Graph Search in the Big Data Era

马帅　　计算机学院

北京航空航天大學
BEIHANG UNIVERSITY

973 Grant on Big Data at Beihang

- 网络信息空间大数据计算的基础研究(2014-2018)
  - Chief Scientist: Prof. Jinpeng Huai.
  - 8 institutes involved
  - Focus on “computing theory and practice on Big Data”

北京航空航天大學 BEIHANG UNIVERSITY

国防科学技术大学 National University of Defense Technology

ISCAS

HUAWEI

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CNCERT/CC 国家计算机网络应急技术处理协调中心

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Baidu 百度

RCBD at Beihang

- International Research Centre on Big Data (RCBD)
  - Founded in September, 2012.
  - Led by **Prof. Wenfei Fan** (ACM Fellow, Fellow of the Royal Society of Edinburgh, Scotland) .
- Research Topics
  - Big Data Analysis: Theory and Applications
  - Data Quality: The Other Side of Big Data
  - Querying Big Data beyond MapReduce
  - Querying Big Social Data

Big Data is a Big Deal

What is Big Data?

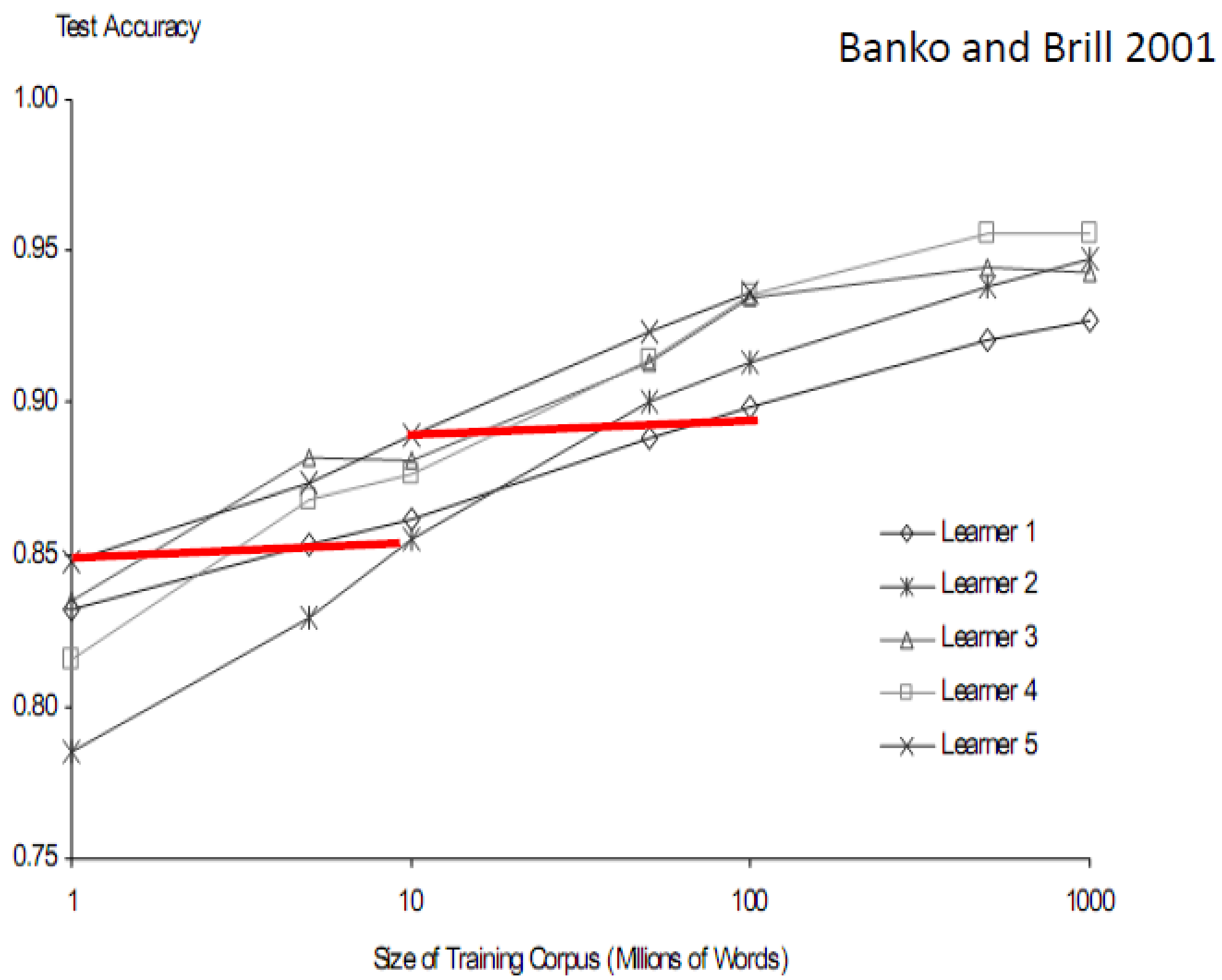
- Big Data refers to datasets that grow so large that it is difficult to capture, store, manage, share, analyze and visualize with those traditional (database) software tools
  - Wikipedia

"Big data" becomes a buzz word, and the focus of both industrial and academic communities!

Human vs. Computer + Big Data

- IBM "Watson" system challenges humans at Jeopardy!
  - In 2011, Watson beat former winners Brad Rutter and Ken Jennings. Watson received the first prize of $1 million.
  - Compared with "Deep Blue", ""Watson" is equipped with Big Data!

More Data Beats Better Algorithms



Kepler's Third Law of Planetary Motion

- The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit

Planet	Period (yr)	Ave. Dist. (au)	T^2/R^3 (yr^2/au^3)
Mercury	0.241	0.39	0.98
Venus	.615	0.72	1.01
Earth	1.00	1.00	1.00
Mars	1.88	1.52	1.01
Jupiter	11.8	5.20	0.99
Saturn	29.5	9.54	1.00
Uranus	84.0	19.18	1.00
Neptune	165	30.06	1.00
Pluto	248	39.44	1.00

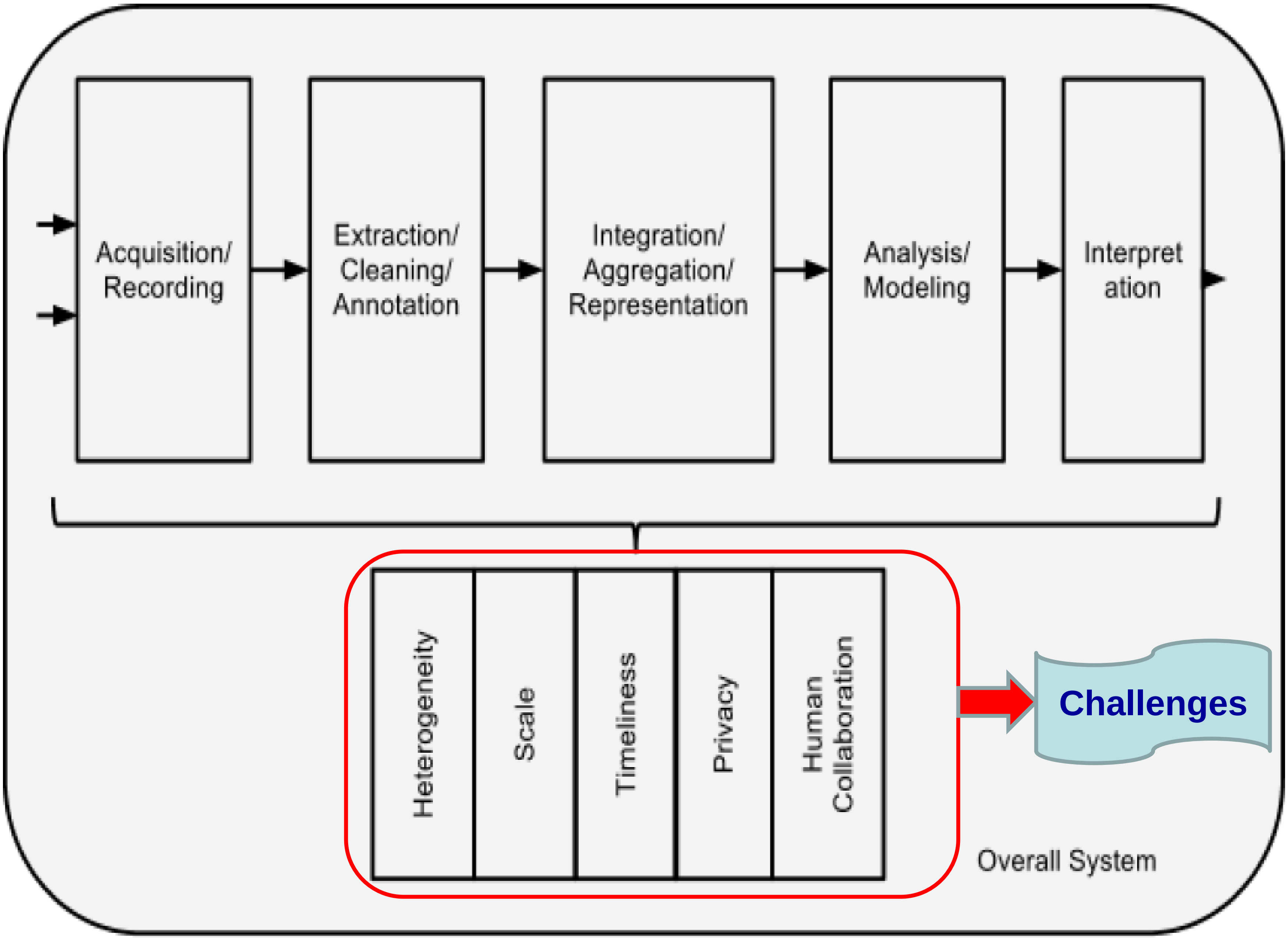
Challenges and Opportunities with Big Data

- A community white paper developed by leading researchers across US

Divyakant Agrawal, UC Santa Barbara
Philip Bernstein, Microsoft
Elisa Bertino, Purdue Univ.
Susan Davidson, Univ. of Pennsylvania
Umeshwar Dayal, HP
Michael Franklin, UC Berkeley
Johannes Gehrke, Cornell Univ.
Laura Haas, IBM
Alon Halevy, Google
Jiawei Han, UIUC
Alexandros Labrinidis, Univ. of Pittsburgh

Sam Madden, MIT
Yannis Papakonstantinou, UC San Diego
Jignesh M. Patel, Univ. of Wisconsin
Raghu Ramakrishnan, Yahoo!
Kenneth Ross, Columbia Univ.
Cyrus Shahabi, Univ. of Southern California
Dan Suciu, Univ. of Washington
Shiv Vaithyanathan, IBM
Jennifer Widom, Stanford Univ

A result of conversation lasted about 3 months (Nov. 2011 ~ Feb. 2012)

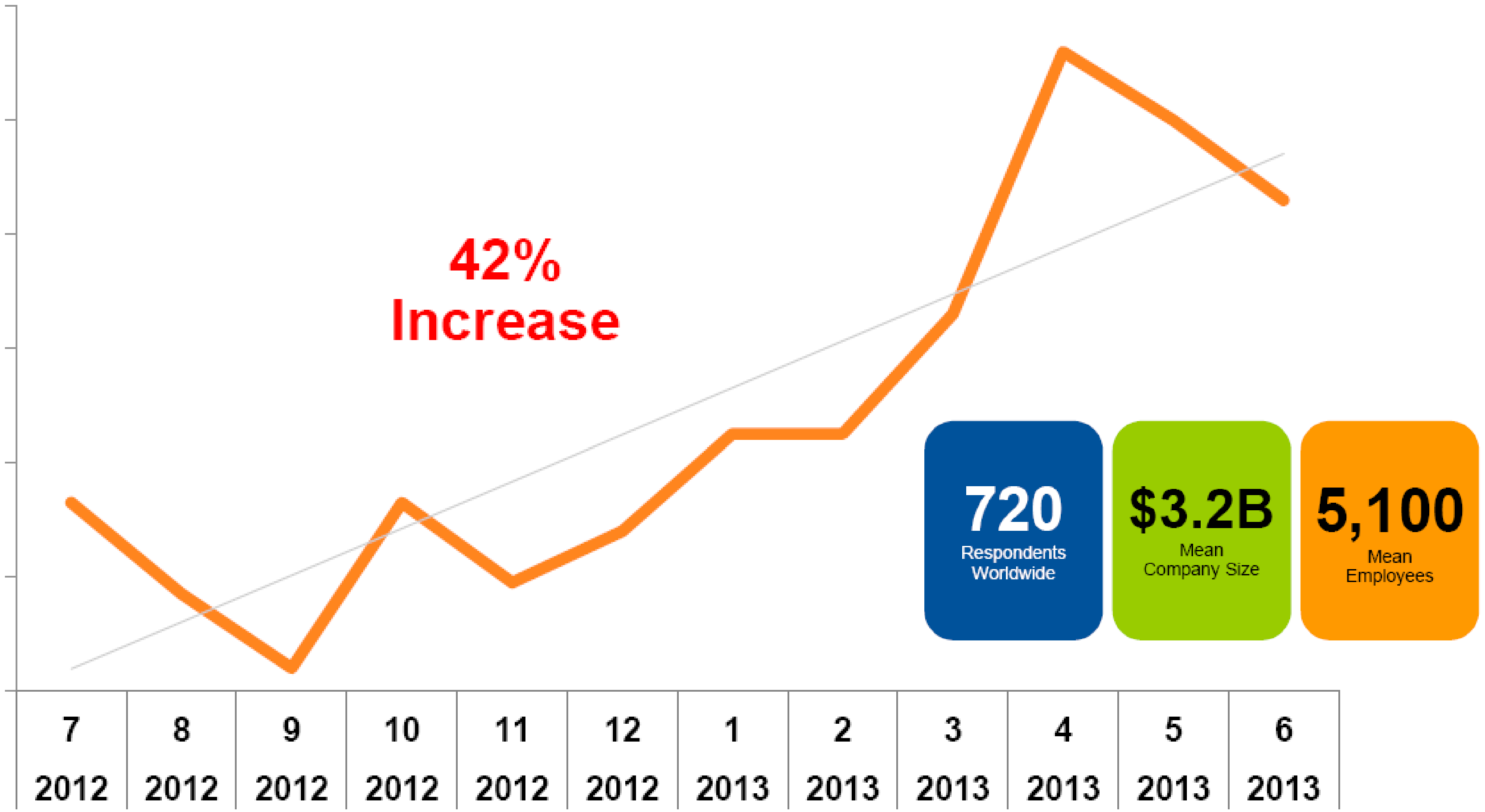


Acquisition/ Recording
Extraction/ Cleaning/ Annotation
Integration/ Aggregation/ Representation
Analysis/ Modeling
Interpret ation
Heterogeneity
Scale
Timeliness
Privacy
Human Collaboration
Challenges
Overall System

Gartner's Most Recent Report

Industries' Interests

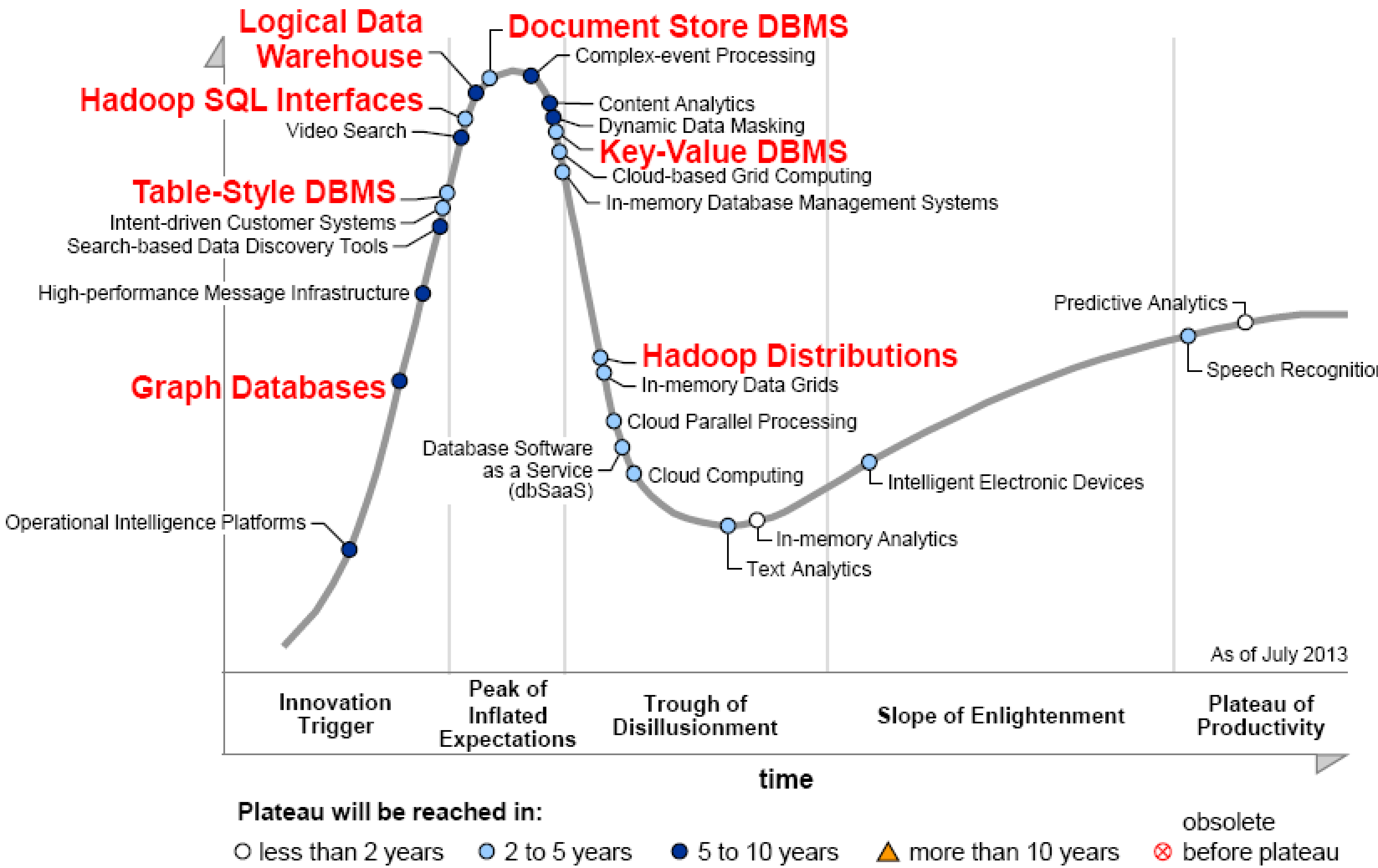
Client Inquiries — Information Management



Source: Information Management Team Inquiry Data, July 2012-June 2013

Gartner.

Big Data Techniques

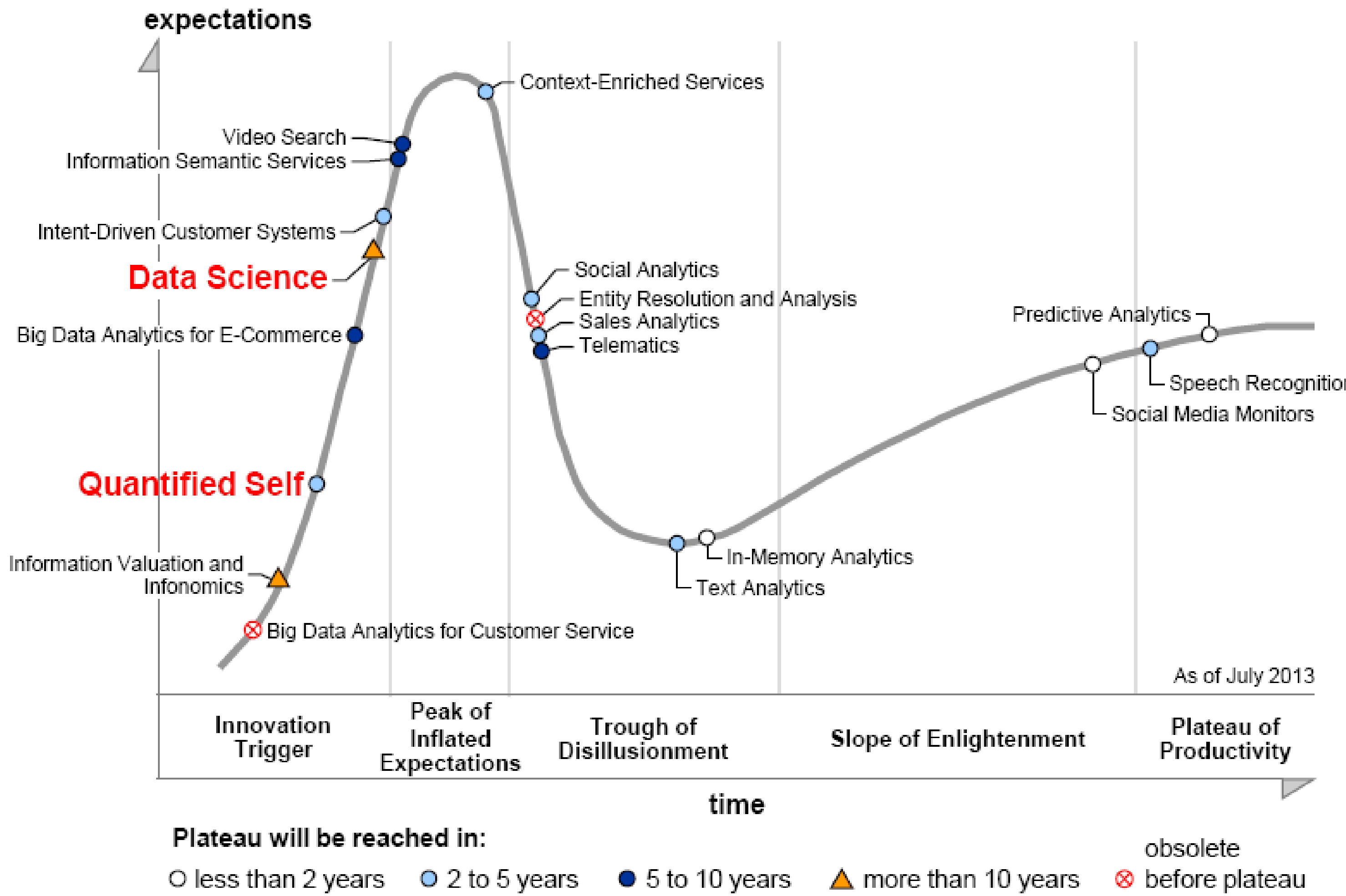


#GartnerSYM

Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

Gartner

Big Data Techniques



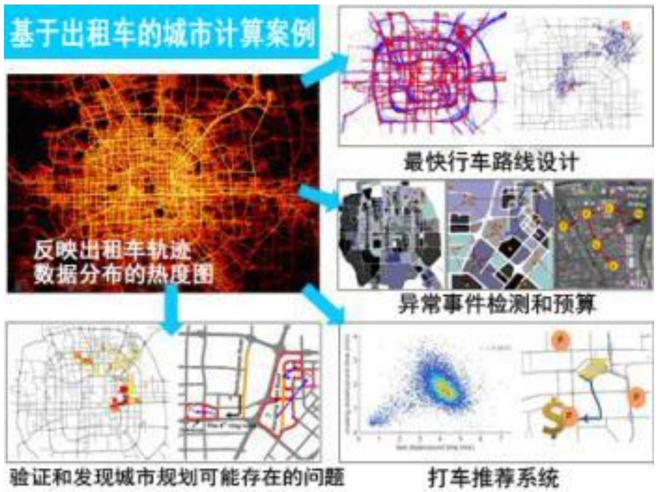
#GartnerSYM

Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

Gartner

Social Networks are Big Graphs

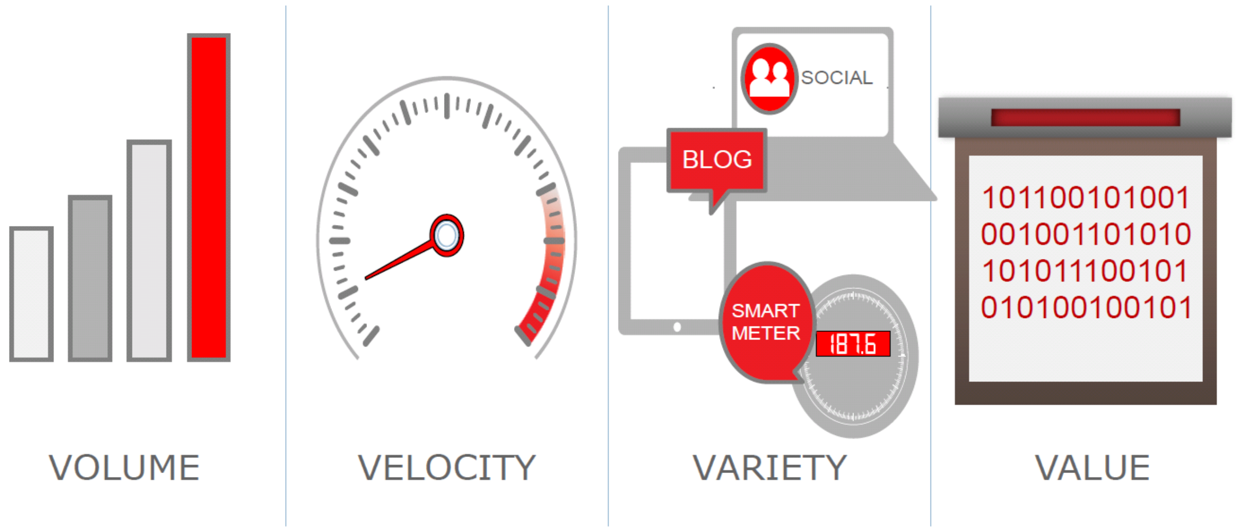
Social Networks are the New Media



金庸“被逝世”

Social networks are becoming an important way to get information in everyday life !

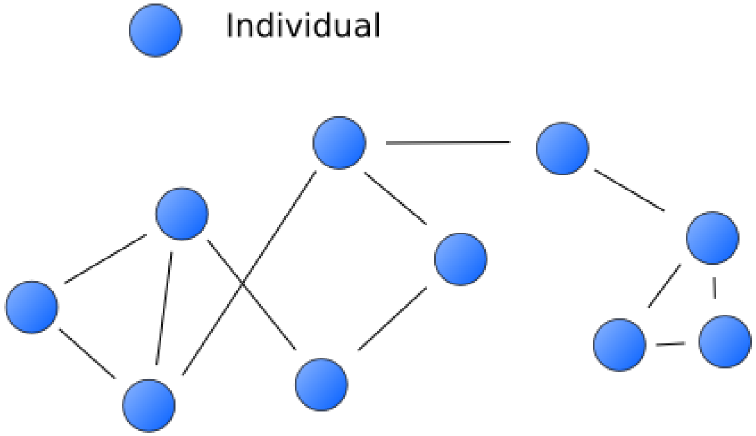
Social Networks are "Big Data"



Facebook:

- Volume： 10×10^8 users, 2400×10^8 photos, $10^4 \times 10^8$ page visits
- Velocity： 7.9 new users per second, over 60 thousands per day
- Variety： text (weibo, blogs) , figures, videos, relationships (topology)
- Value： 1.5×10^8 dollars in 2007, 3×10^8 dollars in 2008, $6 \sim 7 \times 10^8$ dollars in 2009, 10×10^8 dollars in 2010.
- Further, data are often dirty due to data missing and data uncertainty [1, 2]

Social Networks are Big Graphs

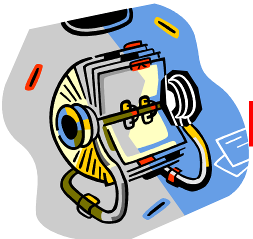


Social networks are graphs

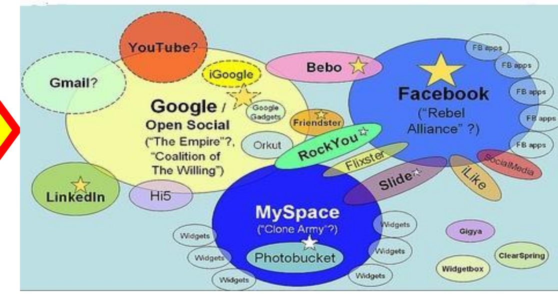
- The nodes are the people and groups
- The links/edges show relationships or flows between the nodes.

The Need for a Social Search Engine

The Need for a Social Search Engine



Graph Searching



File systems **Databases** **World Wide Web** **Social Networks**

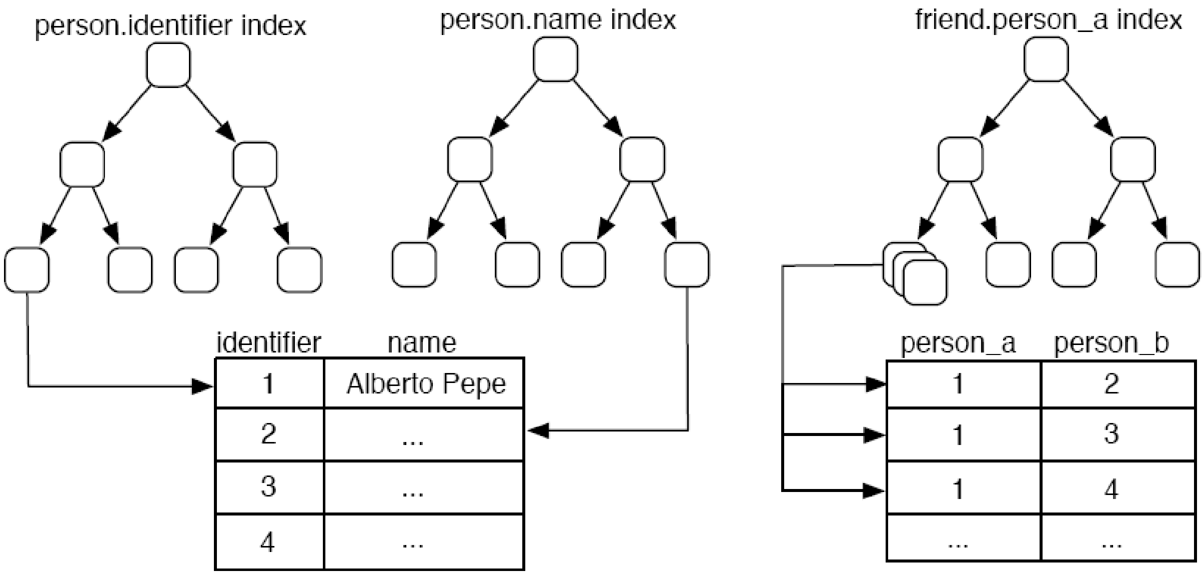
- File systems - 1960's： very simple search functionalities
- Databases - mid 1960's： SQL language
- World Wide Web - 1990's： keyword search engines
- Social networks - late 1990's:

Facebook launched "graph search" on 16th January, 2013

Assault on Google, Yelp, and LinkedIn with new graph search;
Yelp was down more than 7%

Graph search is a new paradigm for social computing!

Graph Search vs. RDBMS[3]



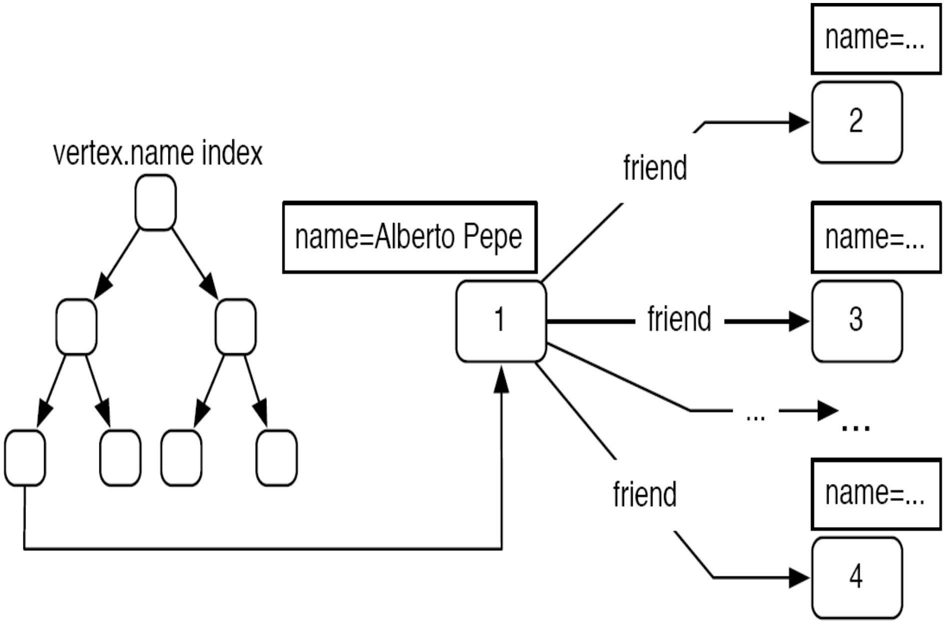
Query:
Find the name of all of Alberto Pepe's friends.

Step 1: The person.name index -> the identifier of Alberto Pepe. [O($\log_2 n$)]

Step 2: The friend.person index -> k friend identifiers. [O($\log_2 x$) : x<<m]

Step 3: The k friend identifiers -> k friend names. [O(k log2n)]

Graph Search vs. RDBMS[3]

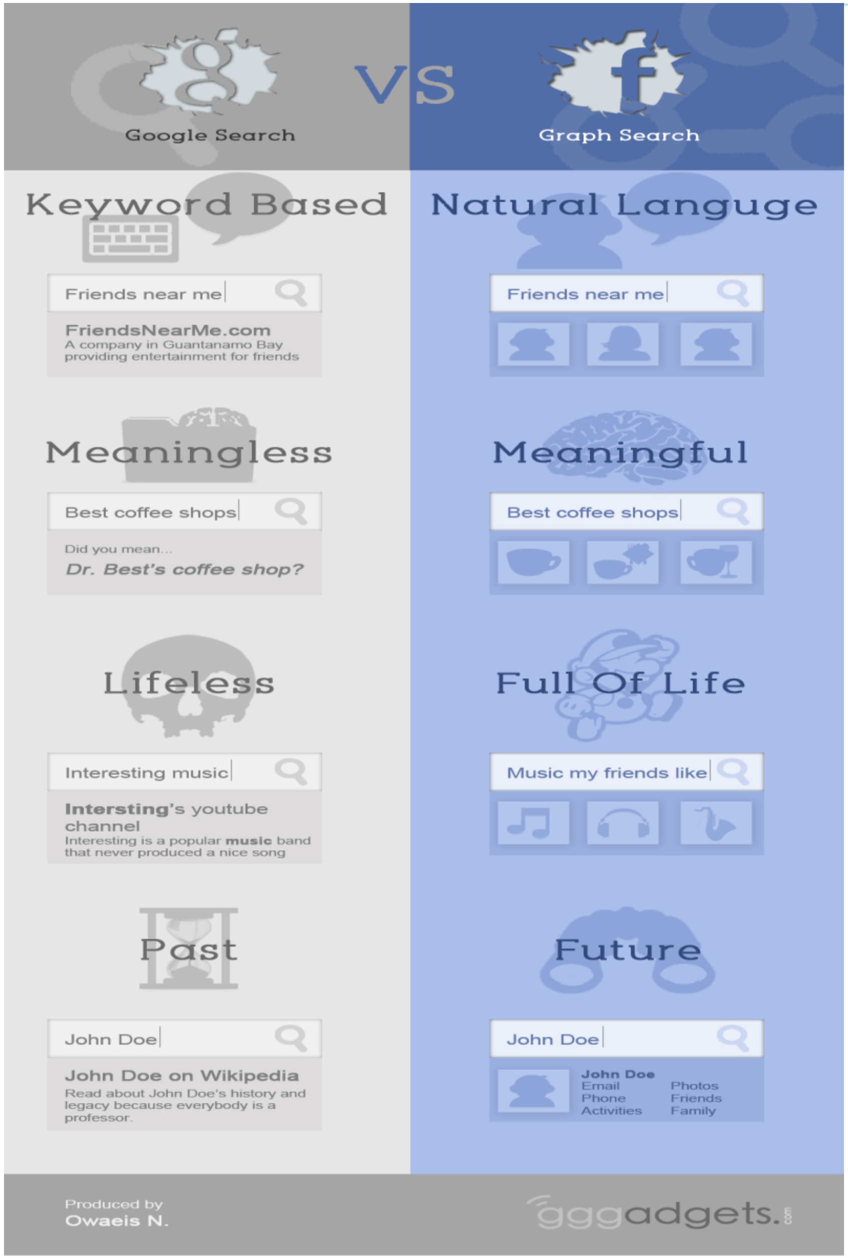


Query:
Find the name of all of Alberto Pepe's friends.

Step 1: The vertex.name index -> the vertex with the name Alberto Pepe. [O(log2n)]

Step 2: The vertex returned -> the k friend names. [O(k + x)]

Social Search vs. Web Search

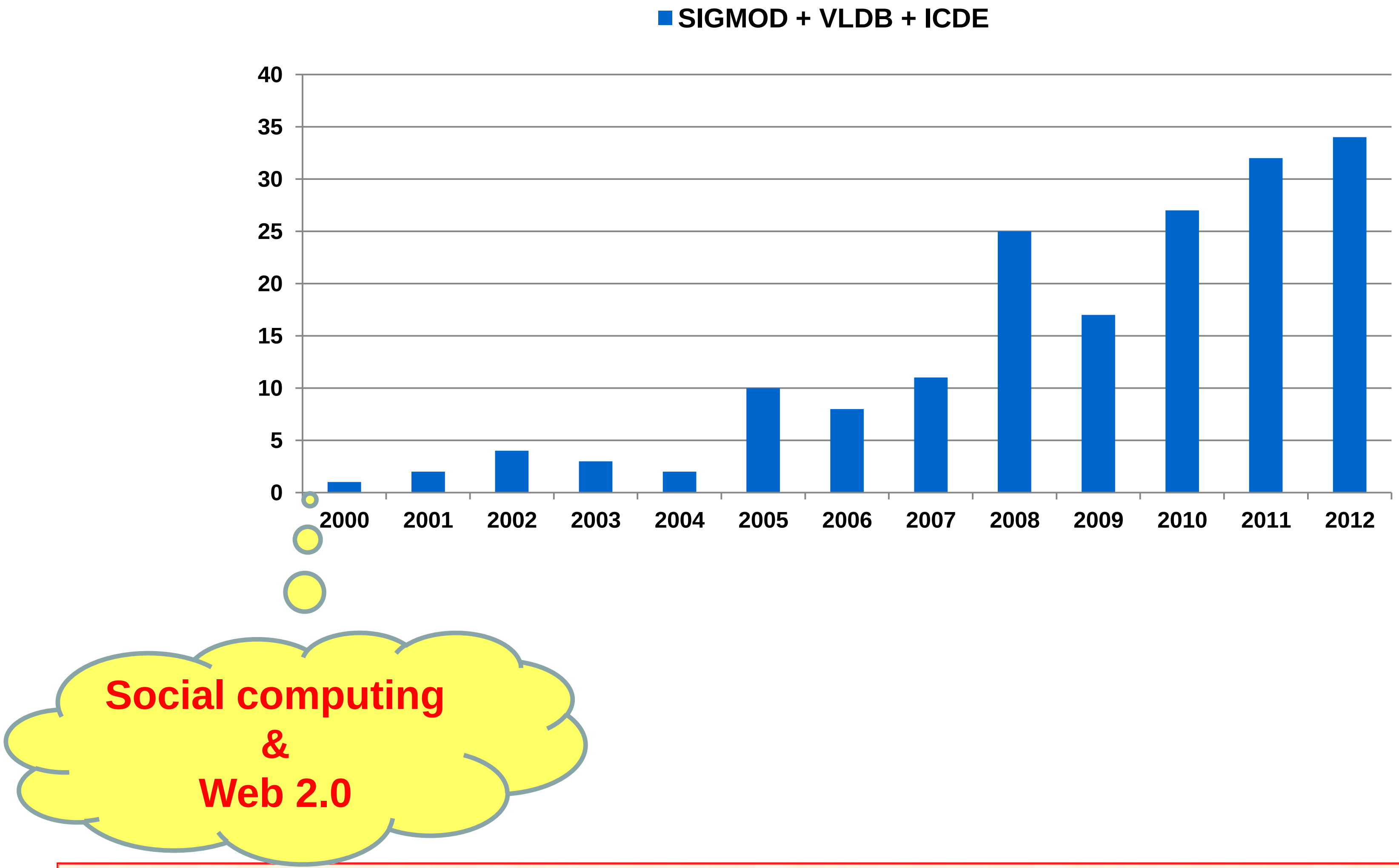


- Phrases、short sentences vs. key words only
- (Simple Web) pages vs. Entities
- Lifeless vs. Full of life
- History vs. Future

it's interesting, and over the last 10 years, people have been trained on how to use search engines more effectively. Keywords & Search In 2013: Interview With A. Goodman & M. Wagner

International Conference on Application of Natural Language to Information Systems (NLDB) started from 1995

Academia's Research Interests



DB people started working on graphs at around the same time!

Applications of Graph Search

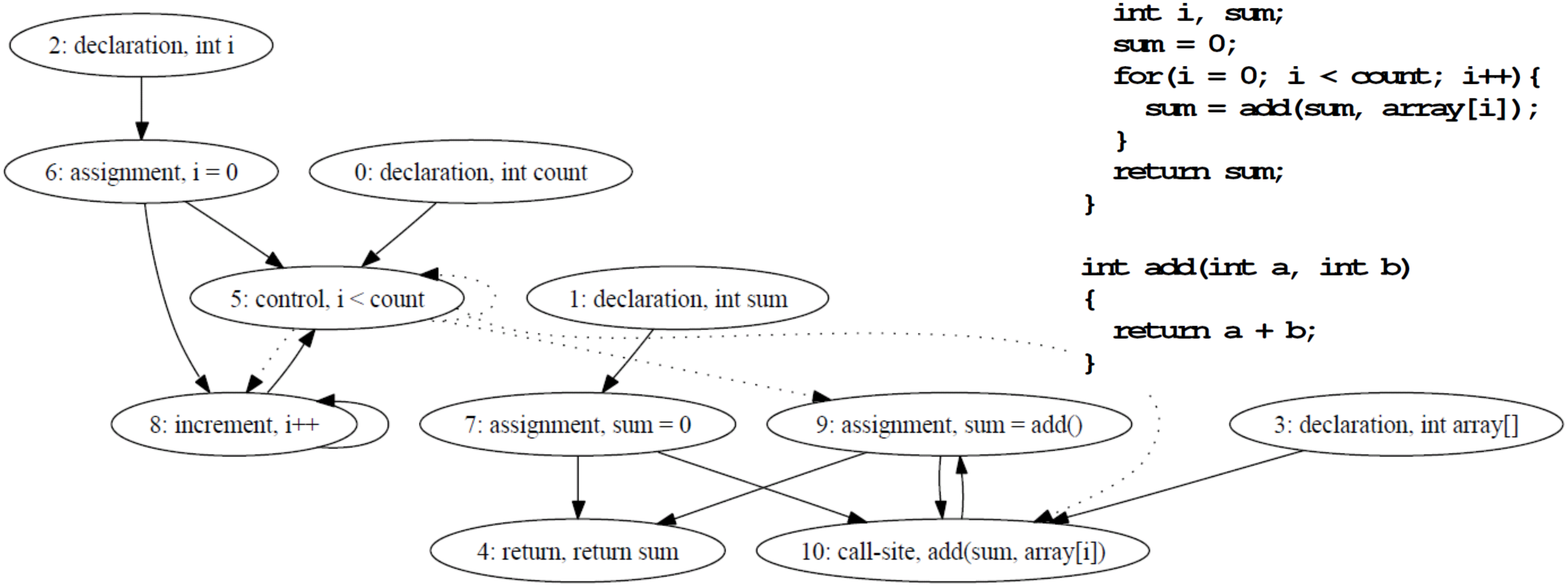
Application Scenarios

Software plagiarism detection [4]

- Traditional plagiarism detection tools may not be applicable for serious software plagiarism problems.
- A new tool based on graph pattern matching
  - Represent the source codes as program dependence graphs [5].
  - Use graph pattern matching to detect plagiarism.

```
int sum(int array[], int count)
{
  int i, sum;
  sum = 0;
  for(i = 0; i < count; i++){
    sum = add(sum, array[i]);
  }
  return sum;
}

int add(int a, int b)
{
  return a + b;
}
```



Application Scenarios

Recommender systems [6]

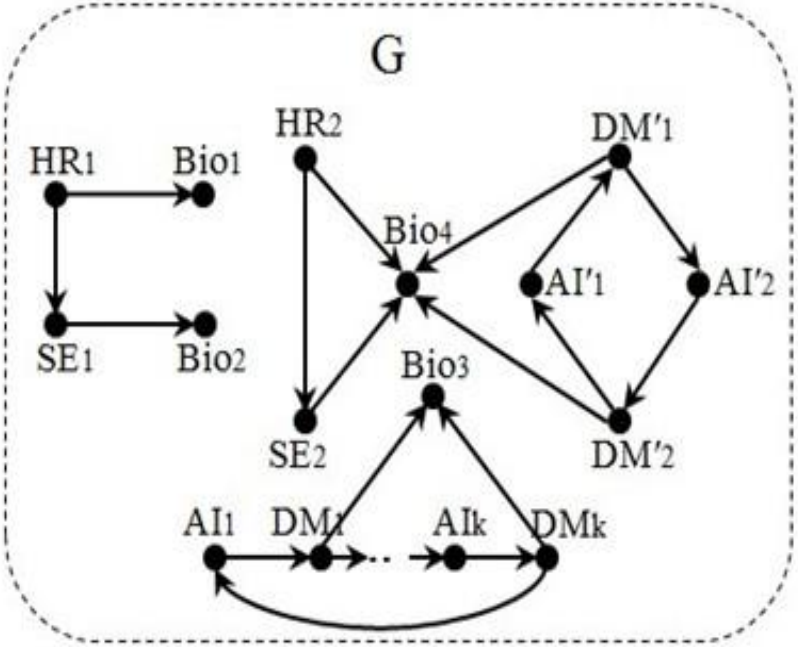
- Recommendations have found its usage in many emerging specific applications, such as social matching systems.
- Graph search is a useful tool for recommendations.

– A headhunter wants to find a biologist (Bio) to help a group of software engineers (SEs) analyze genetic data.

– To do this, (s)he uses an expertise recommendation network G, as depicted in G, where

✓ a node denotes a person labeled with expertise, and

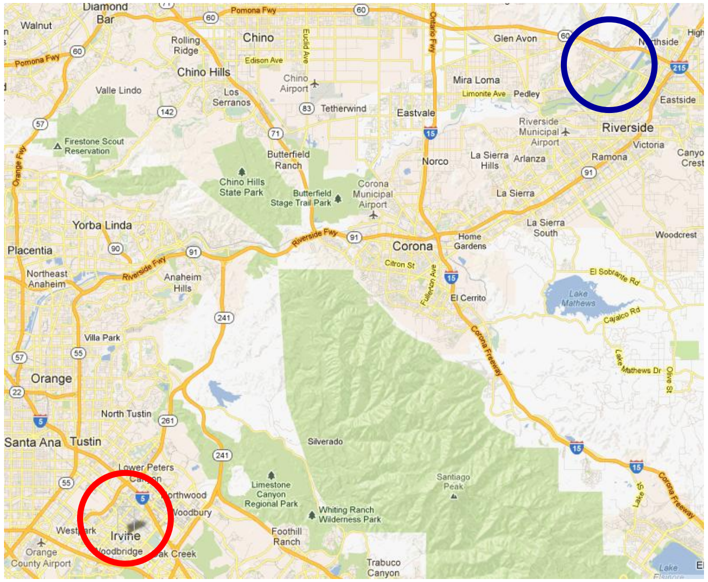
✓ an edge indicates recommendation, e.g., HR_1 recommends Bio_1, and AI_1 recommends DM_1



Application Scenarios

Transport routing [7,10]

- Graph search is a common practice in transportation networks, due to the wide application of Location-Based Services.
- **Example**: Mark, a driver in the U.S. who wants to go from Irvine to Riverside in California.
  - If Mark wants to reach Riverside by his **car** in the shortest time, the problem can be expressed as the shortest path problem. Then by using existing methods, we can get the shortest path from *Irvine, CA* to *Riverside, CA* traveling along State Route 261.
  - If Mark drives a **truck** delivering hazardous materials may not be allowed to cross over some bridges or railroad crossings. This time we can use a pattern graph containing specific route constraints (such as regular expressions) to find the optimal transport routes.



Challenges & Related Techniques

Challenges

- The amount of data has reached hundred millions orders of magnitude.

Graph search with high efficiency, striking a balance between its performance and accuracy.

- The data are updated all the time, and the updated amount of data daily reaches hundred thousands orders of magnitude.

Consider the dynamic changes and timing characteristics of data.

- Same with traditional relational data, there exists data quality problems such as data uncertainty and data missing in the new applications.

Solve the data quality problems.

Distributed Processing

- Real-life graphs are typically way too large:
  - Yahoo! web graph: 14 billion nodes
  - Facebook: over 0.8 billion users

It is NOT practical to handle large graphs on single machines

- Real-life graphs are naturally distributed:
  - Google, Yahoo! and Facebook have large-scale data centers

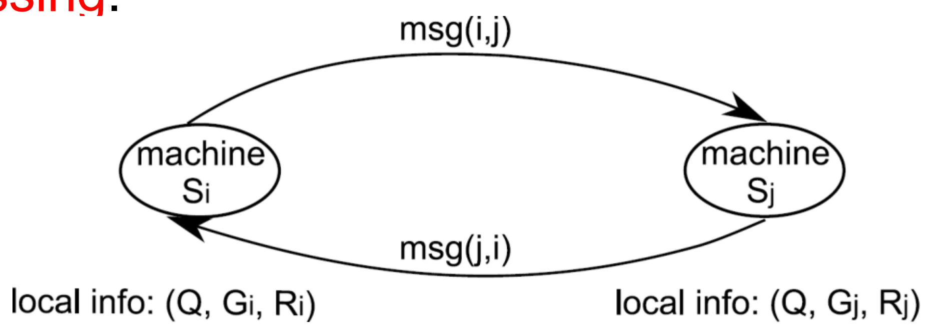
Distributed graph processing is inevitable

It is nature to study "distributed graph search"!

Distributed Processing

Model of Computation [3]:

- A cluster of identical machines (with one acted as coordinator);
- Each machine can directly send arbitrary number of messages to another one;
- All machines co-work with each other by local computations and message-passing.



Complexity measures:

1. **Visit times**: the maximum visiting times of a machine (**interactions**)
2. **Makespan**: the evaluation completion time (**efficiency**)
3. **Data shipment**: the size of the total messages shipped among distinct machines (**network band consumption**)

Incremental Techniques

Google **Percolator** [9]:

- Converting the indexing system to an incremental system,
- Reduce the average document processing latency by a factor of 100
- Process the same number of documents per day, while reducing the average age of documents in Google search results by 50%.

It is a great waste to compute everything from scratch!

Data Preprocessing

- Data Sampling
  - Instead of dealing with the entire data graphs, it reduces the size of data graphs by sampling and allows a certain loss of precision.
  - In the sampling process, ensure that the sampling data obtained can reflect the characteristics and information of the original data graphs as much as possible.
- Data Compression
  - It generates small graphs from original data graphs that preserve the information only relevant to queries.
  - A specific compression method is applied to a specific query application, such that data graph compression is not universal for all query applications.
  - Reachability query, Neighbor query

Data Preprocessing

- Indexing
- There are mainly three standards for measuring the goodness of an indexing method.
  - The space of a graph index
  - Establishing time for a graph index
  - Query time with a graph index
- Data Partitioning
  - Partition a data graph to relatively “small” graphs
  - Hash function is a simple approach for random partitioning.
  - There are well established tools, e.g. Metis [11].

Related Publications

[1] **Shuai Ma**, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Strong Simulation: Capturing Topology in Graph Pattern Matching. ACM TODS, 2014, to appear.

[2] Weiren Yu, Charu Aggarwal, **Shuai Ma***, and Haixun Wang, On Anomalous Hot Spot Discovery in Graph Streams, ICDM 2013.

[3] **Shuai Ma**, Yang Cao, Jinpeng Huai, and Tianyu Wo, Distributed Graph Pattern Matching, WWW 2012.

[4] **Shuai Ma**, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching, VLDB 2012.

[5] Wenfei Fan, Jianzhong Li, **Shuai Ma***, Nan Tang, and Yinghui Wu, Adding Regular Expressions to Graph Reachability and Pattern Queries, ICDE 2011.

[6] Wenfei Fan, Jianzhong Li, **Shuai Ma**, Nan Tang, Yinghui Wu, and Yunpeng Wu, Graph Pattern Matching: From Intractable to Polynomial Time, VLDB 2010.

[7] Wenfei Fan, Jianzhong Li, **Shuai Ma**, Hongzhi Wang, and Yinghui Wu, Graph Homomorphism Revisited for Graph Matching, VLDB 2010.

[8] Wenfei Fan, Jianzhong Li, **Shuai Ma***, Nan Tang, and Yinghui Wu, Adding Regular Expressions to Graph Reachability and Pattern Queries. FCS, Volume 6, Number 3, 313-338, 2012 (Invited).

[9] **Shuai Ma**, Jia Li, Xudong Liu, and Jinpeng Huai, 大数据时代的图搜索技术，《信息通信技术》,第6期, 41-55, 2013. (Invited).

[10] **Shuai Ma**, Yang Cao,Tianyu Wo, and Jinpeng Huai, 社会网络与图匹配查询,《中国计算机学会通讯》第8卷第4期20-24, 2012. (Invited).

[11] **Shuai Ma**, Jianxin Li, and Chunming Hu, 大数据科学与工程的挑战与思考，《中国计算机学会通讯》第8卷第9期, 22-28, 2012. (Invited).

[12] **Shuai Ma**, Jia Li, Xudong Liu, and Jinpeng Huai, 图查询：社会计算时代的新型搜索,《中国计算机学会通讯》第8卷第11期, 26-31, 2012. (Invited).

References

[1] Eytan Adar and Christopher Re, Managing Uncertainty in Social Networks, IEEE Data Eng. Bull., pp.15-22, 30(2), 2007.

[2] Gueorgi Kossinets, Effects of missing data in social networks. Social Networks 28:247-268, 2006.

[3] Marko A. Rodriguez, Peter Neubauer: The Graph Traversal Pattern. Graph Data Management 2011: 29-46.

[4] Chao Liu, Chen Chen, Jiawei Han and Philip S. Yu, GPLAG: detection of software plagiarism by program dependence graph analysis. KDD 2006.

[5] J. Ferrante, K. J. Ottenstein, and J. D. Warren. The program dependence graph and its use in optimization. ACM Trans. Program. Lang. Syst., 9(3):319–349, 1987.

[6] Shuai Ma, Yang Cao, Jinpeng Huai, and Tianyu Wo, Distributed Graph Pattern Matching, WWW 2012.

[7] Rice, M. and Tsotras, V.J., Graph indexing of road networks for shortest path queries with label restrictions,VLDB 2010.

[8] David A. Bader and Kamesh Madduri, A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms. Parallel Computing 2008.

[9] Daniel Peng, Frank Dabek: Large-scale Incremental Processing Using Distributed Transactions and Notifications. OSDI 2010.

[10] C. C. Aggarwal and H. Wang. Managing and Mining Graph Data. Springer, 2010.

[11] Metis. *http://glaros.dtc.umn.edu/gkhome/views/metis.*

Brief Bio

- Dr. Shuai Ma
  - PhDs (Peking University 2004 and Edinburgh University 2010)
  - Bell Labs, USA (Summer Consultant/Intern)
  - Edinburgh University,UK (Post Doc)
  - Microsoft Research Asia (Visiting Researcher)
  - From 2011 Beihang University (Full Professor)

Homepage: http://mashuai.buaa.edu.cn

Email: mashuai@buaa.edu.cn

Address: Room G1122,
New Main Building,
Beihang University

Thanks!