

# EECS E6893 Big Data Analytics Lecture 1:

## *Overview of Big Data Analytics*

Ching-Yung Lin, Ph.D.

Adjunct Professor, Depts. of Electrical Engineering and Computer Science

IEEE Fellow



September 6h, 2019

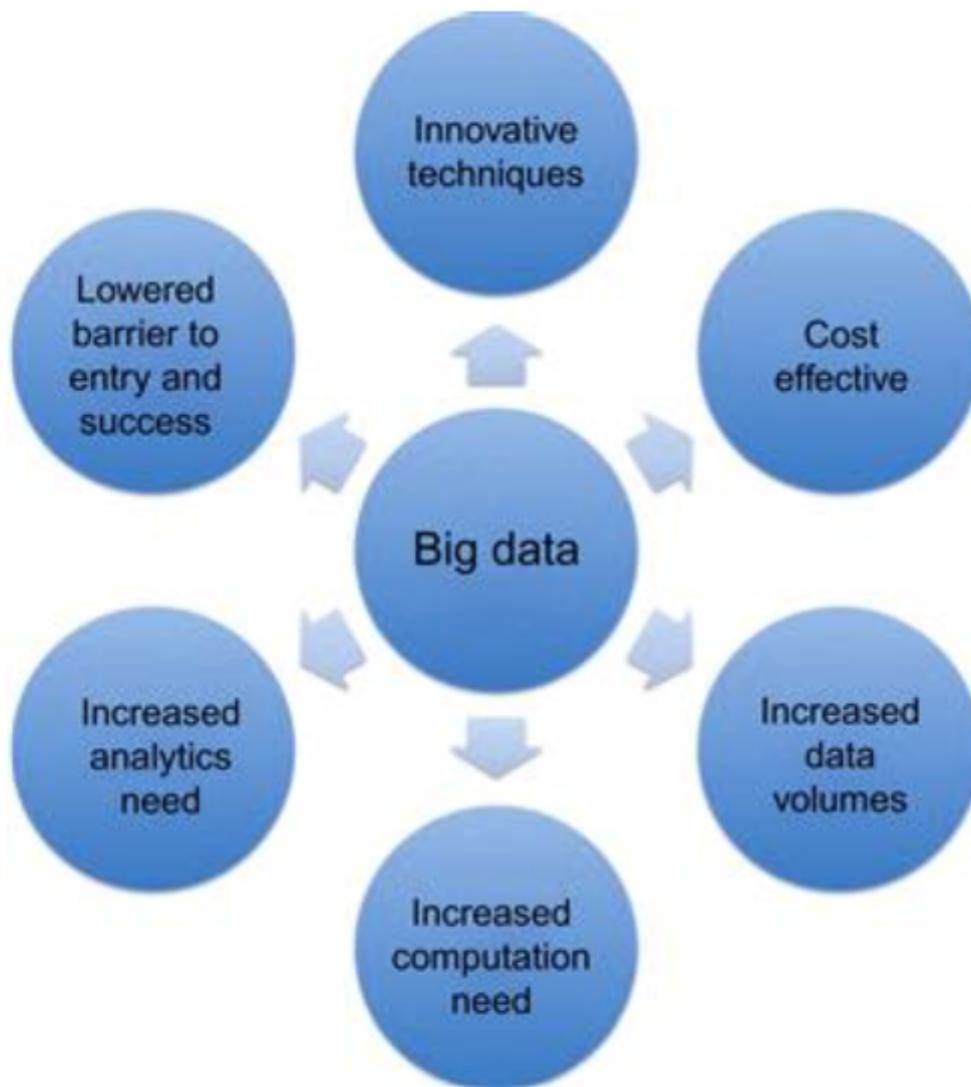
## Definition and Characteristics of Big Data

*“Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making.**” -- Gartner*

which was derived from:

*“While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity and variety.** In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.” – Doug Laney*

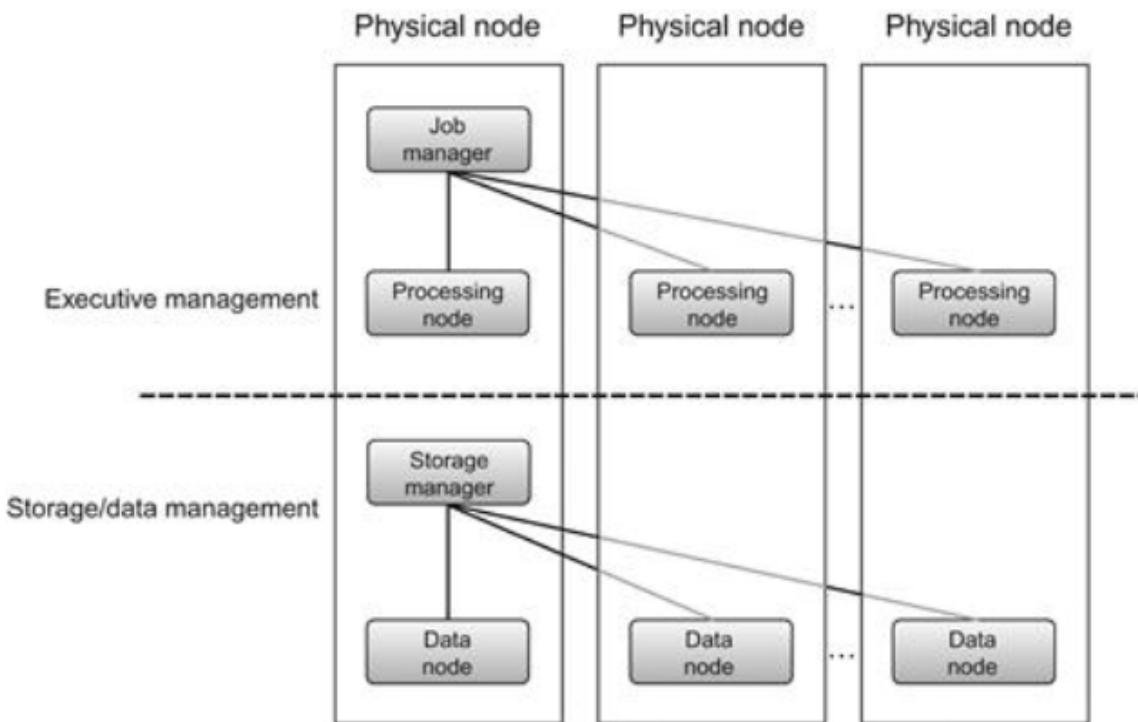
# What made Big Data needed?



“Big Data Analytics”, David Loshin, 2013

# Key Computing Resources for Big Data

- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network

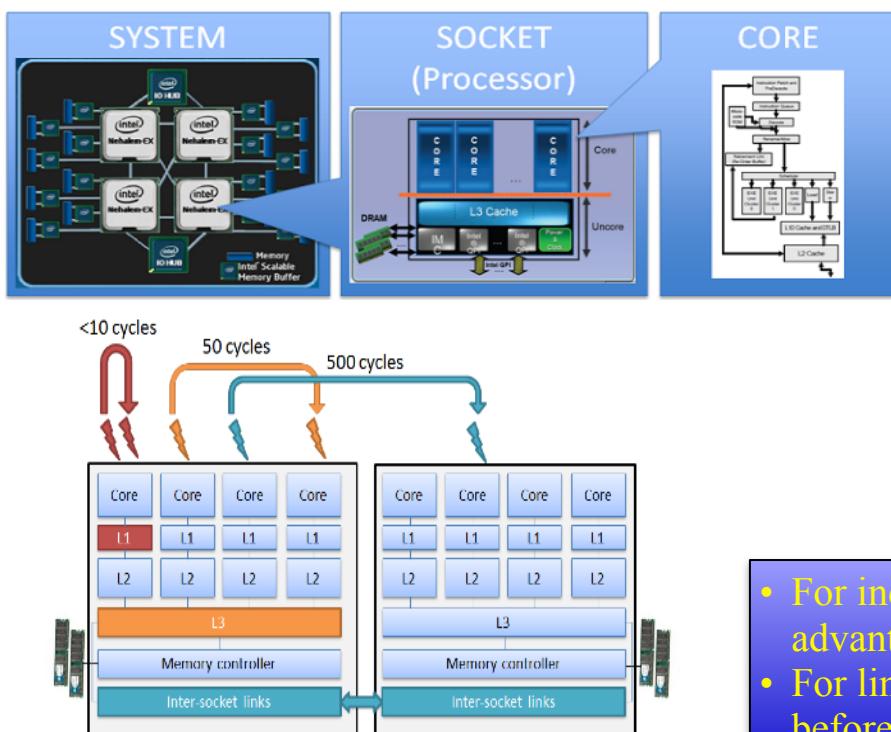


“Big Data Analytics”, David Loshin, 2013

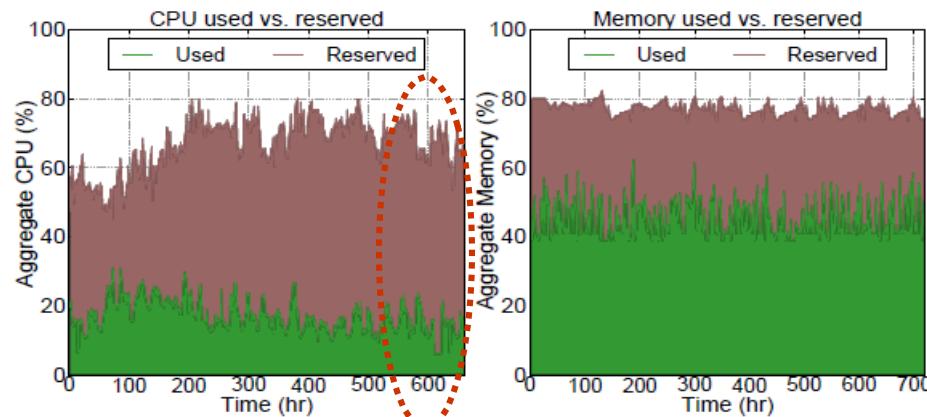
# Scalability — Scale Up & Scale Out



- Scale out
  - Use more resources to distribute workload in parallel
  - Higher data access latency is typically incurred
- Scale up
  - Efficiently use the resources
  - Architecture-aware algorithm design



Example: Resource utilization for a large production cluster at Twitter data center



[www.stanford.edu/~cde1/2014.asplos.quasar.pdf](http://www.stanford.edu/~cde1/2014.asplos.quasar.pdf)

- For independent data ==> scale up may not have obvious advantage than scale out
- For linked data ==> utilizing scale up as much as possible before scale out

Aspect	Typical Scenario	Big Data
Application development	Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning	A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing
Platform	Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices	Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology
Data management	Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts	Alternate models for data management (often referred to as NoSQL or “Not Only SQL”) provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics)
Resources	Requires large capital investment in purchasing high-end hardware to be installed and managed in-house	The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line

“Big Data Analytics”, David Loshin, 2013

# Techniques towards Big Data

---

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

→ Techniques exist for years to decades. Why is Big Data hot now?

# Why Big Data now?

---

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

# Why Big Data now?

---

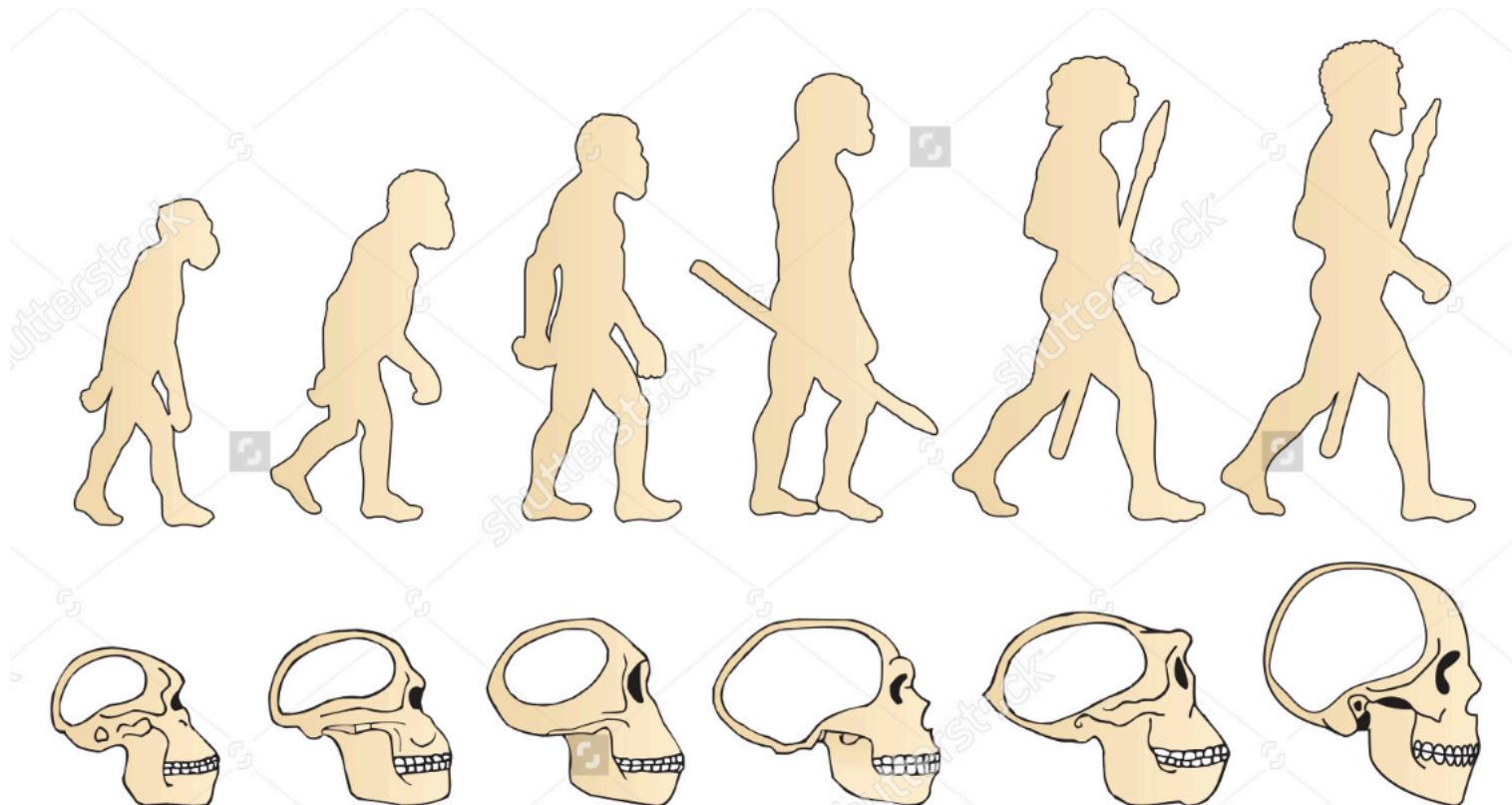
- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

- 
- High-Volume
  - High-Velocity
  - High-Variety

→ Artificial  
Intelligence



<https://www.youtube.com/watch?v=BV8qFeZxZPE>



shutterstock®

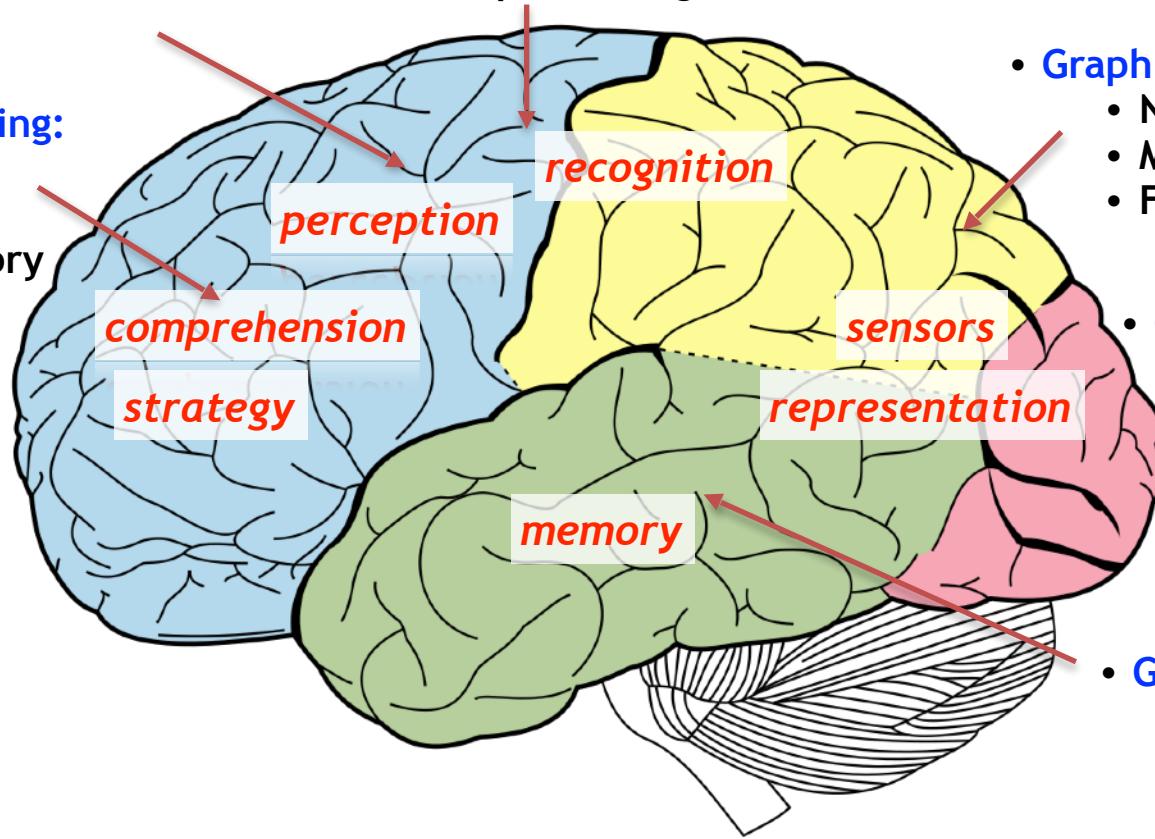
IMAGE ID: 290914883  
[www.shutterstock.com](http://www.shutterstock.com)

Human brain is a graph/network of 100B nodes and 700T edges.

- **Machine Cognition:**
  - Robot Cognition Tools
  - Feeling

- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools

- **Machine Learning:**
  - Machine Learning Tools
  - Deep Learning Tools



- **Graph Analytics:**
  - Network Analysis
  - Matching and Search
  - Flow Prediction
- **Graph Visualization:**
  - Dynamic Graph
  - Big Graph
- **Graph Database:**
  - Large-Scale Native Store



The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

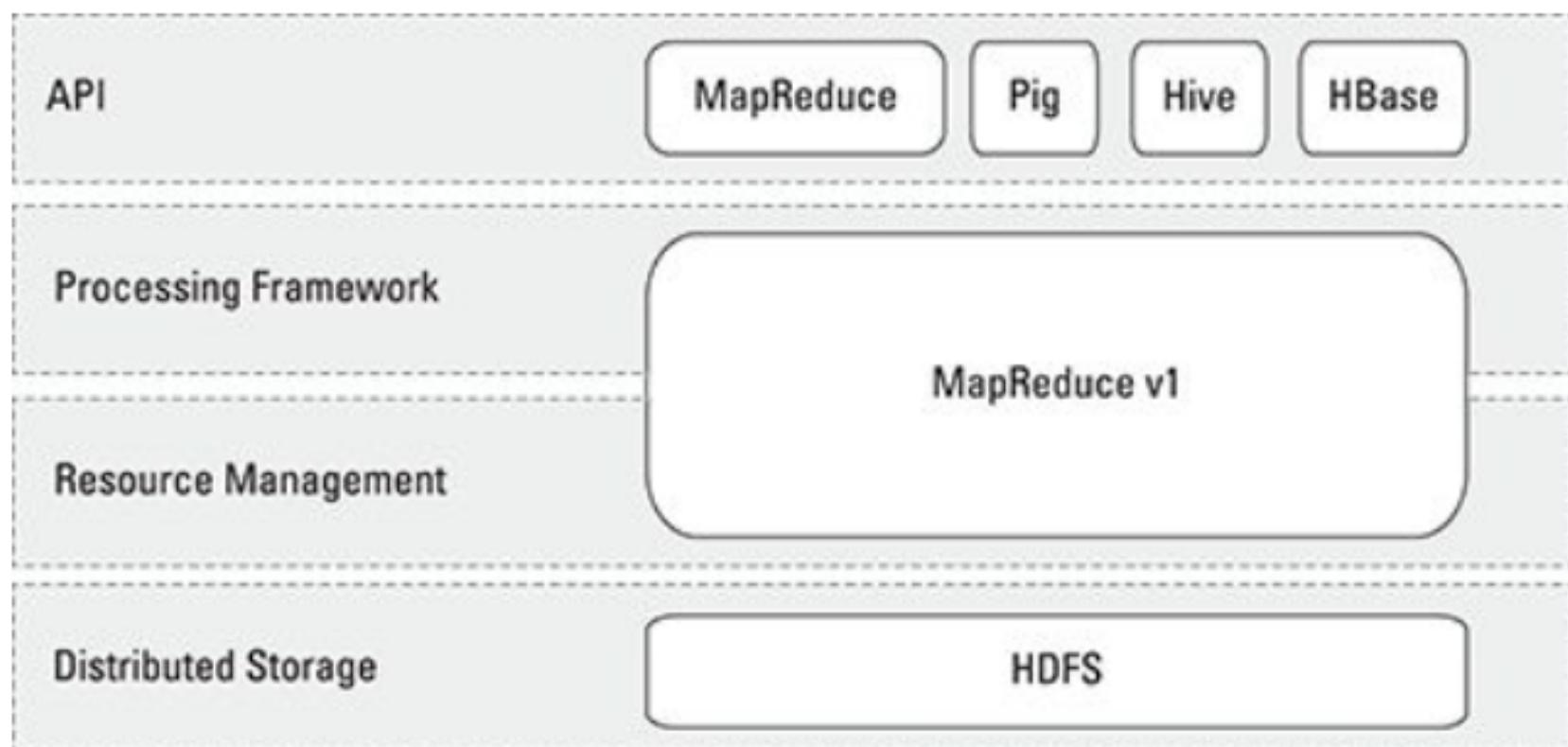
The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

<http://hadoop.apache.org>

# Four distinctive layers of Hadoop





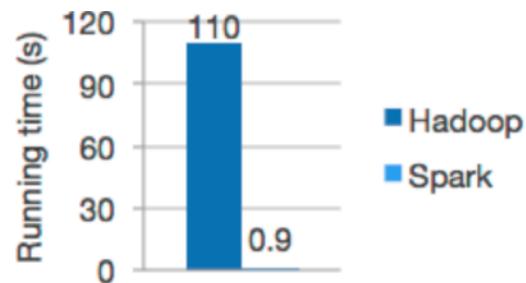
[Download](#)    [Libraries](#) ▾    [Documentation](#) ▾    [Examples](#)    [Community](#) ▾    [Developers](#) ▾

**Apache Spark™** is a unified analytics engine for large-scale data processing.

## Speed

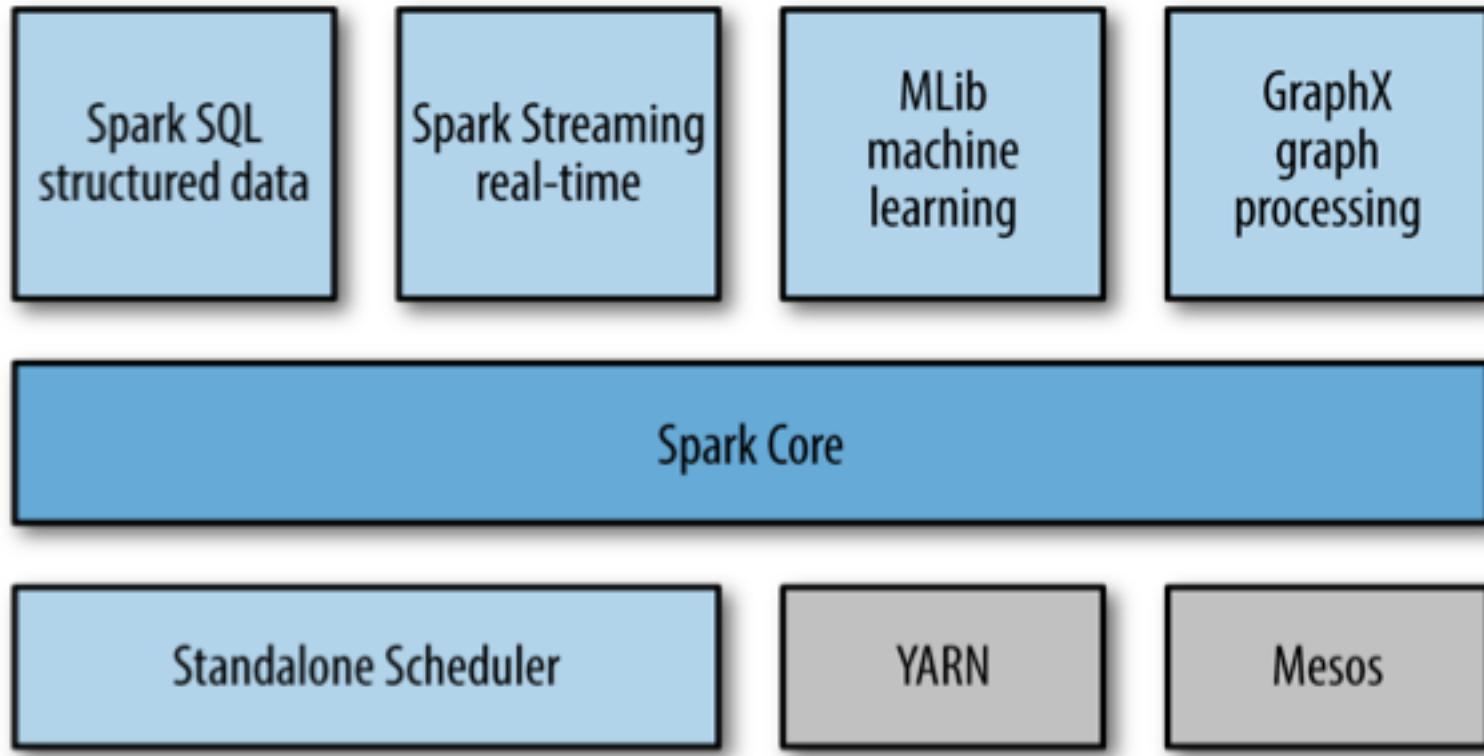
Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

# Main Spark Stack



## Course Main Thrust 3: Linked Big Data — Graph Analysis

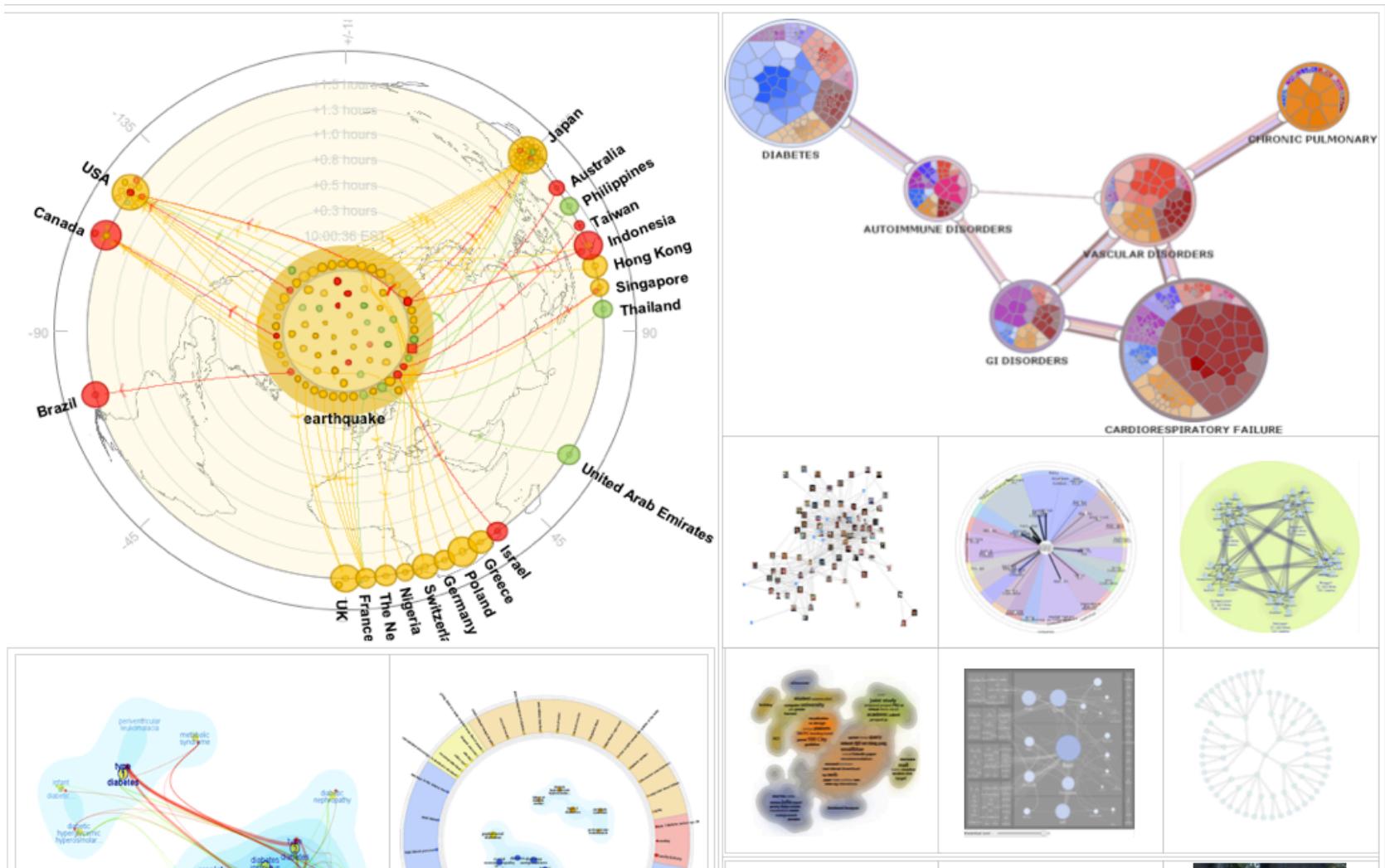


Human brain is a graph of 100B nodes and 700T edges.

# Course Main Thrust 4: Streaming Big Data Analytics



# Course Main Thrust 5: Big Data Visualization



## Course Main Thrust 6: Big Data and AI Solutions

- **Big Data and AI for Finance**
- **Big Data and AI for Healthcare**



# Why you want to take this class

- **Key Differentiator of this class:** Focusing on building a full-spectrum understanding of the latest Big Data Analytics technologies and using them to build real industry real-world solutions.
- **Sapphire Big Data Analytics Open Source Applications:** Create a Big Data open source toolsets for various industries (and disciplines)



- **Dataset and Use Cases:** Welcome!!

# Course Grading

---

- **5 Homeworks: 50%**
  - Individual work (except HW #4); Language Requirement: C/C++, Java, JavaScript, Python)
  - Report and source code
- **HW #0: Big Data Environment Setup and Testing**
- **HW #1: Big Data Analytics and Machine Learning**
- **HW #2: Linked Big Data Analytics**
- **HW #3: Streaming Big Data Analytics**
- **HW #4: Big Data Analytics Visualization (2 students per team)**
- **Final Project: 50%**
  - Teamwork: 2 - 3 students per team (on campus); 1 - 3 students per team for CVN
  - **Proposal** (slides — short presentation in the class, 5 mins presentation with video on YouTube)
  - **Final Report** (paper, up to 10 pages)
  - **Workshop Presentation** (Oral and Demo)
  - **Open Source Codes**
  - **Video Presentation** (on YouTube)

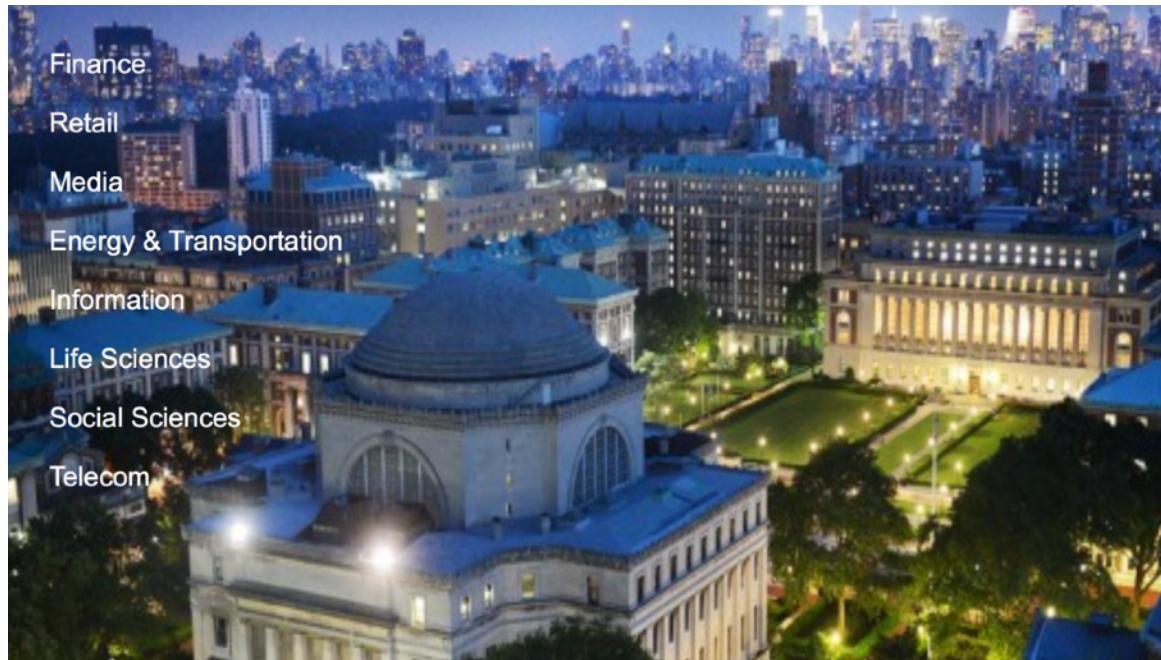
# Course Information

- Website:

<http://www.ee.columbia.edu/~cylin/course/bigdata/>

- Textbook:

-- None, but reference book(s) and/or articles/papers will be provided each lecture.



# Course Outline

<b>Class Date</b>	<b>Class Number</b>	<b>Topics Covered</b>
09/06/19	1	Introduction of Big Data Analytics
09/13/19	2	Big Data Platforms and Data Storage
09/20/19	3	Big Data Analytics Algorithms I
09/27/19	4	Big Data Analytics Algorithms II
10/04/19	5	Linked Big Data Analytics
10/11/19	6	Graph Database and Analytics
10/18/19	7	Streaming Big Data Analytics
10/25/19	8	Big Data Visualization -- I
11/01/19	9	Final Project Proposal Presentation
11/08/19	10	Big Data Visualization -- II
11/15/19	11	Big Data Analytics for AI Finance
11/22/18	12	Final Project Intermediate Presentation
11/29/19		NO CLASS -- Thanksgiving Holiday
12/06/19	13	Big Data Analytics for AI Healthcare
12/13/19	14	Big Data Analytics Workshop

# Assignments and Submissions

<b>Class Date</b>	<b>Assignment</b>	<b>Due</b>
09/06/19	HW #0 Big Data Environment Setup and Testing	
09/13/19		
09/20/19	HW #1 Big Data Analytics and Machine Learning	HW #0
09/27/19		
10/04/19	HW #2 Linked Big Data Analytics	HW #1
10/11/19		
10/18/19	HW #3 Streaming Big Data Analytics	HW #2
10/25/19	HW #4 Big Data Visualization	
11/01/19		HW #3 & Proposal Slides
11/08/19		
11/15/19		HW #4
11/22/18		Intermediate Presentation Slides
11/29/19		
12/06/19		
12/13/19		Final Project Slides

# Other Issues

---

- Professor Lin:
  - Office Hours:

Friday after the class: 9:40pm – 10:00pm (lecture room)  
Or by appointment
  - Contact: [c.lin@columbia.edu](mailto:c.lin@columbia.edu)
- TAs (CAs/IAs/Graders) —
  - Frank Ouyang (ho2271)
  - Tingyu Li (tl2861)
  - Yunan Lu (yl4021)
  - TBDs, probably have 6-7 TAs in total.

# Big Data Analytics

From Strategic Planning to  
Enterprise Integration with Tools,  
Techniques, NoSQL, and Graph



David Loshin

- Chapter 1: Market and Business Drivers for Big Data Analysis
- Chapter 2: Business Problems Suited to Big Data Analytics
- Chapter 3: Achieving Organizational Alignment for Big Data Analytics
- Chapter 4: Developing a Strategy for Integrating Big Data Analytics into the Enterprise
- Chapter 5: Data Governance for Big Data Analytics: Considerations for Data Policies and Processes
- Chapter 6: Introduction to High-Performance Appliances for Big Data Management
- Chapter 7: Big Data Tools and Techniques
- Chapter 8: Developing Big Data Applications
- Chapter 9: NoSQL Data Management for Big Data
- Chapter 10: Using Graph Analytics for Big Data
- Chapter 11: Developing the Big Data Roadmap

# 5 Example Big Data Use Case Categories



## Big Data Exploration

Find, visualize, understand all big data to improve decision making



## Enhanced 360° View of the Customer

Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources



## Security/Intelligence Extension

Lower risk, detect fraud and monitor cyber security in real-time



## Operations Analysis

Analyze a variety of machine data for improved business results



## Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

1. Expertise Location
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Healthcare Analysis
8. Data Exploration and Visualization
9. Personalized Search
10. Anomaly Detection
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring (Smarter another Planet)
14. Cellular Network Monitoring
15. Cloud Monitoring
16. Code Life Cycle Management
17. Traffic Navigation
18. Image and Video Semantic Understanding
19. Genomic Medicine
20. Brain Network Analysis
21. Data Curation
22. Near Earth Object Analysis



# Category 1: 360° View

## Recommendation

amazon.com Ching's Store See All 32 Product Categories Your Account | Cart | Your Lists | Help | Find Gifts

Hello, Ching Yung Lin. We have recommendations for you. (If you're not Ching Yung Lin, click here.) Make this

**BROWSE**

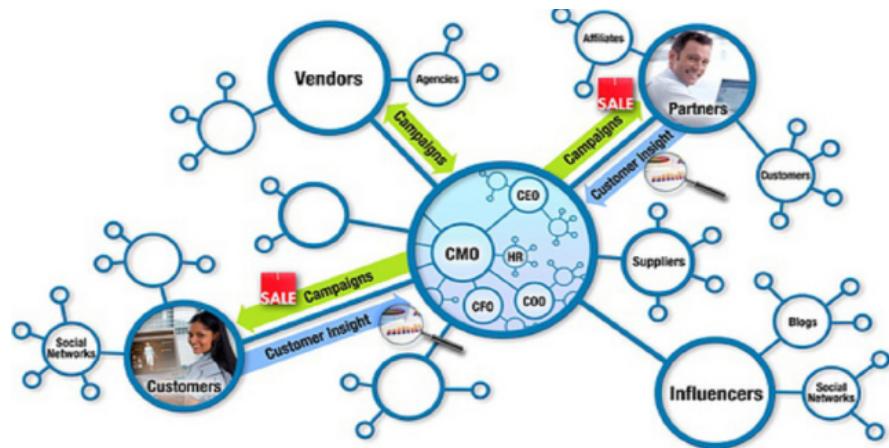
- Your Favorites
  - Books
  - Software
- Featured Stores**
  - Apparel & Accessories
  - Beauty
  - DVD's TV Central

**Recommended for you**

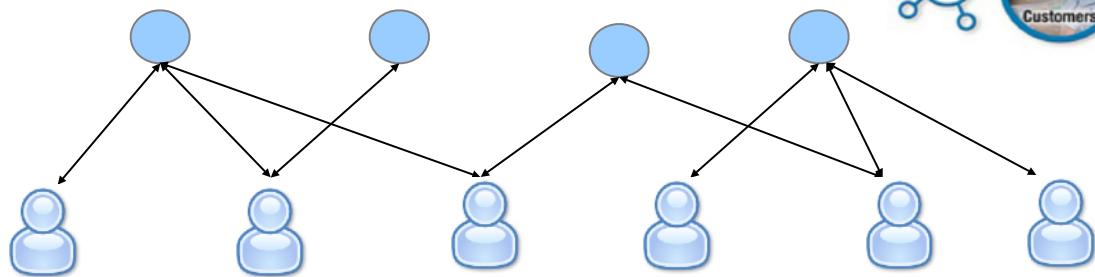
- Spikes [Reprint] Paperback by Fred Rieke
- Spiking Neuron Models Paperback by Wulfram Gerstner
- Methods In Neuronal Modeling - 2nd Edition Hardcover by Christof Koch

(Why is this recommended to me?)

See more Recommendations



Enhancing:



## Graph Visualizations

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

Graph Matching

Graph Sampling

Markov Networks

## Middleware and Database

# Use Case 1: Social Network Analysis in Enterprise for Productivity

Production Live System used by IBM GBS since 2009 – verified ~\$100M contribution

15,000 contributors in 76 countries; 92,000 annual unique IBM users

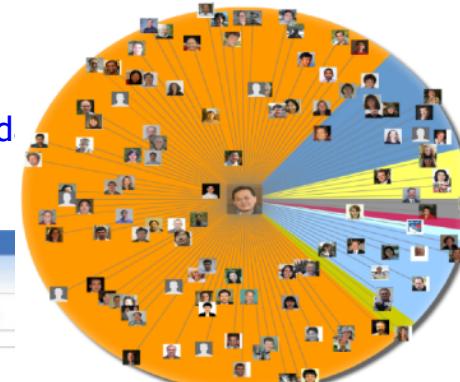
25,000,000+ emails & SameTime messages (incl. Content features)

1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, ..., access d

1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

200,000 people's consulting project & earning d

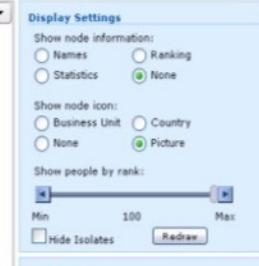
The screenshot shows the SmallBlue Suite interface with a search bar for 'subject keywords' set to 'healthcare'. Below the search bar, there are dropdown menus for 'Country' (all) and 'Division' (Advanced search). A 'Diagram Only' button is also present. On the left, there is a sidebar with user profiles for 1. Patricia (Patti) Okita, 2. Michael Hohenberger, 3. Todd (T.H.) Kalynuk, 4. Susan E. (SUSAN) Rivers, and 5. M.C. (Mark) Effingham. The main area displays a network graph with many nodes connected by lines, representing the social network. A legend on the right indicates node types: Business Unit (blue), Country (green), and Picture (yellow).



Shortest  
Paths

Centralities

Graph  
Search



Dynamic networks  
of 400,000+  
IBMer:

Shortest Paths  
Social Capital  
Bridges  
Hubs  
Expertise Search  
Graph Search  
Graph Recomm.

- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: \$7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal \$228.5 million savings and
- APQC (WW leader in Knowledge Practice) April 2013:

***"The Industry Leader and Best Practice in Expertise Location"***

# Use Case 2: Personalized Recommendation

w3 Search Pages(w3)

Practitioner Portal Translate this page: English Tell a friend How-to videos Portal help Site map Feedback

### People in your network

Network for: Lin, Ching-Yung

81 colleagues are 1 degree from you  
1615 colleagues are 2 degrees from you  
18270 colleagues are 3 degrees from you

Your 1st degree network diagram [\[Show list\]](#)

View networks: Lotus Connections & SmallBlue | ▾

Sort by: Division | Country | Social proximity.



[Edit SmallBlue](#)

[View all tags](#) | [Tags by person](#)

▶ Portlet social rating information

### Buzz in your network

Share your status with your network  Post status

Network buzz for networks: IBM Connections & SmallBlue ▾

Sources:  Profiles  Blogs [\[Go\]](#)

1 of 1 items Network: All Sources: All Sort by: Most recent | Person

Jeffrey Nichols Re: Thoughts (and Questions) on Answers July 09 10:50 AM [Comment](#)

[RSS Feed](#)

▶ Portlet social rating information

### Popular in the Practitioner Portal

Here's what is currently popular in the Practitioner Portal with your colleagues.

- ▼ Top 5 document searches
  - [SAP, cloud pattern, bao\\_signature\\_solutions, bob\\_sc\\_KM and KS case studies](#)
- ▶ Top accessed content
- ▶ Top Bookmarks
- ▶ Portlet social rating information

### Recently shared content in your network

See what content people in your network have been sharing to others. Select the network and sources you are interested in and click go.

Networks: Direct (1st degree) ▾

Sources:  IC Bookmarks  IC Files  IC Wikis  
 Practitioner Portal  Media Library  ILX [GO](#)

5 of top 18 Sort by: Social Proximity | Date | Source

Network: direct Sources: All

[Welcome to Graph Technologies](#) 09 Jul 2013

[Mobile security Workshop \(Bharti Airtel\)](#) 15 Jul 2013

[hci-and-smartphone-data-at-scale-ibm-Jul2013.pptx](#) 30 Jul 2013

### Popular learning

See what education is popular with the people in your network. Select the sources you are interested in and click go.

Sources:  L@IBM  Media Library  ILX [GO](#)

5 of top 30 Sort by: Popularity | Source

Sources: All

Leadership in a Project Team Environment [\[View\]](#)  
★★★★★

PMKN eShareNet June 13, 2013 - Worldwide Project Management Method (WWPMM) 3.0 Release Preview: Improving PM Method Adaptability. Presented by Stacey Lopez and Todd Fredrickson - IBM Rational Asset Manager [\[View\]](#) ★★★★★

New2Blue - Mid-Year Review - Personal Business Commitments (Session Replay) [New Employee Experience 2013 Events] [\[View\]](#) ★★★★★

Junos Pulse for Android Smartphone [\[View\]](#) ★★★★★

Project Management Orientation [\[View\]](#) ★★★★★

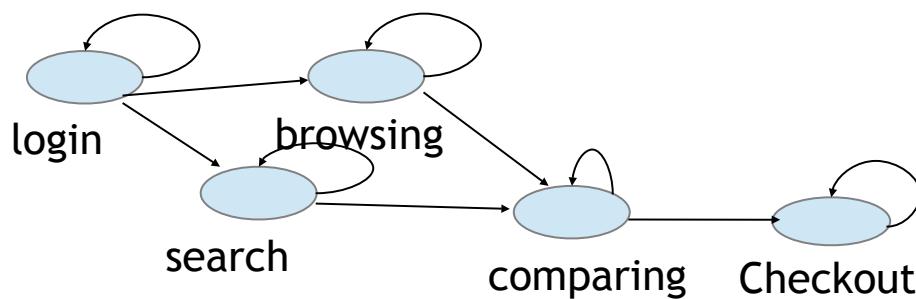
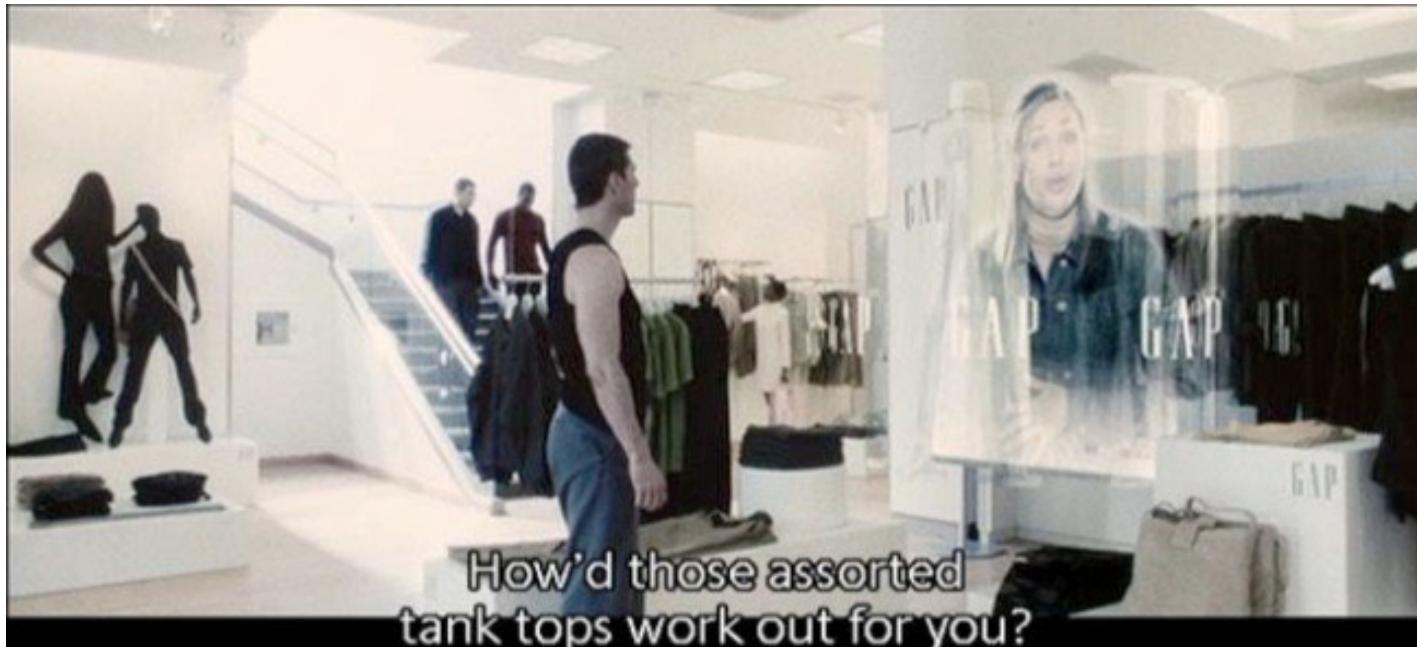
[Show more](#)

# Use Case 3: Customer Behavior Sequence Analytics

Markov  
Network

Latent  
Network

Bayesian  
Network

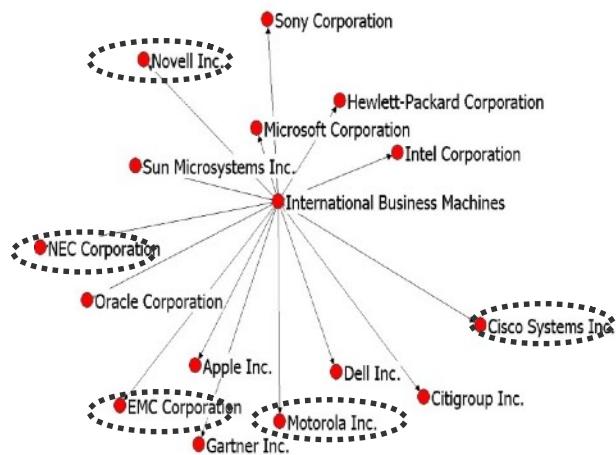


- Behavior Pattern Detection
- Help Needed Detection

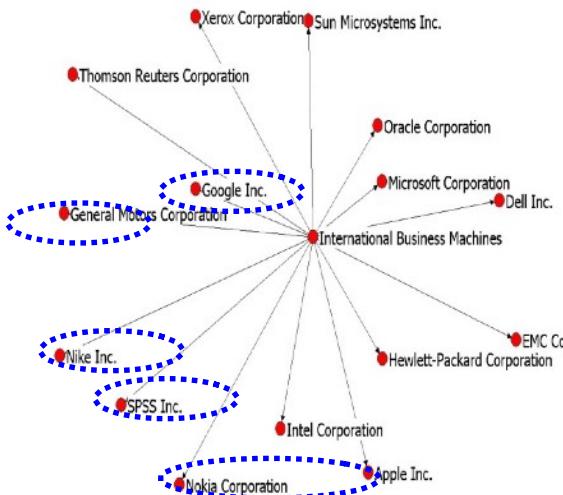
# Use Case 4: Graph Analytics for Financial Analysis

**Goal:** Injecting Network Graph Effects for Financial Analysis. Estimating company performance considering correlated companies, network properties and evolutions, causal parameter analysis, etc.

- IBM 2003



- IBM 2009



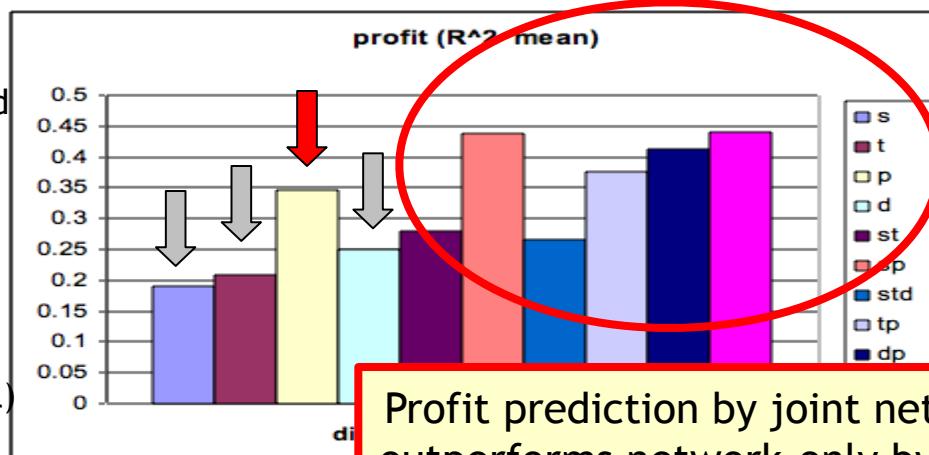
- Data Source:

- Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

Goal: Learn from previous 5 years, and predict next year

Model: Support Vector Regression (RBF kernel)



**Network feature:**  
 s (current year network feature),  
 t (temporal network feature),  
 d (delta value of network feature)

**Financial feature:**  
 p (historical profits and

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 33%.

# Use Case 5: Social Media Monitoring

Home | Live | Forensics | Research Projects | People | New

Select CIO Category(-ies): EXECDB BLADE HRTEANNT IBM SecurityAnalysis SWG WATSON or Word: Egypt GO STOP RESUME language: Arabic

Total Tweets: 231  
Positive: 35 15%  
Negative: 31 13%

EGYPT wearing @RawyaRageh beauty brutality Mor e || اعلى Am Egypt's 12 police هر بريسي hijab Er d dozen Sponge allege Port Egypt than Cairo you my Egyptian مصر Said egypt lady call

Saloom Butilla @SaloomButilla RT @Lion\_King\_Bhr: إعتداء المثلثين الغونة في البحرين على المرافق العامة ورجل الأمن #Bahrain #Egypt #Syria #KSA #UAE #News h .... Translation: RT "@Lion\_King\_Bhr": The traitors in Bahrain Safavid attack on public utilities and security men, 2/19/2013 \*LBahrain\* #Egypt #LSyria\* \*LKSA\* \*LUAE\* \*LNews\* h \*...\* --Wed Feb 20 17:57:58 2013

Zenza Raggi fan-club @Zenzadclub Private Gold 64: Cleopatra 2 // A sect that worships ancient Egypt is attempting to bring Cleopatra back to life... http://t.co/TcvMDiwb --Wed Feb 20 17:57:53 2013

SH\_QalamSara @SH\_QalamSara RT @HebaFaroof: An #Egypt-ian beauty :) ▶ http://t.co/S9BZb5f3 --Wed Feb 20 17:57:53 2013

Mona Metwally @monametwally RT @EgyBloodBank: مريض محتاج متبرعين: دم AB+ مستكتفي الجامدة بالاسمية عليه قصبة دم AB+ 01024705247 #Egypt # مصر http://t.co/5o06mtz5. Translation: . RT \*@EgyBloodBank\*: A

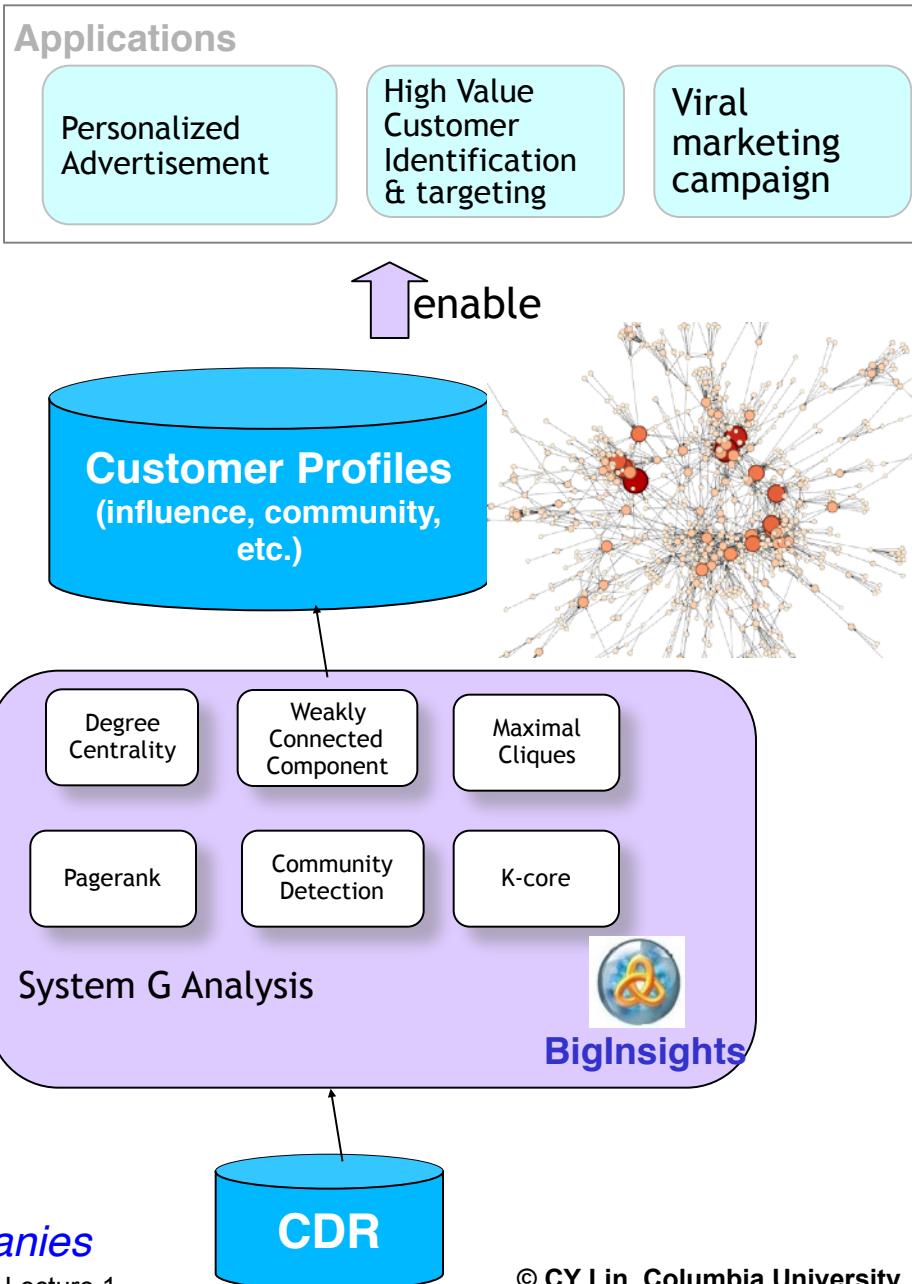
monitoring categories Monitoring filter

Real-Time Translation, Loc Live Tweets, Sentiment, Keywords Graph Zooming / Panning Top Retweets

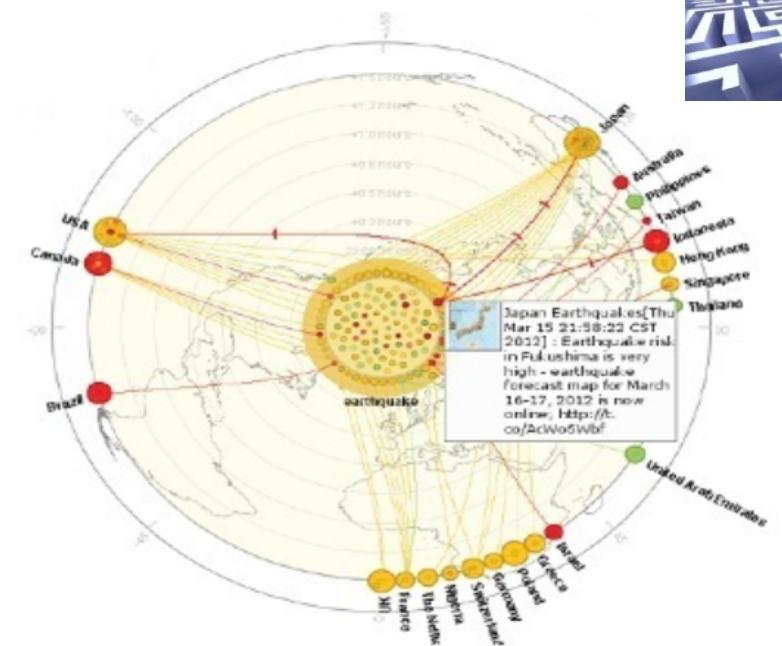
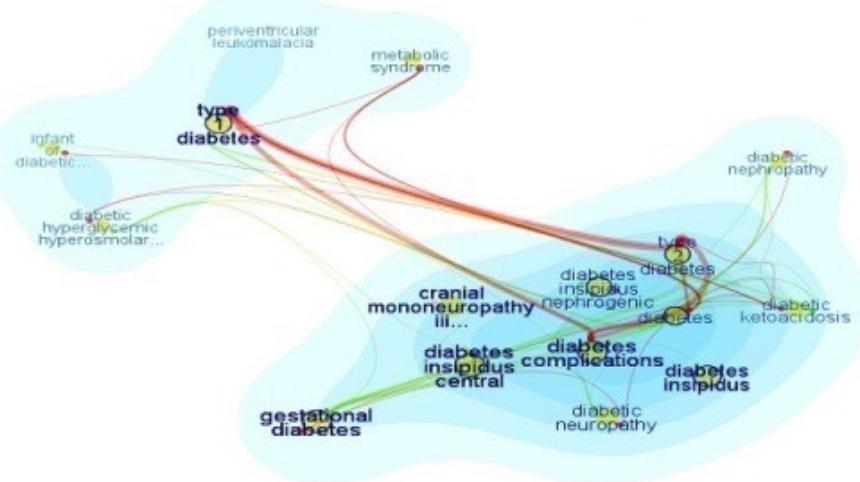
## Use Case 6: Customer Social Analysis for Telco

**Goal:** Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

- Applications based on the extracted social profiles
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign
- Approach
  - Construct social graphs from CDRs based on {caller, callee, call time, call duration}
  - Extract customer social features (e.g. influence, communities, etc.) from the constructed social graph as customer social profiles
  - Build analytics applications (e.g. personalized advertisement) based on the extracted customer social profiles



# Category 2: Data Exploration



Enhancing:



Vivísmo®

cúram®  
SOFTWARE

Huge Network Visualization

Network Propagation

I2 3D Network Visualization

Geo Network Visualization

Graphical Model

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

Graph Matching

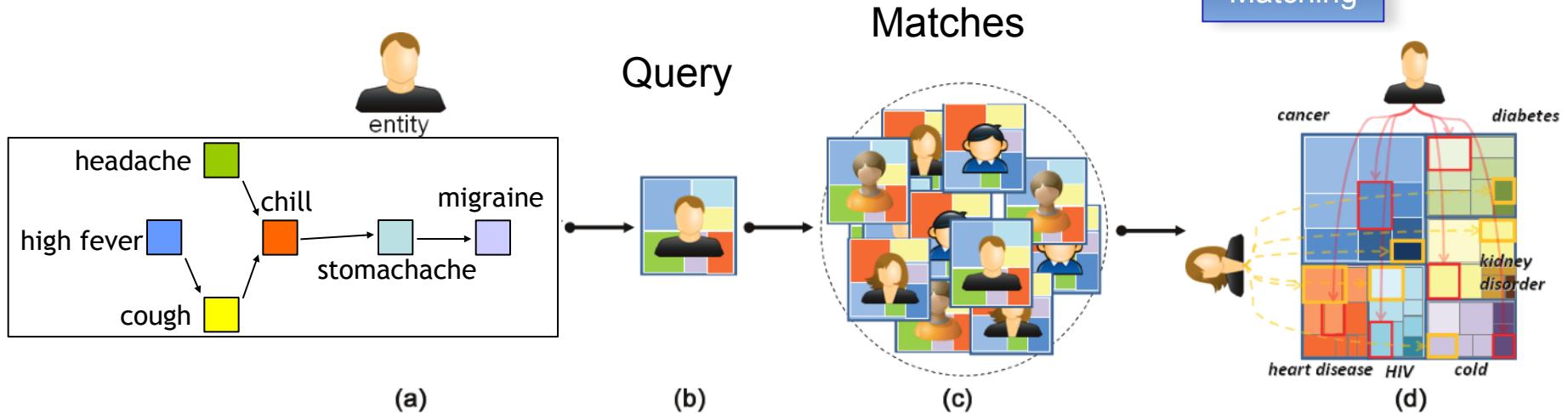
Graph Sampling

Markov Networks

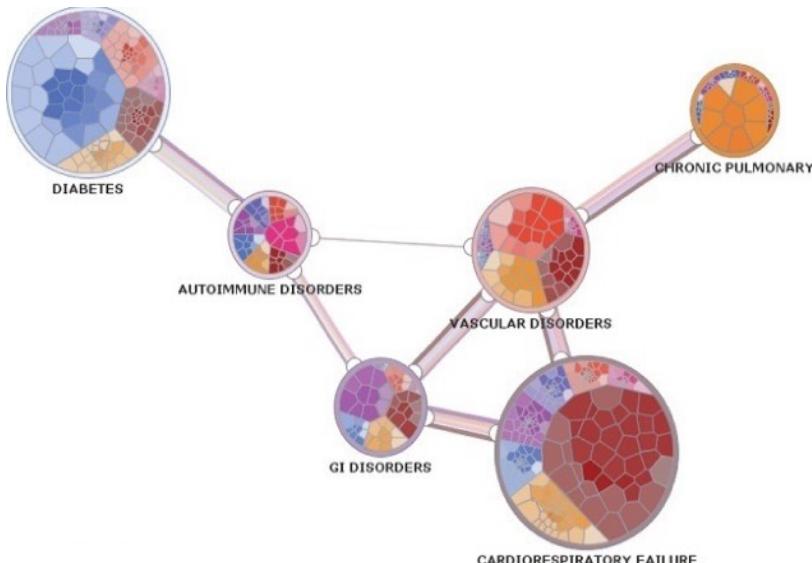
Middleware and Database

# Use Case 7: Graph Analytics and Visualization

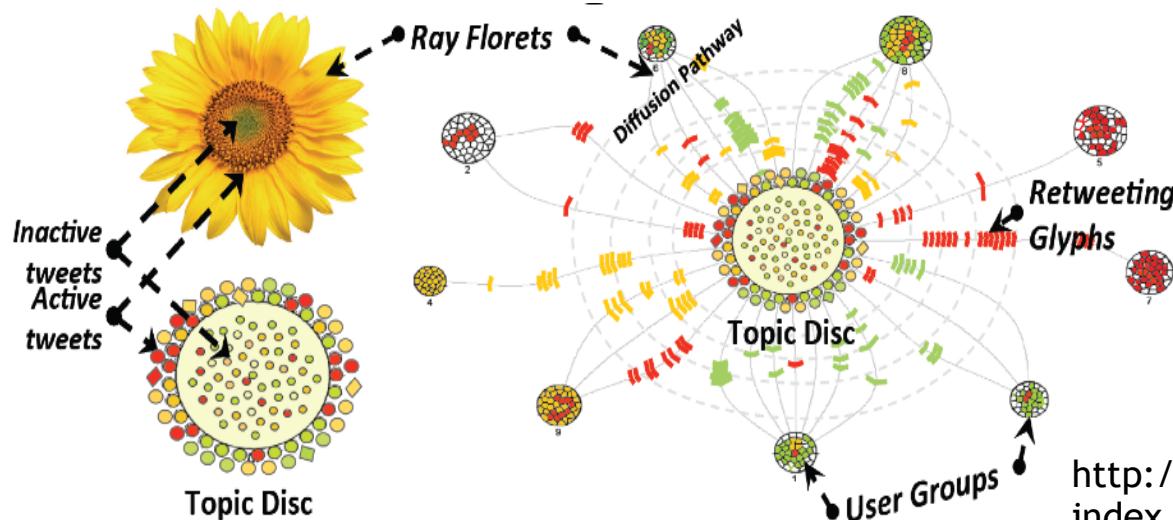
Graph  
Matching



Graph  
Communities



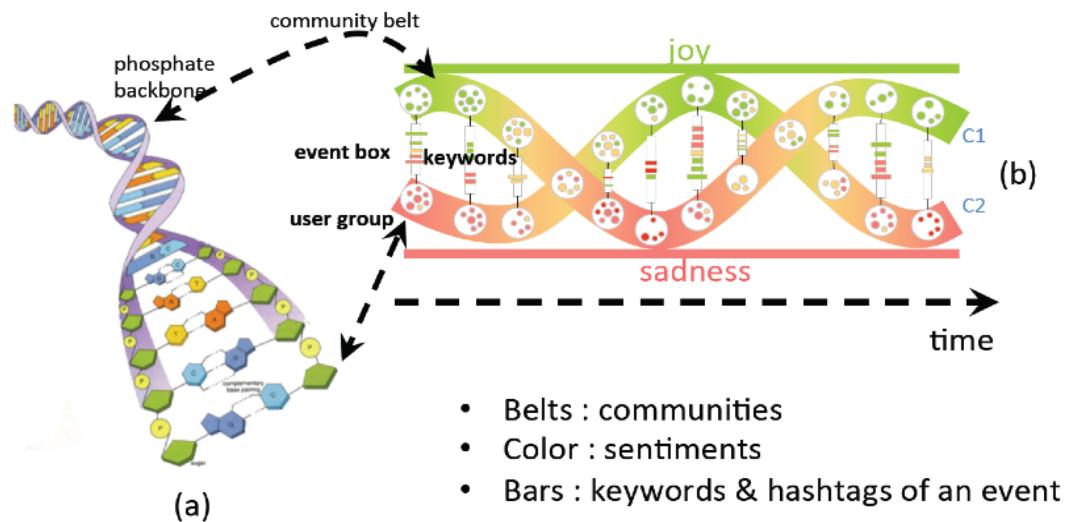
# User Case 8: Visualization for Navigation and Exploration



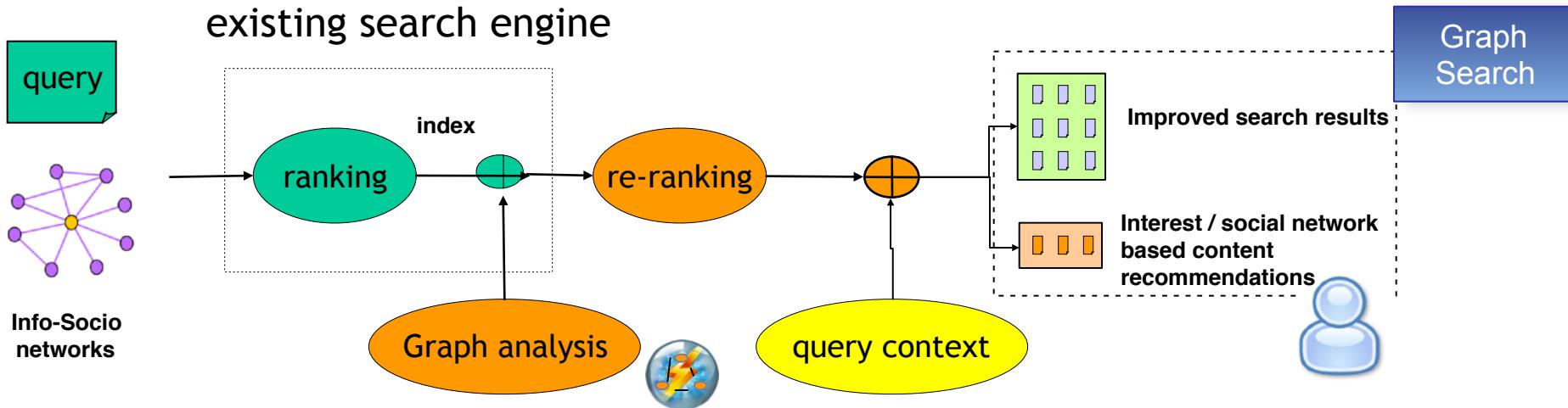
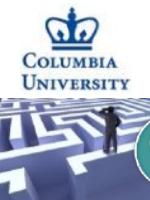
Whisper : Tracing the information diffusion in Social Media

<http://systemg.ibm.com/apps/whisper/index.html>

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media



# Use Case 9: Graph Search



Practitioner Portal

< Return to starting page

Refine Results

- By Tag
  - Select a tag to filter search results ⓘ
  - View as: cloud | list
  - more — less
- 2012 analyst\_report analytics bao baseline csp deliverable europe forrester fccool gartner gba gmu government kh leader\_priority na proposal public\_sector retail sales sales\_tools sandit social social\_business telecommunications
- By Category
  - Select a category to filter search results ⓘ
  - Expand all | Collapse all
  - Asset Type
  - Audience
  - Business Topics
  - Client Value Method (CVM)
  - Geography
  - IBM Business Unit
  - Industry
  - Language

Search criteria

Search terms, pages and tags

Search keywords: social business ⓘ

All results Social network results ⓘ

18,577 results found

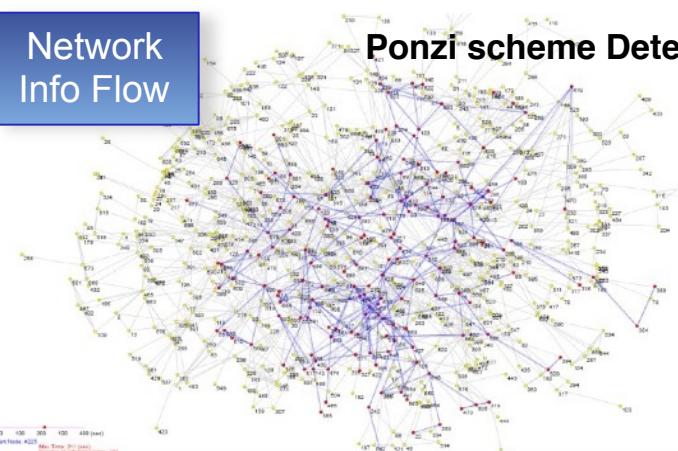
1 to 25 shown

Title	Relevance	Modified	Bookmarks
IBM Social Business Adoption QuickStart (U.S. English) - Proposal Insert in Proposal and Presentation Accelerator (PPX) ⓘ	100 %	29 Aug 2012	0
Drive the successful launch and adoption of social business software throughout your organization with a structured engagement comprised of assessments, planning and design consultation, onsite workshops, and team- and skills-building activities.			
Sales Support Information(SSI) DAGE@stibo.com			

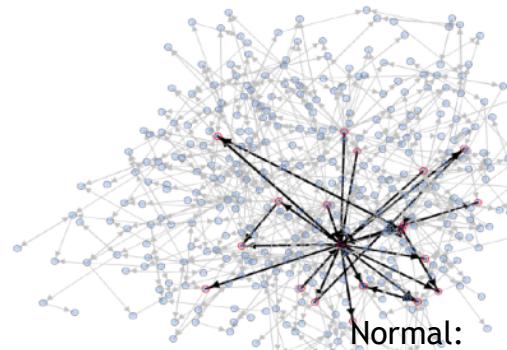
# Category 3: Security

Network  
Info Flow

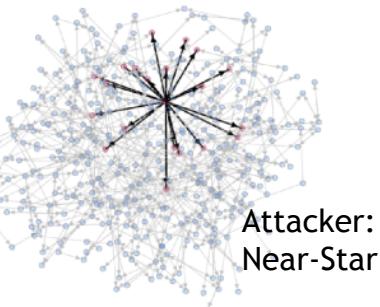
## Ponzi scheme Detection



Ego Net  
Features

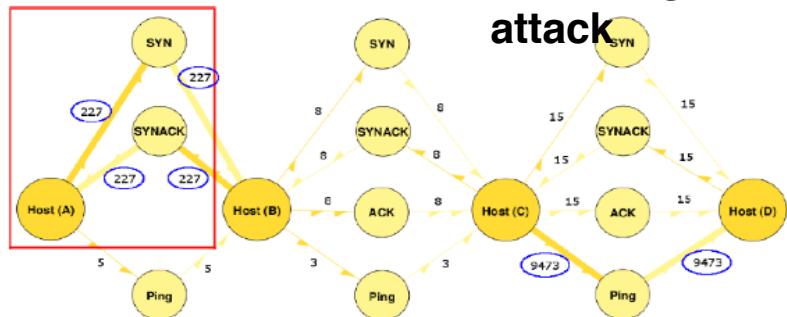


Normal:  
 (1)Clique-like  
 (2)Two-way links

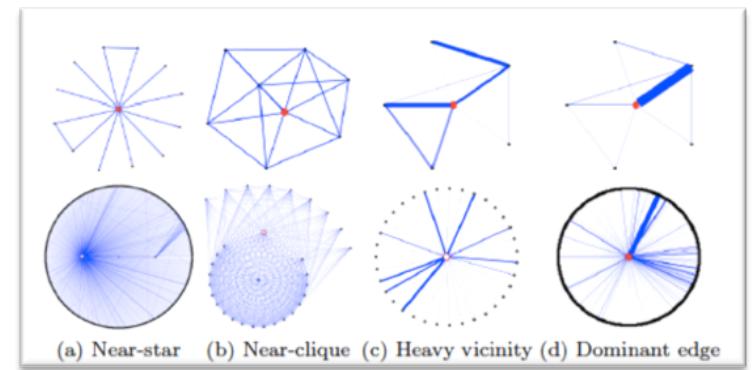


Attacker:  
Near-Star

## Detecting DoS attack



(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.



## Graph Visualizations

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

Graph Matching

Graph Sampling

Markov Networks

## Middleware and Database

# Use Case 10: Anomaly Detection at Multiple Scales

Based on President Executive Order 13587

**Goal:** System for Detecting and Predicting Abnormal Behaviors in Organization, through large-scale social network & cognitive analytics and data mining, to decrease insider threats such as espionage, sabotage, colleague-shooting, suicide, etc.



“Enterprise Information Leakage Impacted economy and jobs” Feb 2013

“What's emerged is a multibillion dollar detective industry”  
*npr Jan 10, 2013*

Emails

Instant Messaging

Web Access

Executed Processes

Printing

Copying

Log On/Off

Social sensors

Click streams capturer

Feed subscription

Database access

Graph analysis

Behavior analysis

Semantics analysis

Psychological analysis

Multimodality Analysis

Detection,  
Prediction  
&  
Exploration  
Interface

Infrastructure + ~ 490 Analytics

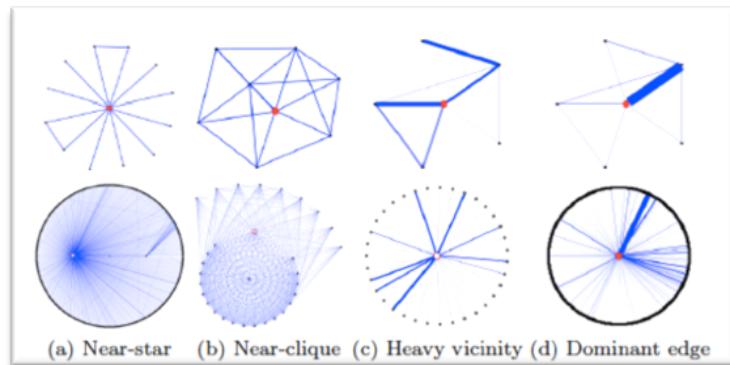
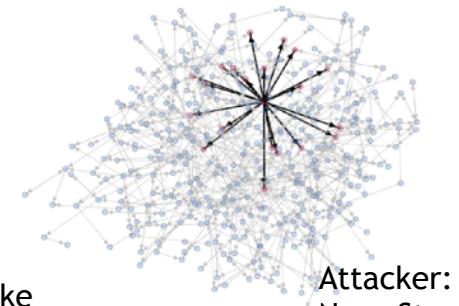
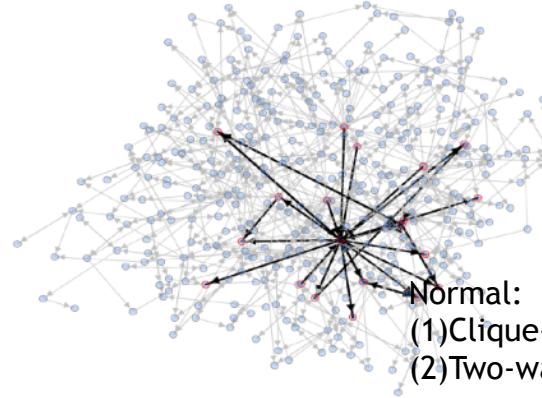
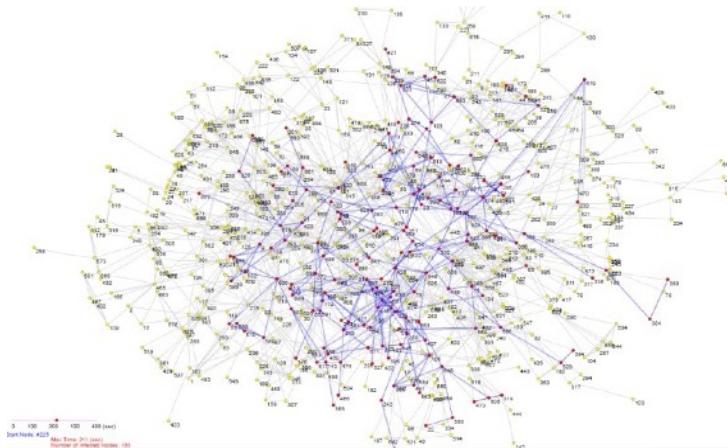
# Use Case 11: Fraud Detection for Bank

Network  
Info Flow

Ego Net  
Features



## Ponzi scheme Detection



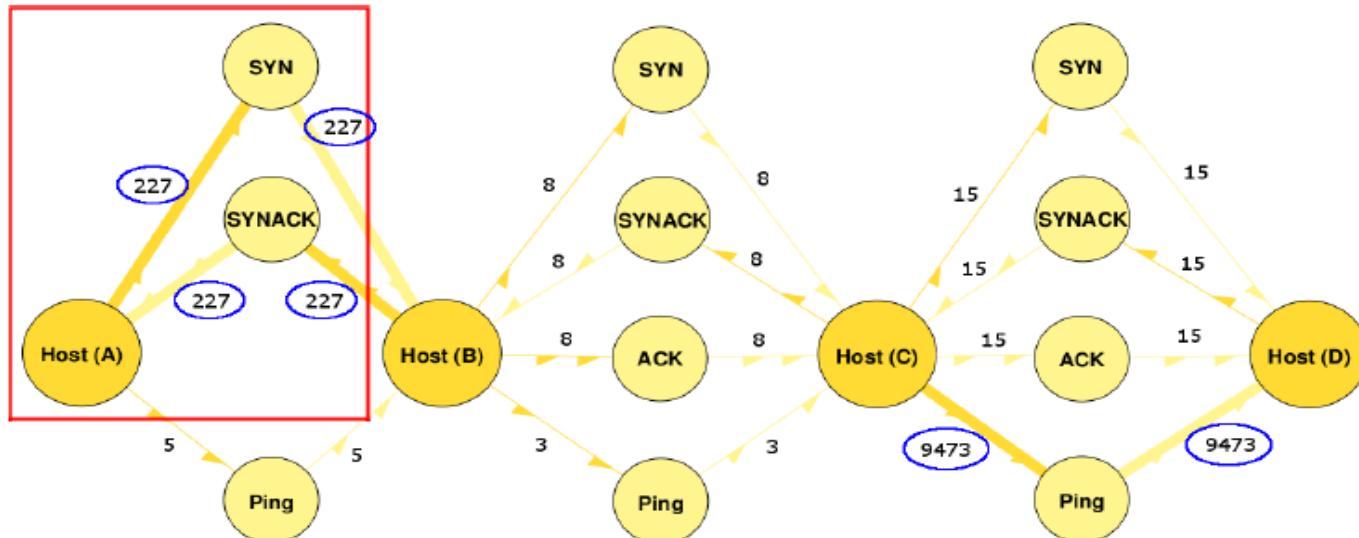
# Use Case 12: Detecting Cyber Attacks

Network  
Info Flow

Ego Net  
Features

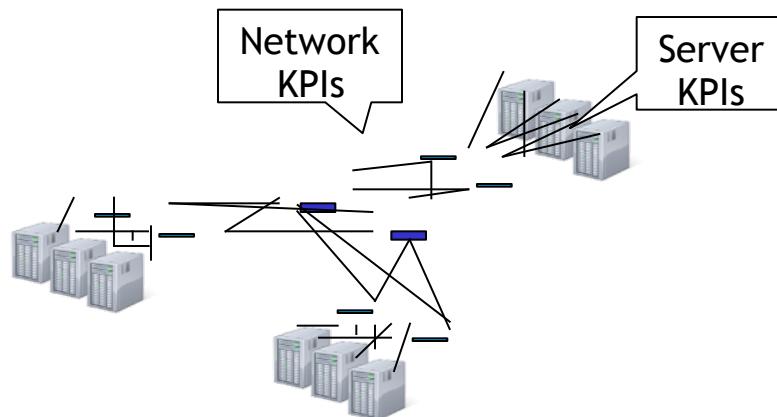


## Detecting DoS attack



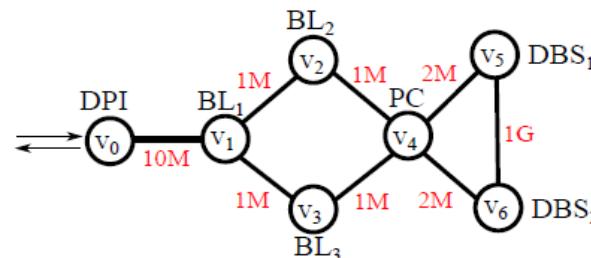
(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

# Category 4: Operations Analysis



## Cloud Service Placement

DPI - Deep Package Inspector    BL - Business Logic  
 PC - Package classifier    DBS - DB Server



Memory requirements

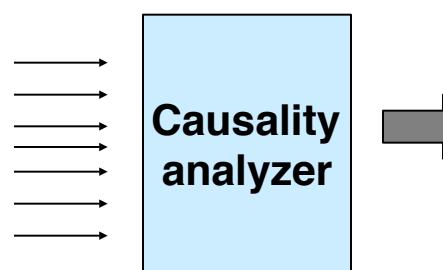
$v_0$	8G
$v_1$	2.5G
$v_2$	2G
$v_3$	2G
$v_4$	12G
$v_5$	20G
$v_6$	32G



Graph Matching

Bayesian Network

KPI time series (e.g., server performance/load, network performance/load)



- KPI (a time series)
- ... (potential) pairwise relationship (e.g., causality)

## Graph Visualizations

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

Graph Matching

Graph Sampling

Markov Networks

## Middleware and Database

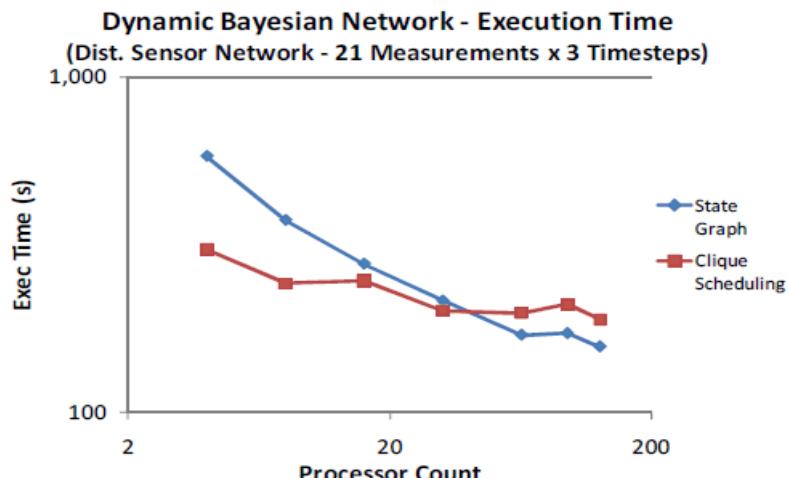
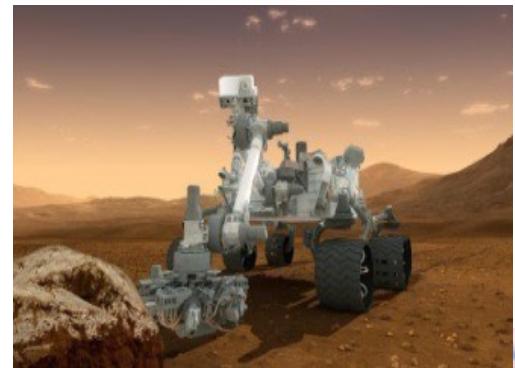
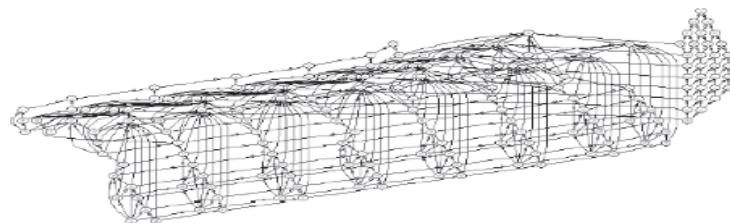
## Use Case 13: Smarter *another* Planet

**Goal:** Atmospheric Radiation Measurement (ARM) climate research facility provides 24x7 *continuous field observations* of cloud, aerosol and radiative processes. **Graphical models** can automate the validation with improvement efficiency and performance.

Bayesian Network



**Approach:** BN is built to represent the dependence among sensors and replicated across timesteps. BN parameters are learned from over 15 years of ARM climate data to support distributed climate sensor validation. Inference validates sensors in the connected instruments.



### Bayesian Network

- \* 3 timesteps      \* 63 variables
- \* 3.9 avg states      \* 4.0 avg indegree
- \* 16,858 CPT entries

### Junction Tree

- \* 67 cliques
- \* 873,064 PT entries in cliques

# Use Case 14: Cellular Network Analytics in Telco Operation

**Goal:** Efficiently and uniquely identify *internal* state of Cellular/Telco networks (e.g., performance and load of network elements/links) using probes between monitors placed at selected network elements & endhosts

- Applied Graph Analytics to telco network analytics based on CDRs (call detail records): estimate traffic load on CSP network with low monitoring overhead

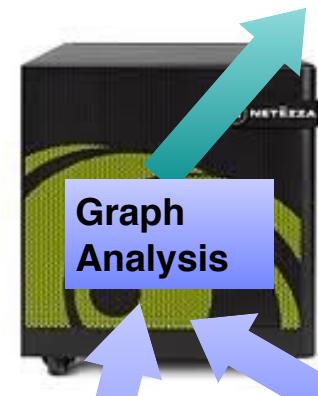
- (1) CDRs, already collected for billing purposes, contain information about voice/data calls
- (2) Traditional NMS\* and EMS\*\* typically lack of end-to-end visibility and topology across vendors
- (3) Employ graph algorithms to analyze network elements which are not reported by the usage data from CDR information

- Approach

- Cellular network comprises a hierarchy of network elements
- Map CDR onto network topology and infer load on each network element using graph analysis
- Estimate network load and localize potential problems

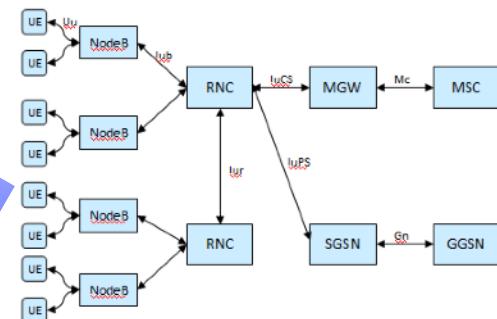


Network load level report



CDR

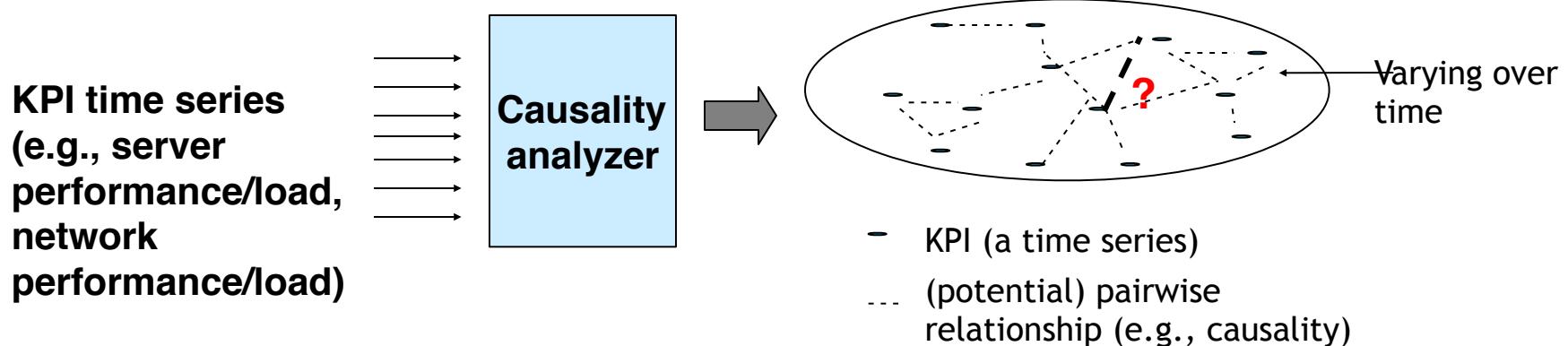
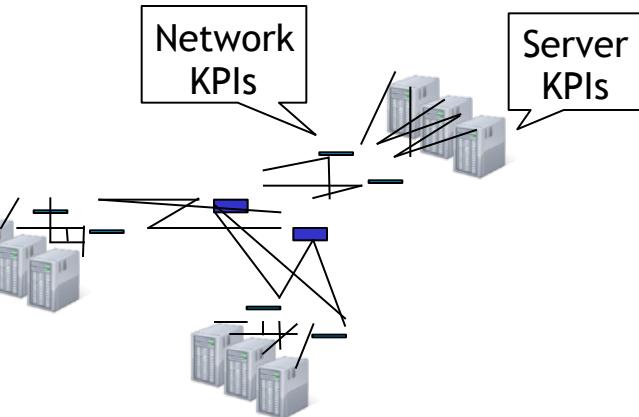
Network topology



# Use Case 15: Monitoring Large Cloud

**Goal:** Monitoring technology that can track the time-varying state (e.g., causality relationships between KPIs) of a large Cloud when the processing power of monitoring system cannot keep up with the scale of the system & the rate of change

- *Causality relationships (e.g., Granger causality) are crucial performance monitoring & root cause analysis*
- *Challenge: easy to test pairwise relationship, but hard to test multi-variate relationship (e.g., a large number of KPIs)*



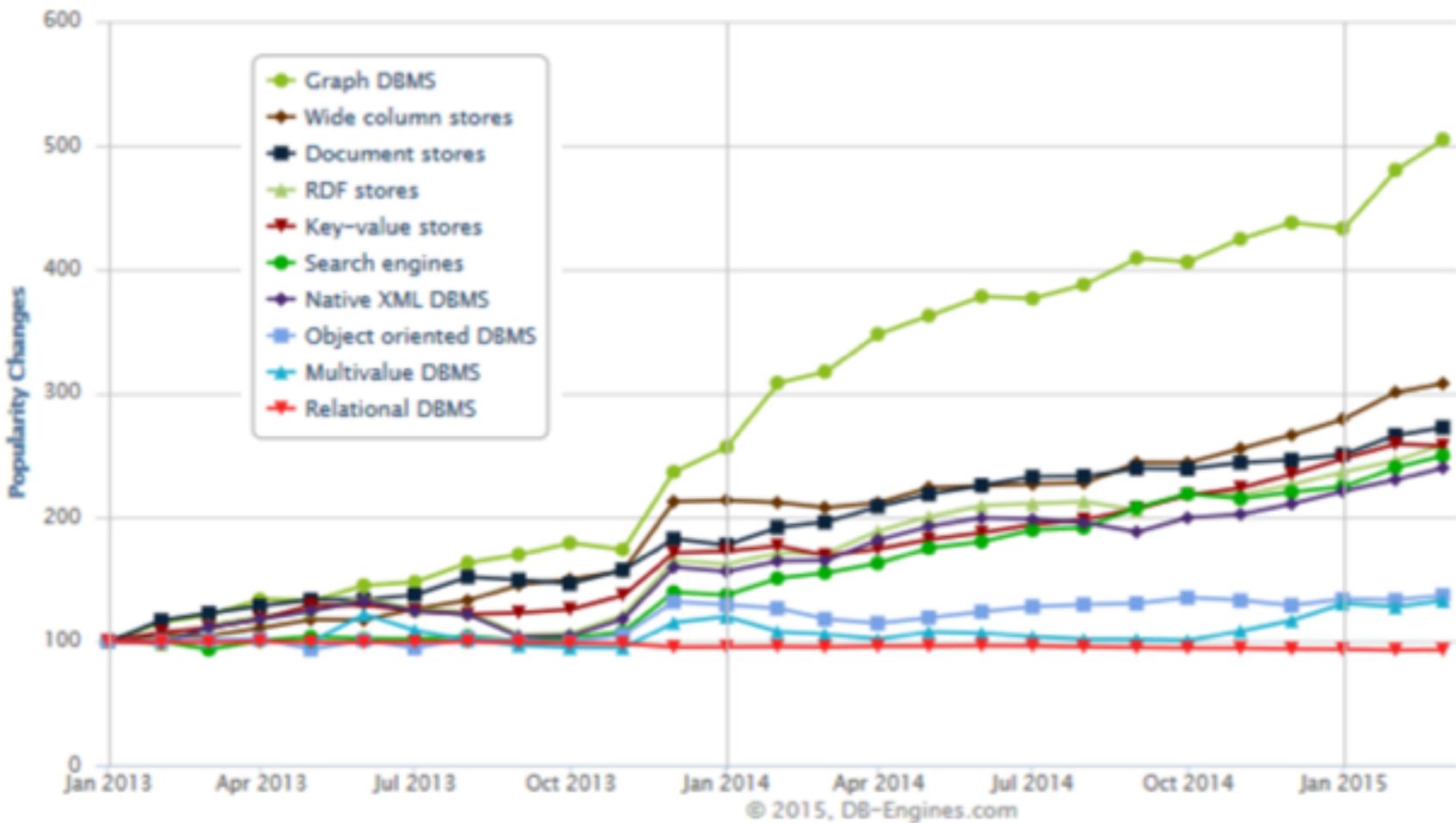
**Our approach:**  
 Probabilistic monitoring via sampling & estimation

Basic analytics engine  
 (e.g., pairwise granger causality)

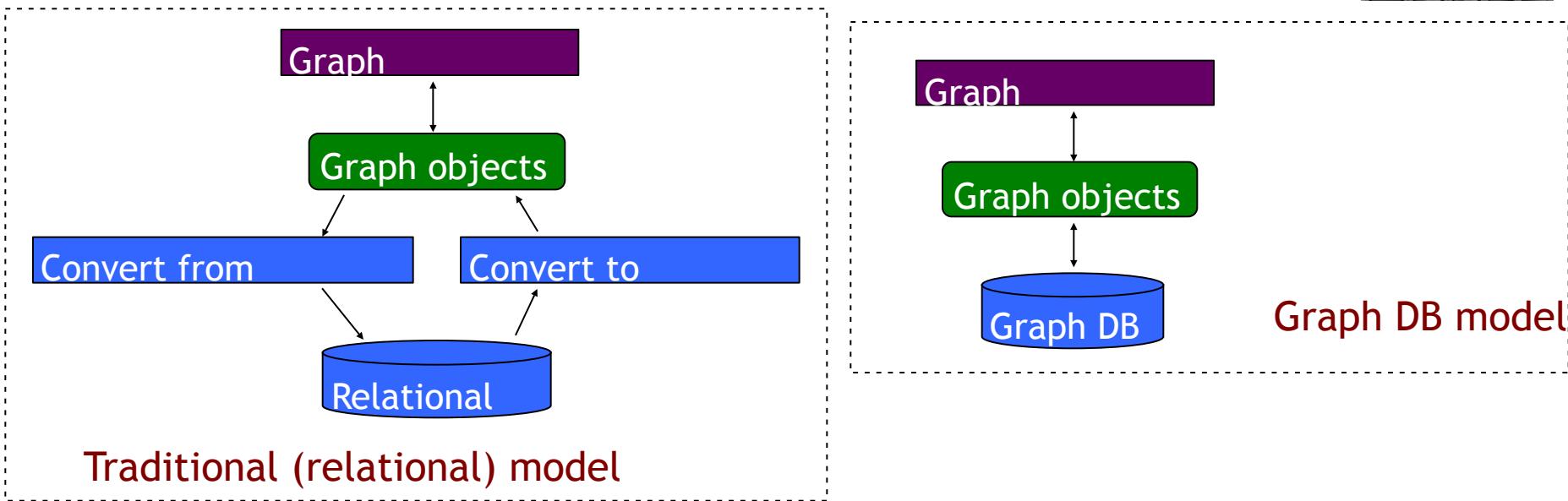
Link sampling & estimation

48 → Overall graph

# Category 5: Data Warehouse Augmentation



# Use Case 16: Code Life Cycle Improvement



- Advantages of working directly with graph DB for graph applications
  - (1) Smaller and simpler code
  - (2) Flexible schema → easy schema evolution
  - (3) Code is easier and faster to write, debug and manage
  - (4) Code and Data is easier to transfer and maintain

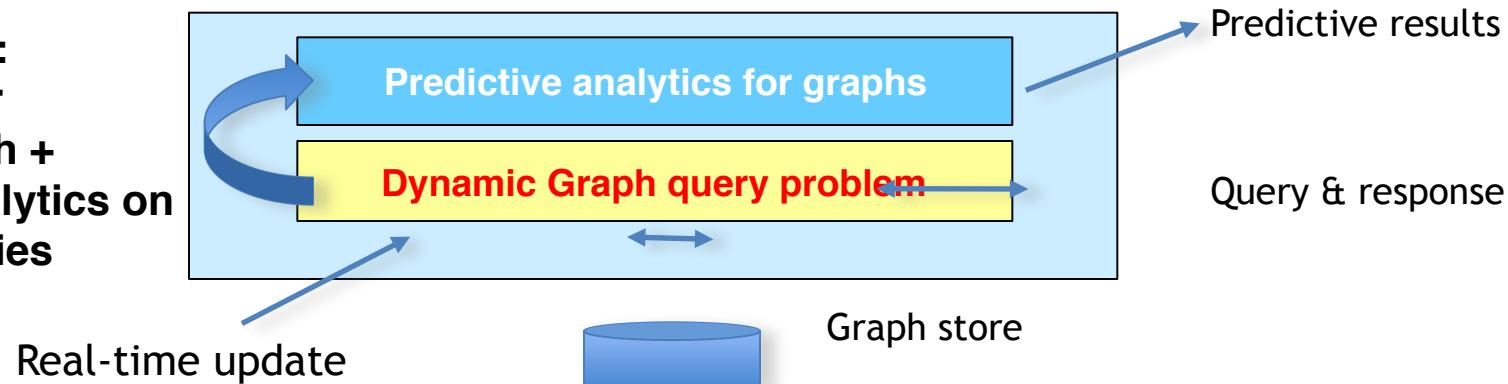
# Use Case 17: Smart Navigation Utilizing Real-time Road Information

**Goal:** Enable unprecedented level of accuracy in **traffic scheduling** (for a fleet of transportation vehicles) and navigation of individual cars utilizing the **dynamic real-time information** of changing road condition and predictive analysis on the data

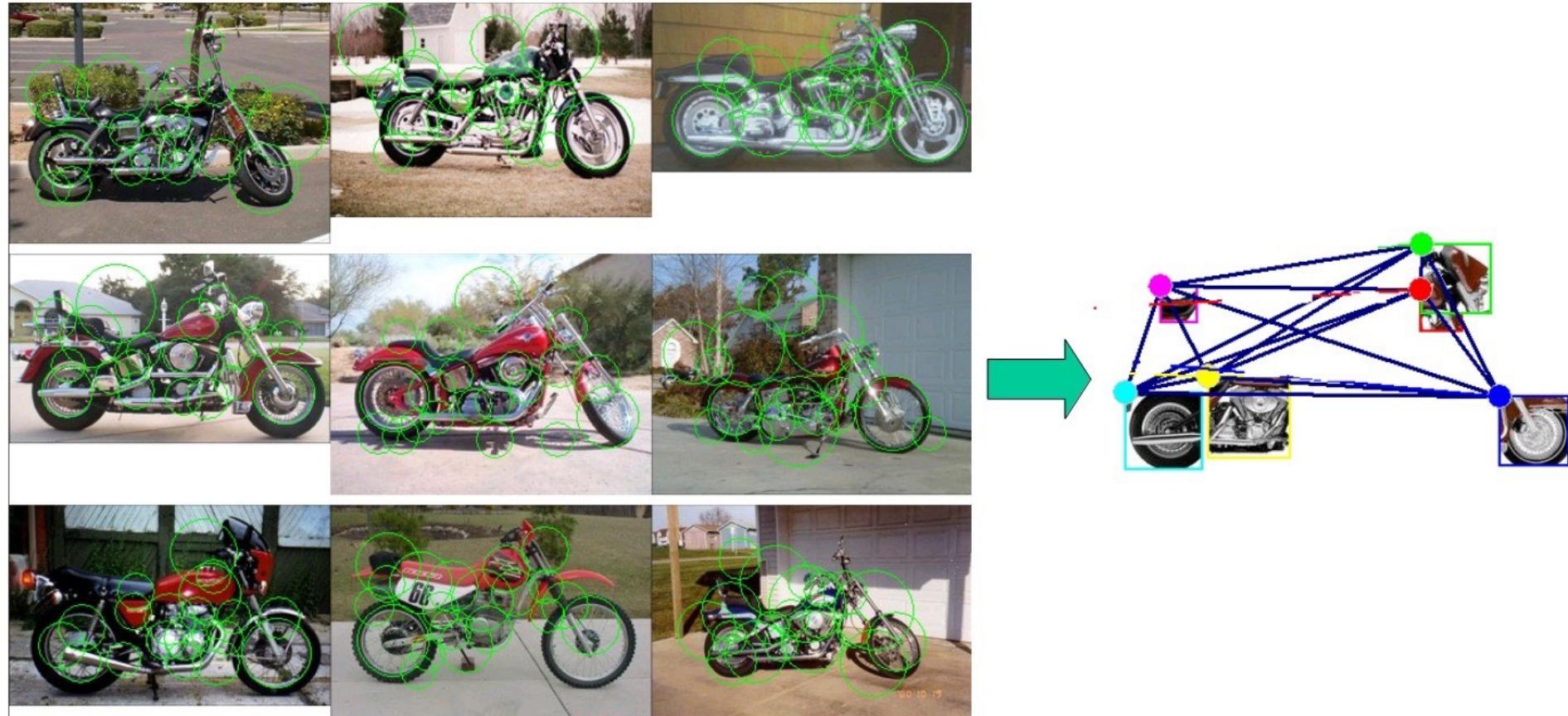
- Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shortest path computation) on time-varying graphs (order of magnitudes improvement over existing solutions)
- High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall



**Our approach:**  
**Querying over dynamic graph + predictive analytics on graph properties**

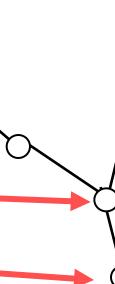


# Use Case 18: Graph Analysis for Image and Video Analysis



ARG s

Vertex  
Correspondence

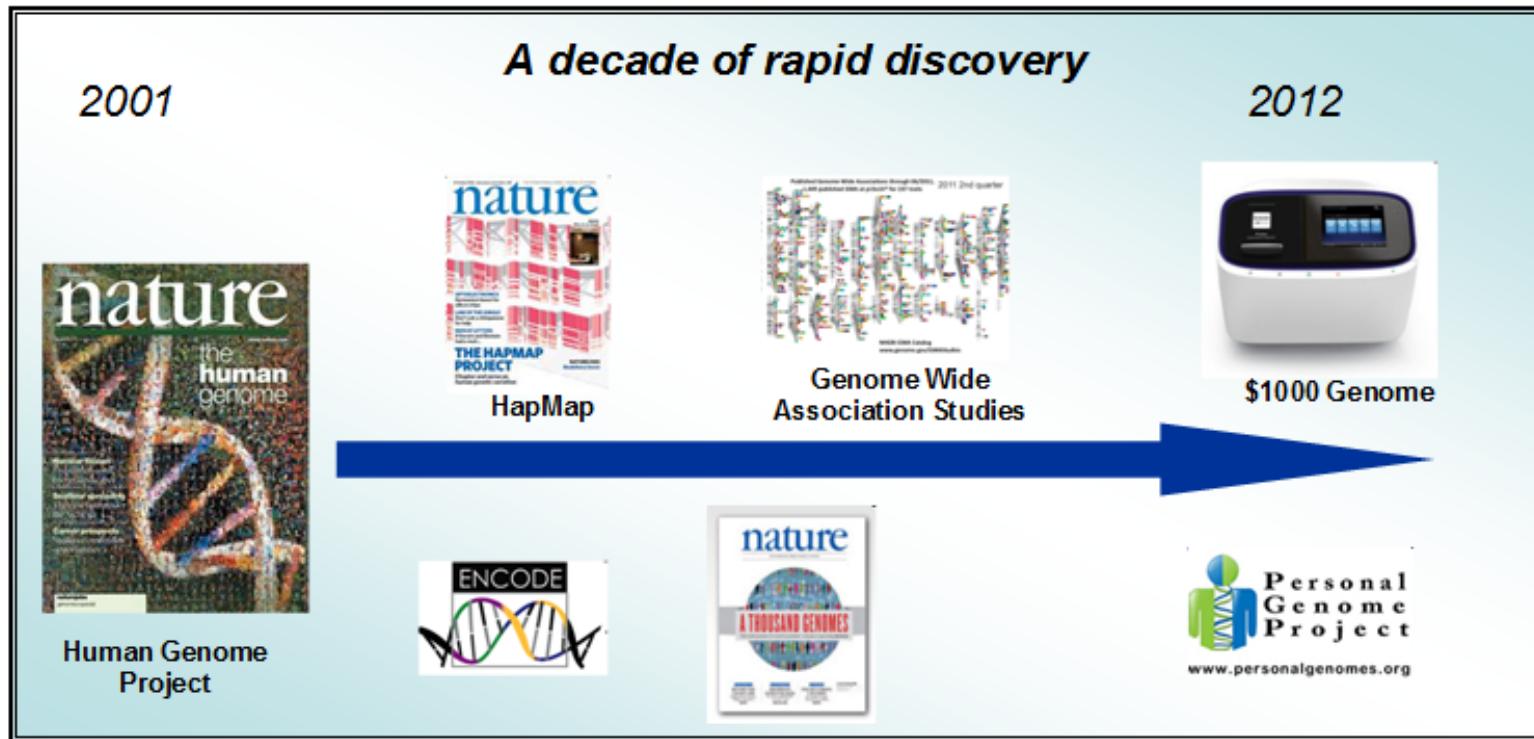


Attribute Transformation

ARG t  
 $Y_t$



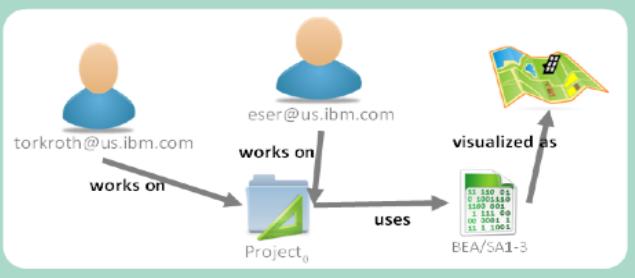
# Use Case 19: Graph Matching for Genomic Medicine



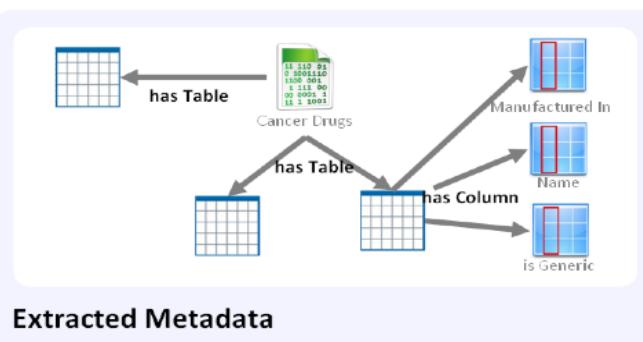
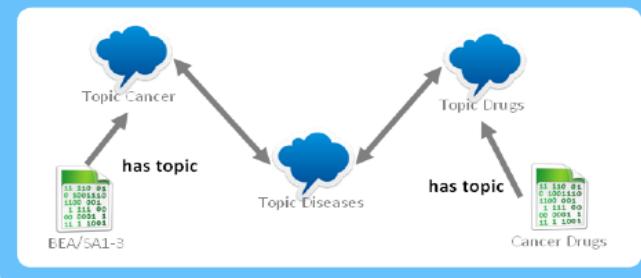
**Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the \$1000 Genome is almost reality.**

# Use Case 20: Data Curation for Enterprise Data Management

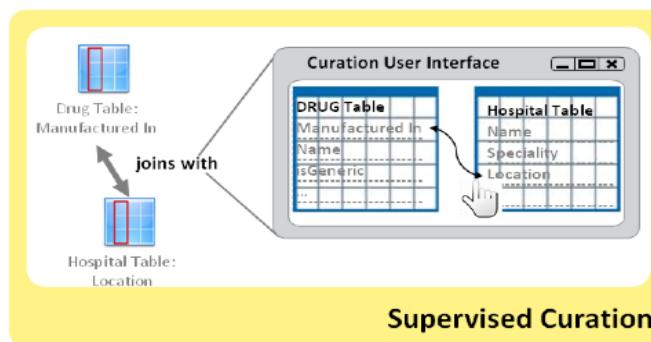
## Prior Collaborative Use



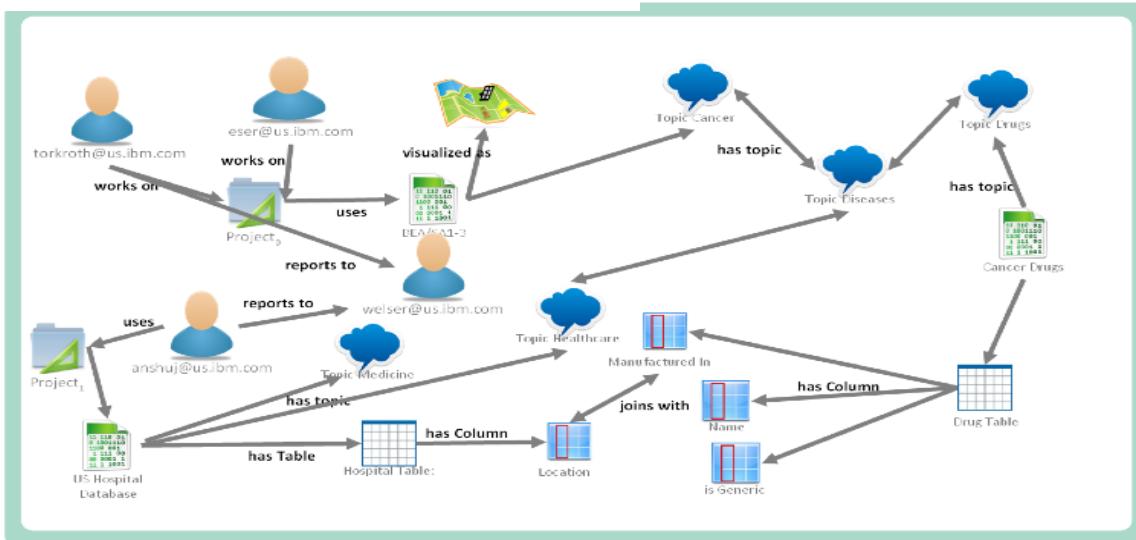
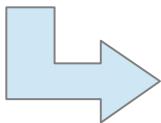
## Semantic Knowledge



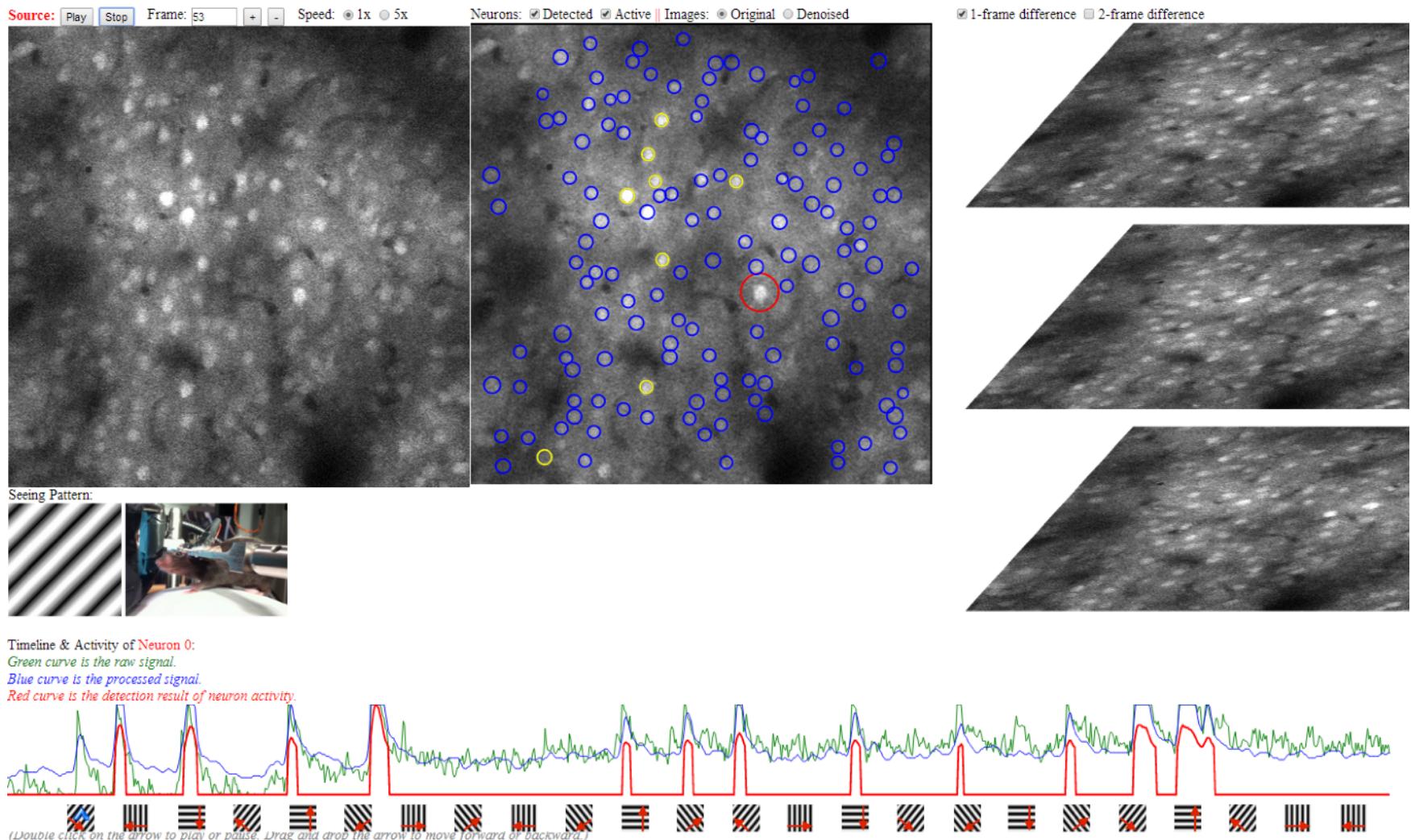
## Extracted Metadata



## Supervised Curation



# Use Case 21: Understanding Brain Network



# Use Case 22: Planet Security

- Big Data on Large-Scale Sky Monitoring



Photograph by Rob Ratkowski for the PS1SC

<b>Dangers from space</b> Learn about the threat to Earth from asteroids & comets and how the Pan-STARRS project is designed to help detect these NEOs. <a href="#">Learn more...</a> 	<b>1,400,000,000 pixels</b> Pan-STARRS has the world's largest digital cameras. <a href="#">Read about them here...</a> 	<b>The PS1 Prototype</b> PS1 goes operational and begins science mission PS1 Science Consortium formed... <a href="#">PS1SC Blog</a> <a href="#">PS1 image gallery</a> 
--	---	--

# Homework #0: Big Data Environment Setup and Test (due September 20, 5pm)

## 1. Warm-Up Exercises:

- Setup Google Cloud account and environment
- Install Google Cloud SDK
- Create a Spark cluster
- Word Count using Google Cloud Storage and Spark
- Hive and BigQuery

## 2. Data Analysis — NYC Bike Expert:

- Load data to a Cloud Storage
- Simple Analyses through BigQuery

## 3. Data Analysis — Understanding Shakespeare:

- Load data to a Cloud Storage
- Simple Analyses through Word Counts
- Analyses after running Natural Language Toolkit

<https://docs.google.com/document/d/1MWBVItrLL0MizDR-9q7-SY986SqCPcib-OrIB1ofYh0/>

## Homework Late Submission Policy

Friday 5pm: submission deadline

Saturday 5pm: 10% penalty

Sunday 5pm: 20% penalty

Monday 5pm: 30% penalty

Any submission after Monday 5pm will not be accepted.