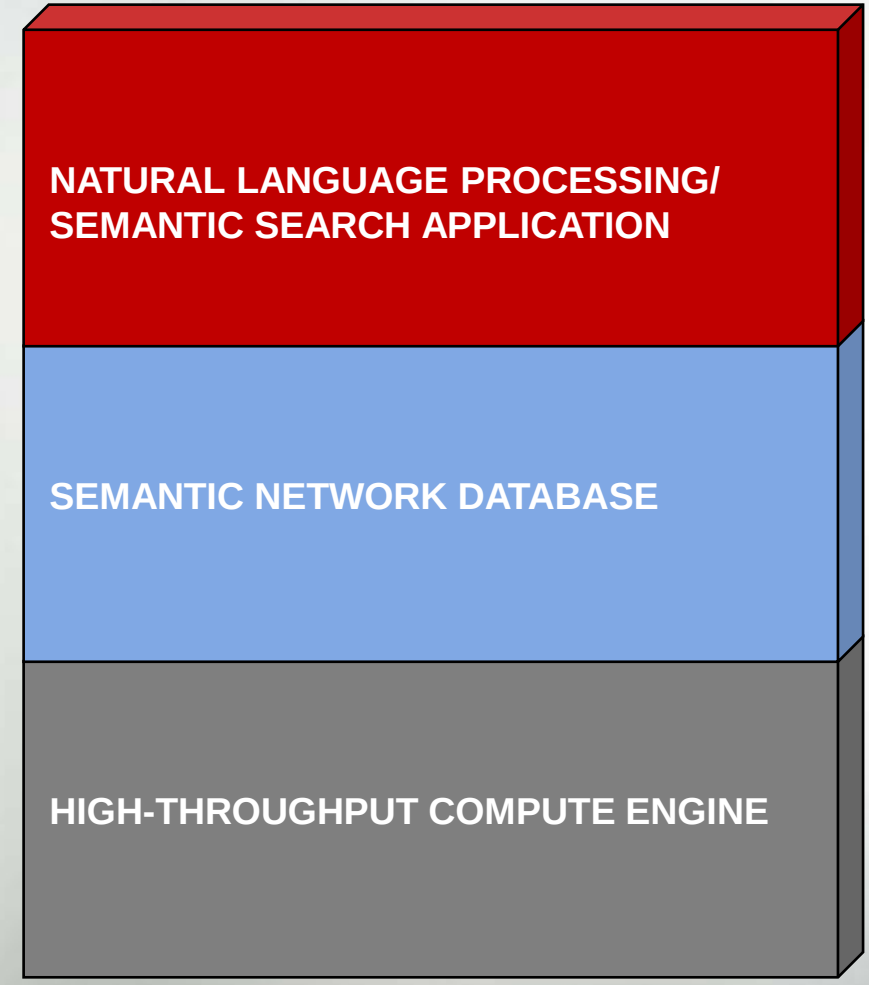
New Paradigm of Unstructured Data Analysis



- Reduction in the database sizes and more accurate search
  - Semantic application extracts relevant information by assigning meaning (semantics) to the data
- Integration of disparate data sources
  - Data structures are based on W3C standards
- Complex query support
  - Support temporal, spatial, reasoning etc.

- Address scaling issues of SNDB
  - SNDB scale poorly on conventional systems
  - Shared memory multithreading architectures with high memory throughput provide good scaling for SNDB.

Which players' scoring averages improved when they were teamed with Magic Johnson?



Which players' scoring averages…

SELECT ?player
WHERE (?player nba:team ?t…

Translate natural language query to SPARQL or equivalent to query the database

```
<rdf:RDF xmlns:csf="http://schemas.microsoft.com/bingservices/pm#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
   <rdf:Description rdf:about="urn:nba_stats">
      <nba:Player>
         <rdf:Bio>
            <nba:name>James Worthy</nba:name>
            <nba:team>Lakers</nba:team>…
```

Represent Internet data as RDF triples

Requirements:
- multi-terabyte shared memory
- high performance on irregular data structures & random reference patterns

Cray Semantic Technology Investigations

- **Semantic search query of interest to government: show pairs of subjects with at least *k* objects in common.**
  - **Preliminary results on multithreaded supercomputer shows a 100x speed-up on 1B triples over a comparable cluster.**
- **Planning to demonstrate "RDF graph closure" using multithreaded supercomputer**
  - Teaming with experts from Pacific Northwest National Lab and Sandia National Lab.

Acknowledgements

- David Bader -- Georgia Tech
- John Feo, Daniel Chavarria – PNNL
- Jon Berry, Bruce Hendrickson, Eric Goodman – Sandia National Labs
- David Mizell -- Cray


CRAY
THE SUPERCOMPUTER COMPANY