

Methods for Enabling Next Generation Knowledge Discovery

Shoaib Mufti

Director Knowledge Management

CRAY
THE SUPERCOMPUTER COMPANY

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?

roster 1985 lakers - Google Search 12:34 7J8D0L1 dmizell ScreenHunter

Web Images Videos Maps News Shopping Gmail more ▾

Google roster 1985 lakers Search Advance

Web Show options... Results

[LakerStats.com - 1984-1985 Team Roster](#)
Repository for the statistical history of the Los Angeles Lakers basketball team.
[www.lakerstats.com/team-roster-8485.html](#) - [Cached](#) - [Similar](#)

[LakerStats.com - Los Angeles Lakers Rosters, Statistics, Records ...](#)
Jun 15, 2009 ... Repository for the statistical history of the Los Angeles Lakers basketball team. ... 1989-1990, 1988-1989, 1987-1988, 1986-1987, 1985-1986, 1984-1985 There are few teams that can go 10-12 deep in their roster and ...
[www.lakerstats.com/](#) - [Cached](#) - [Similar](#)

[Los Angeles Lakers 1985-1986 Roster](#)
Features information about all players that were on the LA Lakers roster in the 1985-1986 NBA season including player position, height, weight, ...
[www.lakersuniverse.com/seasons/1985_1986_roster.htm](#) - [Cached](#) - [Similar](#)

[Los Angeles Lakers 1984-1985 Roster](#)
Features information about all players that were on the LA Lakers roster in the 1984-1985 NBA season including player position, height, weight, ...
[www.lakersuniverse.com/seasons/1984_1985_roster.htm](#) - [Cached](#) - [Similar](#)

Show more results from [www.lakersuniverse.com](#)

[1985-86 Los Angeles Lakers Roster and Statistics | Basketball ...](#)
Your message will replace this ad. 1985-86 Los Angeles Lakers: Select Page, Roster and Statistics, Schedule and Results, Transactions ...
[www.basketball-reference.com > Teams > Los Angeles Lakers](#) - [Cached](#) - [Similar](#)

[1984-85 Los Angeles Lakers Roster and Statistics | Basketball ...](#)
Page Expires: 2009-12-19 Alert Me! 1984-85 Los Angeles Lakers: Select Page, Roster and Statistics, Schedule and Results, Transactions ...
[www.basketball-reference.com > Teams > Los Angeles Lakers](#) - [Cached](#) - [Similar](#)

Show more results from [www.basketball-reference.com](#)

[Los Angeles Lakers 1985-1986 Regular Season and Playoffs, Summary ...](#)
Los Angeles Lakers 1985-86 Summary. Roster, Regular Season, Playoffs. Coach: Pat Riley
Players: Kareem Abdul-Jabbar - James Worth - Magic Johnson - Byron ...

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

12:08 14:19 7J8D0L1 dmizell ScreenHunter

Google player scoring average improved when teamed with Search

Web Show options... Results 1 - 10 of about 179,000 for player sc

[Olympic Spotlight: Basketball's Earvin 'Magic' Johnson, Michael ...](#)
Jul 16, 2008 ... Michigan State's Earvin 'Magic' Johnson was a ...
[Which players' scoring average improved when they were tea...](#) 12:17 7J8D0L1 dmizell ScreenHunter

Ask Which players' scoring average improved when they were tea Search Advanced Search

Web Images News Deals Videos Q&A News More +

[The Golden Era's Greatest Player, Larry Bird Or Magic Johnson?](#)
Twenty years ago, the NBA belonged to Larry Bird and Magic Johnson. Most hoop fans were polarized, supporting either Boston's ... Bird made an immediate impact. The Celtics improved by 32 wins during his rookie season. ... Johnson has to be considered one of the greatest and most unique basketball players of all-time.
[www.hoopsvibe.com/nba-blog/the-golden-era-s-greatest-pl...](#)

[Teamed at Amazon](#)
Save on Teamed Order by Dec 18 - Get it by Xmas
[Amazon.com/sports](#)

[NBA.com: Michael Jordan Bio](#)
Magic Johnson said, ... Ten scoring titles -- an NBA record and seven consecutive matching Wilt Chamberlain; Retired with the NBA's highest scoring average of 30.1 ppg. ... Even in the exhibition season before his rookie campaign, players and coaches were sure that the Rockets and Blazers would regret their picks. King.
[www.nba.com/history/players/jordan_bio.html](#) - [Cached](#)

[NBA.com: Magic Johnson Bio](#)
Few athletes are truly unique, changing the way their sport is played with their singular skills. Earvin 'Magic' Johnson was one of them. ... His 23.9 season average was the highest of his career.
[www.nba.com/history/players/johnson_bio.html](#) - [Cached](#)

[Magic Johnson and Chris Paul « The Wages of Wins Journal](#)
An average mark is 0.100, so D ... As one can see, Chris Paul in 2008-09 compares favorably to Magic Johnson. And that is a point I would emphasize. Chris Paul is developing into one of the all-time great players in NBA history. It's not a stretch to start thinking of him in terms of players like Magic, MJ.
[dbem.wordpress.com/2009/01/27/magic-johnson-and-chris...](#)

[Earvin 'Magic' Johnson - Los Angeles Lakers Players - Lakers Players -](#)
Johnson was a consistent statistical leader, leading the Lakers in scoring three times (1986-87, 1988-89, and 1989-90) and in rebounding twice (1981-82 and 1982-83), as well as leading ... Home > Basketball Tickets > Los Angeles Lakers Tickets > Lakers Players > Earvin 'Magic' Johnson.
[www.barrytickets.com/lakers/lakers-players/magic-johnson...](#) - [Cached](#)

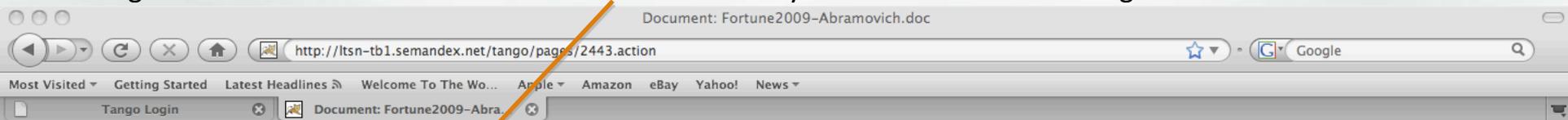
[Top NBA Players Of All-time: Suite101 ranks the be...](#)
Jan 6, 2007 ... And there is no way Larry Bird ranks ahead of I scoring average 50.4. Most points single season 4029 ... N until Phil Jackson built a team around him. ...
[basketball.com/191.../Sports/Basketball/...](#) - [Cached](#) - [Similar](#)

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.

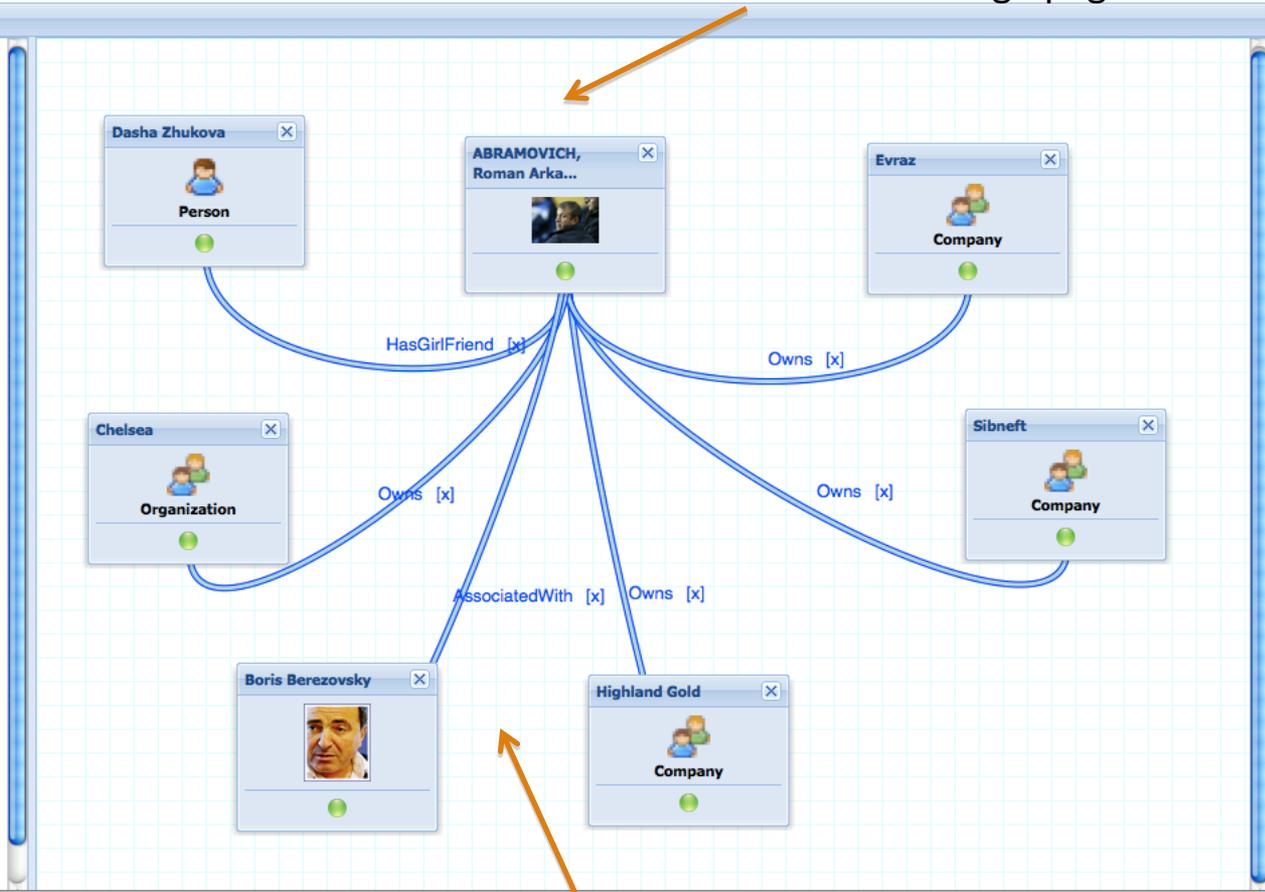
An Example Application (Tango) that Creates Semantic Graphs from Documents and Other Sources

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



3. Each block has an associated Tango page.

Lost more than half his fortune in the past year, mostly due to disastrous performance of Russia's second biggest steel company, **Evraz**, in which he has a large stake. **Russian** government had to loan **Evraz** money so it could pay a tax bill in the fourth quarter of 2008; its stock has dropped almost 90% in the past year. Owns U.K. soccer team **Chelsea**; has spent \$1.5 billion trying to boost team but hasn't helped. Divorced two years ago, recently helped fund his girlfriend **Dasha Zhukova**'s art gallery. She just named editor of Pop magazine. Reportedly bought two homes last year in Colorado ski village, Snowmass. Orphaned as a child, **Abramovich** dropped out of college, then made fortune in a series of controversial oil export deals in early 1990s. Teamed up with **Boris Berezovsky** to take over oil giant **Sibneft** at a fraction of its market value. In 2003 to 2004, sold stake in **Russian** Aluminum to Oleg Deripaska. Sold 73% stake in **Sibneft** to gas titan Gazprom for \$13 billion in 2005. Later bought stake in **Evraz**. Sold part of his stake in **Highland Gold**, a U.K. mining company



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

❖ 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:

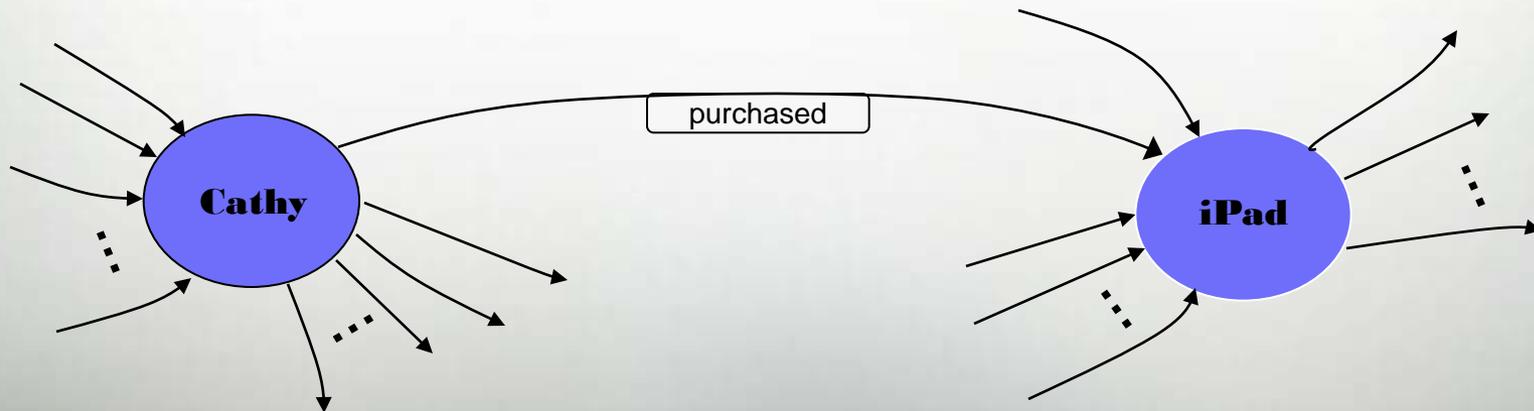
```
title
=====
http://www.recshop.example/cd/Empire Burlesque
http://www.recshop.example/cd/Blonde On Blonde
http://www.recshop.example/cd/Blood on the Tracks
http://www.recshop.example/cd/Nashville Skyline
http://www.recshop.example/cd/John Wesley Harding
http://www.recshop.example/cd/Highway 61 Revisited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

Semantic databases look like graphs...

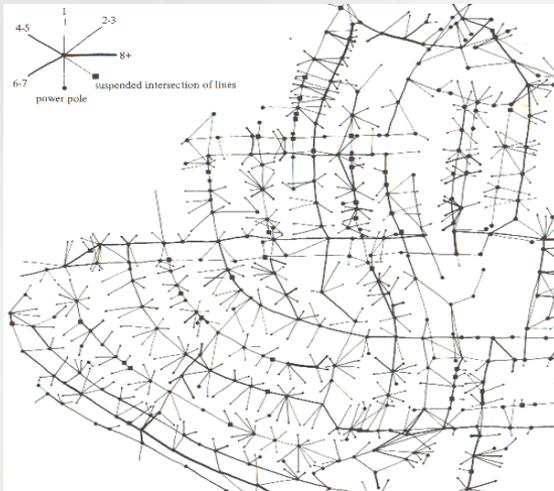
❖ Semantic network databases!



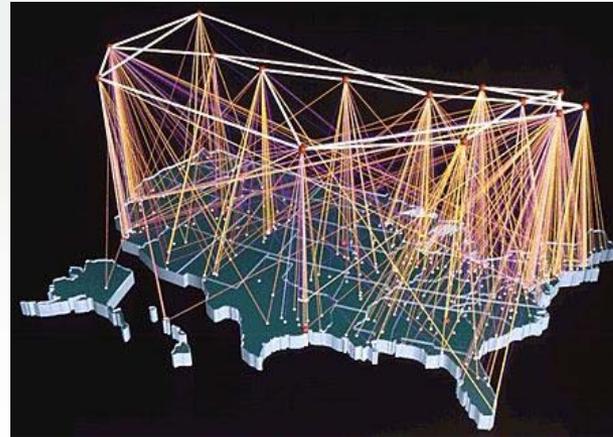
❖ Each RDF triple is a pair of vertices and an edge in a directed graph:



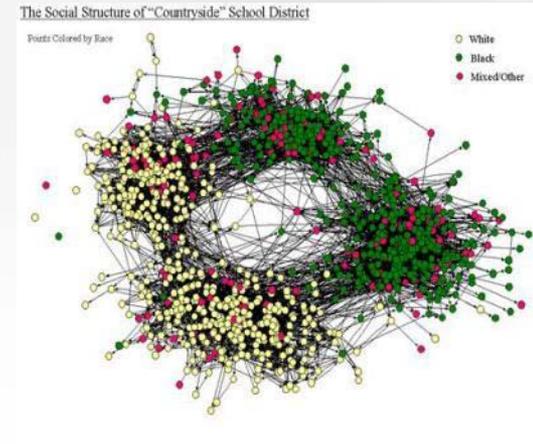
Power Distribution Networks



Internet backbone



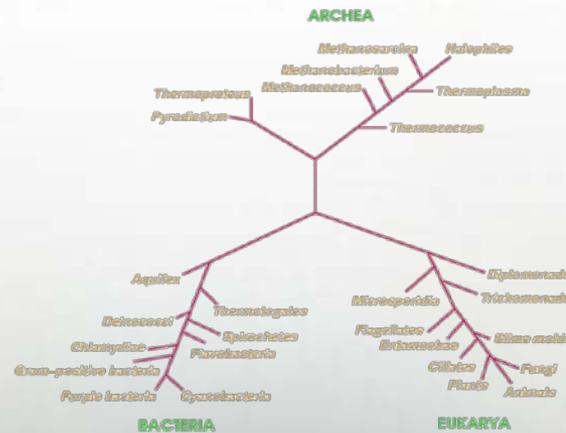
Social Networks



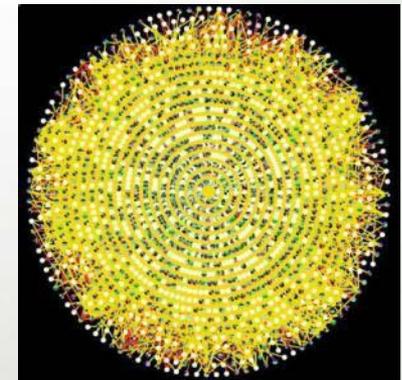
Graphs are everywhere!!!



Ground Transportation



Tree of Life



Protein-interaction networks

Relational Databases vs Semantic Network Databases

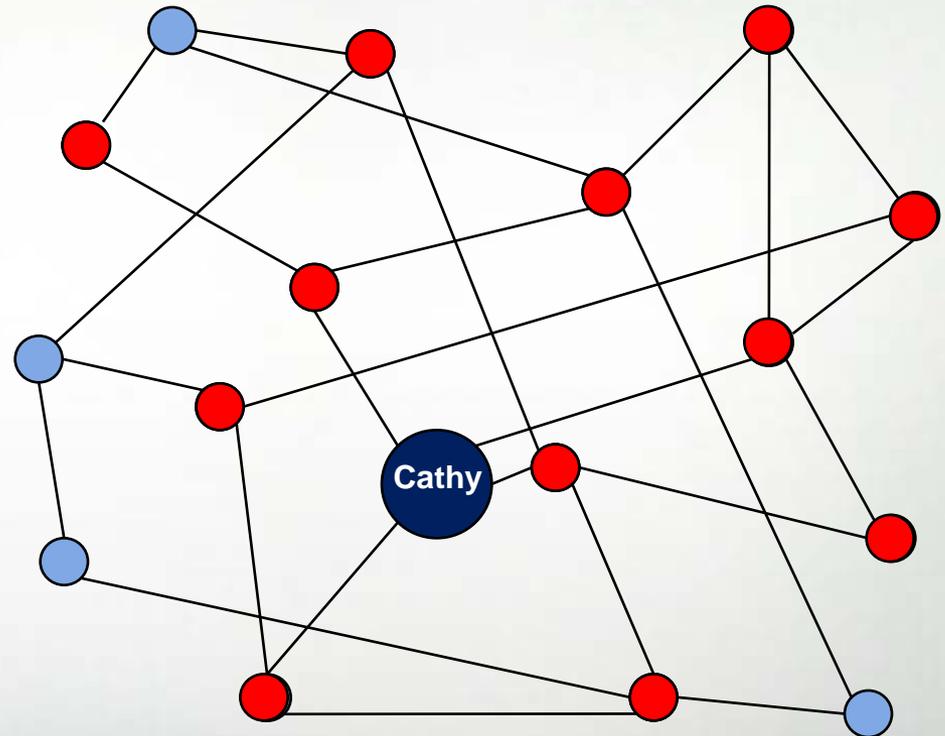
This type of query is easiest for RD:

“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:

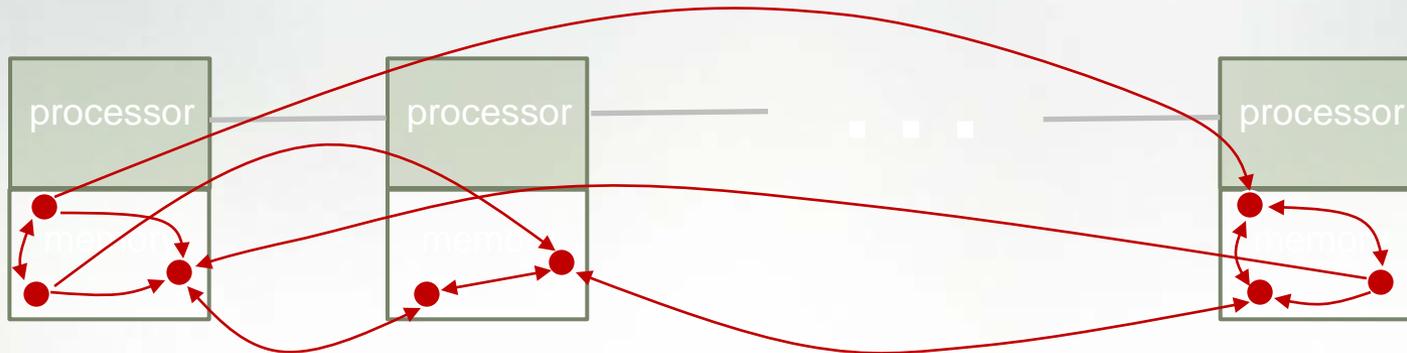
“Show everyone who has talked with Cathy or have talked with someone who talked with Cathy”



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

- ❖ **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



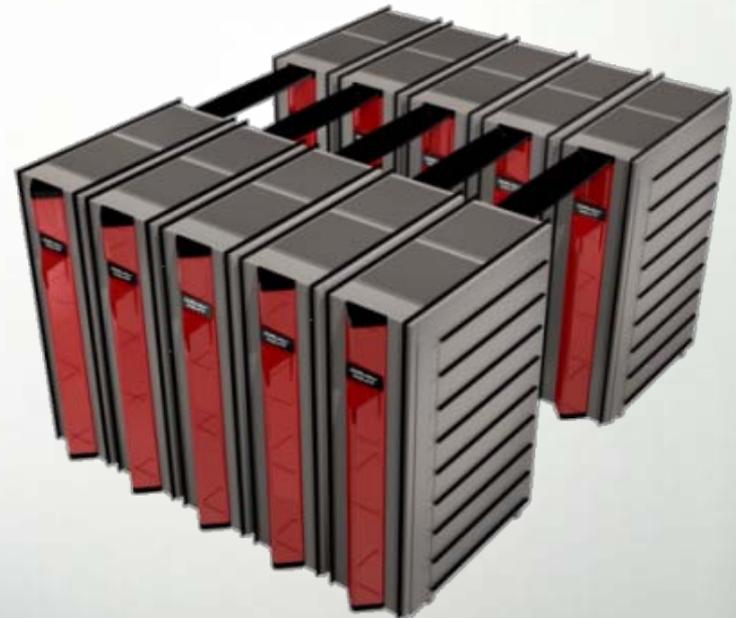
- ❖ **Large delays when one processor needs data from another's memory**
 - In graph analytics/semantic networks, this happens *almost all the time*
 - No way to “partition” the data so that references are usually local
- ❖ **Limited scaling – adding more boxes doesn't improve performance**
- ❖ **Architecture 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**

❖ 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

❖ Application Significance:

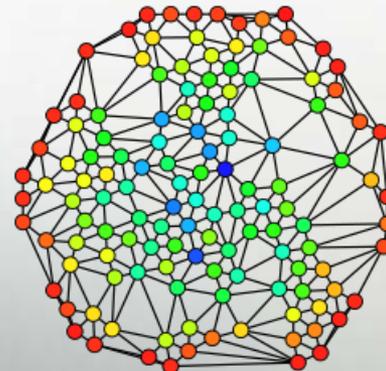
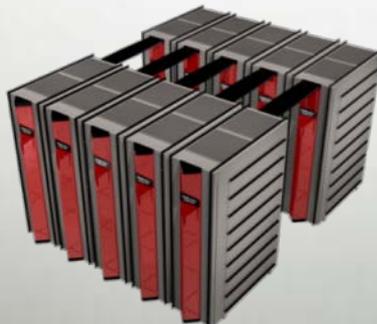
- ✓ Betweenness is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

❖ XMT vs. X86 Cluster:

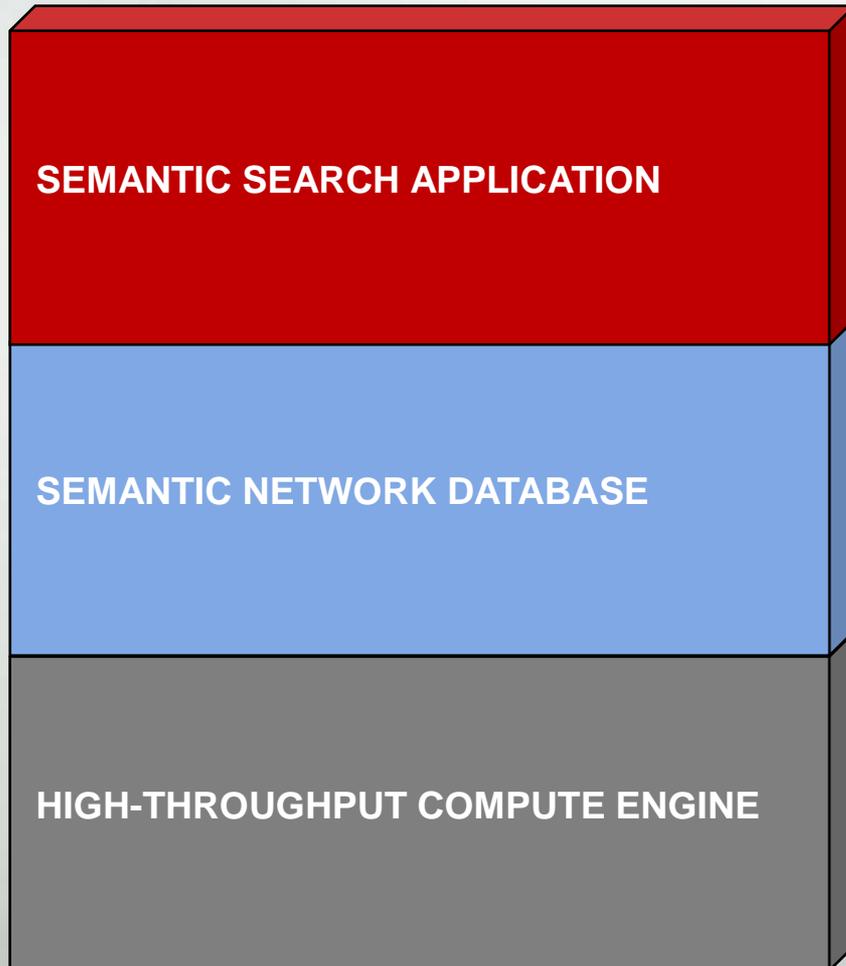
- ✓ 64 processor Cray XMT vs. 64 processor X86 Cluster

❖ XMT processed the graph in **10.6 seconds** versus **65 minutes** with the cluster

- *Batch has become interactive*

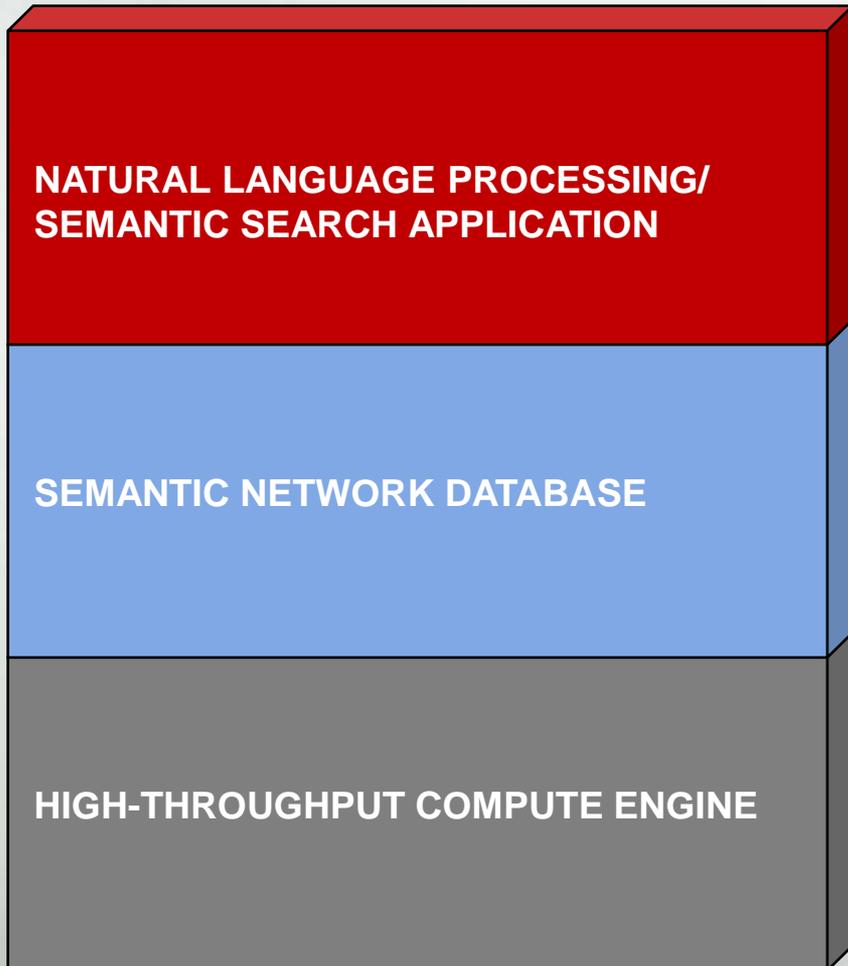


New Paradigm of Unstructured Data Analysis



- **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

- ❖ **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