Big Data Motivation and Big Graph Challenges

Rakesh Nagi

Willet Professor and Head

nagi@illinois.edu

CSE Big Data Research Workshop May 28-30, 2014

Industrial and Enterprise Systems Engineering

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Brief Outline

- Motivation: Big Data Introduction
- Big Graphs
  - Link Analysis
  - Graph Matching
  - Graph Association

Introduction

– Big Data: What is it?

Metric prefixes						
Prefix	**Symbol**	1000^m	10^n	**Decimal**	**Short scale**	**Long scale**
yotta	Y	1000^8	10^{24}	1 000 000 000 000 000 000 000 000	septillion	quadrillion
zetta	Z	1000^7	10^{21}	1 000 000 000 000 000 000 000	sextillion	trilliard
exa	E	1000^6	10^{18}	1 000 000 000 000 000 000	quintillion	trillion
peta	P	1000^5	10^{15}	1 000 000 000 000 000	quadrillion	billiard
tera	T	1000^4	10^{12}	1 000 000 000 000	trillion	billion
giga	G	1000^3	10^9	1 000 000 000	billion	milliard
mega	M	1000^2	10^6	1 000 000	million	
kilo	k	1000^1	10^3	1 000	thousand	
hecto	h	$1000^{2/3}$	10^2	100	hundred	
deca	da	$1000^{1/3}$	10^1	10	ten	
		1000^0	10^0	1	one	

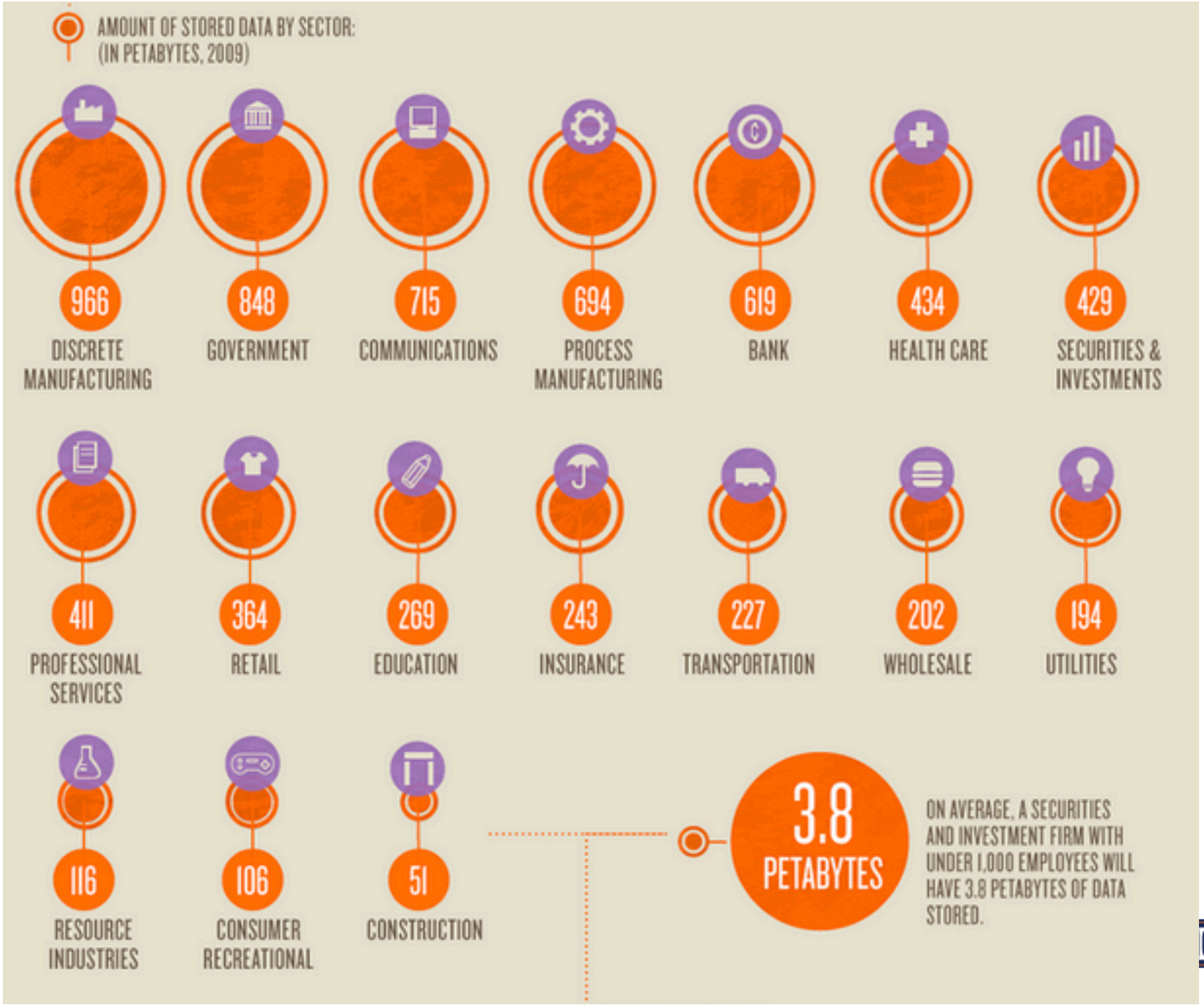
Some Big Data Statistics

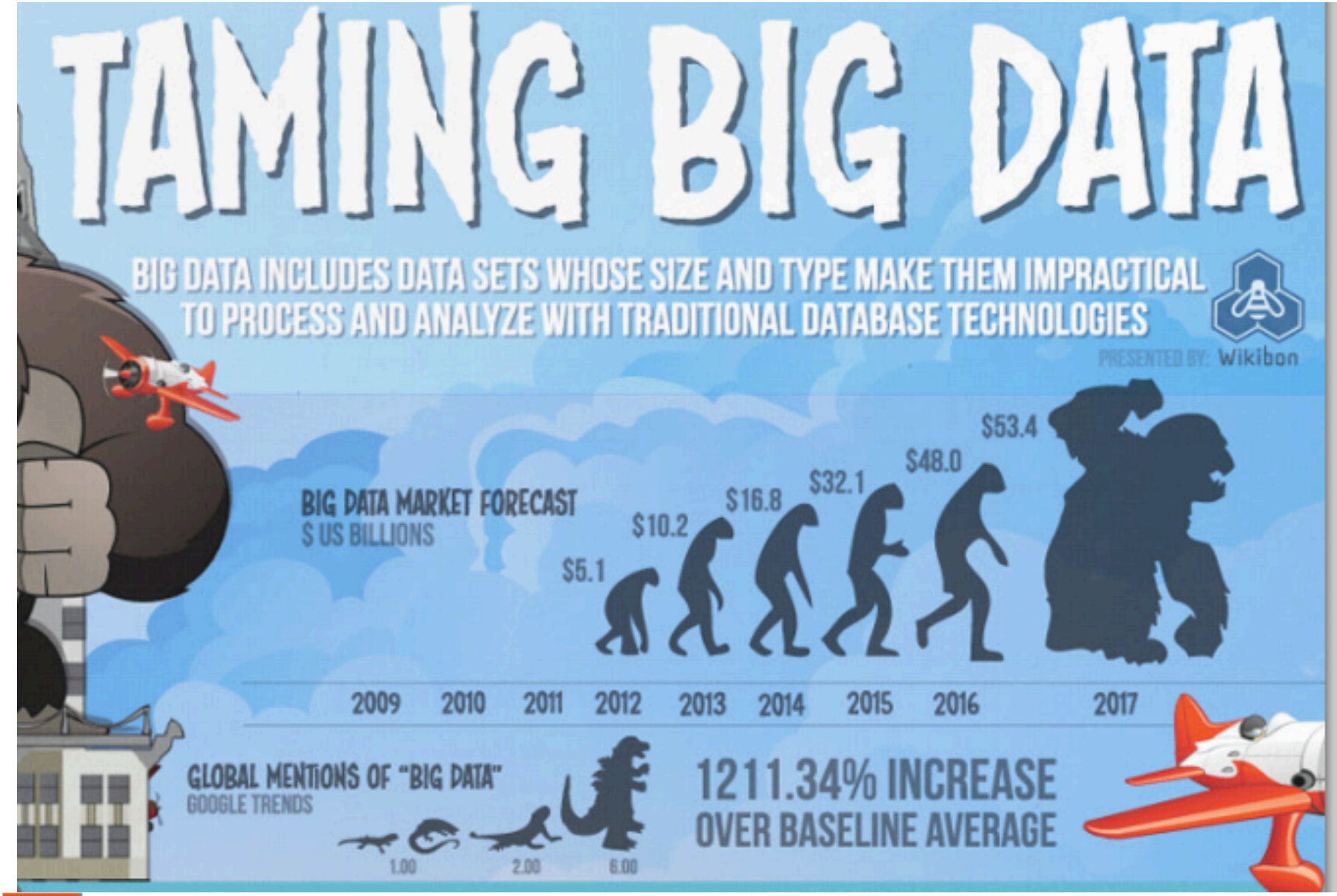
- http://wikibon.org/blog/big-data-statistics/

*Source: http://www.baselinemag.com/analytics-big-data/slideshows/surprising-statistics-about-big-data.html

Business and Technology Environment

- 2.7 Zettabytes of data exist in the digital universe today. http://www.marketingtechblog.com/ibm-big-data-marketing/
- 235 Terabytes of data has been collected by the U.S. Library of Congress in April 2011. http://blog.getsatisfaction.com/2011/07/13/big-data/?view=socialstudies
- The Obama administration is investing $200 million in big data research projects. http://wikibon.org/blog/taming-big-data/
- IDC Estimates that by 2020,business transactions on the internet- business-to-business and business-to-consumer – will reach 450 billion per day. http://wikibon.org/blog/unstructured-data/


AMOUNT OF STORED DATA BY SECTOR:
(IN PETABYTES, 2009)
966
DISCRETE MANUFACTURING
848
GOVERNMENT
715
COMMUNICATIONS
694
PROCESS MANUFACTURING
619
BANK
434
HEALTH CARE
429
SECURITIES & INVESTMENTS
411
PROFESSIONAL SERVICES
364
RETAIL
269
EDUCATION
243
INSURANCE
227
TRANSPORTATION
202
WHOLESALE
194
UTILITIES
116
RESOURCE INDUSTRIES
106
CONSUMER RECREATIONAL
51
CONSTRUCTION
3.8 PETABYTES
ON AVERAGE, A SECURITIES AND INVESTMENT FIRM WITH UNDER 1,000 EMPLOYEES WILL HAVE 3.8 PETABYTES OF DATA STORED.


TAMING BIG DATA
BIG DATA INCLUDES DATA SETS WHOSE SIZE AND TYPE MAKE THEM IMPRACTICAL TO PROCESS AND ANALYZE WITH TRADITIONAL DATABASE TECHNOLOGIES
PRESENTED BY: Wikibon
BIG DATA MARKET FORECAST
$ US BILLIONS
$5.1
$10.2
$16.8
$32.1
$48.0
$53.4
2009
2010
2011
2012
2013
2014
2015
2016
2017
GLOBAL MENTIONS OF "BIG DATA"
GOOGLE TRENDS
1.00
2.00
6.00
1211.34% INCREASE
OVER BASELINE AVERAGE


facebook
stores, accesses and analyzes
30+ PETABYTES
of user generated data
LinkedIn
processes and mines
PETABYTES
of user data to power
"People You May Know"
amazon
crunches click-stream and historical user data to recommend products
Akamai analyzes
75 MILLION
events per day
to better target advertisements
JPMorgan Chase & Co. analyzes web logs, transaction data, and social media to detect fraudulent activity
Treato
the voice of the patient
taps Big Data to help researchers and physicians better determine patient treatments
The New York Times processed 4TB worth of raw images
into 11 MILLION finished PDFs in 24 HOURS
DECODING THE HUMAN GENOME USED TO TAKE TEN YEARS.
IT CAN NOW BE DONE IN 7 DAYS.
THE OBAMA ADMINISTRATION IS INVESTING
$200 MILLION IN BIG DATA RESEARCH PROJECTS.
massively parallel processing, columnar architecture, and data compression to ingest and analyze Big Data in near real-time
hadoop
open source framework for storing, processing and analyzing massive amounts of distributed, multi-structured data
MPP Analytic Database

Business and Technology Environment

- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.
- Akamai analyzes 75 million events per day to better target advertisements.
- 94% of Hadoop users perform analytics on large volumes of data not possible before; 88% analyze data in greater detail; while 82% can now retain more of their data. http://www.sys-con.com/node/1920943
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data. http://www.sas.com/resources/whitepaper/wp_46345.pdf
- More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide. [Above Ref.]

Business and Technology Environment

- Decoding the human genome originally took 10 years to process; now it can be achieved in one week. http://www.economist.com/node/15557443
- In 2008, Google was processing 20,000 terabytes of data (20 petabytes) a day. http://techcrunch.com/2008/01/09/google-processing-20000-terabytes-a-day-and-growing/
- The largest AT&T database boasts titles including the largest volume of data in one unique database (312 terabytes) and the second largest number of rows in a unique database (1.9 trillion), which comprises AT&T's extensive calling records. http://www2.research.att.com/~daytona/

Unstructured Data

- YouTube users upload 48 hours of new video every minute of the day.
- 571 new websites are created every minute of the day.
- Brands and organizations on Facebook receive 34,722 Likes every minute of the day.
- 100 terabytes of data uploaded daily to Facebook.

[R1] http://wikibon.org/blog/big-data-infographics/

- According to Twitter's own research in early 2012, it sees roughly 175 million tweets every day, and has more than 465 million accounts.
  http://www.mediapost.com/publications/article/173109/a-conversation-on-the-role-of-big-data-in-marketin.html#axzz2GfHngeVn

Unstructured Data

- 30 Billion pieces of content shared on Facebook every month. [R2] http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
- Data production will be 44 times greater in 2020 than it was in 2009. [R1]
- In late 2011, IDC Digital Universe published a report indicating that some 1.8 zettabytes of data will be created that year. http://siliconrepublic.com/strategy/item/22420-amount-of-data-in-2011-equa

  In other words, the amount of data in the world today is equal to:
  - Every person in the US tweeting three tweets per minute for 26,976 years.
  - Every person in the world having more than 215m high-resolution MRI scans a day.
  - More than 200bn HD movies – which would take a person 47m years to watch.

The Market(ers') Challenge with Big Data

- Big data is a top business priority and drives enormous opportunity for business improvement. Wikibon's own study projects that big data will be a $50 billion business by 2017. http://siliconangle.com/blog/2012/07/13/studies-confirm-big-data-as-key-business-priority-growth-driver/
- As recently as 2009 there were only a handful of big data projects and total industry revenues were under $100 million. By the end of 2012 more than 90 percent of the Fortune 500 will likely have at least some big data initiatives under way. http://www.smartplanet.com/blog/business-brains/big-data-market-set-to-explode-this-year-but-what-is-8216big-data/22126
- Market research firm IDC has released a new forecast that shows the big data market is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015. http://statspotting.com/2012/03/big-data-statistics-16-9-billion-market-by-2015/

The Market(ers') Challenge with Big Data

- In the developed economies of Europe, government administrators could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using big data, not including using big data to reduce fraud and errors and boost the collection of tax revenues. [R2]
- Poor data across businesses and the government costs the U.S. economy $3.1 trillion dollars a year. http://www-new.insightsquared.com/2012/01/7-facts-about-data-quality-infographic/
- 140,000 to 190,000. Too few people with deep analytical skills to fill the demand of Big Data jobs in the U.S. by 2018. [R2]
- 14.9 percent of marketers polled in Crain's BtoB Magazine are still wondering "What is Big Data?". http://www.forbes.com/sites/lisaarthur/2012/04/17/b2b-marketers-use-big-data-new-tools-to-evaluate-execute-evolve/

The Market(ers') Challenge with Big Data

- 39 percent of marketers say that their data is collected "too infrequently or not real-time enough." http://www.prnewswire.com/news-releases/study-finds-marketers-struggle-with-the-big-data-and-digital-tools-of-today-142312475.html
- 29 percent report that their marketing departments have "too little or no customer/consumer data." When data is collected by marketers, it is often not appropriate to real-time decision making. [Same source as above]

Big Data & Real Business Issues

- According to estimates, the volume of business data worldwide, across all companies, doubles every 1.2 years. http://knowwpcarey.com/article.cfm?cid=25&aid=1171
- Poor data can cost businesses 20%–35% of their operating revenue. http://www.fathomdelivers.com/big-data-facts-and-statistics-that-will-shock-you/
- Bad data or poor data quality costs US businesses $600 billion annually.
- According to execs, the influx of data is putting a strain on IT infrastructure. 55 percent of respondents reporting a slowdown of IT systems and 47 percent citing data security problems, according to a global survey from Avanade. http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEOv.pdf
- In that same survey, by a small but noticeable margin, executives at small companies (fewer than 1,000 employees) are nearly 10 percent more likely to view data as a strategic differentiator than their counterparts at large enterprises.

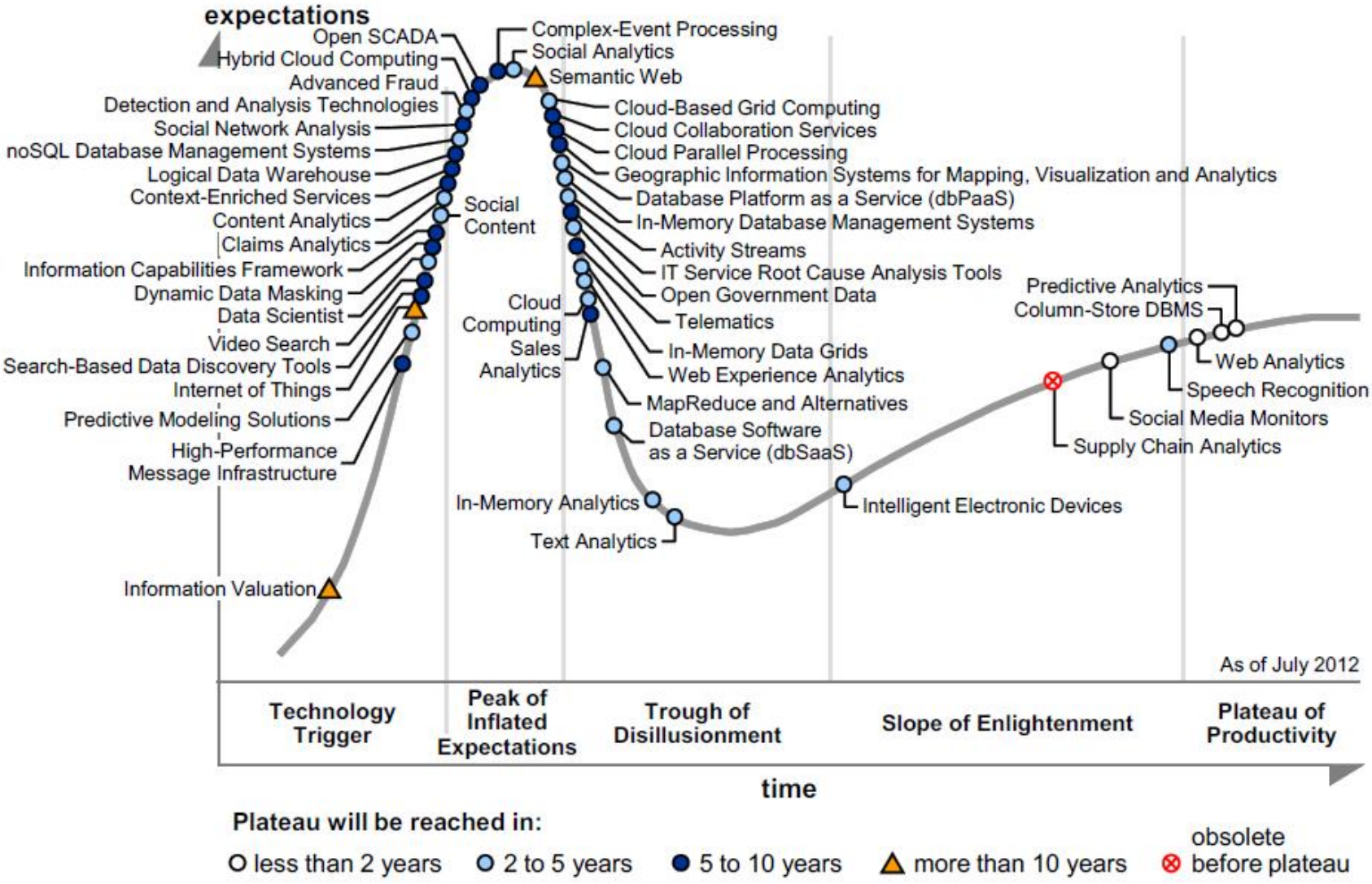
Big Data & Real Business Issues

- Three-quarters of decision-makers (76 per cent) surveyed anticipate significant impacts in the domain of storage systems as a result of the "Big Data" phenomenon. http://www.btplc.com/news/articles/showarticle.cfm?articleid=%7B74889611-be1c-4f91-bbed-ab9e72e25918%7D
- A quarter of decision-makers surveyed predict that data volumes in their companies will rise by more than 60 per cent by the end of 2014, with the average of all respondents anticipating a growth of no less than 42 per cent. [same as above]
- 40% projected growth in global data generated per year vs. 5% growth in global IT spending. [R2]

Why should Operations Researchers care?

- The IBM Business Analytics and Optimization for the Intelligent Enterprise study revealed that businesses can gain 20 time more profits and a 30 percent higher return by using big data and a mature business analytics and optimization strategy.
- In addition, over 150 CIO and managers that target the federal government market participated in a study by storage provider NetApp and government portal maker MeriTalk released in May. The results showed that 60 percent of the agencies they serve are using big data, and 40 reported their federal customers are using big data for decision making.
- A whooping 96 said they believed the use of big data will increase in the next two years. Fifty nine percent identified improving efficiency as the key benefit of big data.

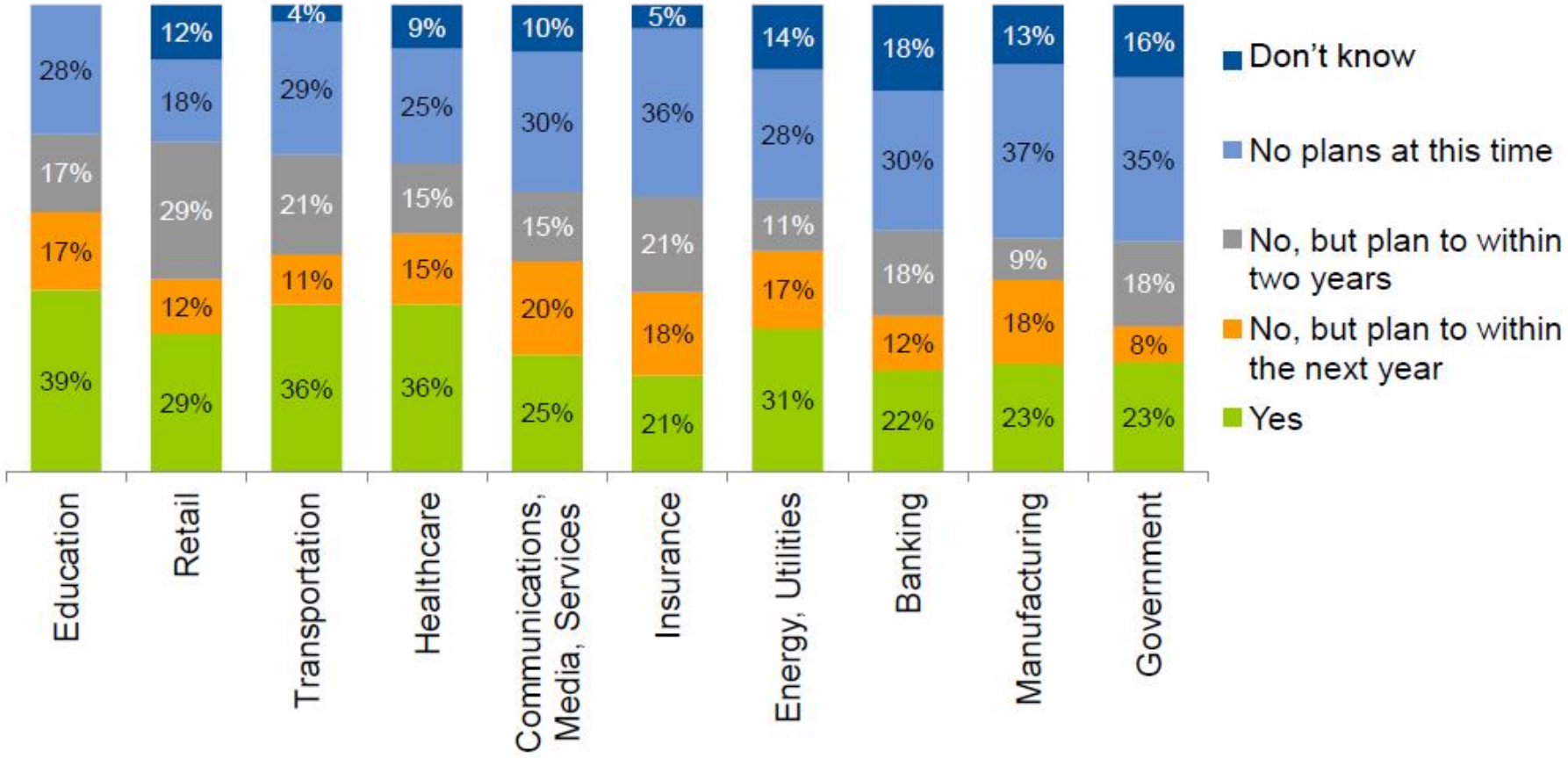
Figure 1. Hype Cycle for Big Data, 2012



Source: Gartner (July 2012)

Big Data Investments by Industry

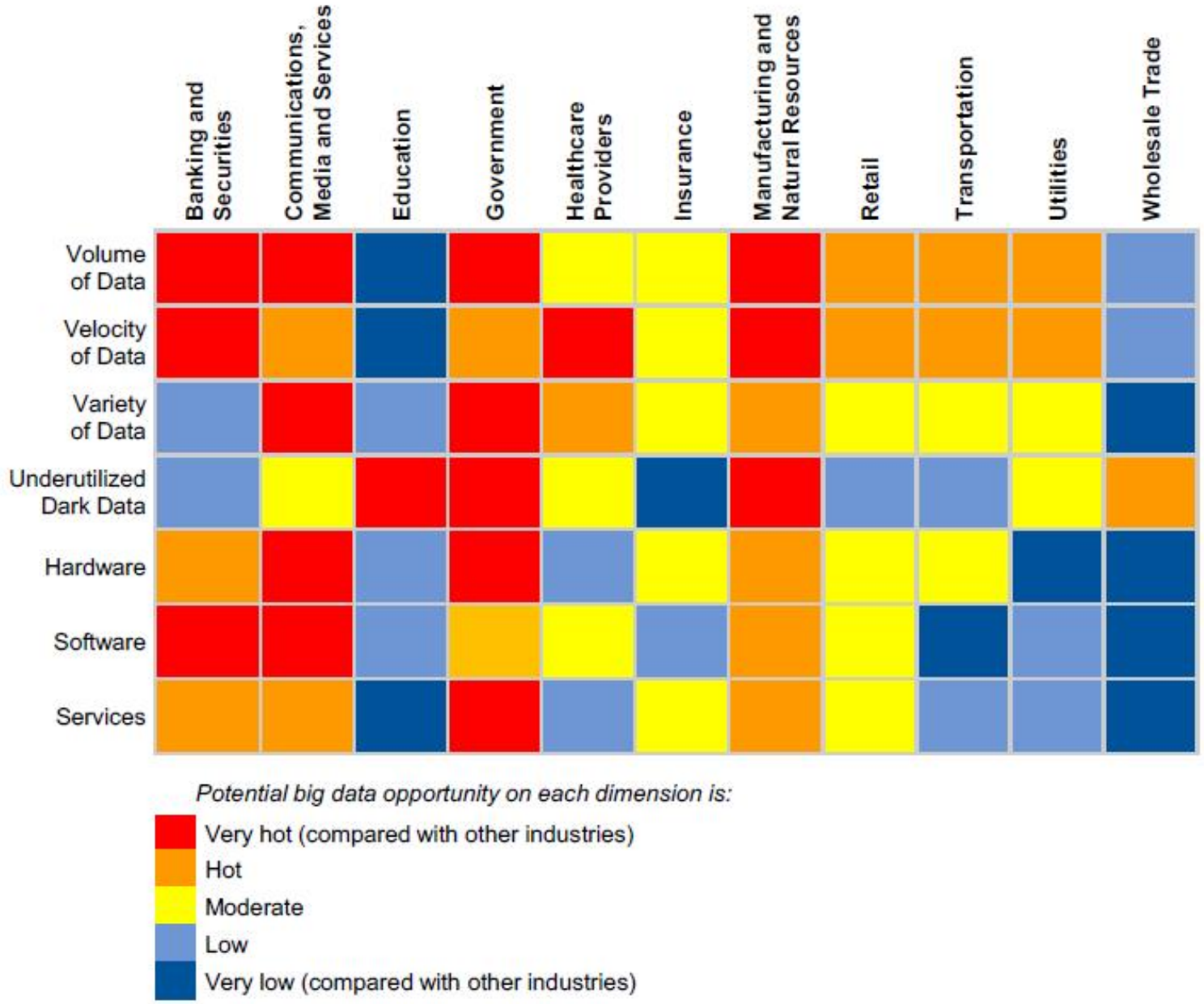
Has your organization already invested in technology specifically designed to address the big data challenge?



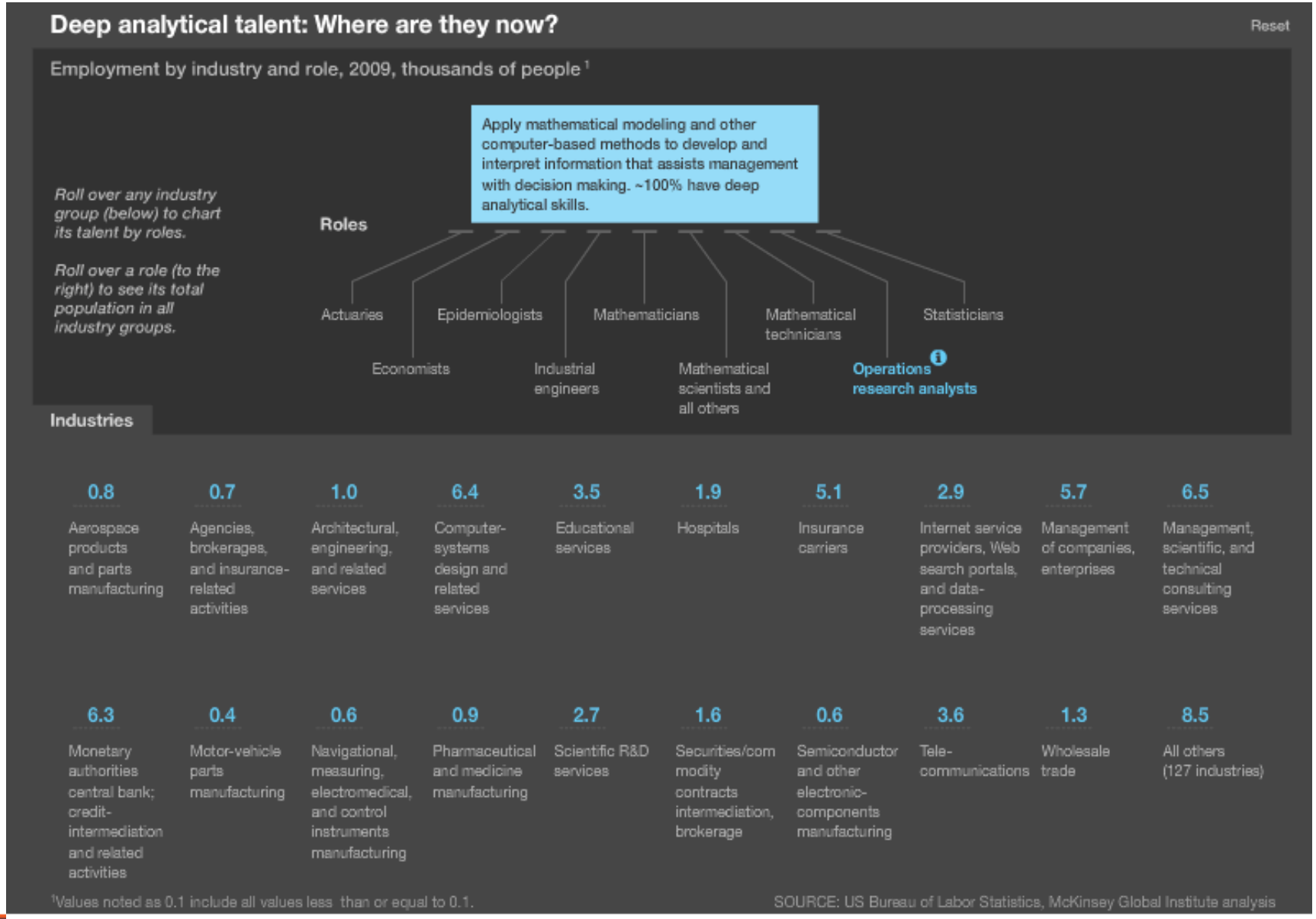
Source: Gartner (July 2012)

Gartner

Figure 2. Big Data Opportunity Heat Map by Industry



Source: Gartner (July 2012)



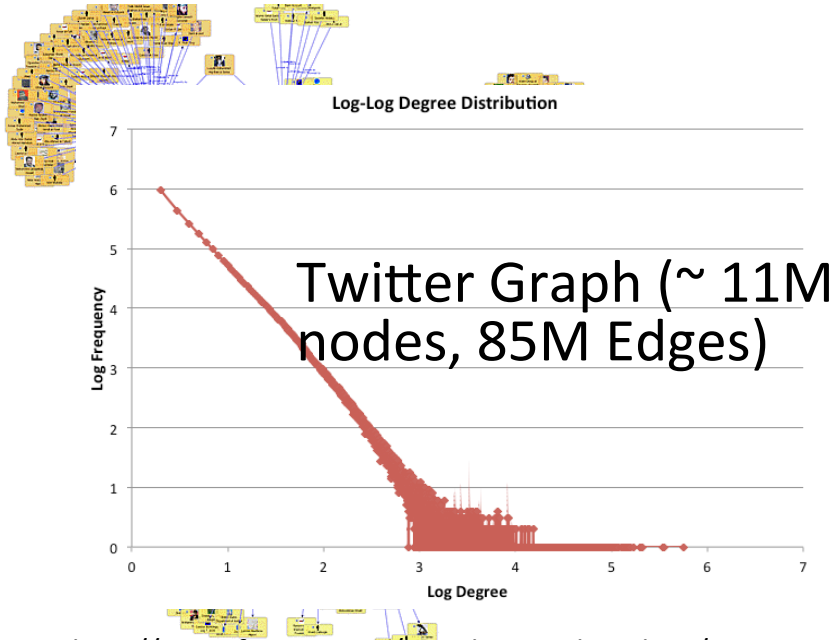
Interactive Popout:
http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

References

- Full Report from MGI
- http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Big Graphs

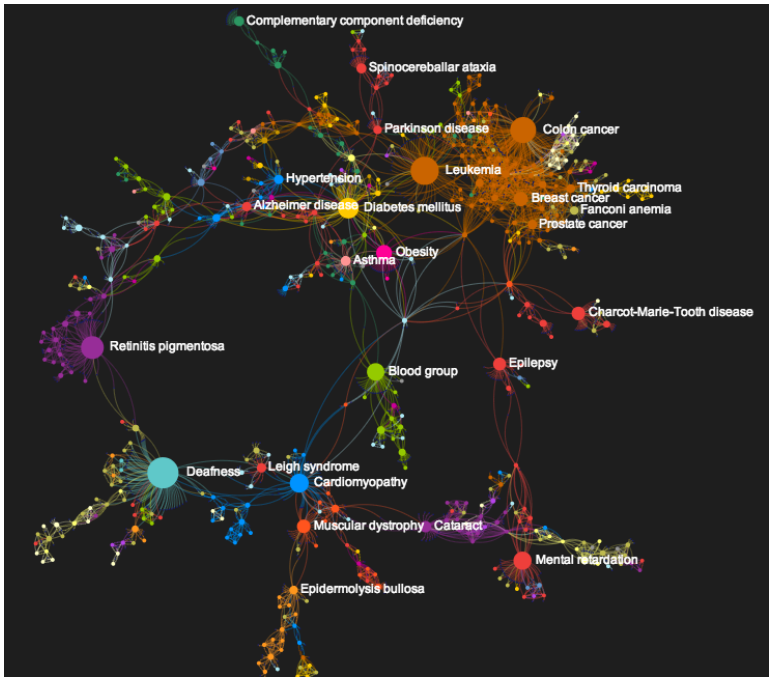
Social Networks



Twitter Graph (~ 11M nodes, 85M Edges)

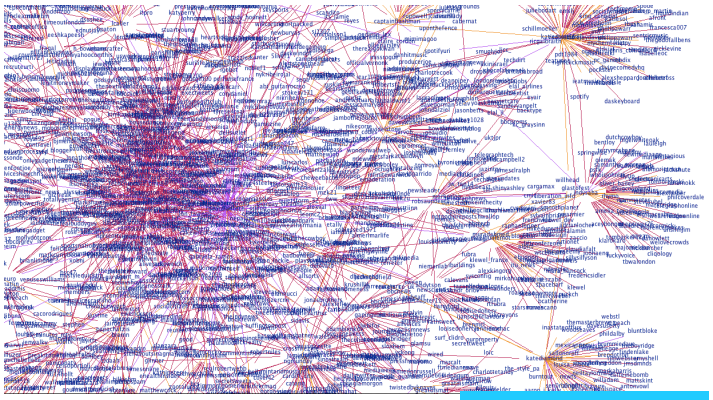
http://www.fmsasg.com/SocialNetworkAnalysis/

Human Disease Networks



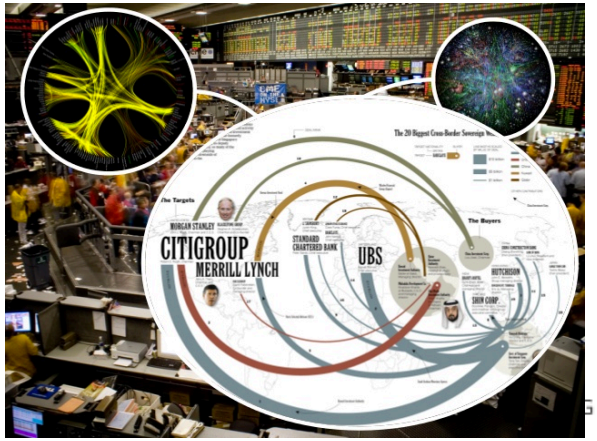
http://www.visualizing.org/visualizations/human-disease-network-graph

Twitter



http://www.catehuston.com/

Financial Networks

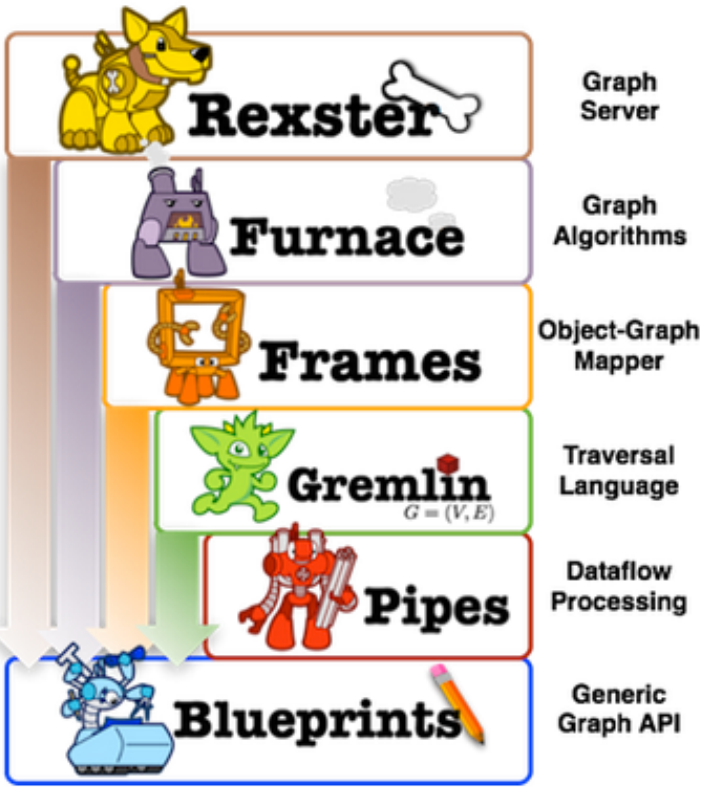


Big Graph Tools

Neo4j
the graph database

TITAN
Distributed Graph Database

TinkerPop



ORACLE BERKELEY DB 12c

APACHE HBASE

cassandra

GraphLab
Carnegie Mellon

(1) Link Analysis

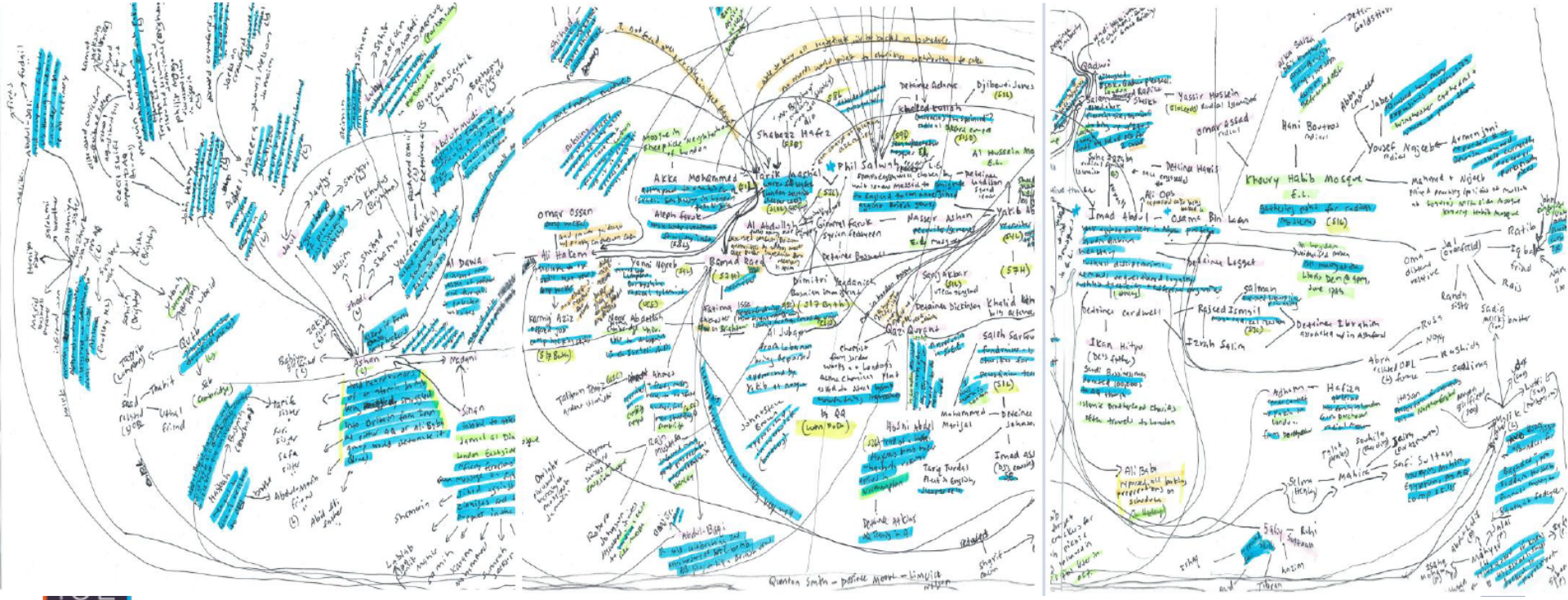
Milan Patel, I2WD, Aberdeen MD
Geoff Gross, ggross@gmail.com

Link Analysis – Motivation

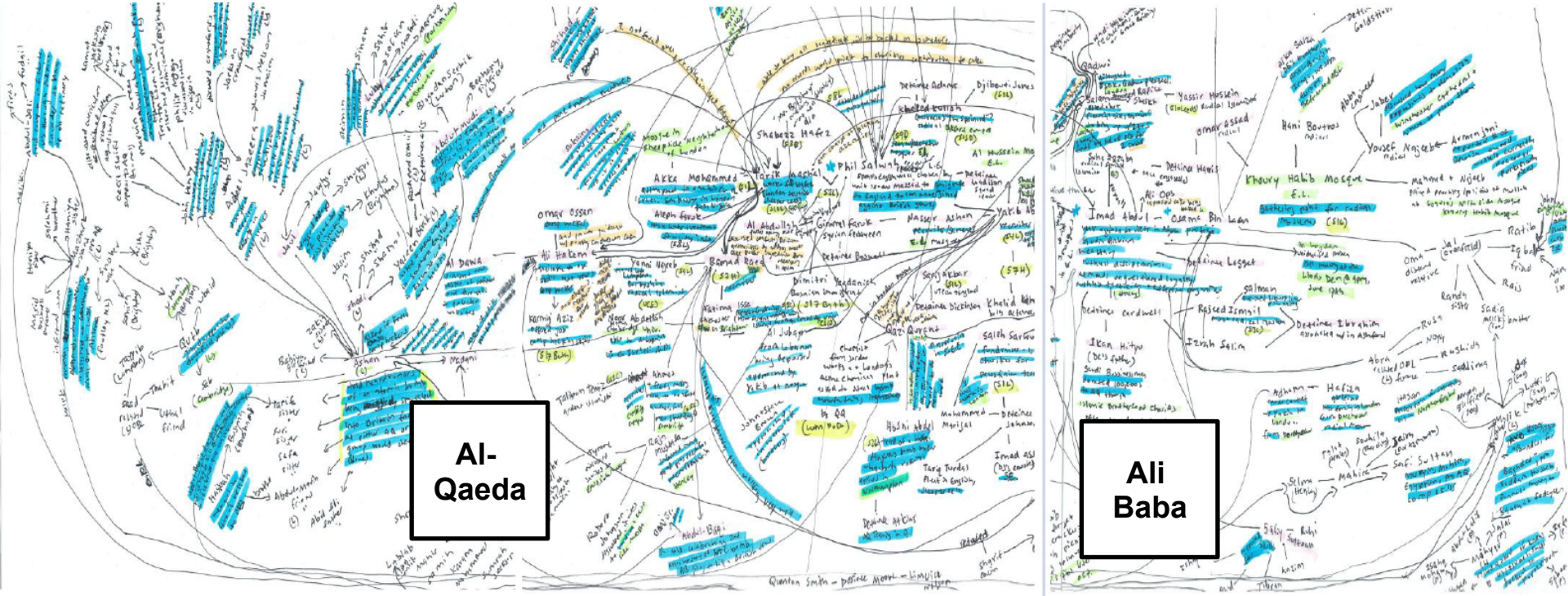
- Analysts are often interested in loose (many-hop) connections between entities in evidential data
  - e.g., distant familial relationships
- Graph traversal queries are time consuming in large (distributed) graphs
  - Requires scalable (parallel) link analysis solution to handle the magnitude of such queries
- Intermediate data can experience exponential path number explosion
- Interested in assessing the ability of two data access methodologies (PostgreSQL and HBase) to support graph analytic techniques such as link analysis

Link Analysis Augmentation Utility

- Link analysis tool developed to automatically identify potentially meaningful connections across entities (people, organizations, locations, etc.) of interest within global graph
- Can be used to augment existing link diagrams created by a human analyst, or forage other relationships



Link Analysis Augmentation Utility Execution



1. Specify two or more entities of interest
2. Filter results (if desired)
   - Can be filtered automatically* or manually
3. Link analysis algorithm executed

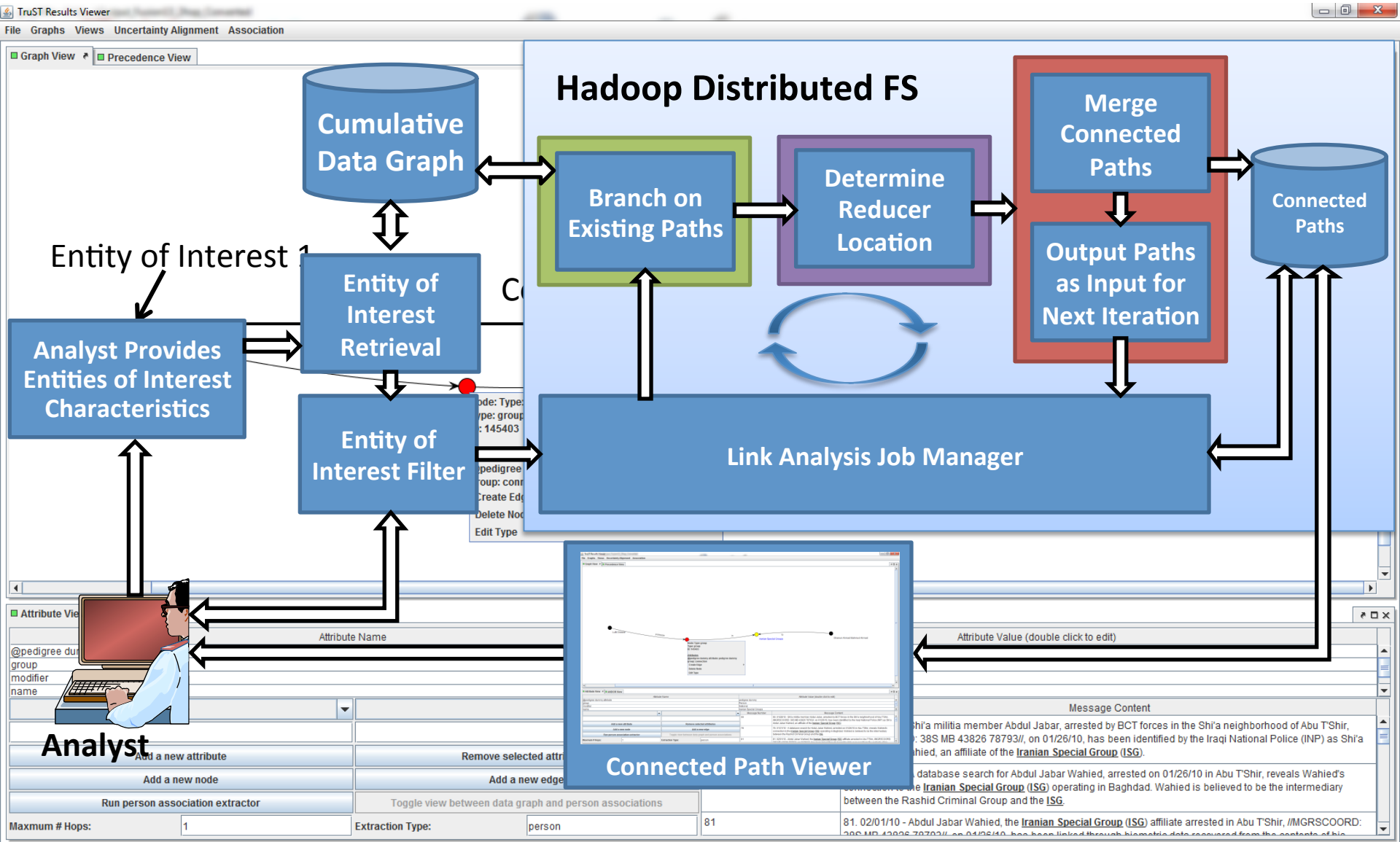
* - not currently implemented

Link Analysis - Methodology

Task Description

- Given two (or more) entities of interest (EOI), identify all paths connecting these nodes within the global graph
  - Termination Criteria
    - Number of hops
    - Number of solutions
    - Runtime
    - Analyst interrupt

Link Analysis Framework (Enhance Enterprises LLC)

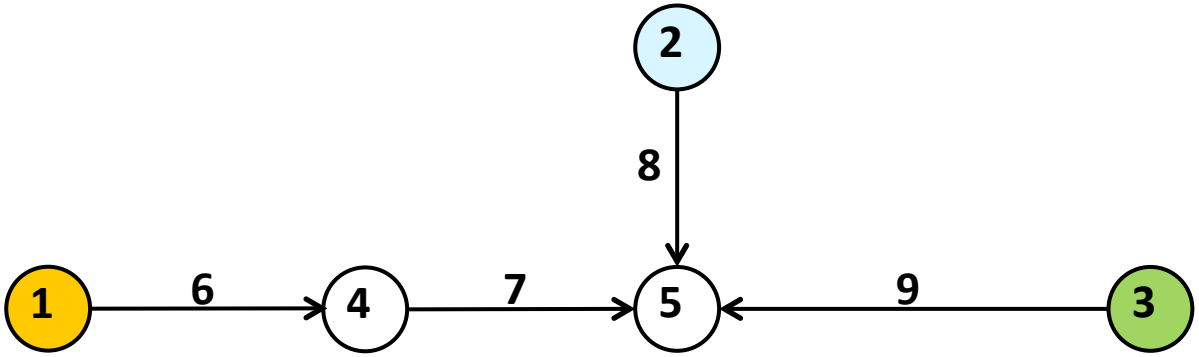


Link Analysis - Methodology

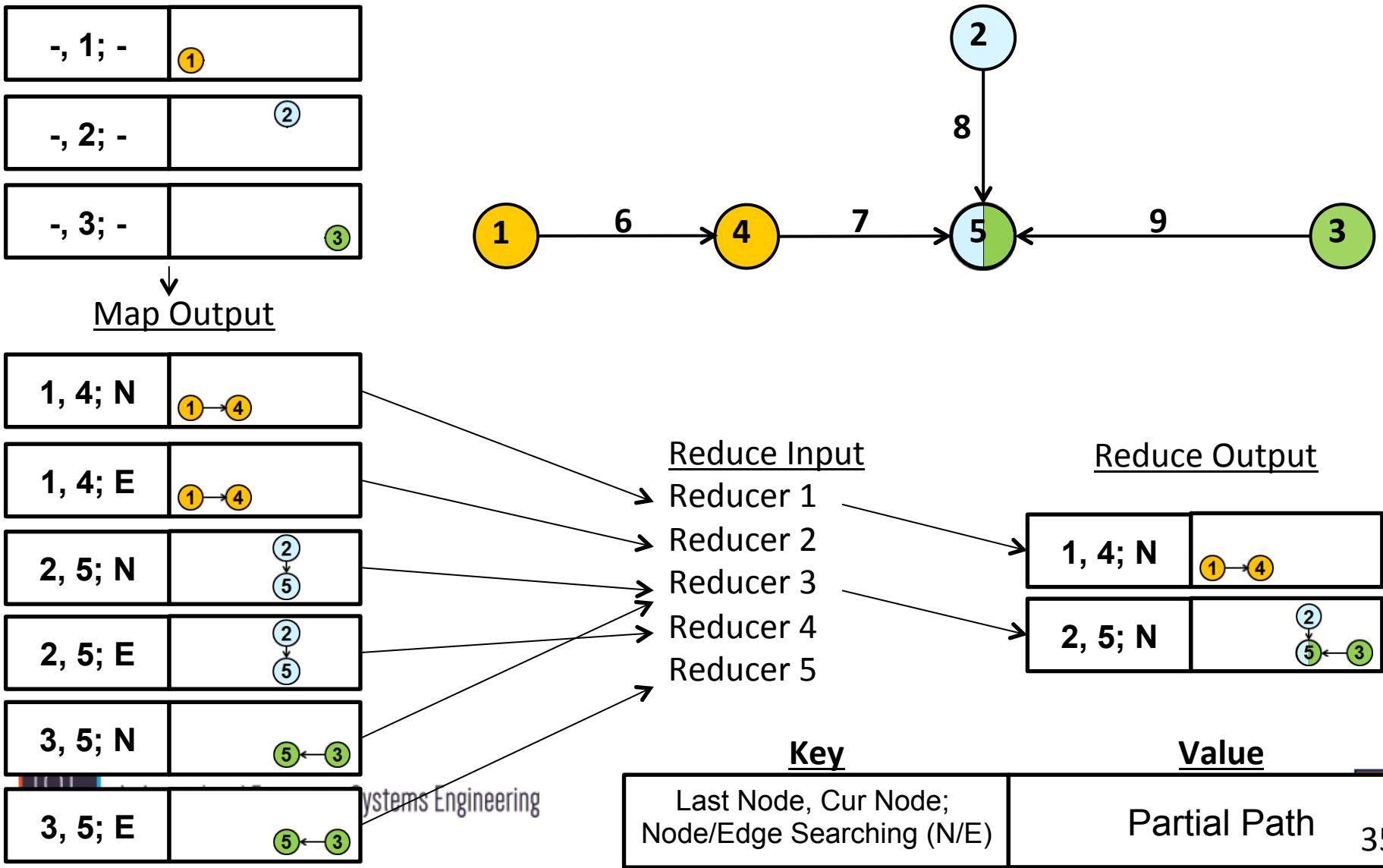
Map-Reduce Implementation Overview

1. Identify instances of the entities (root nodes) of interest within the global graph
2. Branch from the existing frontier (root nodes only for first iteration)
3. Determine overlap of newly reached nodes with paths from other root nodes
   - Output connected paths
4. Check termination criteria, if not met return to Step 2

Step 1 – EOI Identification (Nodes 1, 2 and 3)

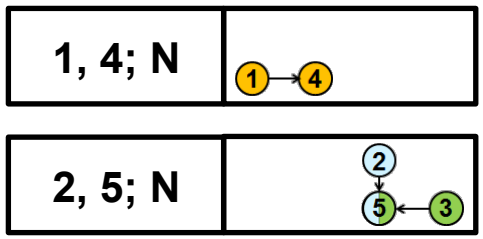


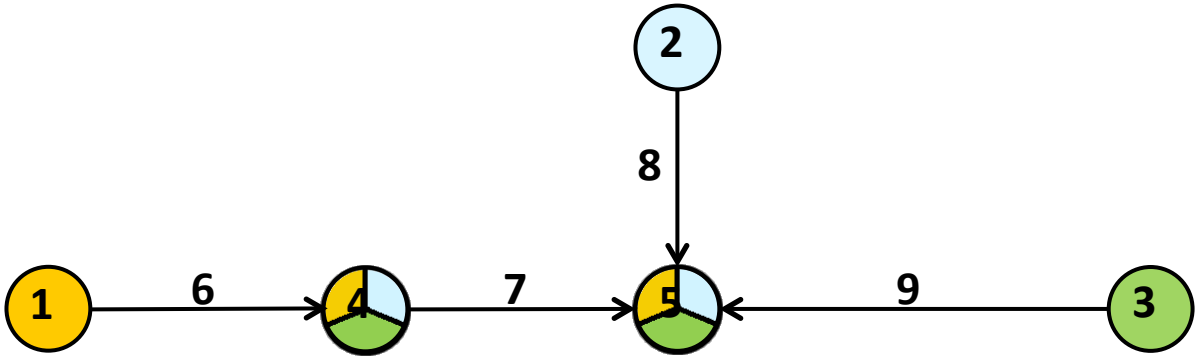
Step 2 – 1st Frontier Expansion



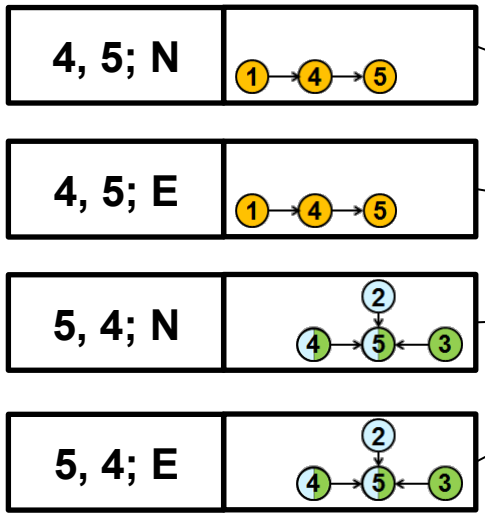
Step 3 – 2nd Frontier Expansion

Map Input





Map Output



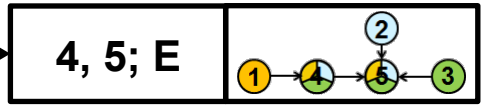
Reduce Input

Reducer 1

Reducer 2

Reducer 3

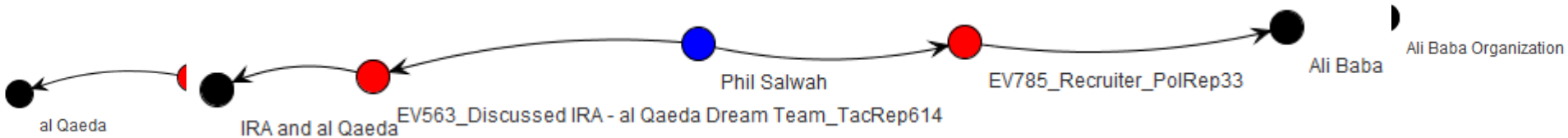
Reduce Output

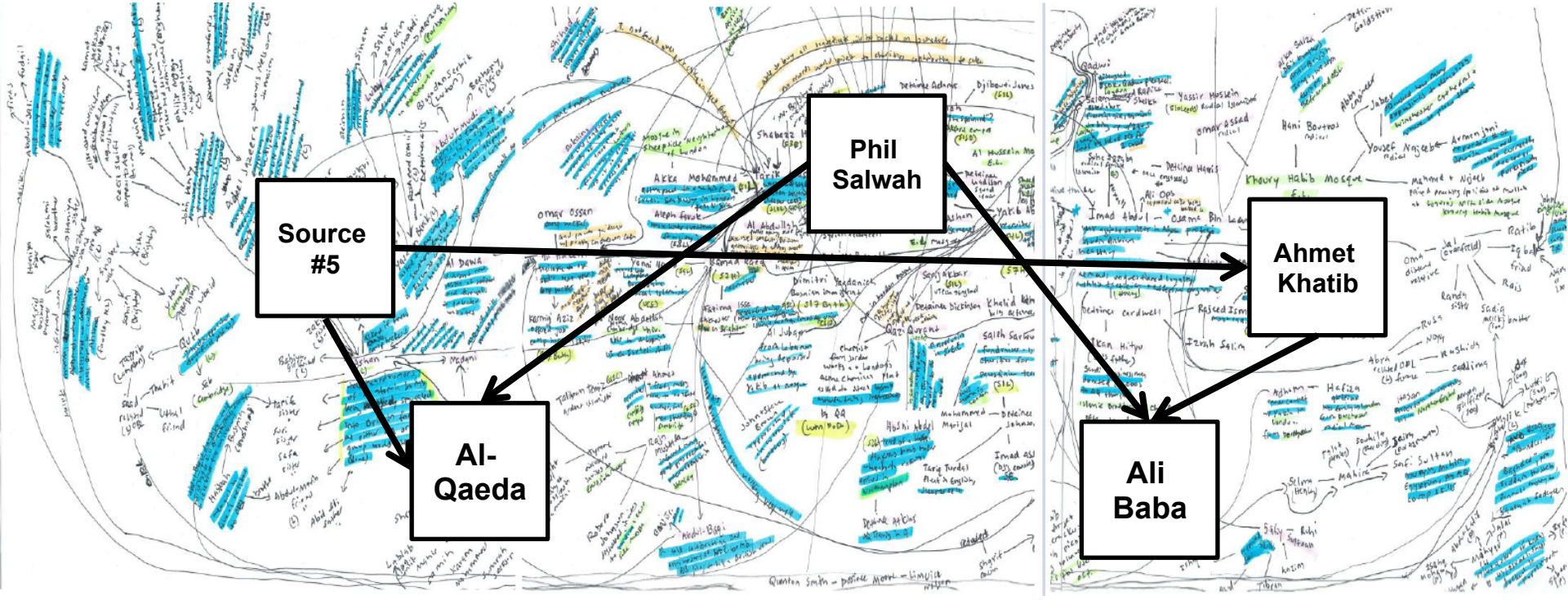


Key	Value
Last Node, Cur Node; Node/Edge Searching (N/E)	Partial Path

Link Diagram Augmentation – Algorithm Output Analysis

= entity of interest





Link Analysis Testing

- Test graph consisted of 10 million nodes and 19 million edges
- Sequential link analysis implementation followed NIST recommendation for all simple paths approach – depth first search
  - Sample link analysis job required 31,886 seconds, requesting 609,442 adjacencies
- 15 link analysis queries utilized as testbed
  - 5 link analysis jobs with 100, 200 and 300 starting points

Link Analysis Testing Results

Adjacency Requests by Query and Iteration Number

Query Number	Iteration Number 1	2	3
1	100	230	22,129
2	200	471	43,531
3	300	777	66,001
4	100	232	18,044
5	200	523	47,601
6	300	700	53,892
7	100	251	19,174
8	200	509	36,489
9	300	753	64,046
10	100	263	19,484
11	200	515	49,456
12	300	757	63,608
13	100	270	22,619
14	200	517	41,349
15	300	792	67,573

Link Analysis Testing Results

- Driving runtime factor is adjacency retrieval time in iteration 3
- Runtime advantage (at ideal cluster configuration settings) of Postgres data access leads to lower link analysis job runtime

Average Adjacency Retrieval Time (ms) by Data Access Method and Iteration

		Iteration		
		1	2	3
Data Access Method	HBase	30.58	31.00	11.09
	Postgres	7.38	44.17	2.43

Link Analysis Testing Results

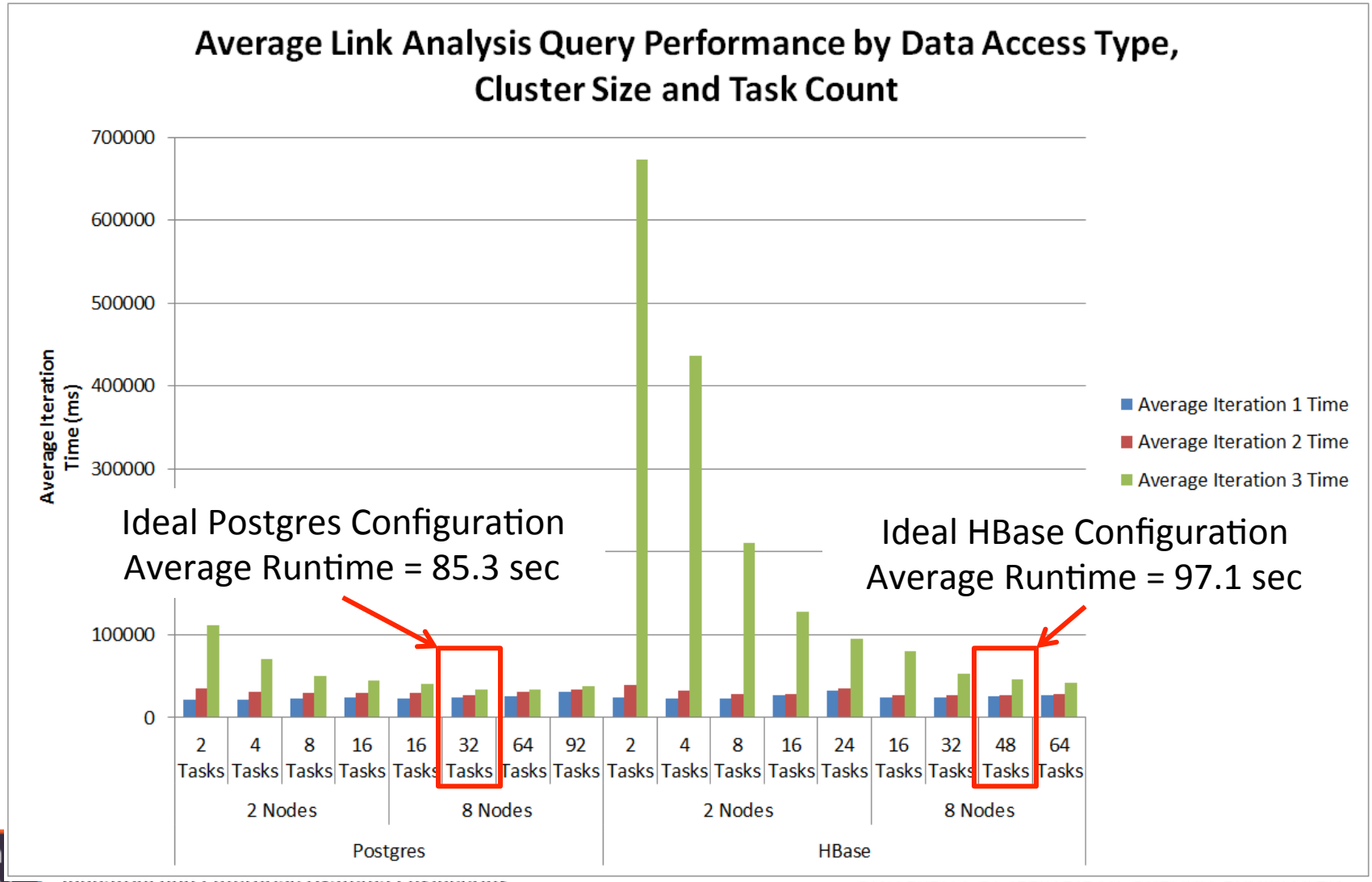
- Iteration completion blocked by mapper and reducer completion times
  - Maximum mapper/reducer limit iteration

Average Maximum Mapper/Reducer Task Times by Data Access Method, Iteration Number and Task Count

		Average Maximum Mapper Times											
		Iteration 1				Iteration 2				Iteration 3			
	Task Count	16	32	64	92	16	32	64	92	16	32	64	92
Data Access Method	HBase	2756	2383	3150	3304	3942	3318	3501	4259	66123	37249	21738	19212
	Postgres	1837	1703	2587	3221	7726	5176	5961	5866	18123	11100	9978	9695

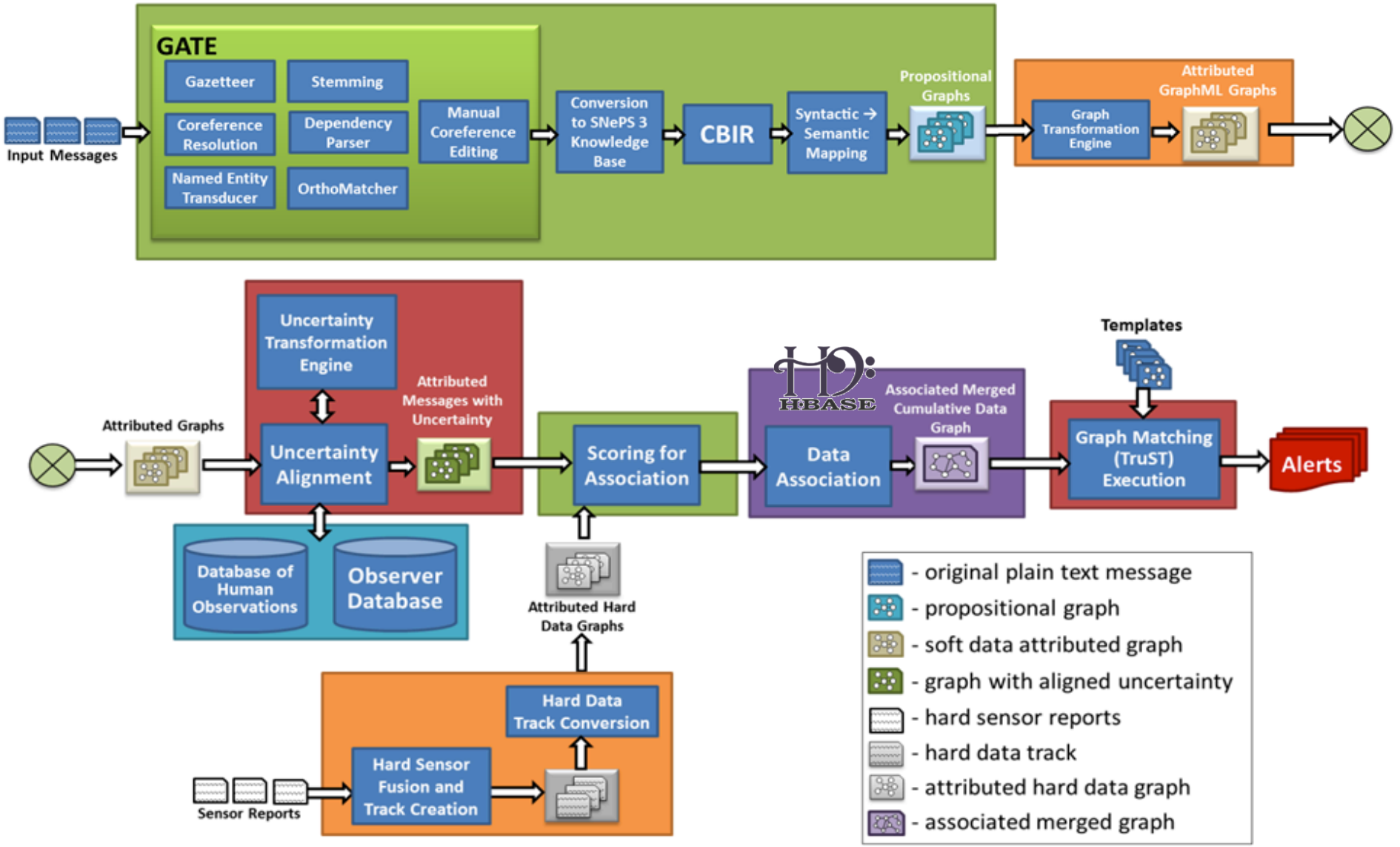
		Average Maximum Reducer Times											
		Iteration 1				Iteration 2				Iteration 3			
	Task Count	16	32	64	92	16	32	64	92	16	32	64	92
Data Access Method	HBase	674	728	989	1454	972	953	1033	1355	1440	1173	1274	1587
	Postgres	615	614	869	940	1145	891	1039	1444	1530	1358	1591	1434

Link Analysis Testing Results



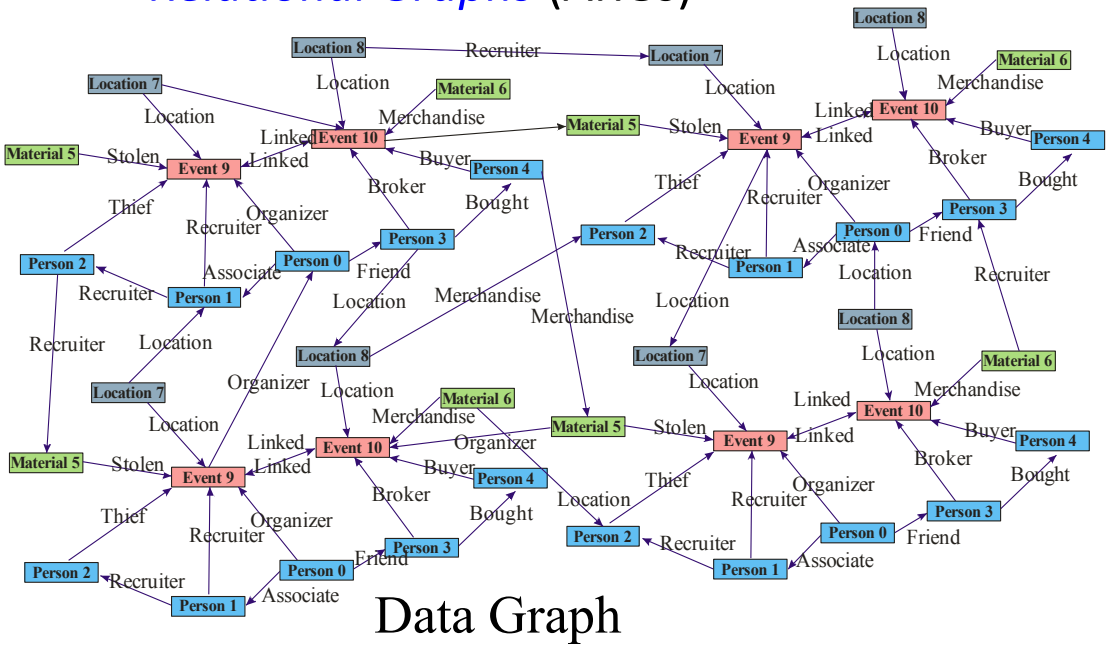
(2) Graph Matching

Hard-Soft Information Fusion MURI Architecture

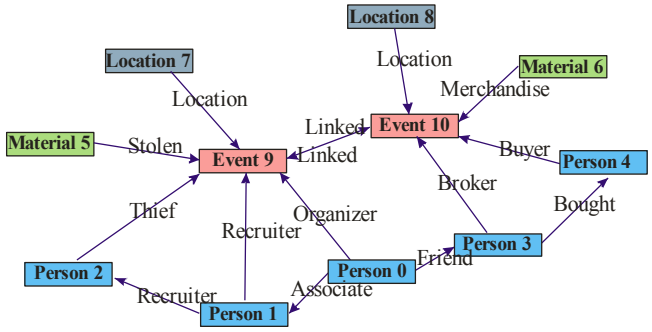


Data Graph and Template Graph

- Decision-maker often encounter complex uncertain situations and they need to develop *relationship* between them.
- One of the techniques they use often is to simply draw *Attributed Relational Graphs* (ARGs)



Data Graph



Template Graph

Definitions and Notation

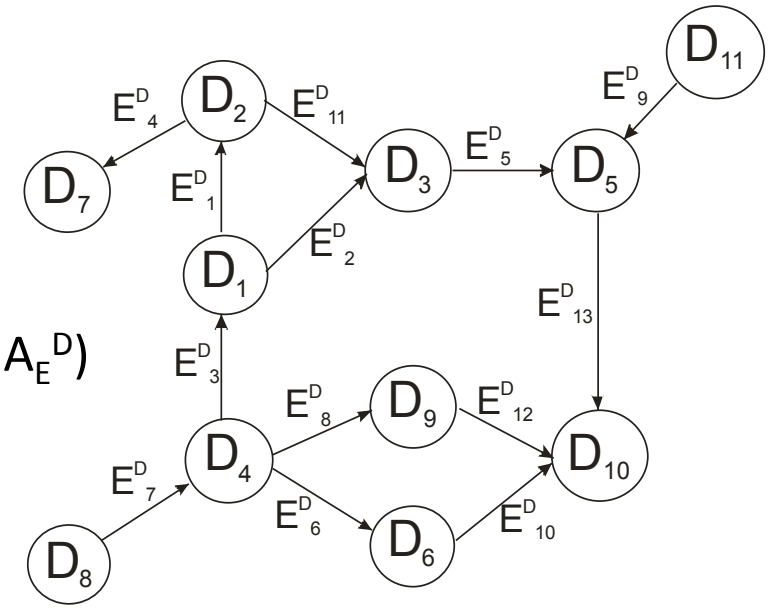
- Attributed Graph Structure

$$G = (V, E, A_v, A_E)$$

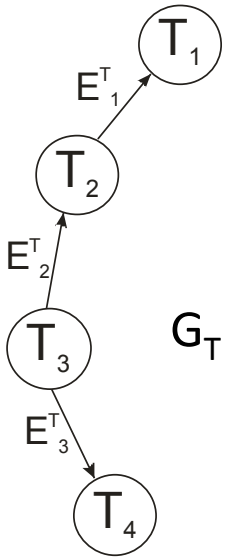
where V - the set of nodes; E - the set of arcs;

A_v - the set of node attributes; A_v - the set of arc attributes.

$G_D = (V_D, E_D, A_V^D, A_E^D)$



Data Graph



Template

$G_T = (V_T, E_T, A_V^T, A_E^T)$

Similarity Functions for Triplets

- Triplets:

Template Data Graph

- One-to-one Scores of Nodes and Arcs:

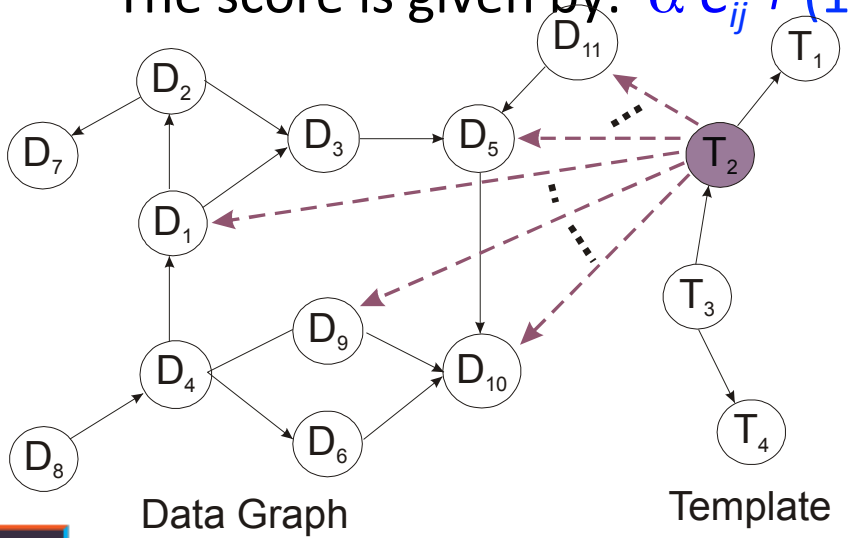
$S(v_i, v_j) = 0.9$ $S(v_h, v_k) = 0.7$ $S(e_{ih}, e_{jk}) = 0.8$

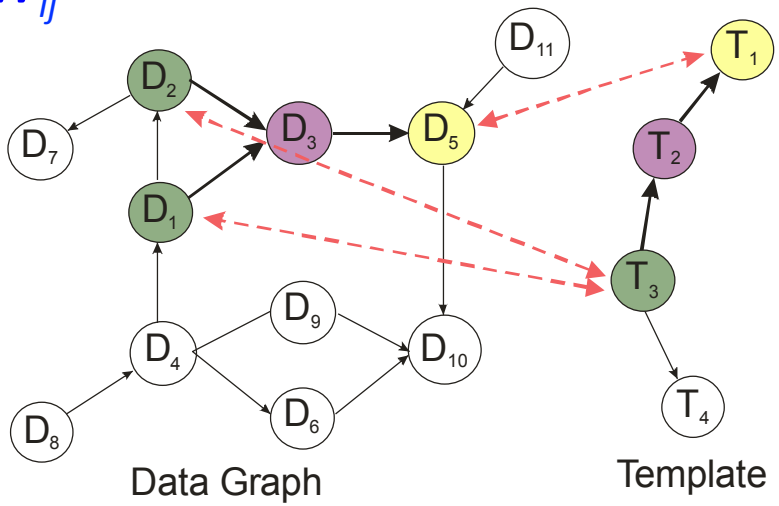
- Similarity Functions for Triplets:

$$f_{ihjk} = f(S(v_i, v_j), S(v_h, v_k), S(e_{ih}, e_{jk})) = 0.8$$

1-Hop Neighborhood Matching

- Algorithm
  - Step 1: Compute a node score, denoted as C_{ij} , for each node i in the template graph to each node j in the data graph.
  - Step 2: Compute the scores, denoted as W_{ij} , for the 1-Hop neighbors of each root node pair.
- The score is given by: $\alpha\, C_{ij} + (1-\alpha)\, W_{ij}$



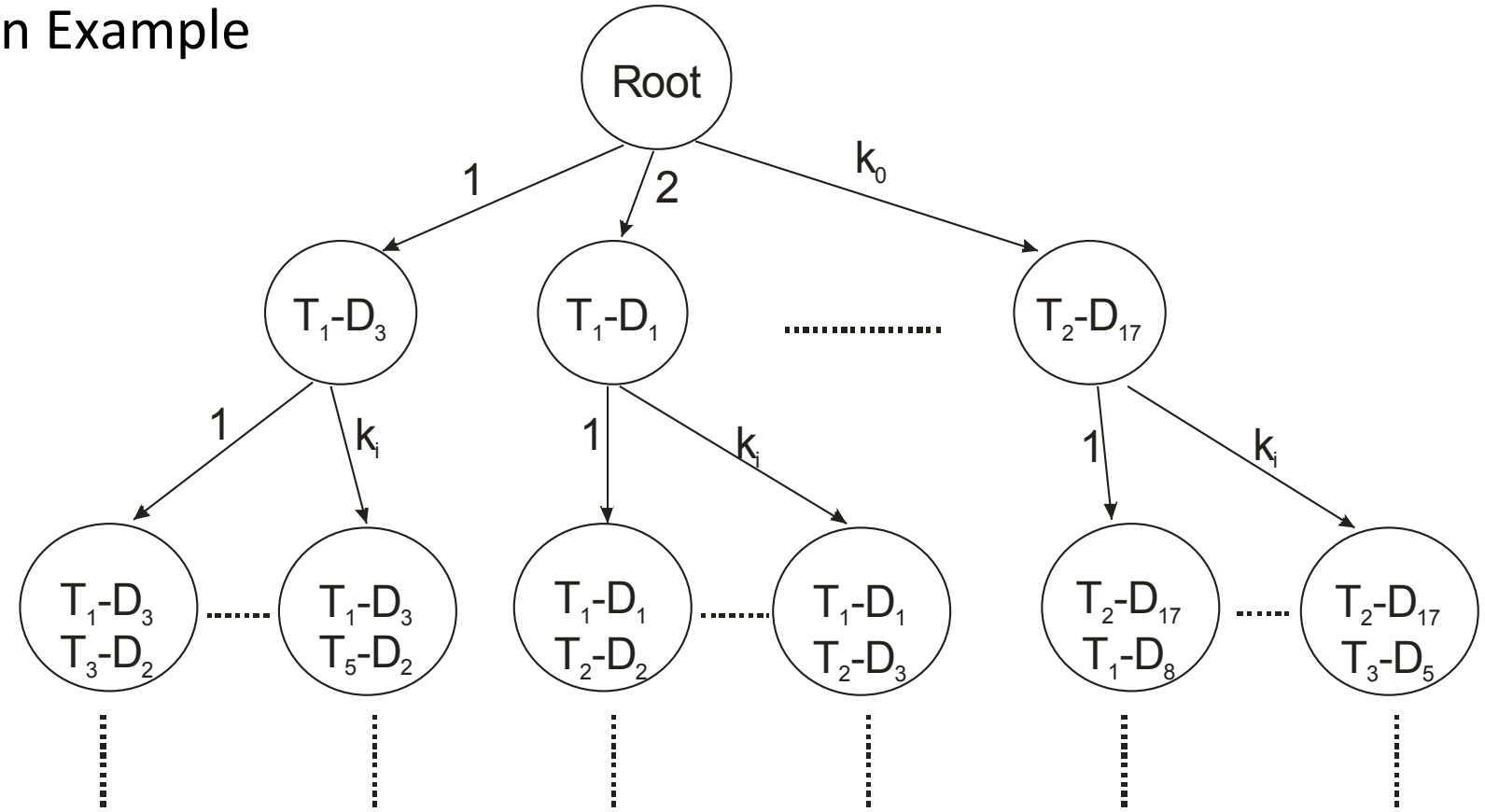


TruST (Truncated Search Tree) Algorithm

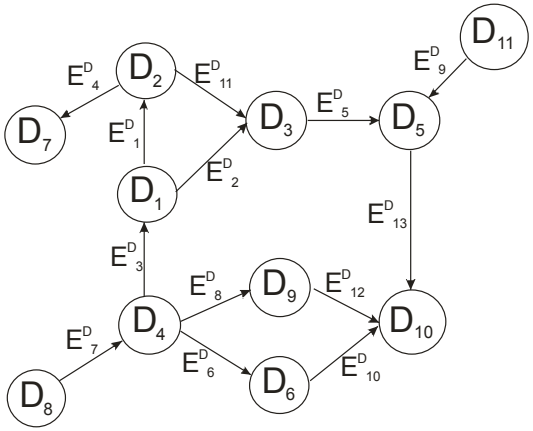
- Advantages:
  - Easier to *implement*, faster to *execute* and requires *less computing resources*.
- Disadvantages:
  - *Not* guaranteed to yield an *optimal solution*.
- Selection strategy:
  - Beam Search (Breadth-first with limited size)
- Parameters of controlling state space
  - k_0 : The number of child problems of the root.
  - k_i : The number of child problems of the problem at level i-1.
  - β : The total number of problems.
  - δ : The total number of levels.

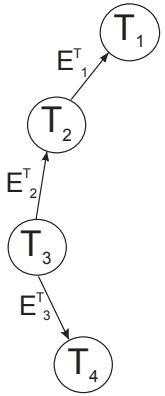
TruST Algorithm (cont'd)

- An Example

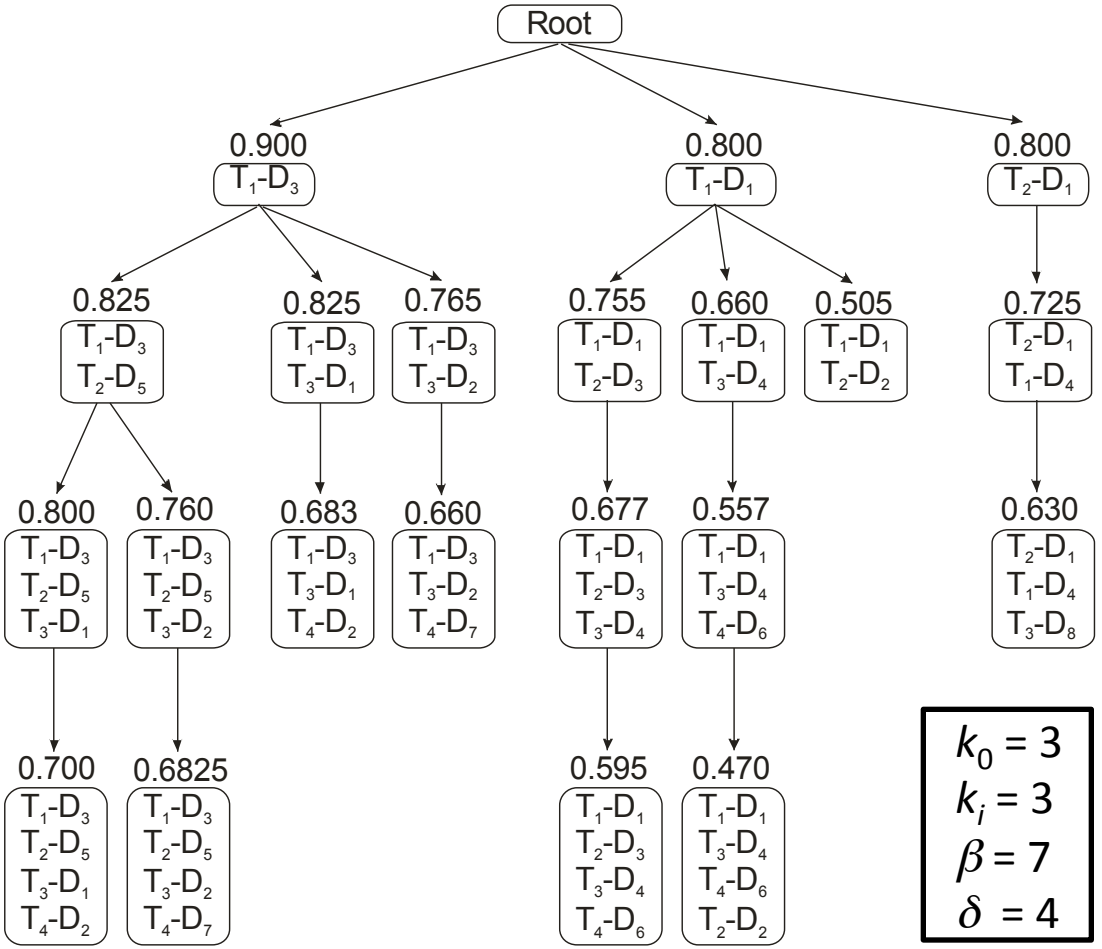


An Illustrative Example





T_1		T_2		T_3		T_4	
D_3	0.9	D_1	0.8	D_6	0.77	D_{10}	0.6
D_1	0.8	D_5	0.75	D_1	0.75	D_7	0.45
D_9	0.75	D_3	0.71	D_2	0.63	D_2	0.4
D_6	0.73	D_7	0.68	D_4	0.52	D_6	0.35
D_5	0.7	D_9	0.55	D_8	0.44	D_4	0.28
D_4	0.65	D_6	0.53	D_9	0.27	D_8	0.18
D_2	0.4	D_4	0.23	D_3	0.23	D_1	0.15
D_7	0.17	D_2	0.21	D_7	0.17	D_3	0.14
D_{10}	0.15	D_{10}	0.13	D_{11}	0.11	D_5	0.12
D_{11}	0.11	D_8	0.09	D_5	0.09	D_9	0.1
D_8	0.09	D_{11}	0.05	D_{10}	0.05	D_{11}	0.07

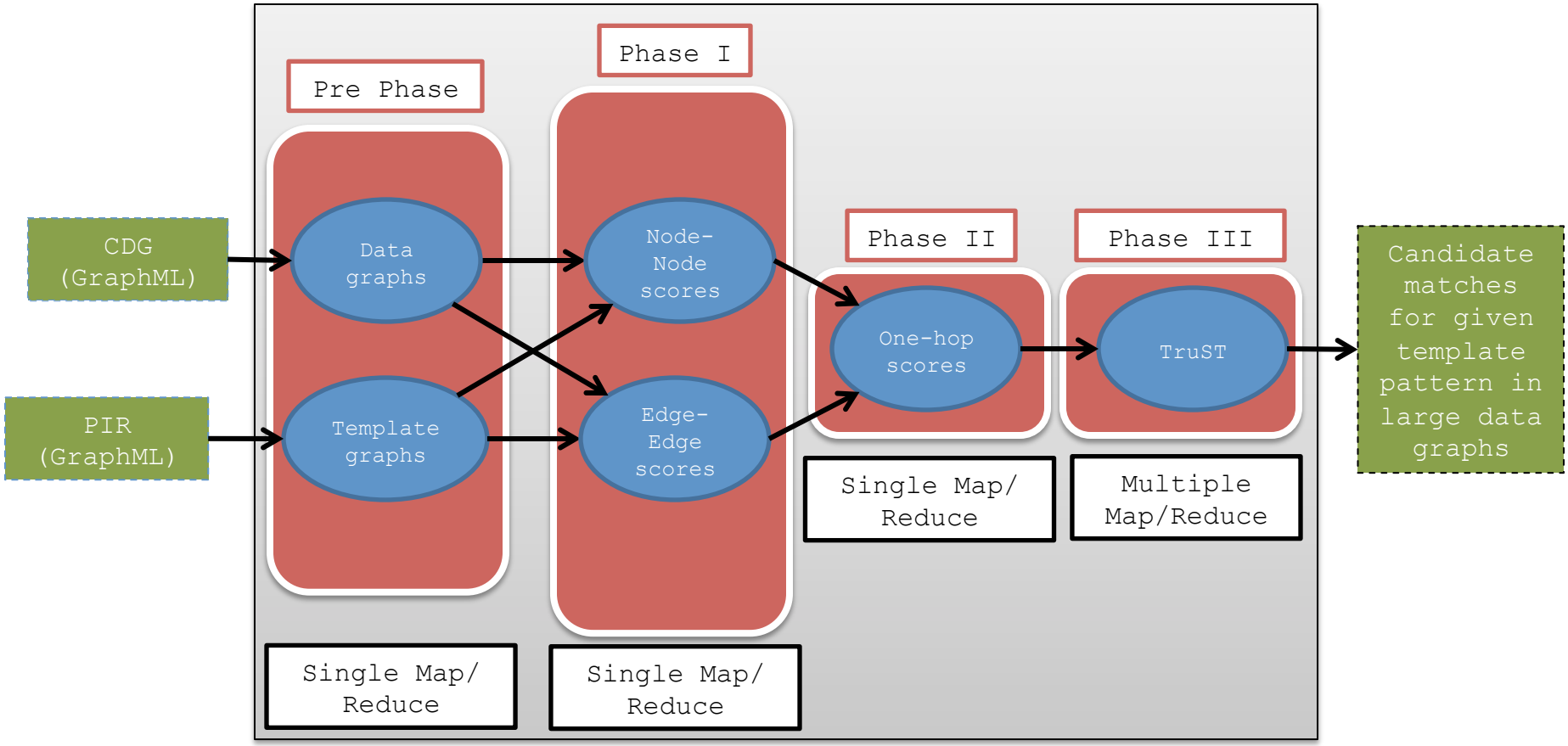


$k_0 = 3$
$k_i = 3$
$\beta = 7$
$\delta = 4$

Technical Approach
Parallel Graph Analytics
TruST Phases

Phase	Significance	Job Type	Input	Output
Pre Phase	Converts Cumulative Data Graph (CDG) and PIR to data graph and template graph representations stored in form of HBase tables.	Single Map Reduce job for the two importers	• CDG • PIR	• Data graph table • Template graph table
Phase I	Calculates similarity between the template and data graph nodes and edges using attribute information.	Single Map Reduce job	• Data graph table • Template graph table	• Similarity scores table
Phase II	Calculates neighborhood between the template and data graph nodes using adjacency information and node and edge similarity scores.	Single Map Reduce job	• Data graph table • Template graph table	• Expanded Similarity scores table containing one hop scores.
Phase III	Actual trust algorithm (template graph pattern matching with data graph).	Multiple Map Reduce jobs run iteratively	• Initial set of associations candidates. • Data graph table • Template graph table • Similarity scores table	• Expanded set of associations candidates formed using one hop scores and data graph and template graph adjacency information.

Technical Approach
Parallel Graph Analytics
TruST Architecture



(3) Graph Association

Graph Association

Data Association

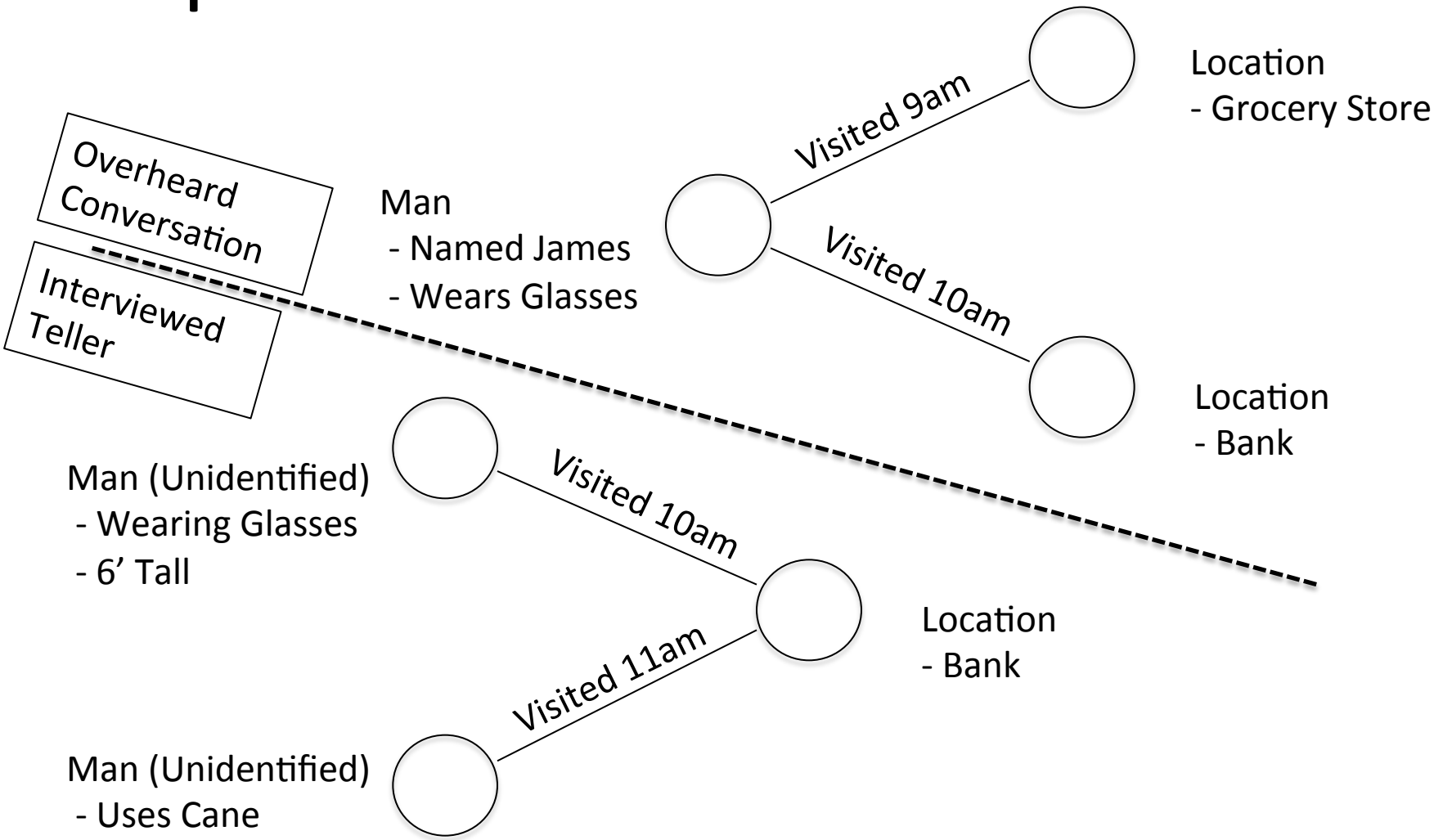
The problem of data association is to take information from multiple (noisy) data sources and identify any overlapping information.

Motivation for Data Association:

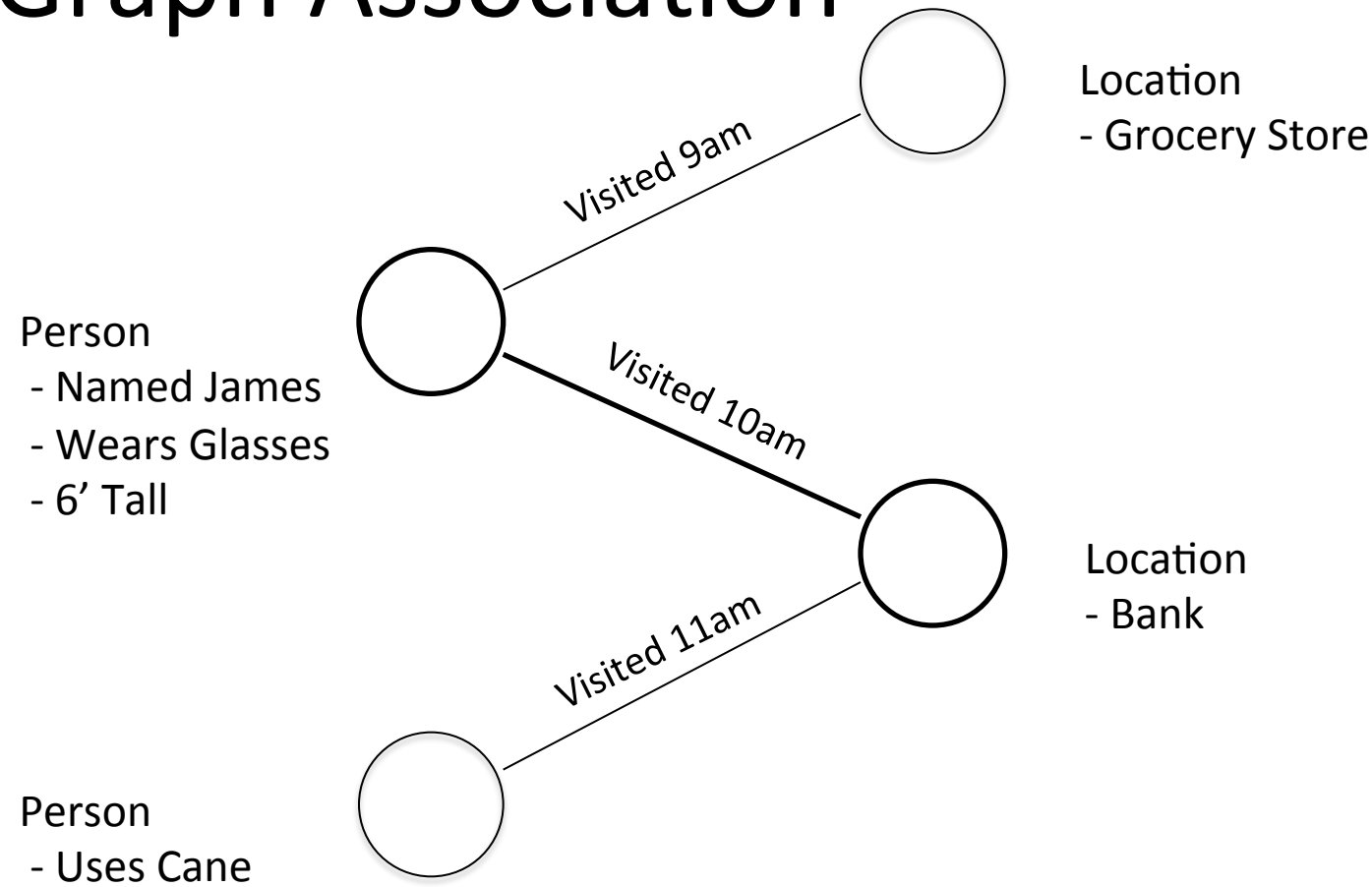
- Information Gain
  - Provides a means to find answers that span multiple sources.

- Dimensionality Reduction
  - Reduction in duplicate entities improves the efficiency of analytics.

Graph Association



Graph Association



Relational information helps improve the confidence that the "James" and "Bank" entities are the same.

References

- Tauer, G. and Nagi, R. "A Map-Reduce Lagrangian Heuristic for Multidimensional Assignment Problems with Decomposable Costs," *Parallel Computing*, November 2013, Vol. 39(11), pp. 653-668.
- Gross, G.A., Nagi, R. and Sambhoos, K. "A Fuzzy Graph Matching Approach in Intelligence Analysis and Maintenance of Continuous Situational Awareness," *Information Fusion*, July 2014, Vol. 18, pp. 43-61.
- Tauer, G., Nagi, R. and Sudit, M. "The Graph Association Problem: Mathematical Models and a Lagrangian Heuristic," *Naval Research Logistics*, April 2013, Vol. 60(3), pp. 251–268.
- Sambhoos, K., Nagi, R., Sudit, M. and Stotz, A. "Enhancements to High Level Data Fusion using Graph Matching and State Space Search," *Information Fusion*, 2010, Vol. 11(4), pp. 351-364.
- Date, K., Gross, G.A., Khopkar, S., Nagi, R. and Sambhoos, K., "Data Association and Graph Analytical Processing of Hard and Soft Intelligence Data," *16th International Conference on Information Fusion*, Istanbul, Turkey, 9-12 July 2013.

Concluding thoughts

- Big Data is everywhere
- Significant business value to be derived
- Algorithms/Methods for Big Data change
- Big Graph Analysis remains challenging
  - Graph Stores
  - Hardware Architectures
  - New algorithms