Columbia University

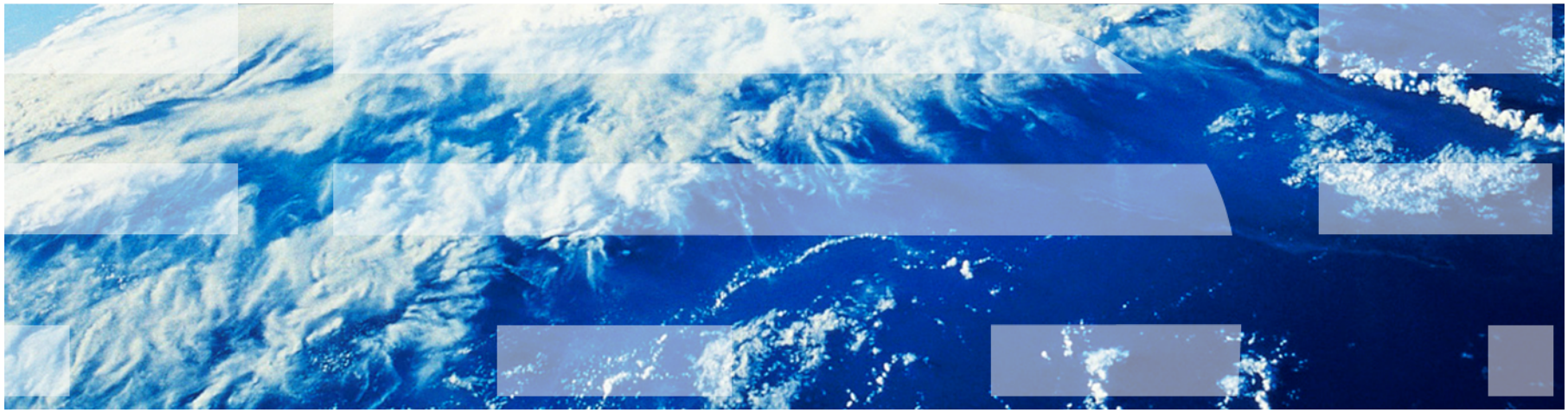
# EECS E6893 Big Data Analytics Lecture 1:

## *Overview of Big Data Analytics*

**Ching-Yung Lin**, Ph.D.

Adjunct Professor, Depts. of Electrical Engineering and Computer Science

IEEE Fellow



September 6h, 2018

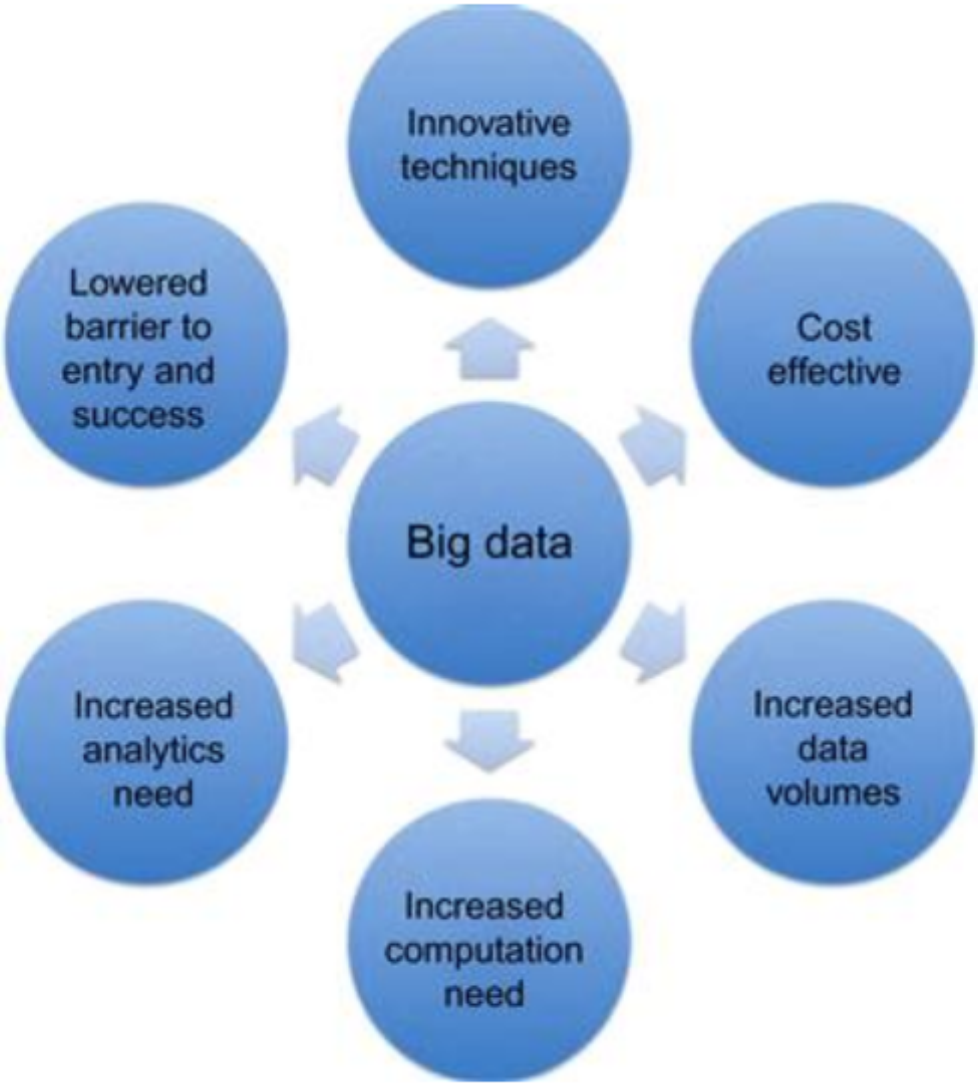
## Definition and Characteristics of Big Data

"*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

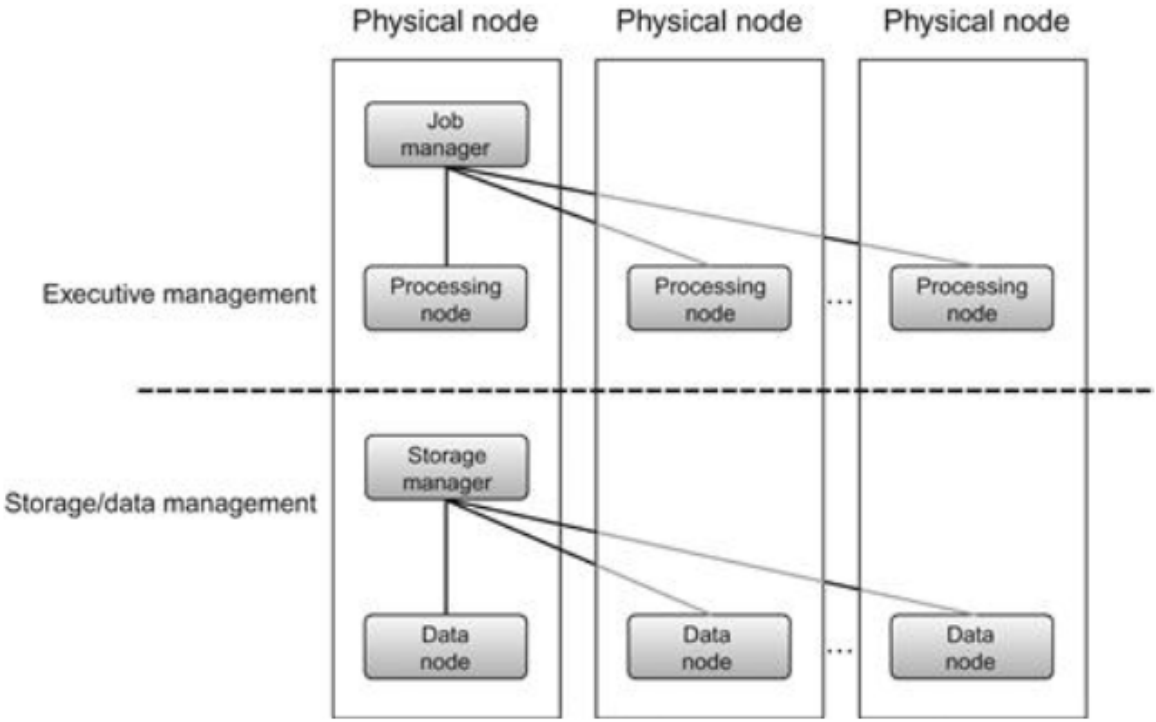
# What made Big Data needed?




"Big Data Analytics", David Loshin, 2013

# Key Computing Resources for Big Data

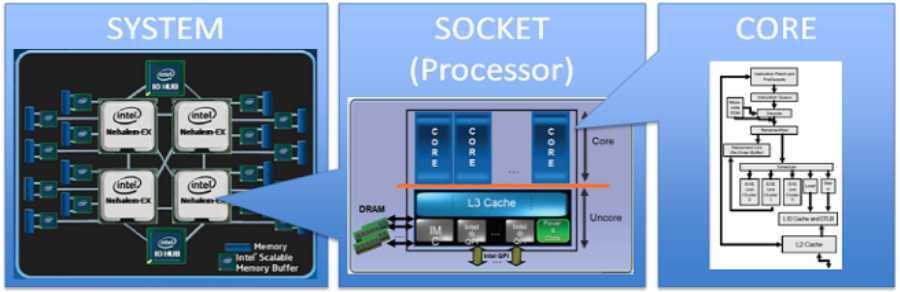
- Processing capability:  CPU, processor, or node.
- Memory
- Storage
- Network

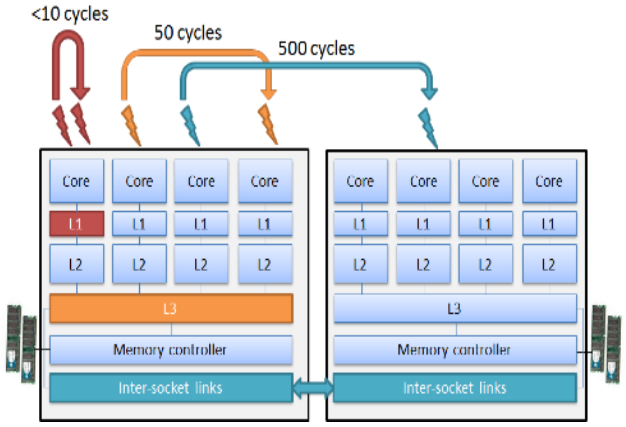



"Big Data Analytics", David Loshin, 2013

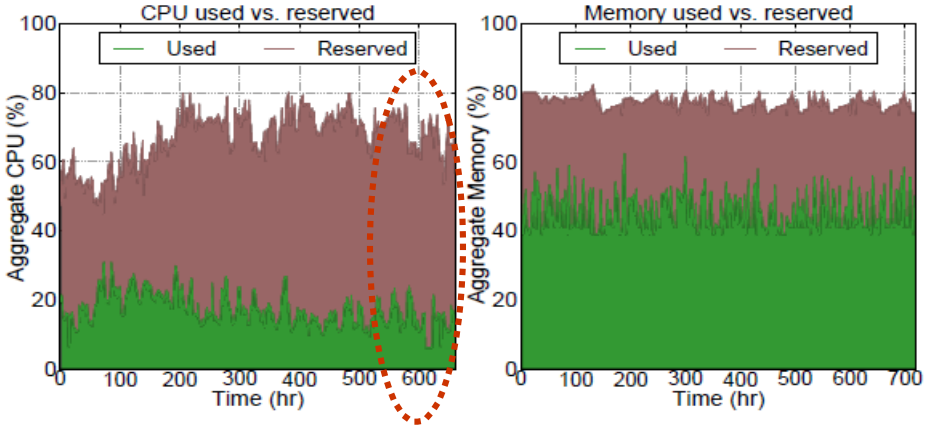
# Scalability — Scale Up & Scale Out

- 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/~cdel/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

# Contrasting Approaches in Adopting High-Performance Capabilities

| 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

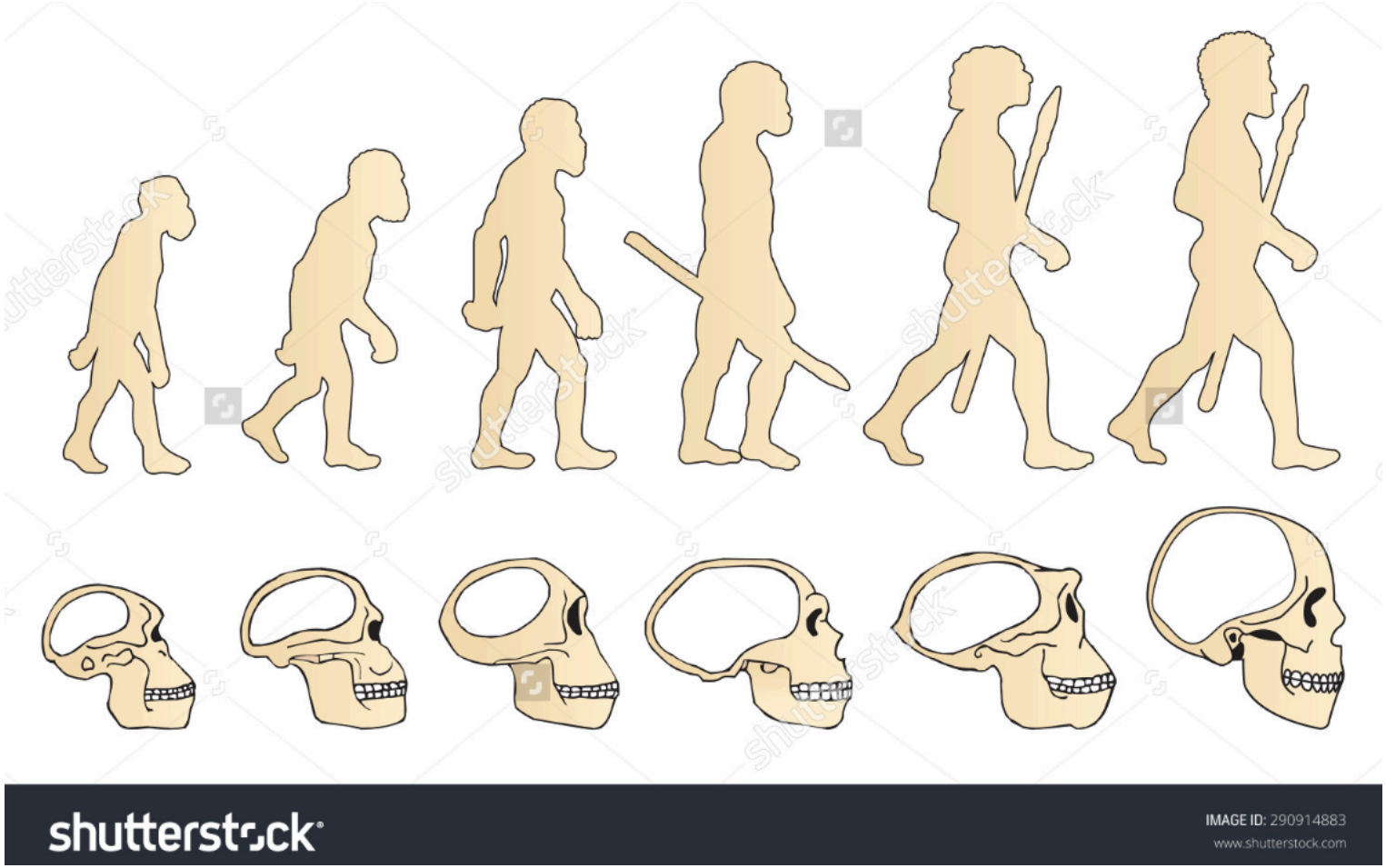
→ Artificial Intelligence


1997
IBM
GARRY
KASPAROV
DEEP
BLUE


2011
THINK
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600


2015
ALPHAGO
01:59:50
AlphaGo
Lee Sedol

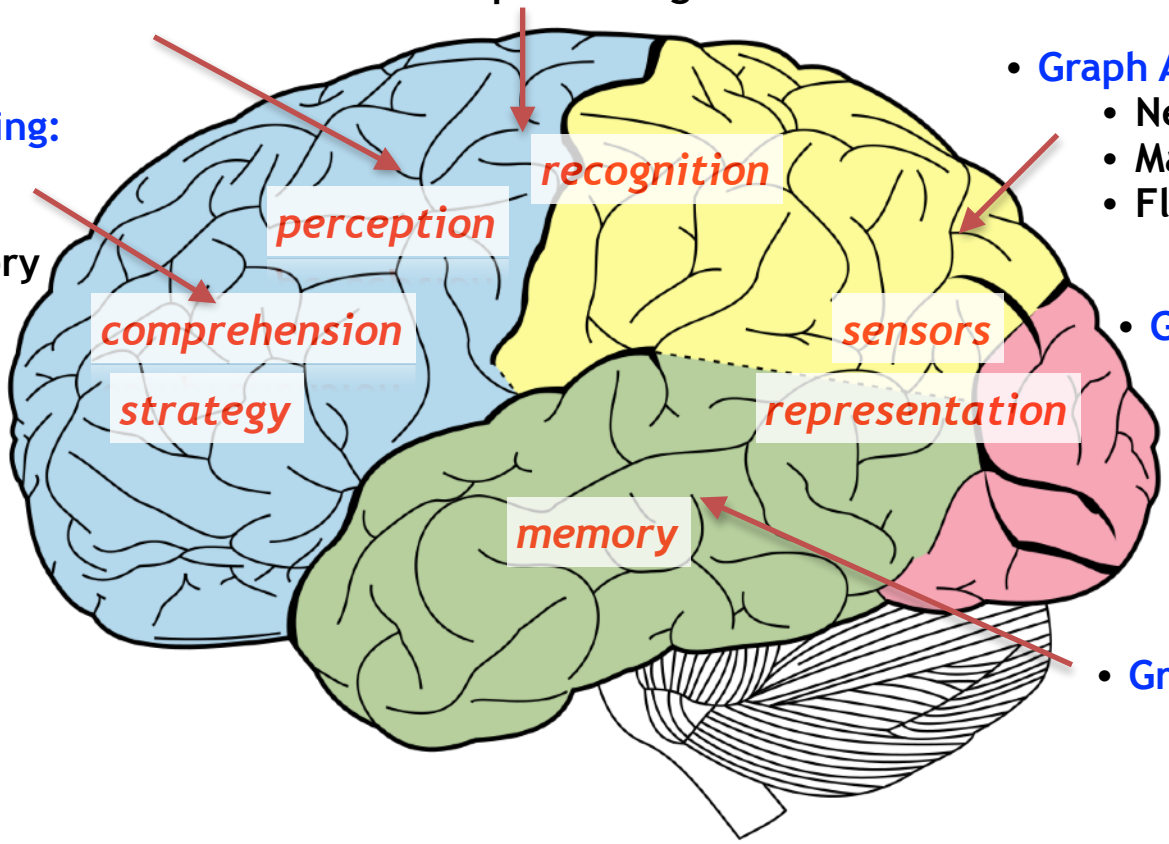

https://www.youtube.com/watch?v=BV8qFeZxZPE



# 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

# 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 and Artificial Intelligence 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 Main Thrust 1: Apache Hadoop


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

# Course Main Thrust 2: Apache Spark

Columbia University

Lightning-fast unified analytics engine

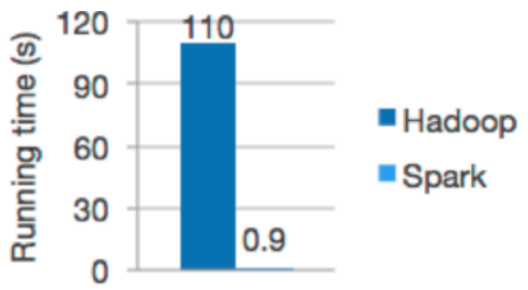
Download | Libraries | Documentation | Examples | Community | Developers

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

## Speed

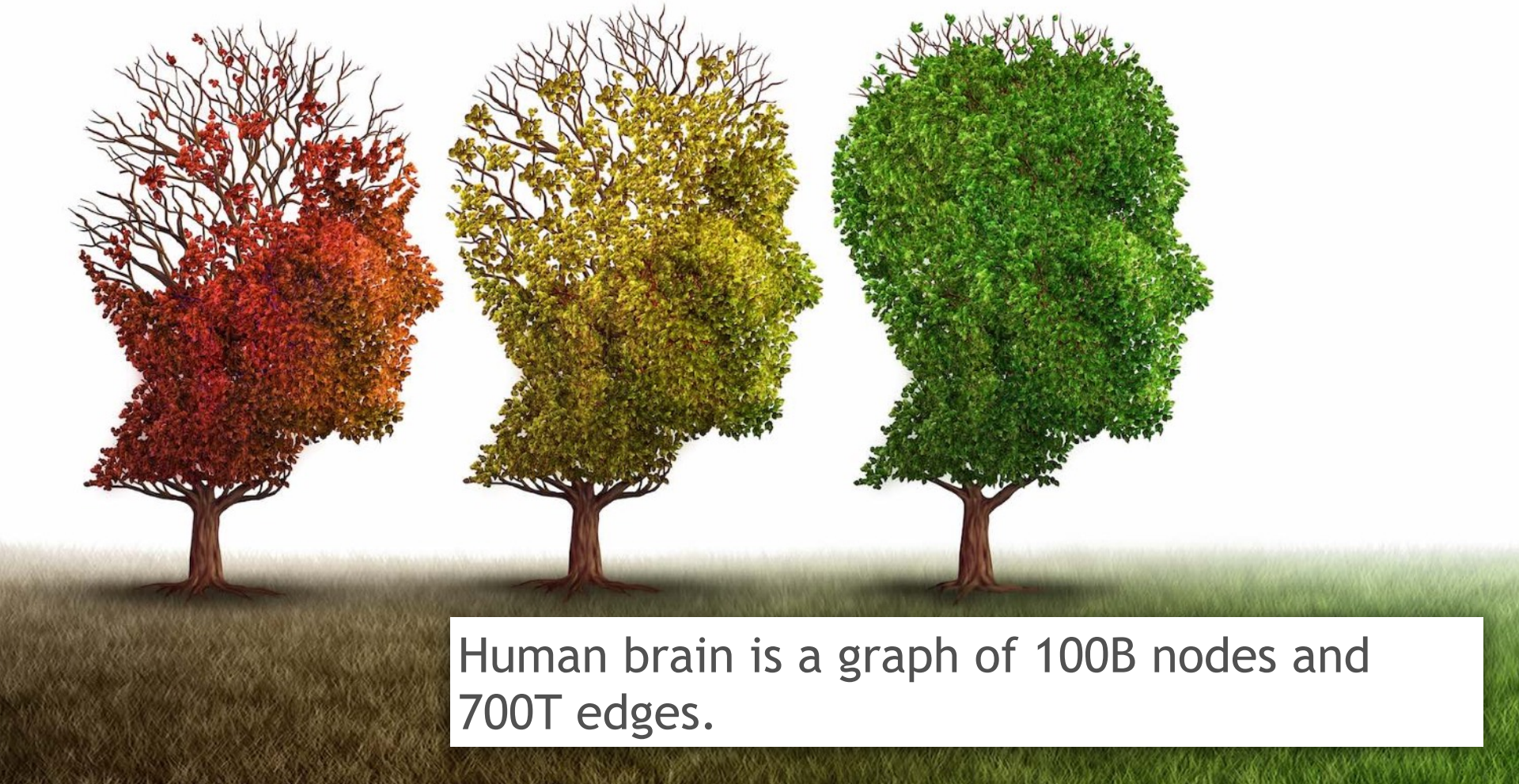
Run workloads 100x faster.

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.



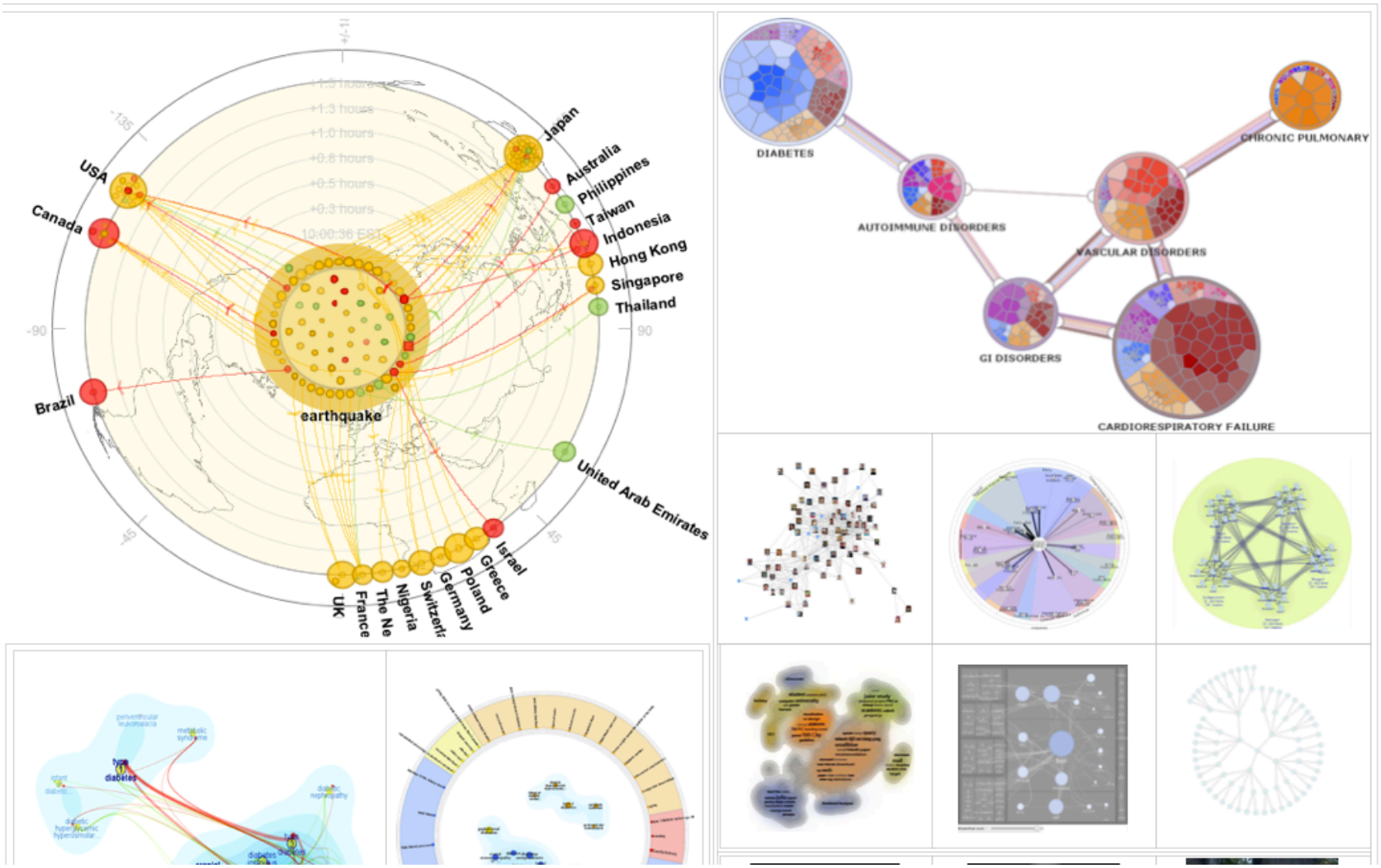

Logistic regression in Hadoop and Spark

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



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

# Course Main Thrust 4: Big Data Visualization



# Course Main Thrust 5: Big Data and AI Solutions

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

# Course Grading

- 4 Homeworks: 50%
  - -- **Individual work** (except HW #4); Language Requirement: C/C++, Java, JavaScript, Python)
  - -- Report and source code
  - **HW #1: Big Data System Installation and Testing**
  - **HW #2: Big Data Analytics using Hadoop and Spark**
  - **HW #3: Big Data Analytics using Spark and Graphs**
  - **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)

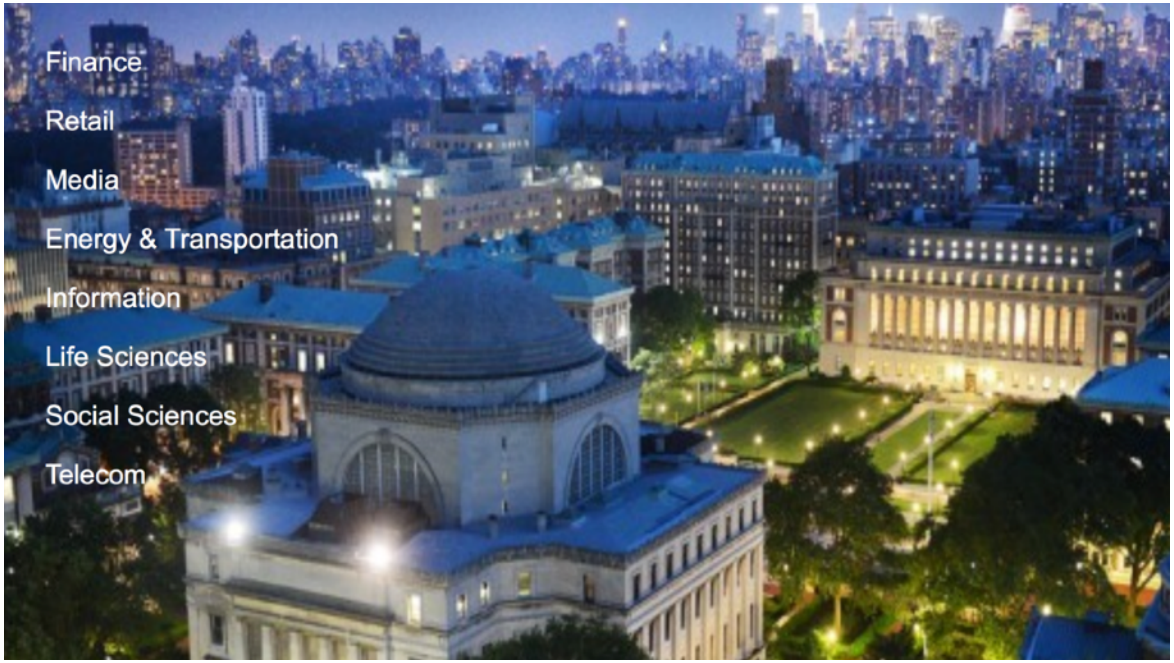
# Course Information

- 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

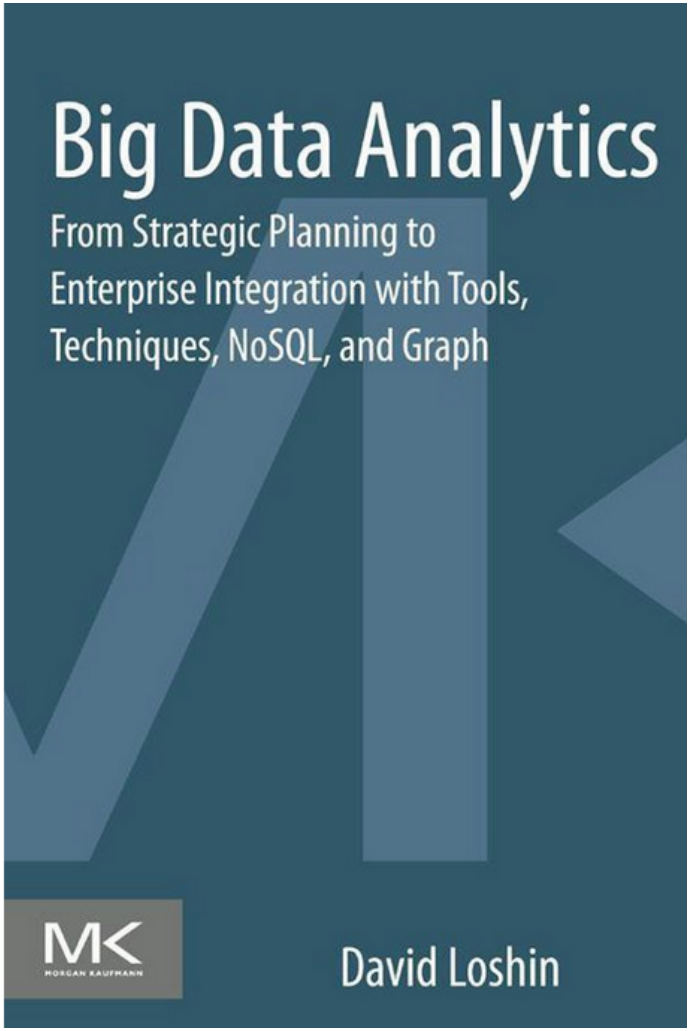
**Course Outline**

| Class Date | Class Number | Topics Covered | Assignment | Due |
|---|---|---|---|---|
| 09/06/18 | 1 | Introduction to Big Data Analytics | HW #1 Big Data System Installation and Testing | |
| 09/13/18 | 2 | Big Data Platforms and Data Storage | | |
| 09/20/18 | 3 | Big Data Analytics Algorithms I | HW #2 Analytics using Hadoop and Spark | HW #1 |
| 09/27/18 | 4 | Big Data Analytics Algorithms II | | |
| 10/04/18 | 5 | Big Data Analytics Algorithms III | HW #3 Analytics using Spark and Graphs | HW #2 |
| 10/11/18 | 6 | Linked Big Data Graph Database and Analytics | | |
| 10/18/18 | 7 | Big Data Visualization -- I | HW #4 Big Data Analytics Visualization | HW #3 |
| 10/25/18 | 8 | Big Data Visualization -- II | | |
| 11/01/18 | 9 | Final Project Proposal Presentations | | HW #4 & Proposal Slides |
| 11/08/18 | 10 | Big Data Analytics and AI for Finance I | | |
| 11/15/18 | 11 | Big Data Analytics and AI for Finance II | | |
| 11/22/18 | | NO CLASS -- Thanksgiving Holiday | | |
| 11/29/18 | 12 | Big Data Analytics and AI for Healthcare | | |
| 12/06/18 | 13 | Big Data Analytics and AI for Security | | |
| 12/13/18 | 14 | Big Data Analytics Workshop | | Final Project Slides |

# Other Issues

- Professor Lin:
  - Office Hours:
    Thursday after the class: 9:40pm – 10:00pm (SIPA 417, lecture room)
  - Contact: c.lin@columbia.edu
- TAs (CAs/IAs/Graders) —
  - Vishal Anand (va2361)
  - Yanbei Pang (yp2442)
  - Chao-Yang Lo (cl3636)
  - Pratyus Pati (pp2636)
  - Zekun Gong (zg2273)
  - TBDs, probably have 8-10 TAs in total.

# Reading Reference for Lecture 1




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 360º 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

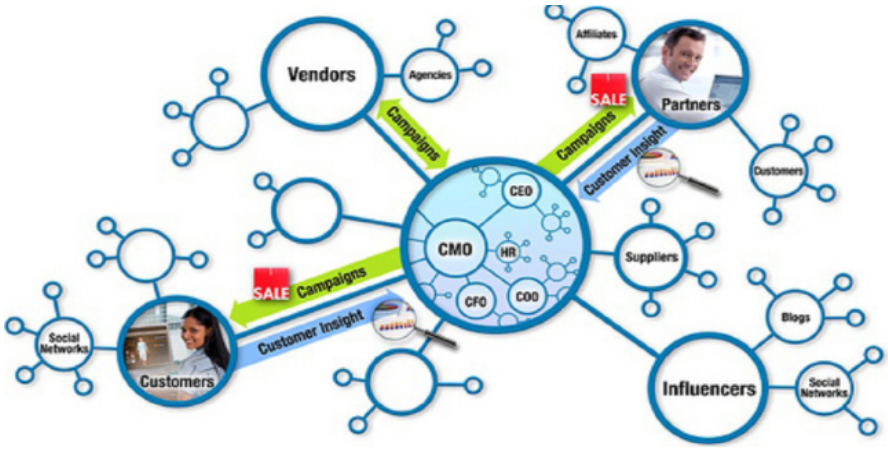
**Data Warehouse Augmentation**
Integrate big data and data warehouse capabilities to increase operational efficiency

# Big Data Examples -- Application Use Cases

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

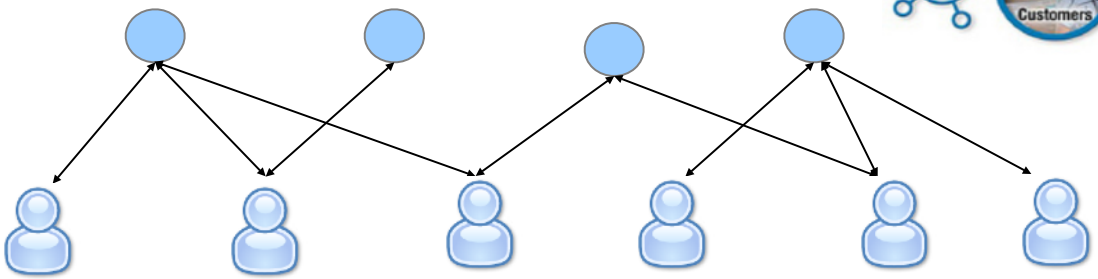
# Category 1: 360° View

*Recommendation*


item

user



*Enhancing:*

Coremetrics An IBM Company

unica

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

COLUMBIA UNIVERSITY

**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

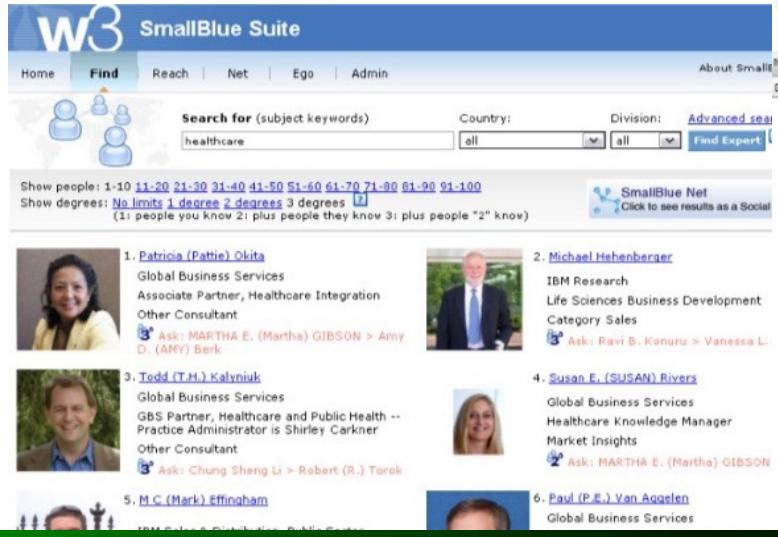
1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

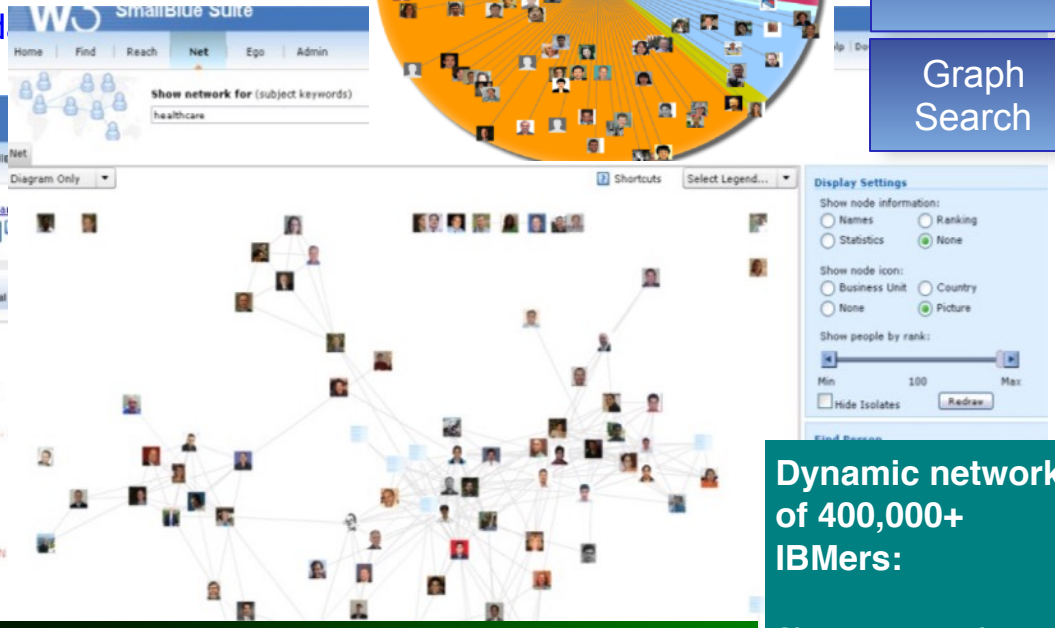
200,000 people's consulting project & earning d

Shortest Paths

Centralities

Graph Search



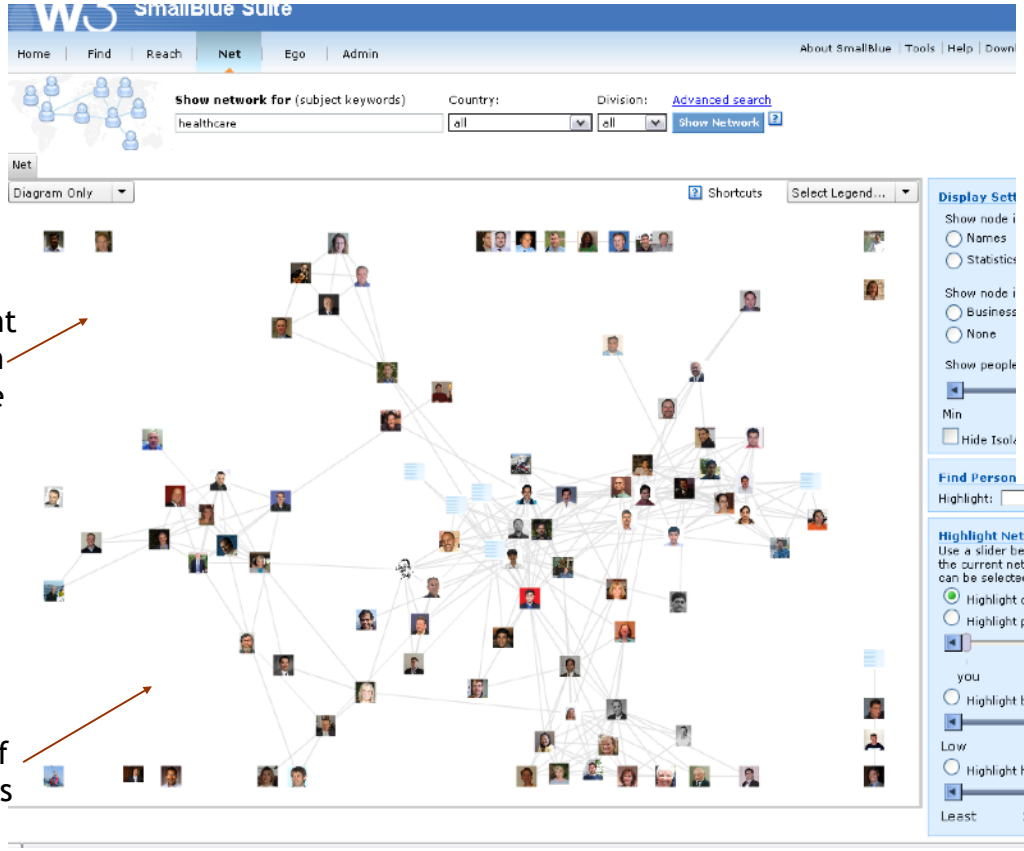

- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: $7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings an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

**Dynamic networks of 400,000+ IBMers:**

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

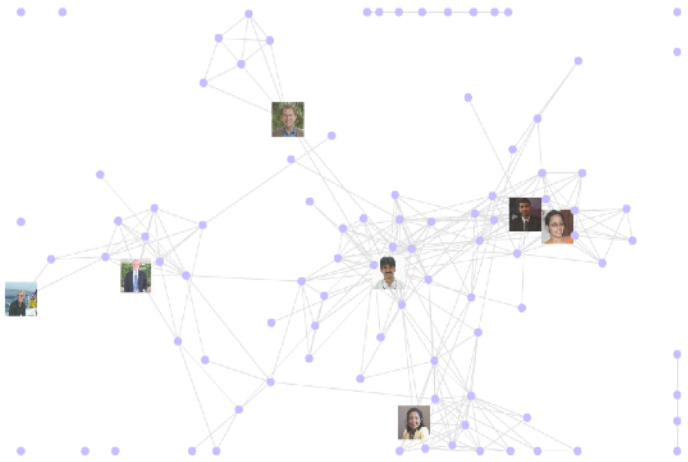
# Finding and Ranking Expertise – Social Network Analysis

- Decades of Social Science studies demonstrates that (social) network structure is the key indicator determining a person's influence, organizational operation efficiency, social capital to get help, potential to be successful, etc.
- Who are the key bridges? Who have the most connections? How do these experts cluster?
- Analogy – Google founders utilized the concept of network analysis on webpages to create ranking.




Independent experts on healthcare

A cluster of XYZ experts



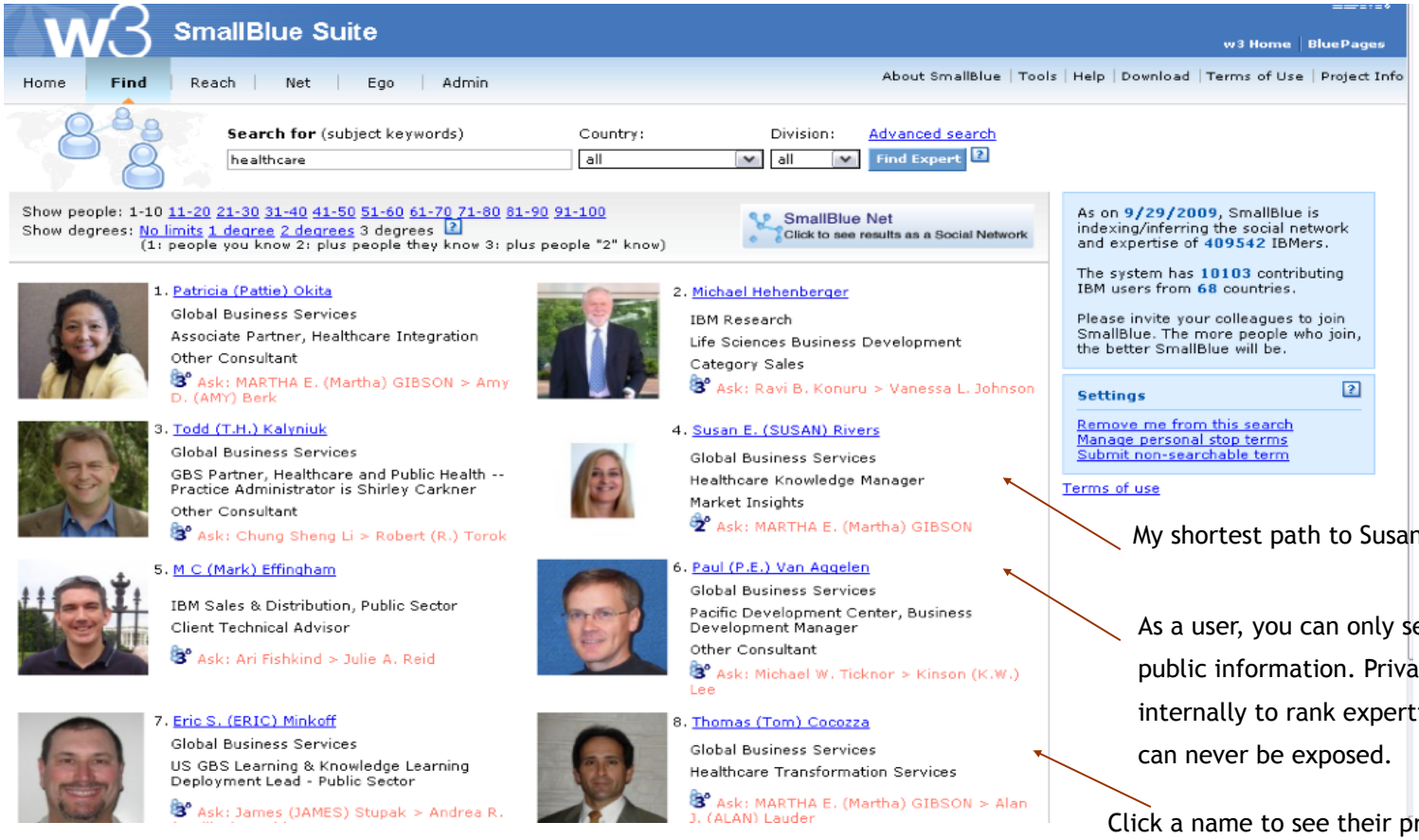
Influencers are the one with high 'Betweeness' and 'Degree' value

UI to highlight experts based on my social proximity, the number of experts she connects, or the 'social bridges' importance

**SmallBlue analyzes underlining dynamic network structure in enterprise**

# User Interface of finding knowledgeable and influential colleagues

- Search for the most knowledgeable colleagues within organization or my 3-degree network for who knows topic XYZ (or within a country, a division, a job role, or any group/community)
- Based on IBM HR requirements, adding the 'sponsored search' for business department needs
- IBM HR gives a list of about 10,000 IBMers whose name should not be listed in the search result – mostly high level managers, lawyers, people involving acquisition, etc.
- A list of 2,000+ words that are inappropriate to search in enterprise.

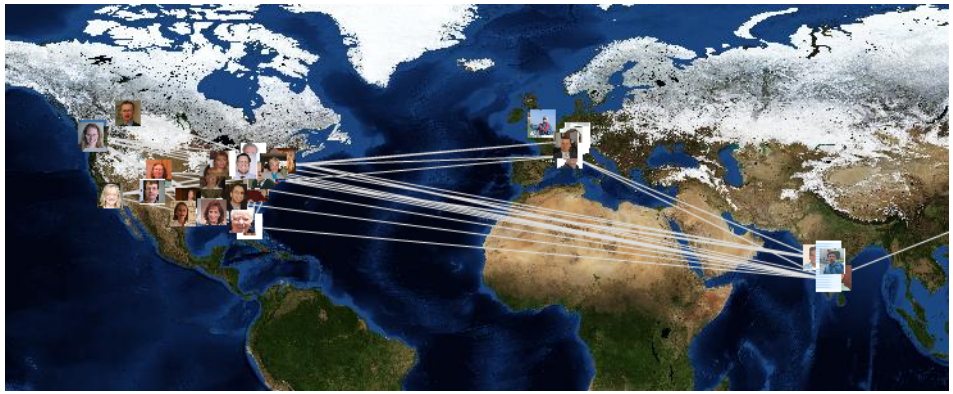



My shortest path to Susan

As a user, you can only see their public information. Private info is used internally to rank expertise but private data can never be exposed.

Click a name to see their profile (SmallBlue Reach)

# Visualize social roles of individuals in company

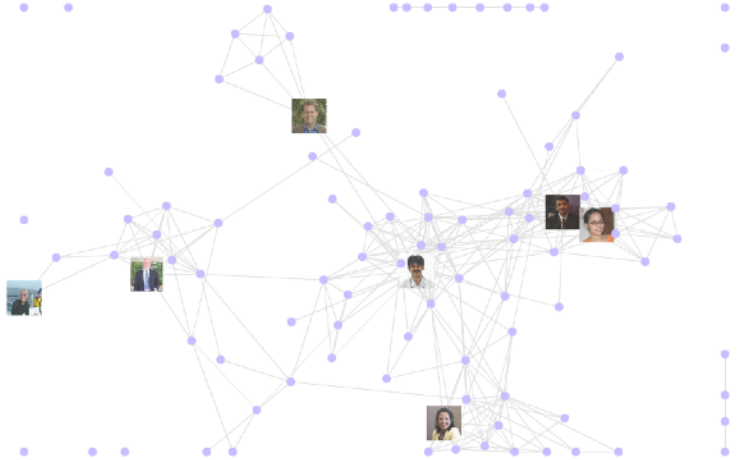


Example: Healthcare experts in the world

Example: Healthcare experts in the U.S.



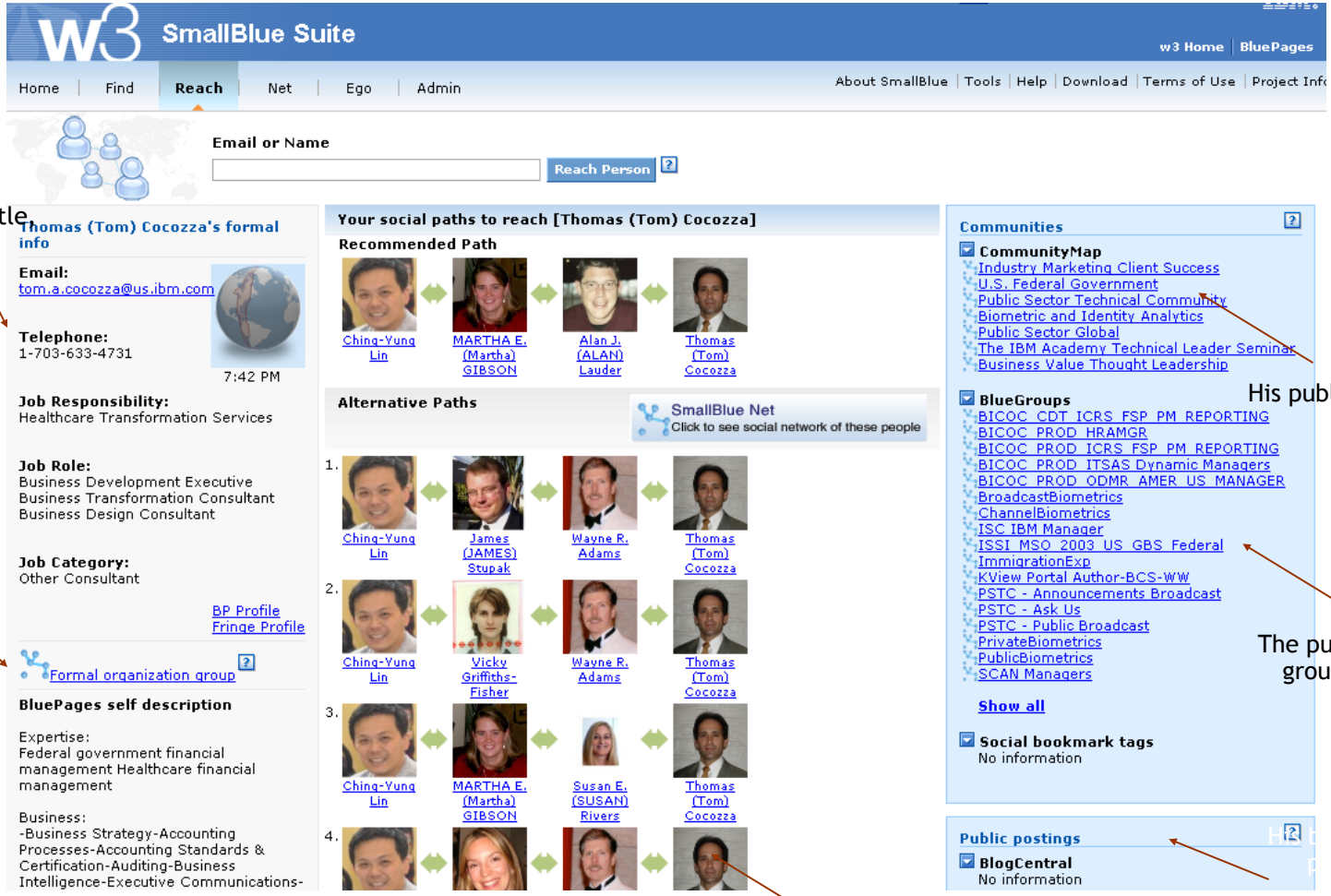
Connections between different divisions



Key social bridges

# Shortest Paths between two people in enterprise

- **Example: Is Tom a right person to me?**




His official job role, title, contact info

His self-described expertise

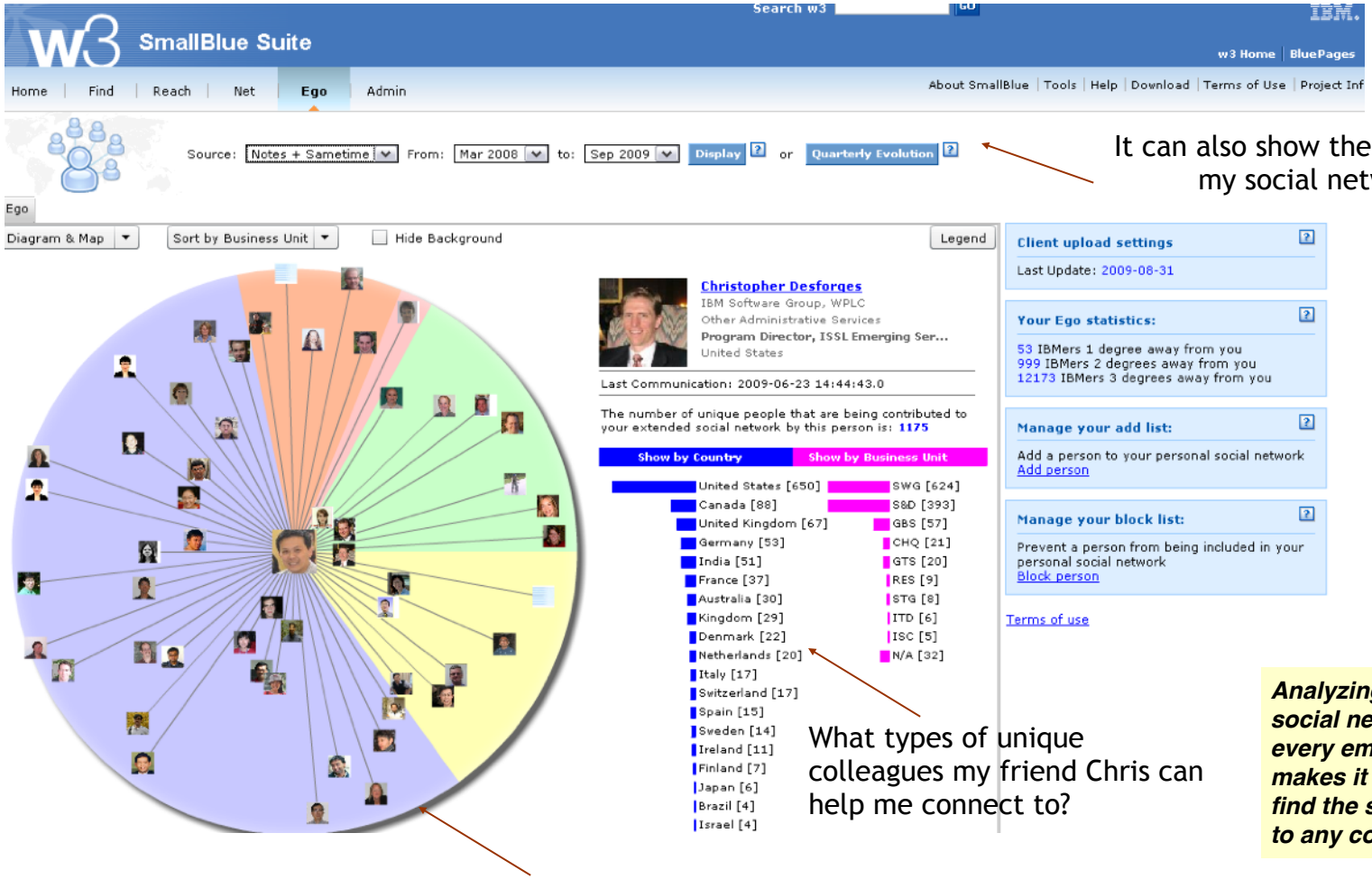
His public communities

The public interest groups he is in

My various paths to Tom. SmallBlue can show the paths to any colleagues up to 6-degree away

# Personal social network capital management

- What is a friend's social capital to me? Am I losing an 'important' friend?



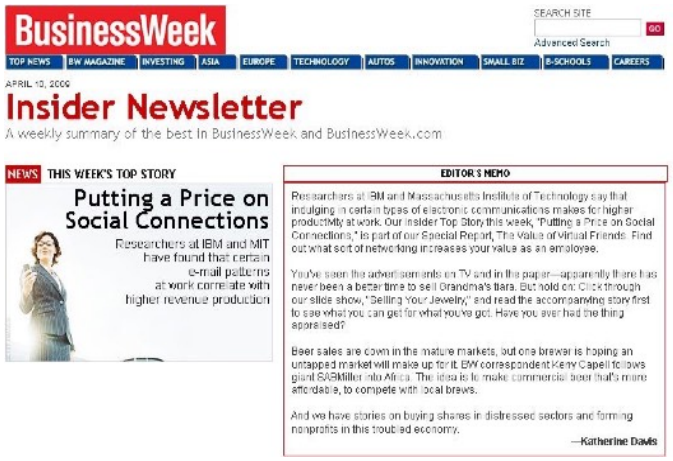

It can also show the evolution of my social network..

What types of unique colleagues my friend Chris can help me connect to?

***Analyzing existing social networks of every employee That makes it possible to find the shortest path to any colleague..***

Evolutionalry personal social network

# Network Value Analysis – First Large-Scale Economical Social Network Study

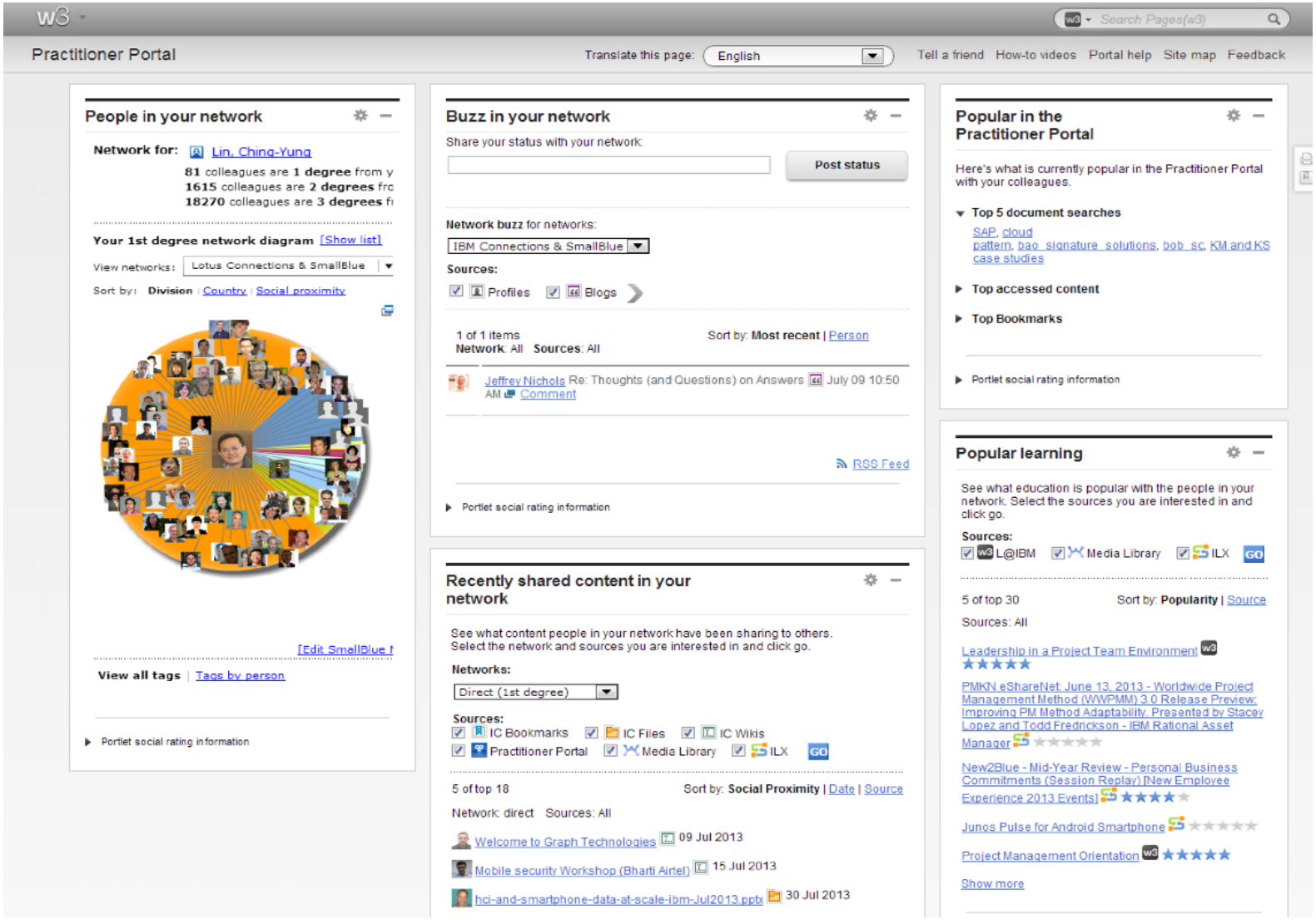




Productivity effect from network variables

- An additional person in network size ~ $986 revenue per year
- Each person that can be reached in 3 steps ~ $0.163 in revenue per month
- A link to manager ~ $1074 in revenue per month
- 1 standard deviation of network diversity (1 - constraint) ~ $758
- 1 standard deviation of btw ~ -$300K
- 1 strong link ~ $-7.9 per month

- Structural Diverse networks with abundance of structural holes are associated with higher performance.
  - *Having diverse friends helps.*
- Betweenness is negatively correlated to people but highly positive correlated to projects.
  - *Being a bridge between a lot of people is bottleneck.*
  - *Being a bridge of a lot of projects is good.*
- Network reach are highly corrected.
  - *The number of people reachable in 3 steps is positively correlated with higher performance.*
- Having too many strong links — the same set of people one communicates frequently is negatively correlated with performance.
  - *Perhaps frequent communication to the same person may imply redundant information exchange.*

# Use Case 2: Recommendation



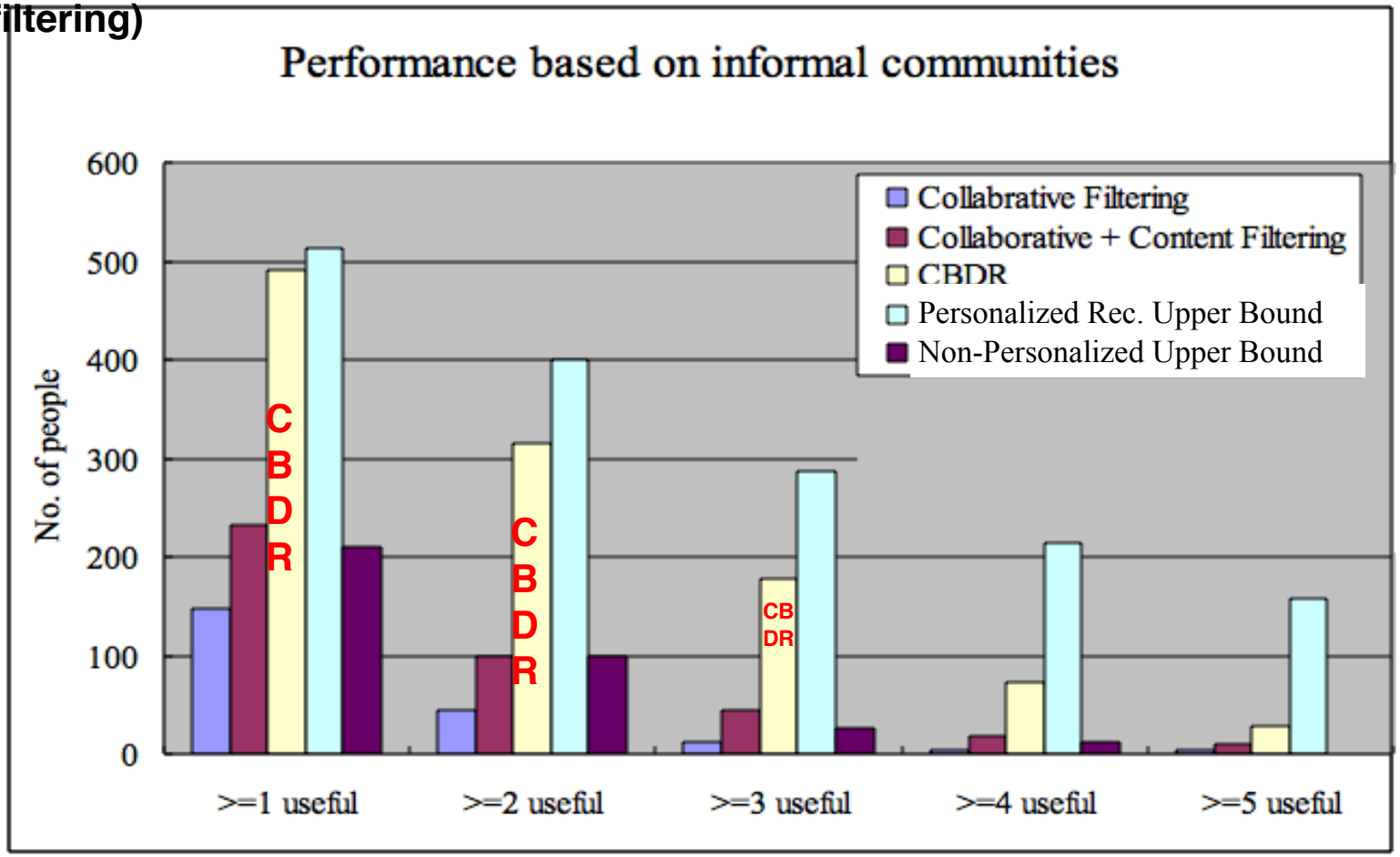

- Integrated Practitioner Portal, KnowledgeView, Media Library, Lotus Connections, and Learning@IBM and for a personalized ranking

# Improving Recommendation Quality by Graph Community Analytics

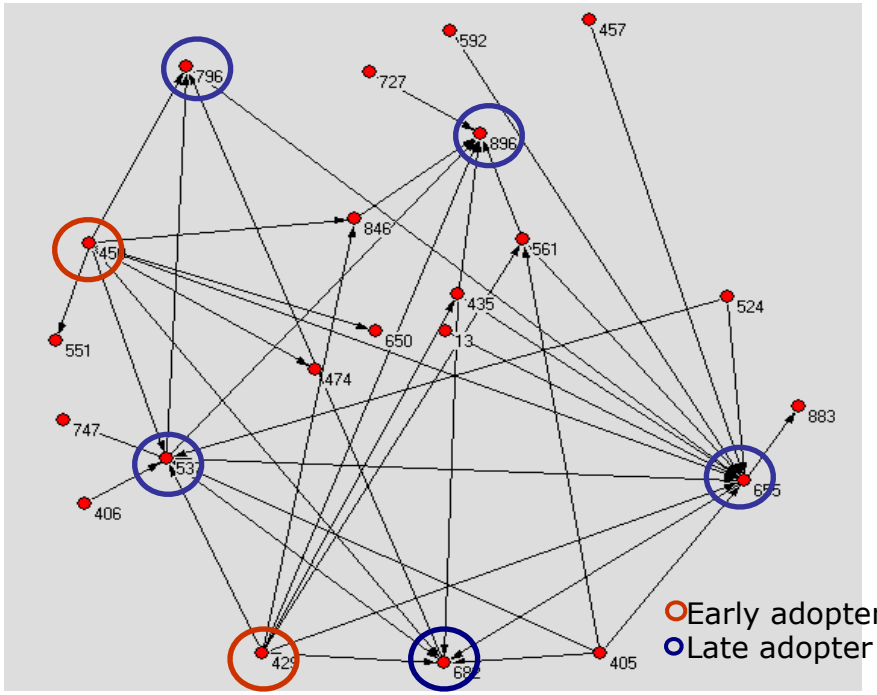
Graph Communities

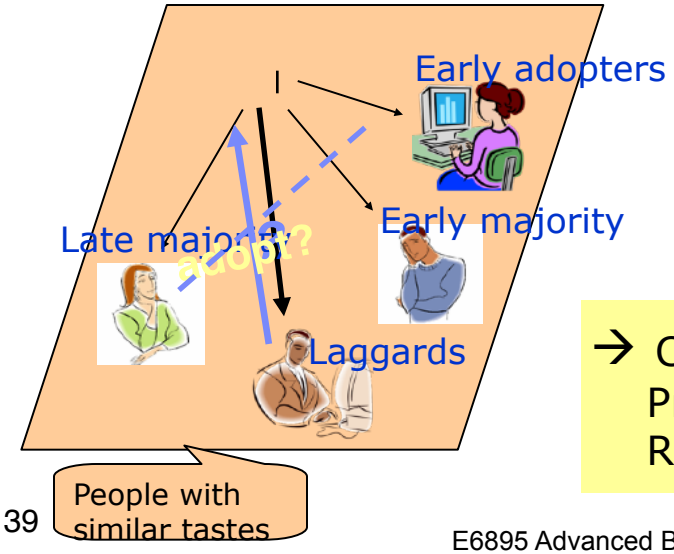
- A 3rd party Knowledge Repository: 30K users and 20K documents. Study the most active 697 users who have at least 20 download in a year.

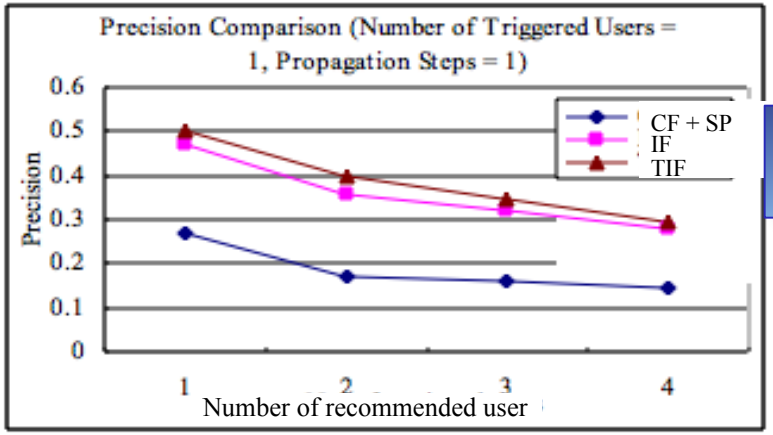
- **Results: beyond Collaborative Filtering: (1) Collaborative + Content Filtering (53% improvement); (2) CBDR: Collaborative + Content Filtering + Graph Community Analytics (259% accuracy improvement over collaborative filtering)**



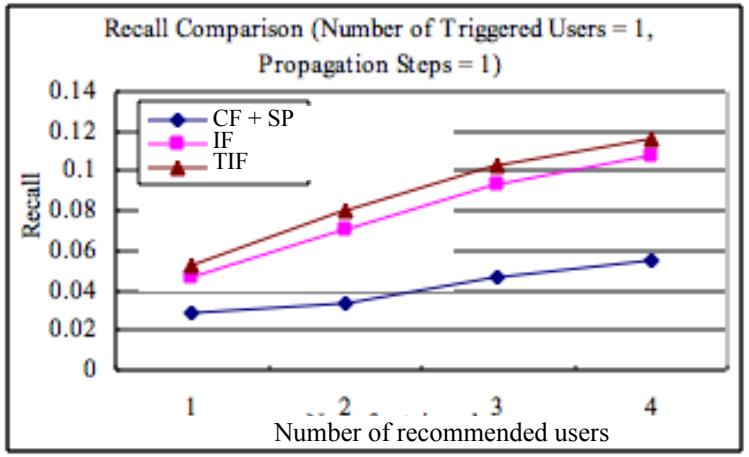
# Use Case 3: Recommendation for Commerce




Network Info Flow



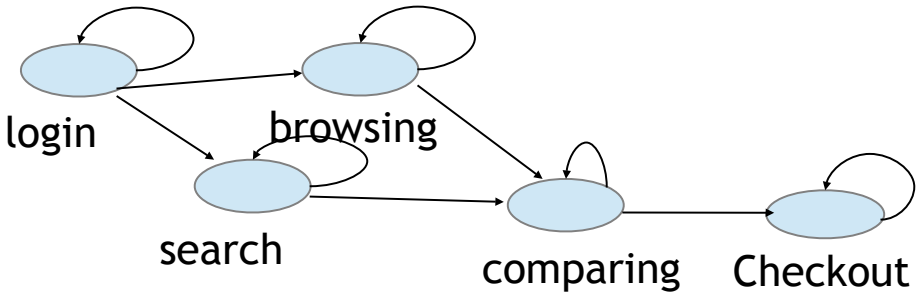

Tests:
- 1 month
- 586 new docs
- 1,170 users

IF: Graphical Information Flow Model
TIF: Joint Topic Detection + Information Flow Model

→ Comparing to Collaborative Filtering (CF) + Similar People
Precision: IF is 91% better, TIF is 108% better
Recall: IF is 87% better, TIF is 113% better

# Customer Behavior Sequence Analytics

Markov Network | Latent Network | Bayesian Network

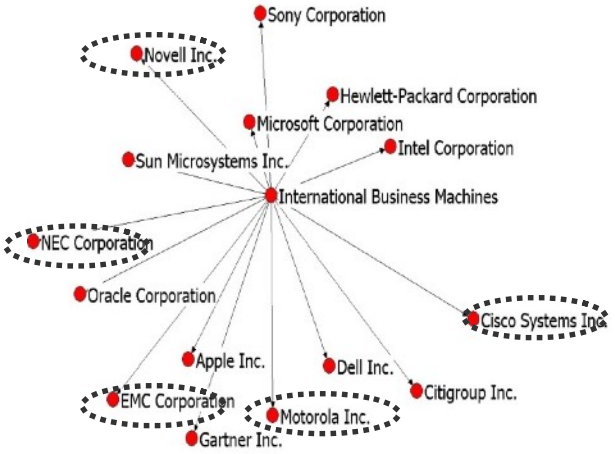




- Behavior Pattern Detection
- Help Needed Detection

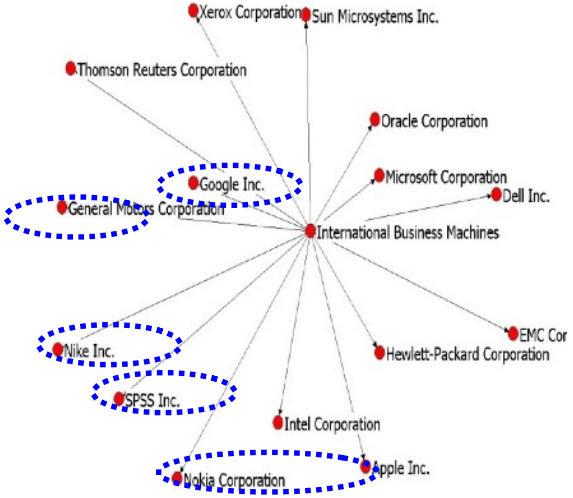
# Use Case 4: Graph Analytics for Financial Analysis

**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

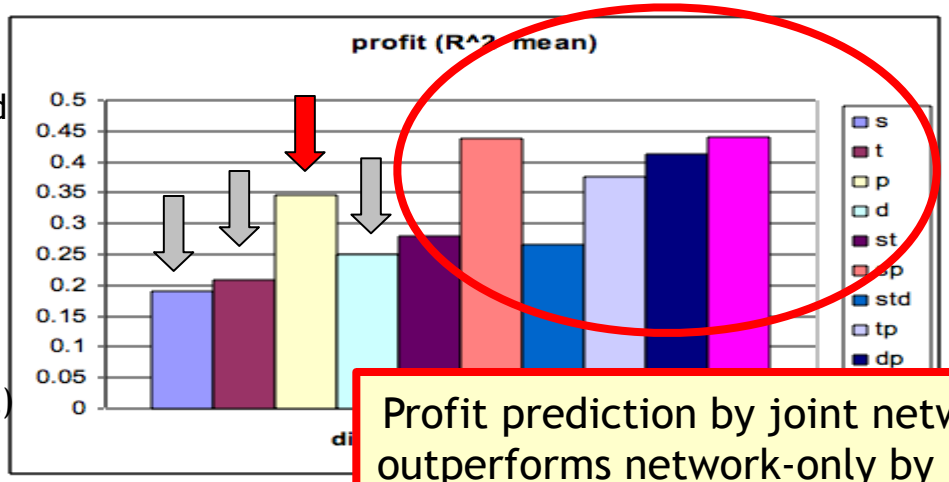



- Data Source:
  - Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

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

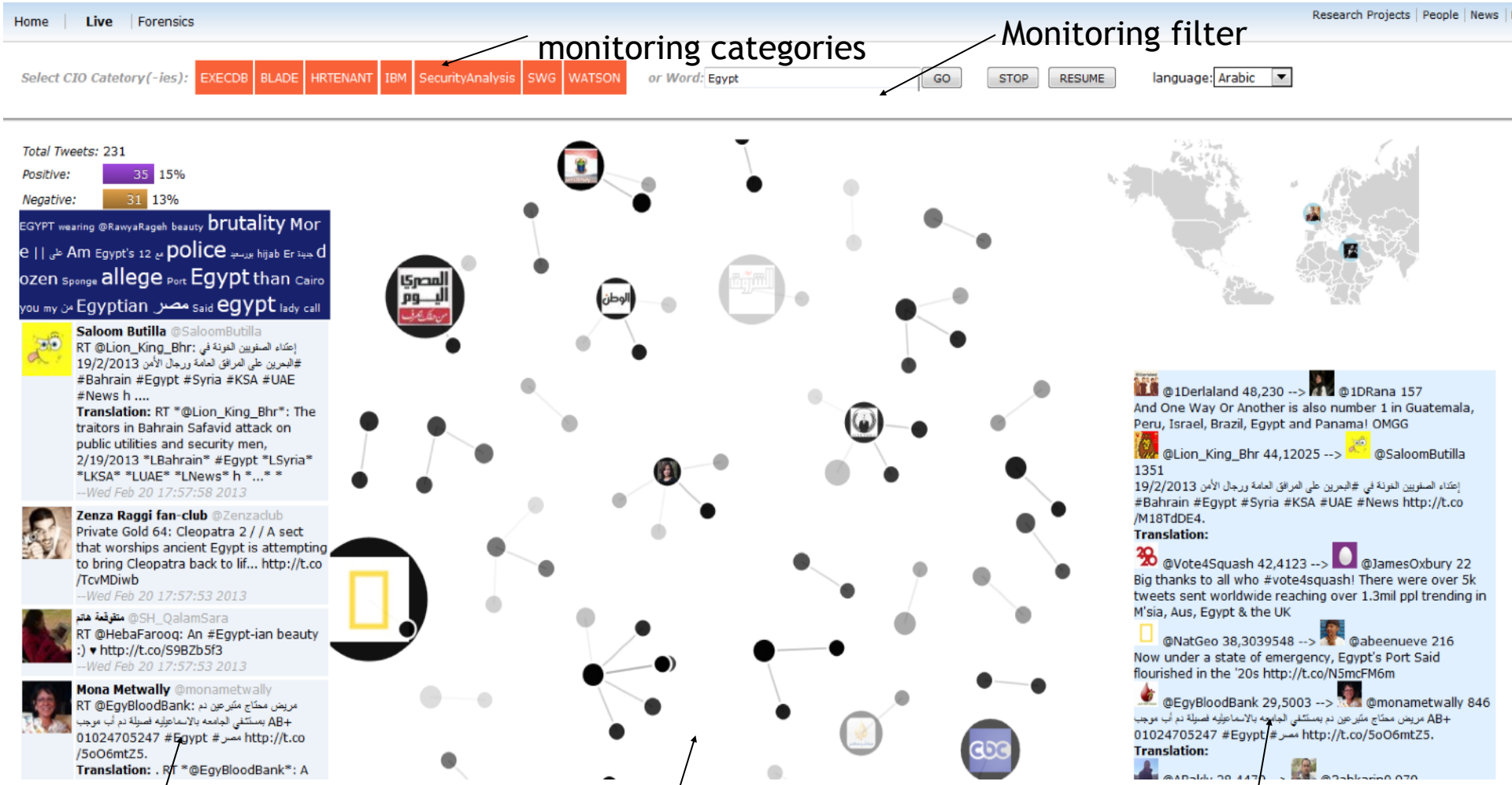
Model: Support Vector Regression (RBF kernel)




**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



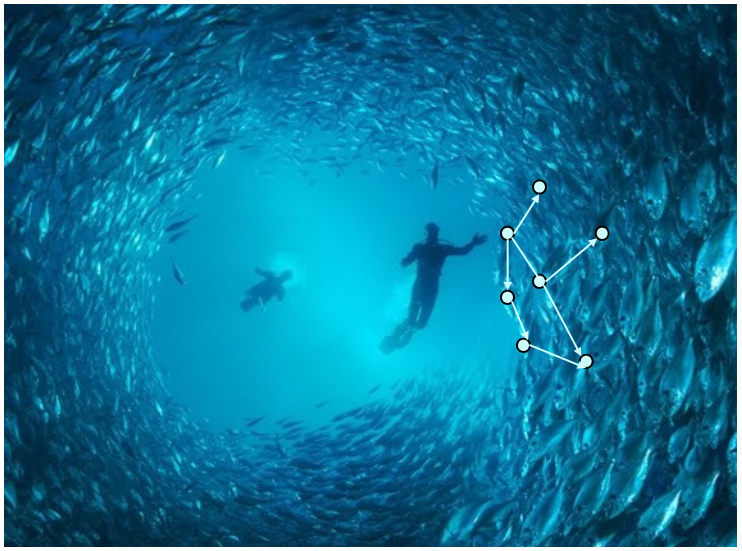
# IBM System G Social Media Solution Research Tasks

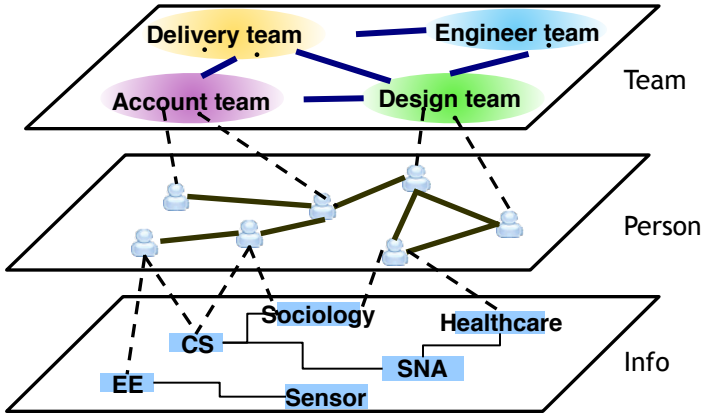
**Thrust 1. Modeling Information Dissemination:**
Task 1.1. Computational Modeling of User Dynamic Behavior
Task 1.2. Computational Models of Trust and Social Capital
Task 1.3. Information Morphing Modeling
Task 1.4. Persuasiveness of Memes
Task 1.5. The Observability of Social Systems
Task 1.6. Culture-Dependent Social Media Modeling
Task 1.7. Dynamics of Influence in Social Networks
Task 1.8. Understanding the Optimal Immunization Policy
Task 1.9. Modeling and Identification of Campaign Target Audience
Task 1.10. Modeling and Predicting Competing Memes

**Thrust 2. Detecting and Tracking Information Dissemination:**
Task 2.1. Real-Time and Large-Scale Social Media Mining
Task 2.2. Role and Function Discovery
Task 2.3. Detecting Malicious Users and Malware Propagation
Task 2.4. Emergent Topic Detection and Tracking
Task 2.5. Detecting Evolution History and Authenticity of Multimedia Memes
Task 2.6. Synchronistic Social Media Information and Social Proof Opinion Mining
Task 2.7. Community Detection and Tracking
Task 2.8. Interplay Across Multiple-Networks
Task 2.9: Assessing Affective Impact of Multi-Modal Social Media

**Thrust 3. Affecting Information Dissemination:**
Task 3.1. Crowd-sourcing Evidence Gathering to Formulate Counter-messaging Objectives
Task 3.2. Delivery and Evaluation of a Counter-messaging Campaign
Task 3.3. Optimal Target People Selection
Task 3.4. Automated Generation of Counter Messaging
Task 3.5. User Interfaces for Semi-Automatic Counter Messaging
Task 3.6. Controlling the Dynamics of Influence in Social Networks
Task 3.7. Influencing the Outcome of Competing Memes and Counter Messaging

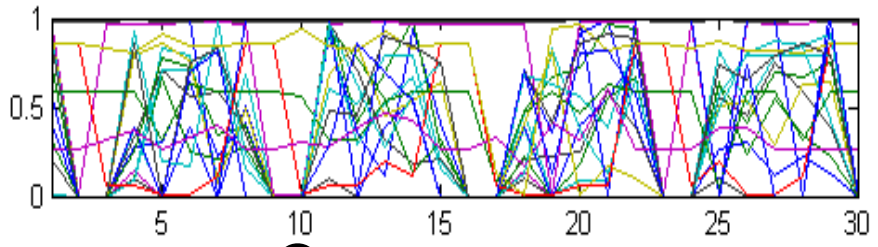
## *Heterogeneous Synchronicity Networks Predict Performance*

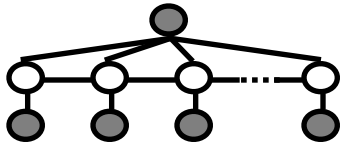





Outperform existing approaches by up to 18% (SDM 13)

## *One-class HCRF to detect temporal anomalies*



**Detected as top 1 anomaly in Sandy Tweets**

Outperform existing approaches by up to 180% (IJCAI 13)

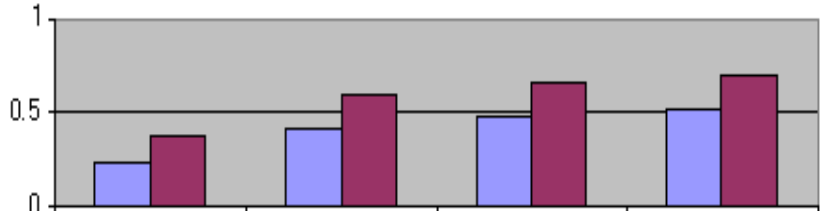
- **Motivation:**
  - –Info morph: new links keep emerging to give new meaning to existing phrases
- **Approach:**
  - –Compare characteristics of meta-paths between nodes in heterogeneous networks

***Peace West King*** from ***Chongqing fell from power***, still need to ***sing red songs***?

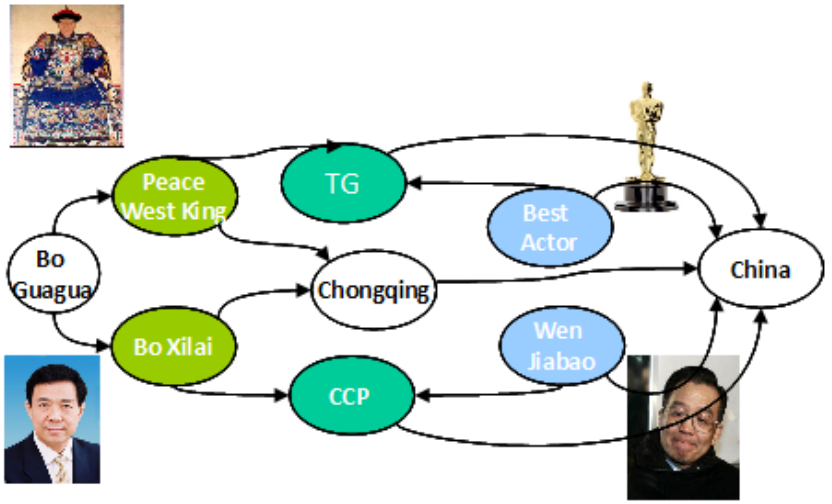
中国日报 CHINA DAILY
THE NATIONAL ENGLISH LANGUAGE NEWSPAPER

- ***Bo Xilai*** led ***Chongqing*** city leaders and 40 district and county party and government leaders to ***sing red songs***

**Entity morph resolution accuracy (ACL 2013)**




| | Acc@1 | Acc@5 | Acc@10 | Acc@20 |
|---|---|---|---|---|
| NLP | 0.234 | 0.416 | 0.477 | 0.519 |
| NLP+Graph | 0.379 | 0.594 | 0.659 | 0.701 |




$$\sum_{i=1}^{N} p_m(x_i) \log \frac{p_m(x_i)}{p_e(x_i)} + p_e(x_i) \log \frac{p_e(x_i)}{p_m(x_i)}$$

# Measuring Human Essential Traits in Social Media

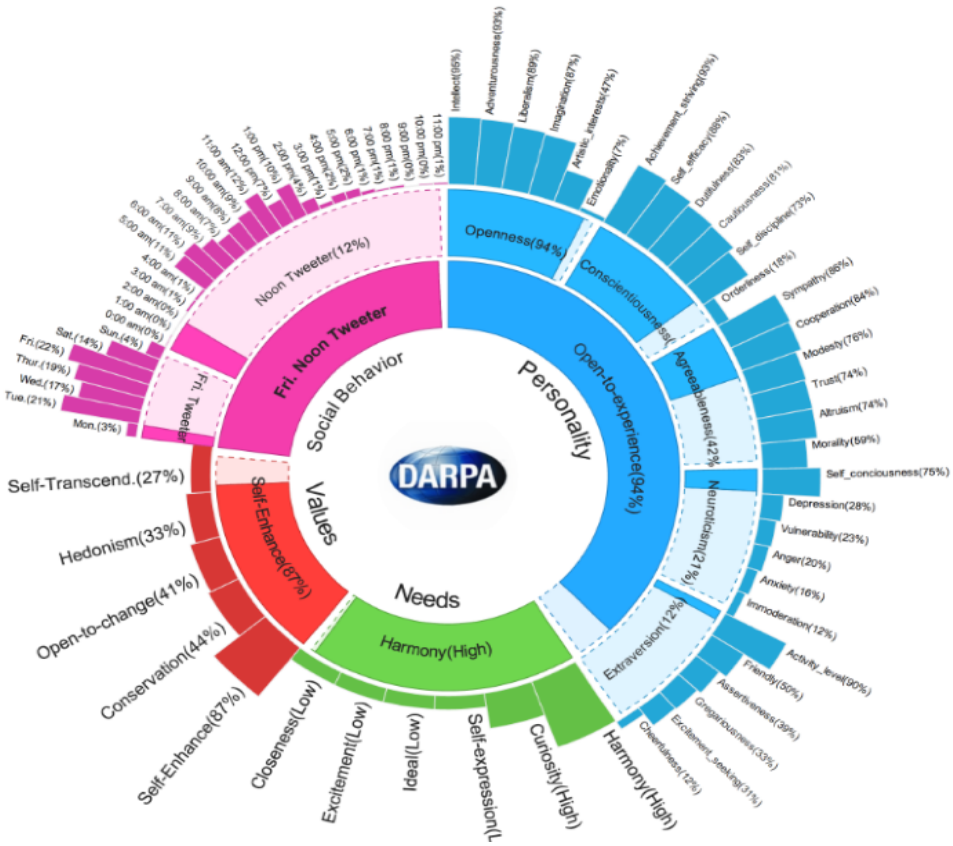
– **Personality:** Mapping personal/ organizational social media postings to scores of BIG 5 Personality (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neurocism*)

– **Needs:** Mapping personal/organizational social media postings to scores of *Harmony, Curiousity, Self-expression, Ideal, Excitement, and Closeness*.

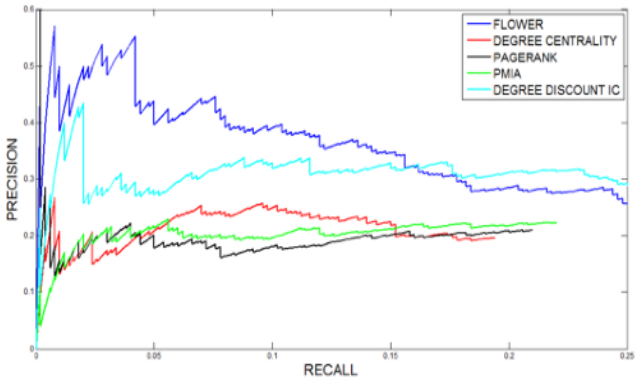
– **Values:** Mapping personal/organizational social media postings to scores of *Self-Enhance. Conservation, Open-to-Change, Hedonism, and Self-Transcend*.

– **Trustingness and Trustworthness:** Deriving from *interaction and propagation* history between the user and his followers and the people he follows.

– **Influence:** Total *attention* received by user as leader across all discovered flows.



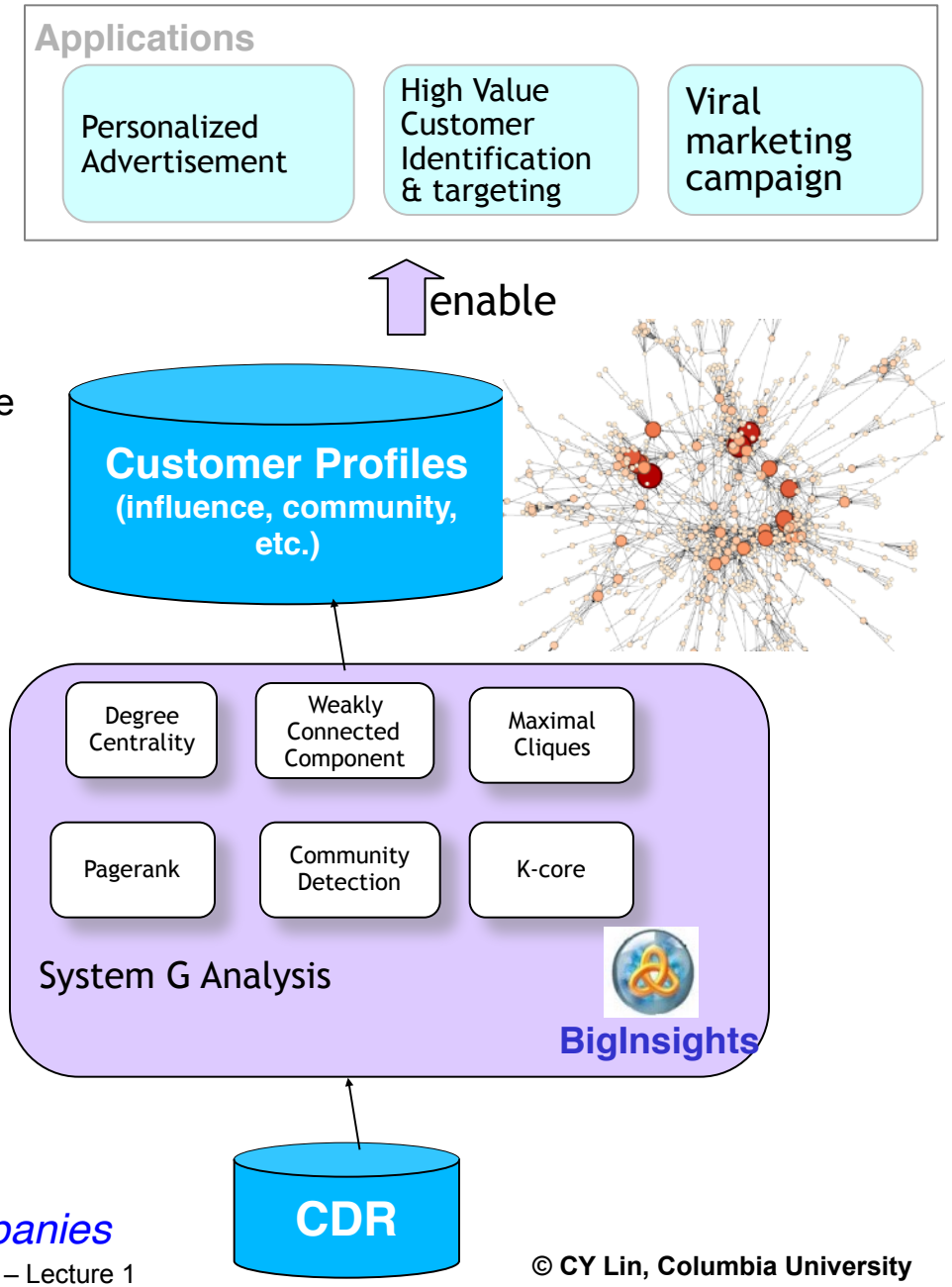

Precision-Recall performance of predicting info propagation by different features *(Our proposed influence index: FLOWER)*



# Use Case 6: Customer Social Analysis for Telco

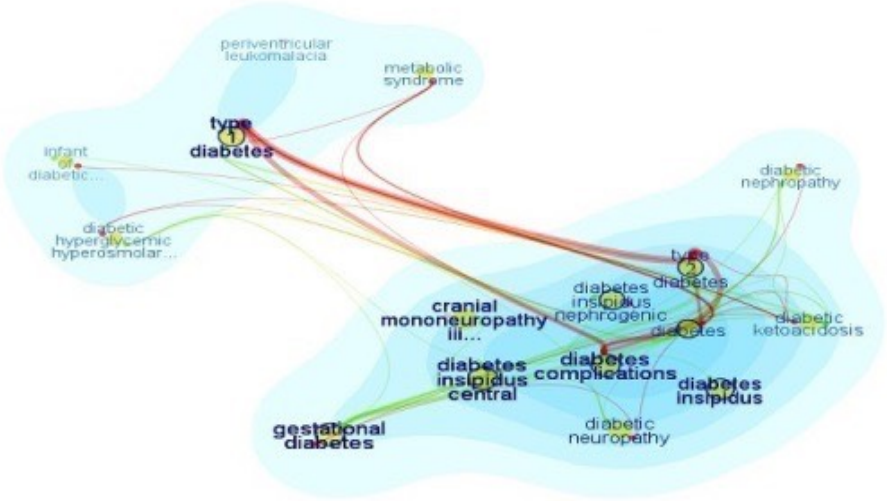
**Goal:** Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

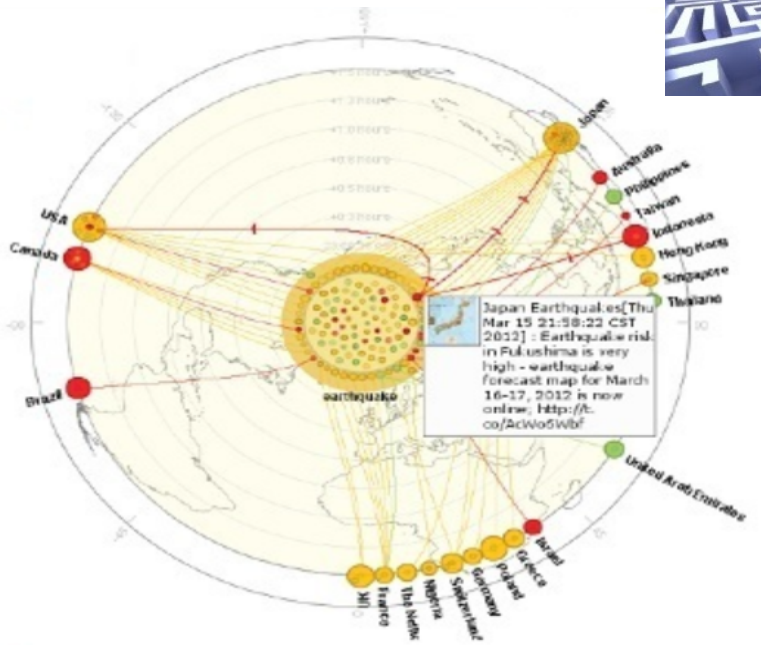
- Applications based on the extracted social profiles
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign
- Approach
  - Construct social graphs from CDRs based on {caller, callee, call time, 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




*PoCs with Chinese and Indian Telecomm companies*

# Category 2: Data Exploration




*Enhancing:*


Vivísimo

CÚRAM SOFTWARE

Huge Network Visualization | Network Propagation | I2 3D Network Visualization | Geo Network Visualization | Graphical Model

Communities | Graph Search | Network Info Flow | Bayesian Networks

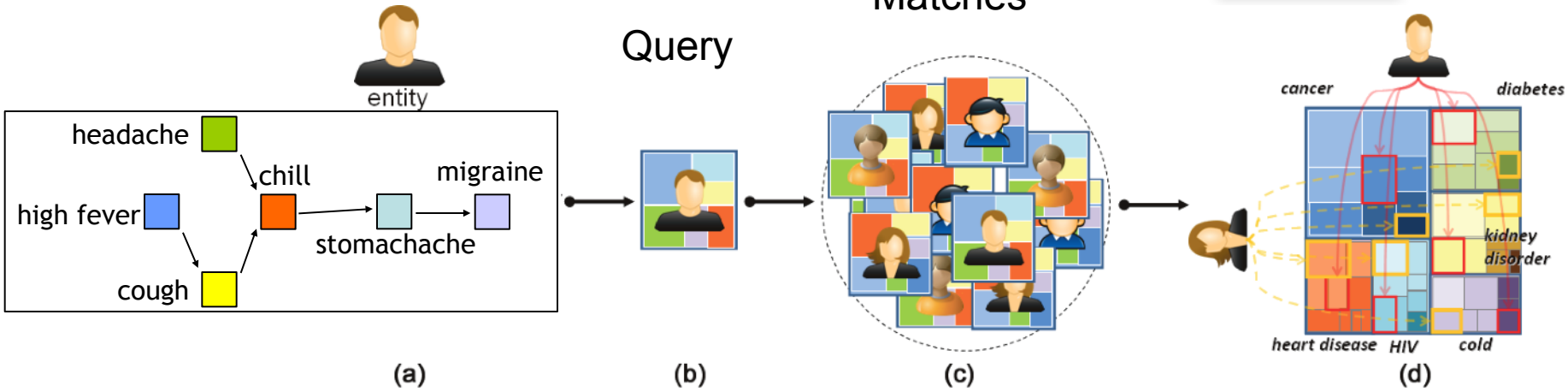
Centralities | Graph Query | Shortest Paths | Latent Net Inference

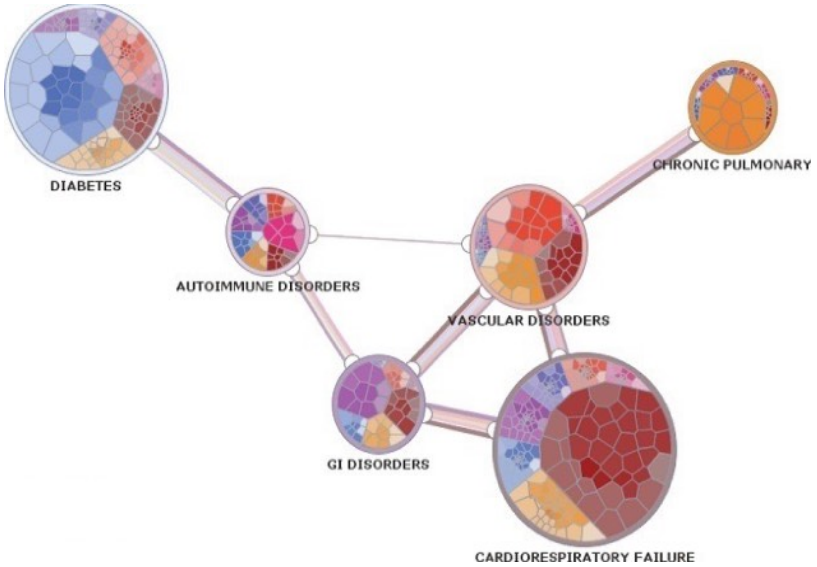
Ego Net Features | Graph Matching | Graph Sampling | Markov Networks

Middleware and Database

# Use Case 7: Graph Analytics and Visualization for Watson

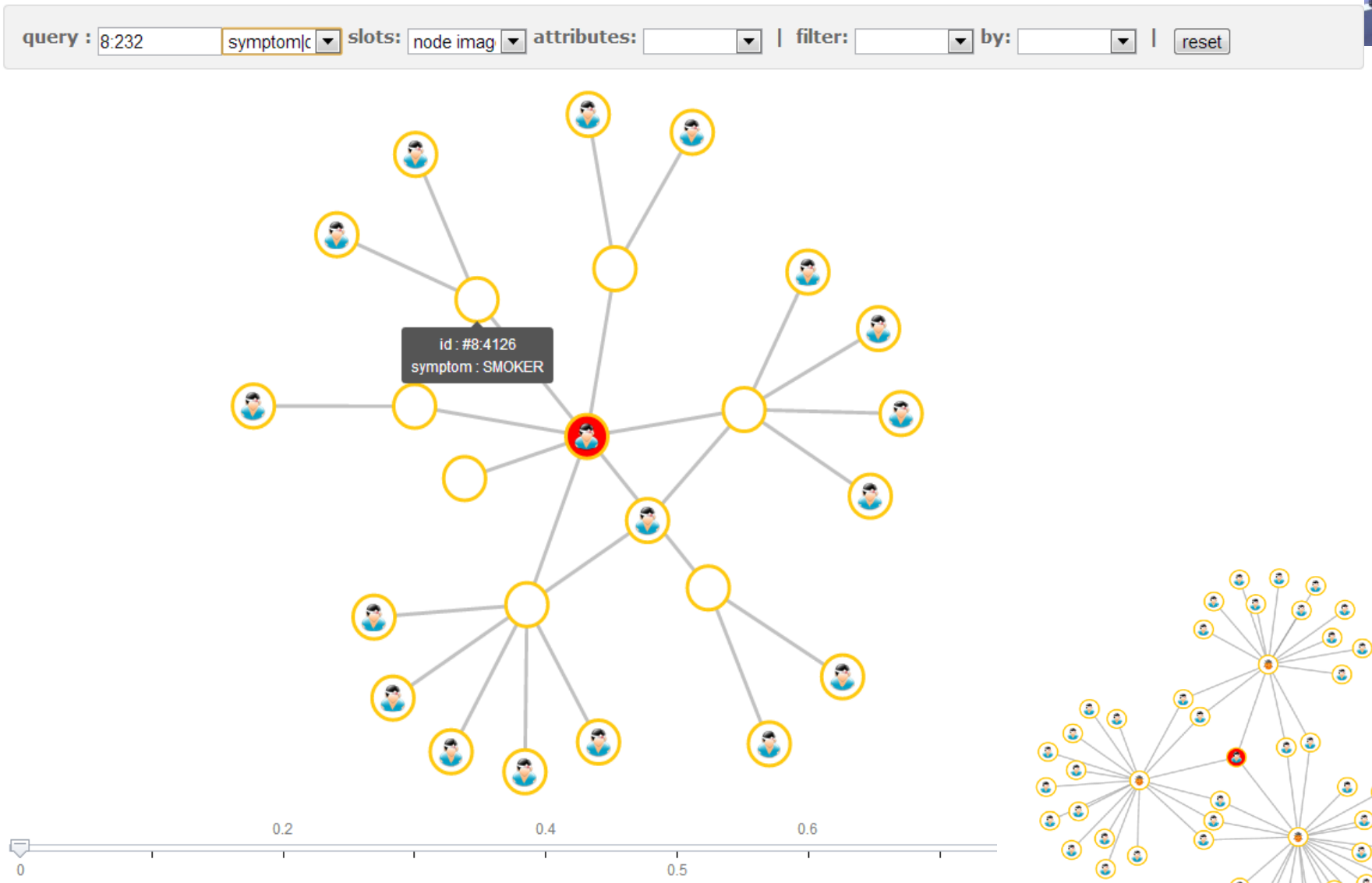
Graph Matching




Graph Communities

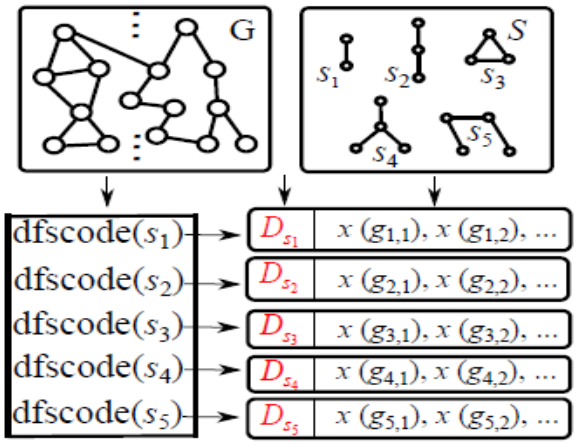
# Graph Analytics for Watson



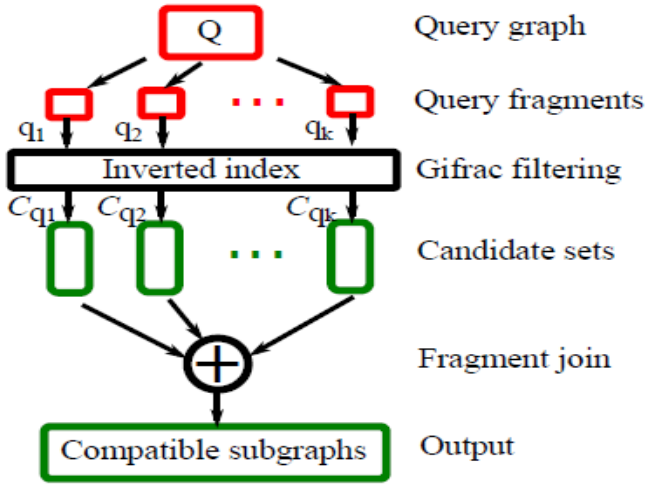
# Fast Graph Matching Algorithm

- Data: (CAIDA) 26.5K nodes and 106.8K edges
- Index construction: 13-20 times faster than the prior state-of-the-art
- Query time: close to UpdAll (upper bound) and ~8x faster than UpdNo and NaiveGrid

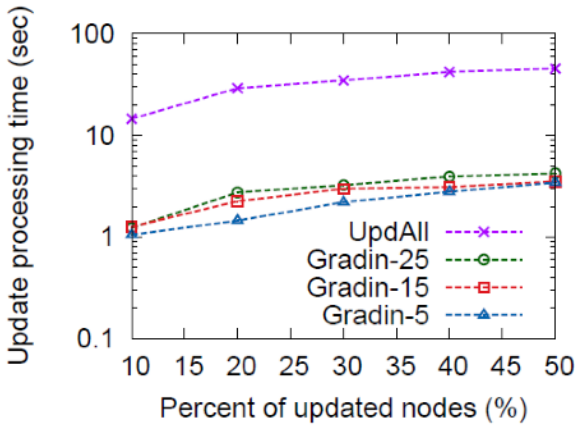
Graph Matching



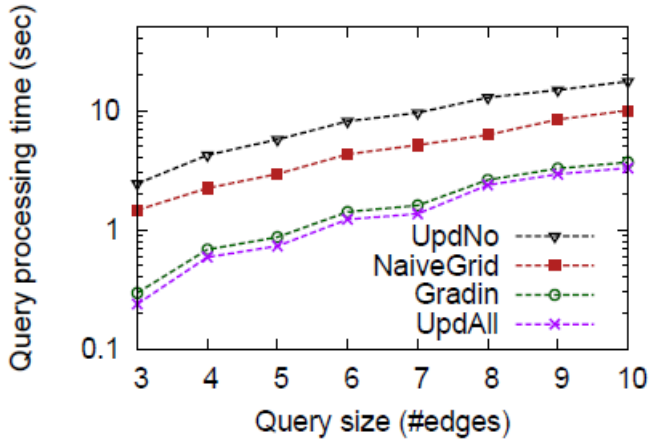

(a) Offline index building




(b) Online query processing

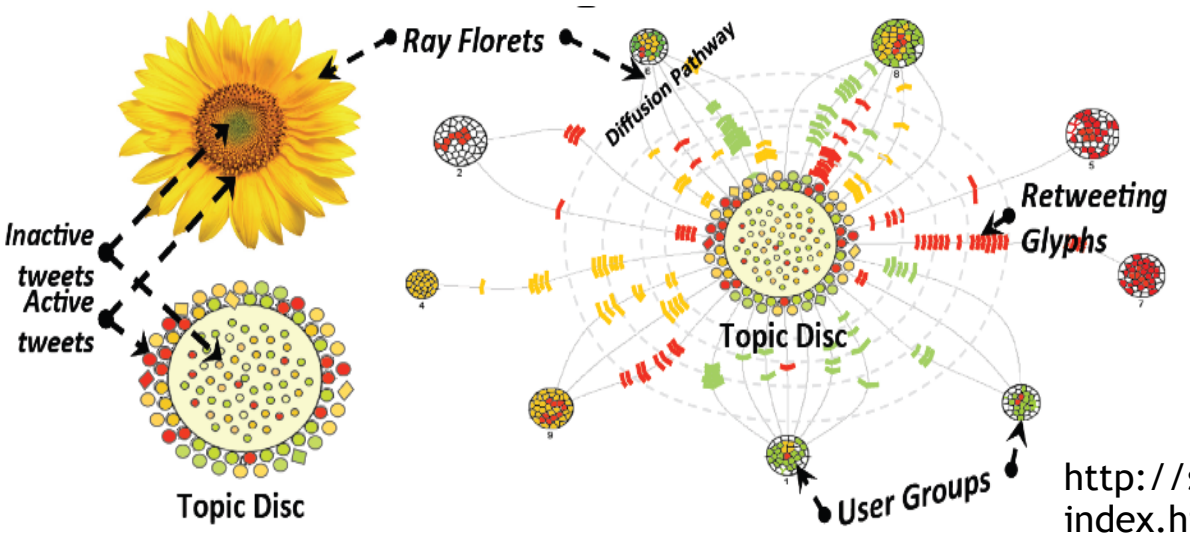



Indexing time




Query processing time

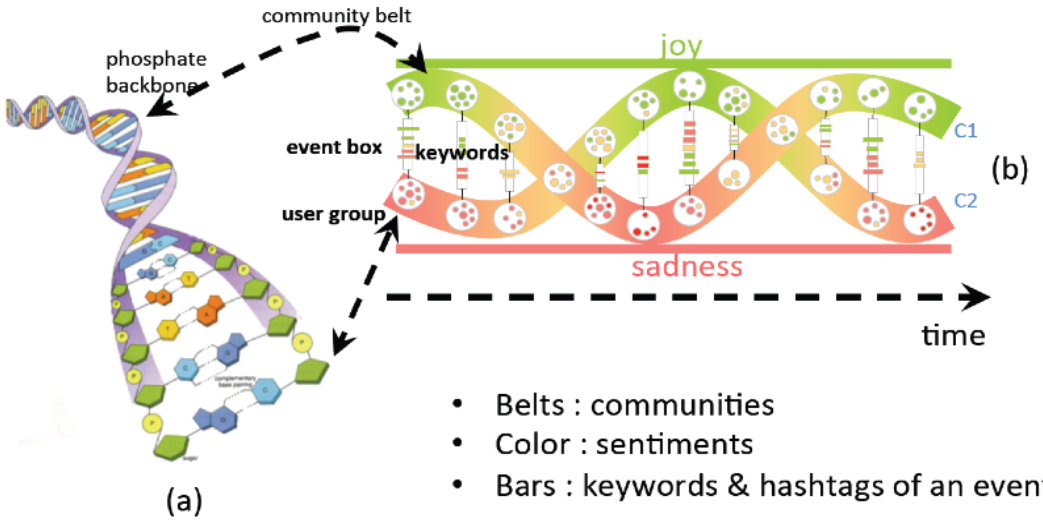
# User Case 8: Visualization for Navigation and Exploration




Whisper : Tracing the information diffusion in Social Media

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

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media

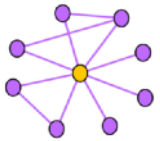


# Use Case 9: Graph Search

Graph Search

existing search engine

query



Info-Socio networks

index

ranking

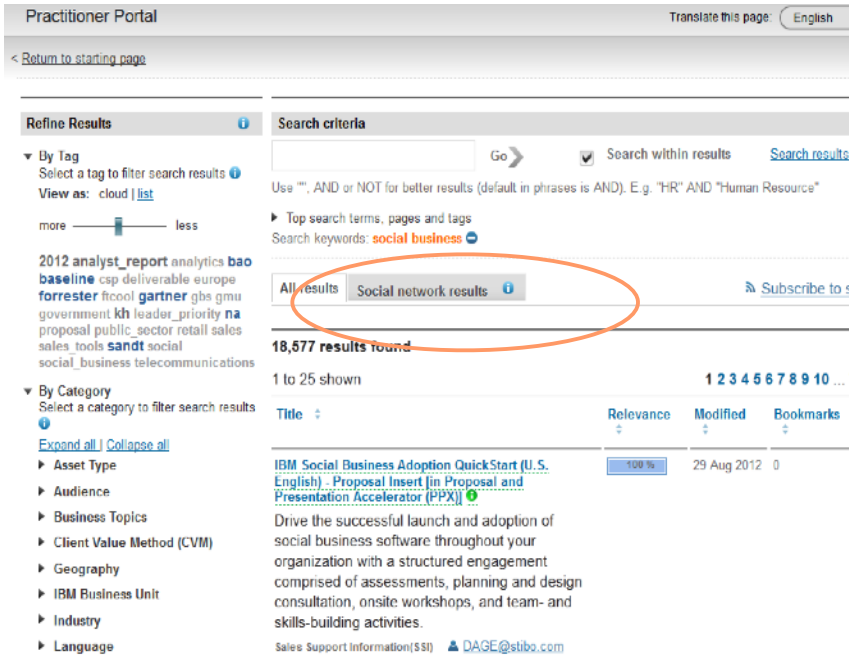
re-ranking

Improved search results

Interest / social network based content recommendations

Graph analysis

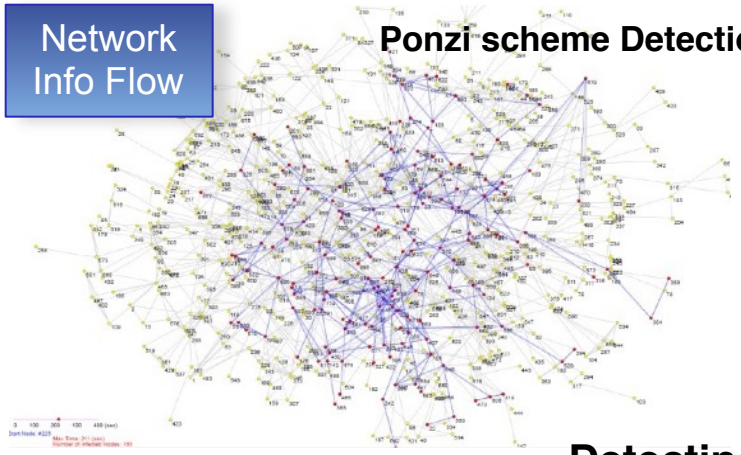
query context



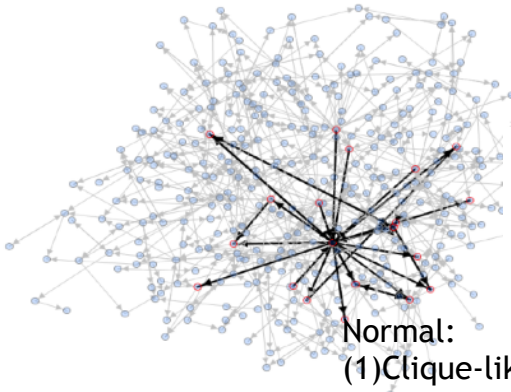
# Category 3: Security

Network Info Flow

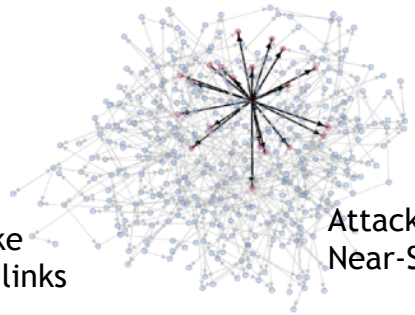
**Ponzi scheme Detection**



Ego Net Features

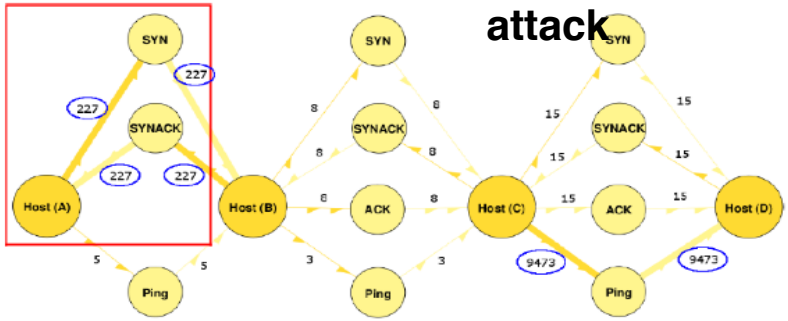


Normal:
(1)Clique-like
(2)Two-way links

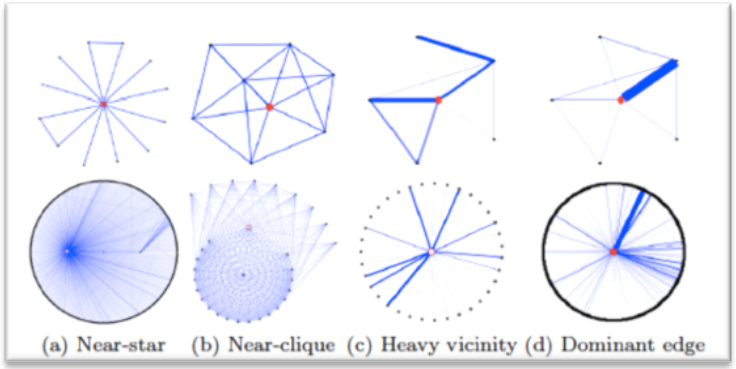


Attacker:
Near-Star

**Detecting DoS attack**




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



(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

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

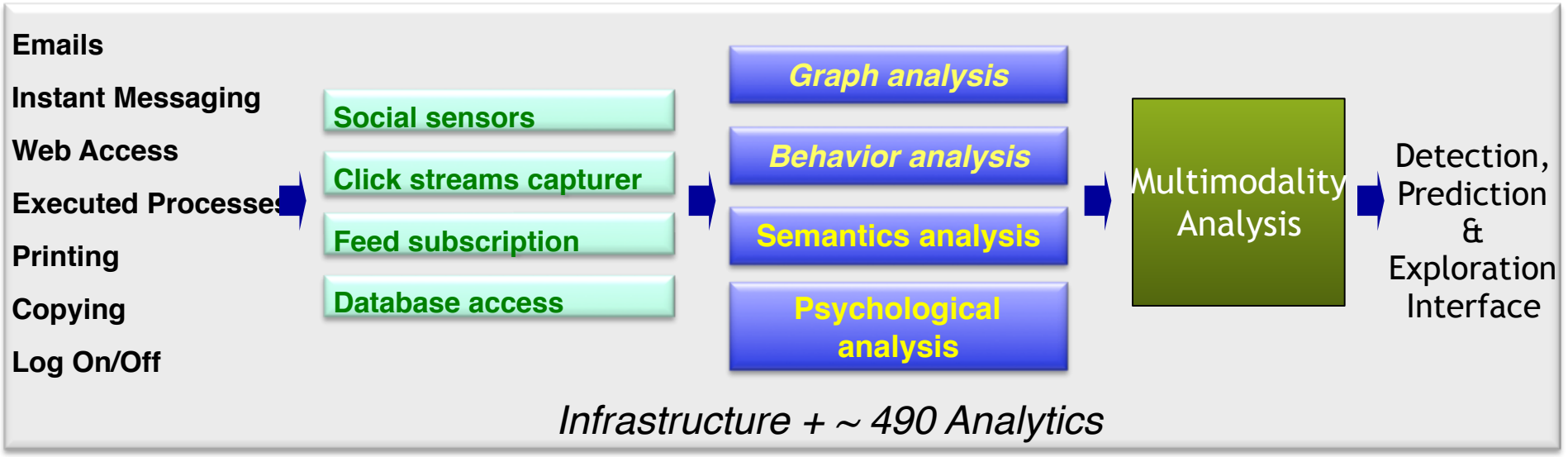
# Use Case 10: Anomaly Detection at Multiple Scales

**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

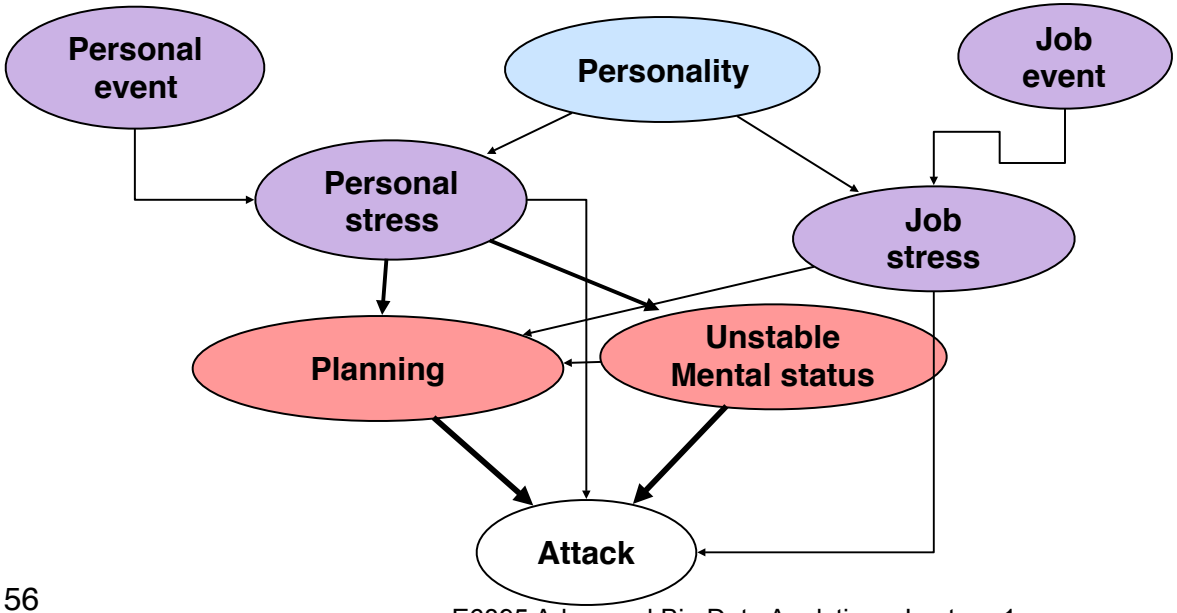
"What's emerged is a multibillion dollar detective industry"
*npr Jan 10, 2013*



# Story – Espionage Example

- Personal stress:
  - Gender identity confusion
  - Family change (termination of a stable relationship)
- Job stress:
  - – Dissatisfaction with work
    - Job roles and location (sent to Iraq)
    - long work hours (14/7)

- Unstable Mental Status:
  - Fight with colleagues, write complaining emails to colleagues
  - Emotional collapse in workspace (crying, violence against objects)
  - Large number of unhappy Facebook posts (work-related and emotional)
- Planning:
  - Online chat with a hacker confiding his first attempt of leaking the information




(1) Attack:

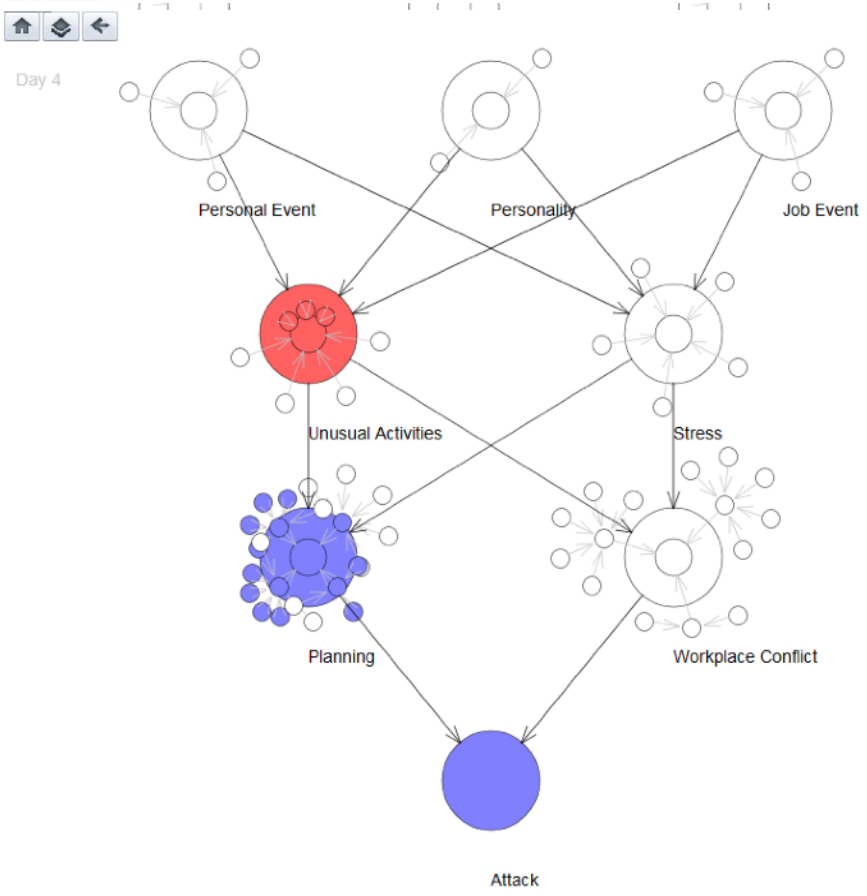
– Brought music CD to work and downloaded/ copied documents onto it with his own account

# Example of Graphical Analytics and Provenance

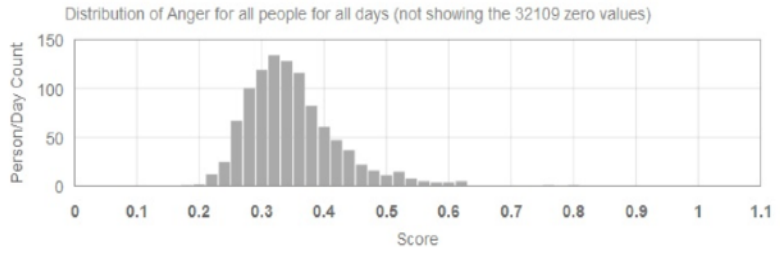
Markov Network | Latent Network | Bayesian Network

Sam Chang
Security Specialist
IT Department
samchang@company.com

Timeline: start stop




**Email (Anger)**




## Top Person/Days -- Anger, group Self

| Userid | Date | Anger |
|---|---|---|
| 0A40E96890D1ED5E62F9C7F19191108D | 2012-07-18 | .81 |
| 7B131181E78AF8C08F3A9E605E58227B | 2012-07-17 | .77 |
| 41C50A4433E2CC2C34E6DCDFD1290C3F | 2012-07-10 | .63 |
| 369730CC2965FD97E8435A4365AF6FF6 | 2012-07-17 | .63 |
| 2FF547A39E2D3E4D0B5DC6ACB1165441 | 2012-07-27 | .62 |
| 6826DFDB7FFE74364C4FF7C58089E795 | 2012-07-10 | .62 |
| CBD7C1659A41C74998837310639E2817 | 2012-07-24 | .62 |
| C00A0A9DB436F130DCC5671D2936CD45 | 2012-07-13 | .62 |
| 9C2568C8AF6809EAAAAC53E7B267AB00 | 2012-07-10 | .62 |
| 215291F411748EDD54E2AA43966615A8 | 2012-07-24 | .61 |
| BCE7365D13A8396290BFB4CCB1D9DE36 | 2012-07-25 | .60 |

# Evaluations on the Real-World Data in Vegas Lab (Oct 2013)

- Each month, 3 cases were inserted (1 abnormal person per case) in the real data.
- Each performer system retrieved top abnormal people out of the 5,500 people per month.
- This chart showed where the 3 IBM systems (Sabotage, Espionage, and Fraud) ranked the abnormal person in each case. "All" is a combined rank list of the 3 systems. (Oct 2013 review on 12/12 ~ 03/13 data)

| | | Sabotage | Espionage | Fraud | All |
|---|---|---|---|---|---|
| Dec | Sabotage (Scenario 12) | 4 | 241 | 1667 | 9 |
| | Espionage (Scenario 8) | 981 | 1 | 120 | 1 |
| | Fraud (Scenario 13) | 1526 | 454 | 1 | 2 |
| Jan | Fraud (Scenario 13) | 4230 | 3367 | 1 | 2 |
| | Espionage (Scenario 14) | 11 | 44 | 574 | 30 |
| | Fraud (Scenario 5) | 4230 | 1462 | 3 | 8 |
| Feb | Espionage (Scenario 14) | 1936 | 73 | 232 | 203 |
| | Espionage (Scenario 4) | 4101 | 9 | 803 | 26 |
| | Sabotage (Scenario 15) | 65 | 4101 | 654 | 181 |
| Mar | Sabotage (Scenario 16) | 1 | 1690 | 294 | 1 |
| | Fraud (Scenario 5)) | 1544 | 9 | 5 | 10 |
| | Espionage (Scenario 4) | 4325 | 11 | 46 | 27 |

12. **Layoff Logic Bomb:** An engineer is worried about rumors of impending layoffs feels that he needs some kind of an "insurance policy", in case he gets laid-off or fired. He creates a "logic bomb" which will delete all files from a number of company Linux systems in five days, unless he resets the timer before then.

13. **Outsourcer's Apprentice:** (http://www.bbc.co.uk/news/technology-21043693) A software developer outsources his job to China and spends his workdays surfing the web. Most surfing occurs on a second laptop. He pays just a small fraction of his salary to a Chinese company to do his job. The developer provides his VPN credentials to the company and enabling Terminal Services on his workstation. The Chinese consulting firm sends the developer PayPal invoices.

8. **Anomalous Encryption:** A Subject wishes to pass sensitive information to a foreign government in exchange for that government setting him up with his own business. Subject researches NSA monitoring capabilities, generates a long random passphrase and then tests encrypting and mails data to personal account. The subject encrypts documents and emails the key.

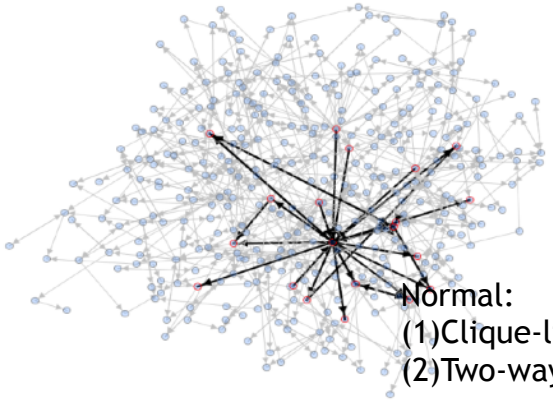
**1 in Top #21-#50, and 2 in Top #51-#100. Performer 2 did not report results. Performer 3 reported: 3 of the 12 cases Top #50-#100, 6 cases Top #101-#500, and 3 cases beyond Top #501.**

# Use Case 11: Fraud Detection for Bank

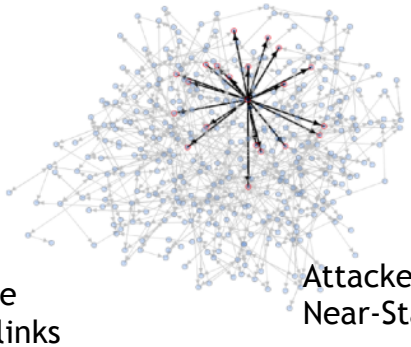
Network Info Flow

Ego Net Features

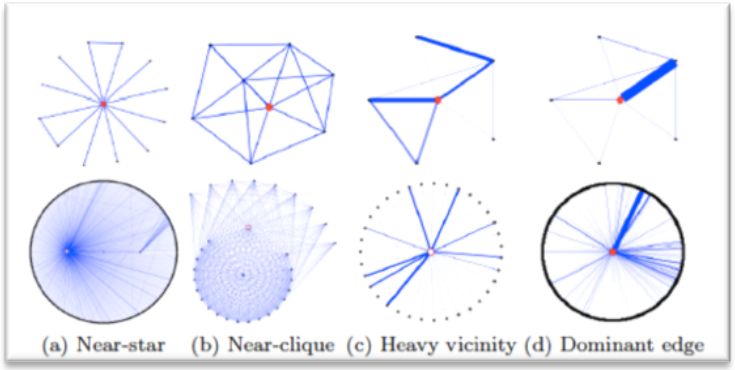
**Ponzi scheme Detection**



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



Attacker:
Near-Star

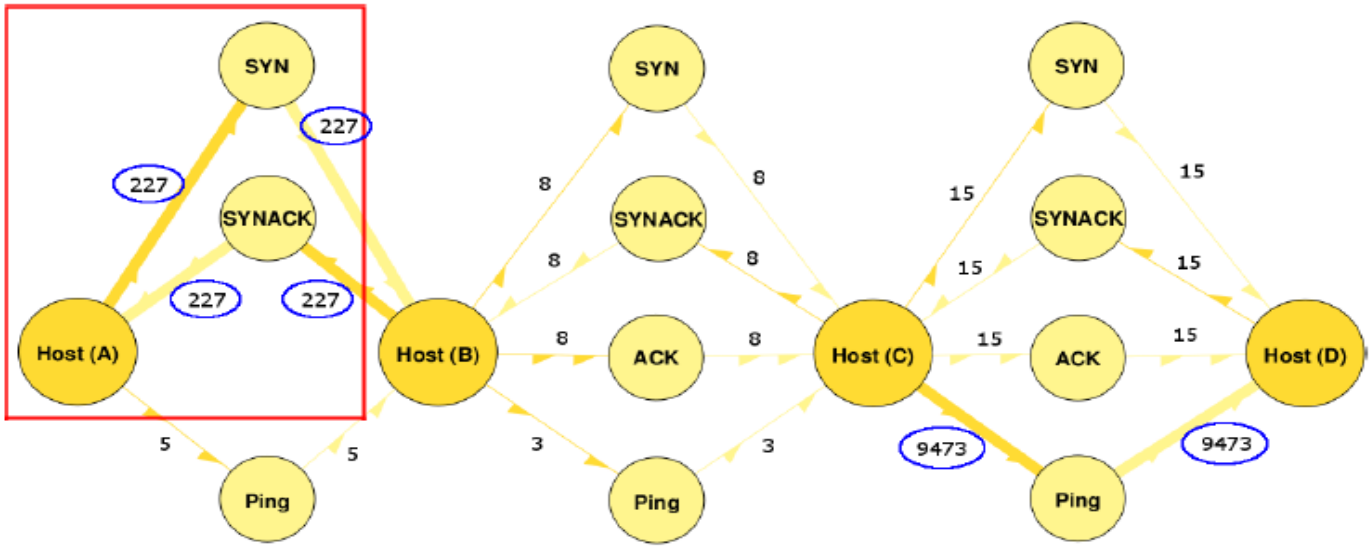


# Use Case 12: Detecting Cyber Attacks

Network Info Flow

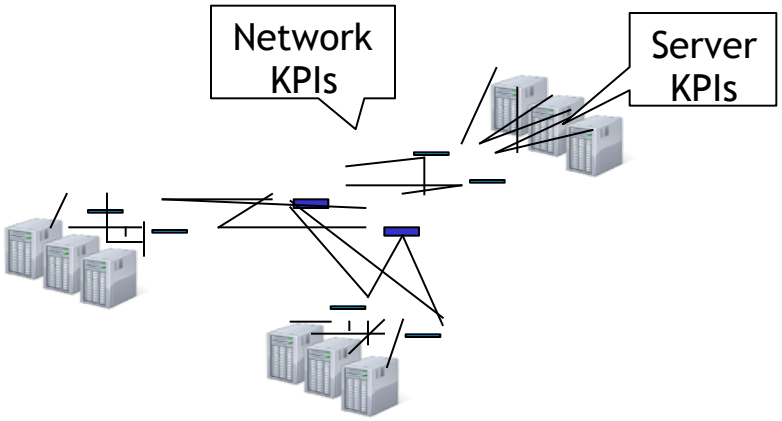
Ego Net Features

**Detecting DoS attack**

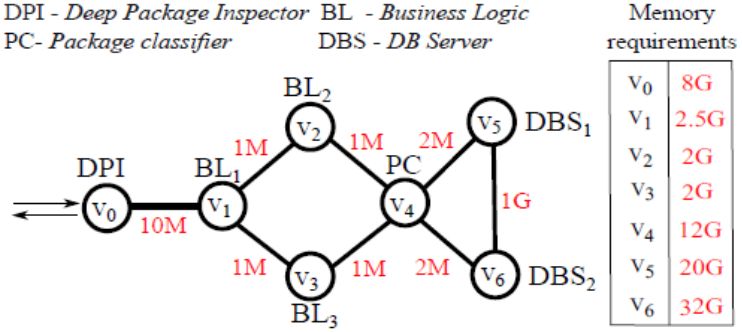



(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

# Category 4: Operations Analysis




**Cloud Service Placement**

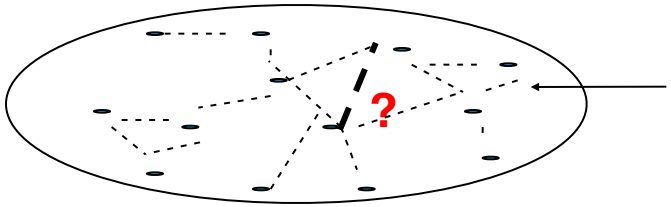



Graph Matching

Bayesian Network

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

**Causality analyzer**




Varying over time

- 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 |

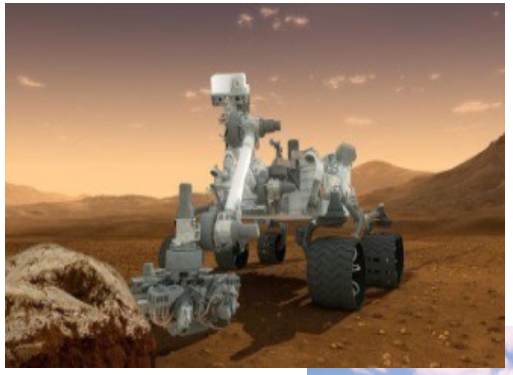
Middleware and Database

# Use Case 13: Smarter *another* Planet

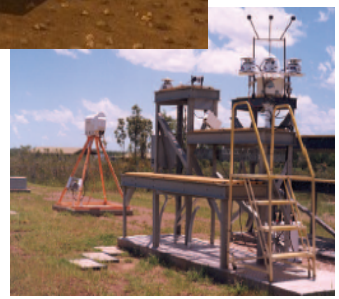
Columbia University

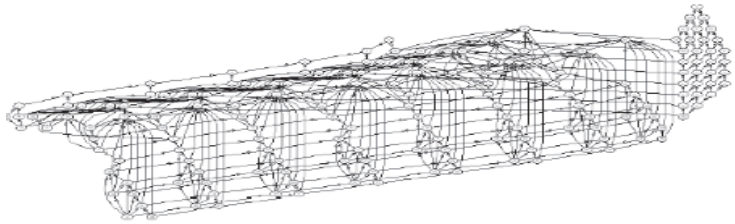
Bayesian Network

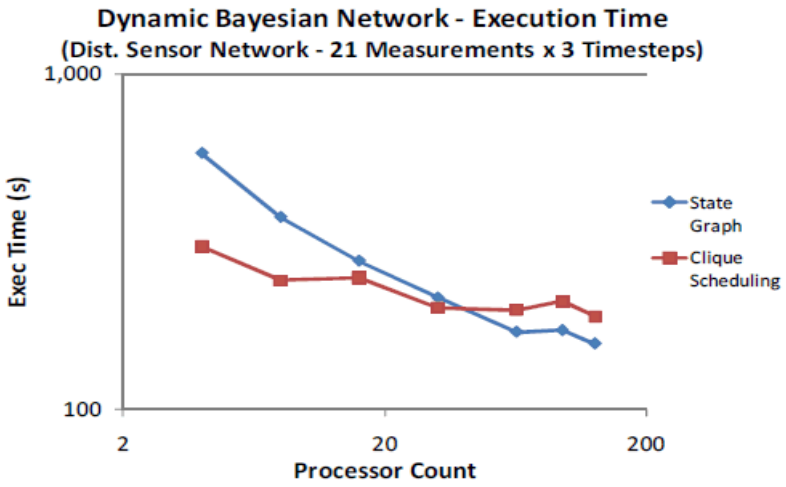
**Goal:** Atmospheric Radiation Measurement (ARM) climate research facility provides *24x7 continuous field observations* of cloud, aerosol and radiative processes. **Graphical models** can automate the validation with improvement efficiency and performance.



**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**

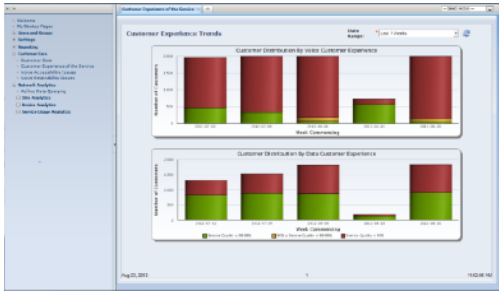
* 3 timesteps * 63 variables
* 3.9 avg states * 4.0 avg indegree
* 16,858 CPT entries

**Junction Tree**

* 67 cliques
* 873,064 PT entries in cliques

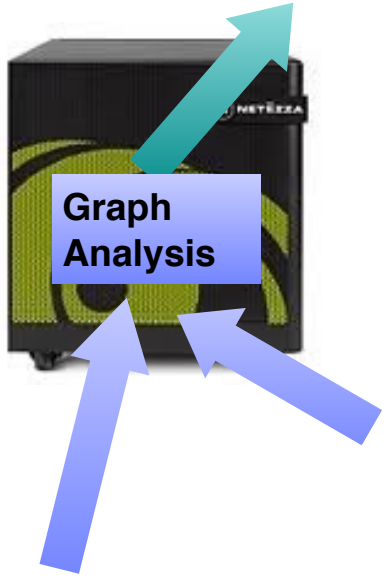
# Use Case 14: Cellular Network Analytics in Telco Operation

**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

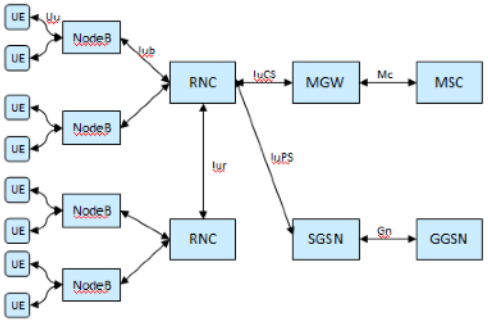


Network load level report

- 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



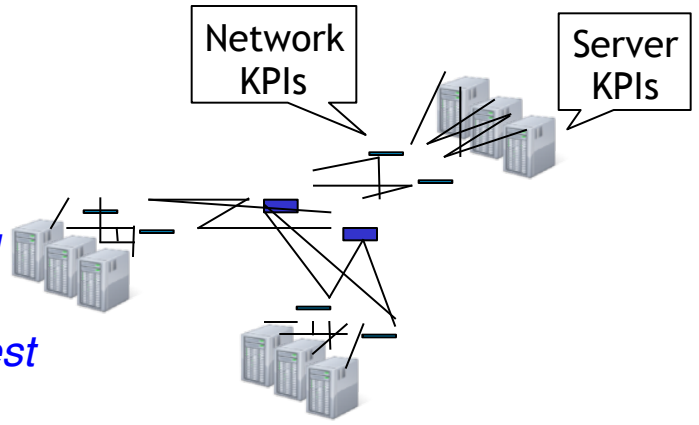

Network topology



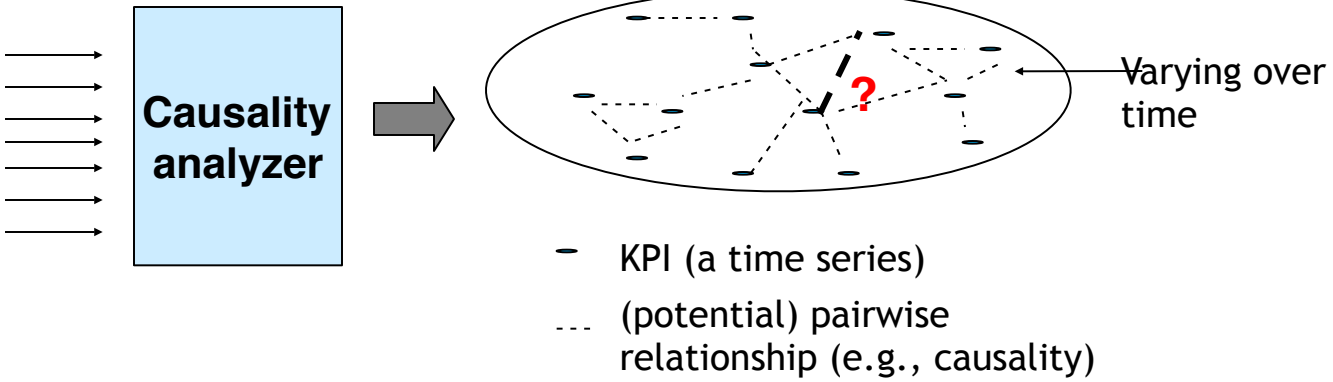
# Use Case 15: Monitoring Large Cloud

**Goal:** *M*onitoring 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

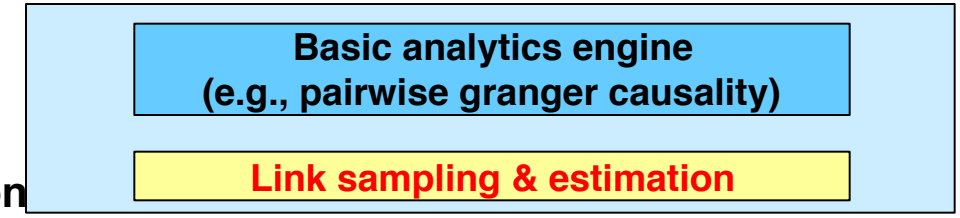
- *Causality relationships (e.g., Granger causality) are crucial performance monitoring & root cause analysis*
- *Challenge: easy to test pairwise relationship, but hard to test multi-variate relationship (e.g., a large number of KPIs)*




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

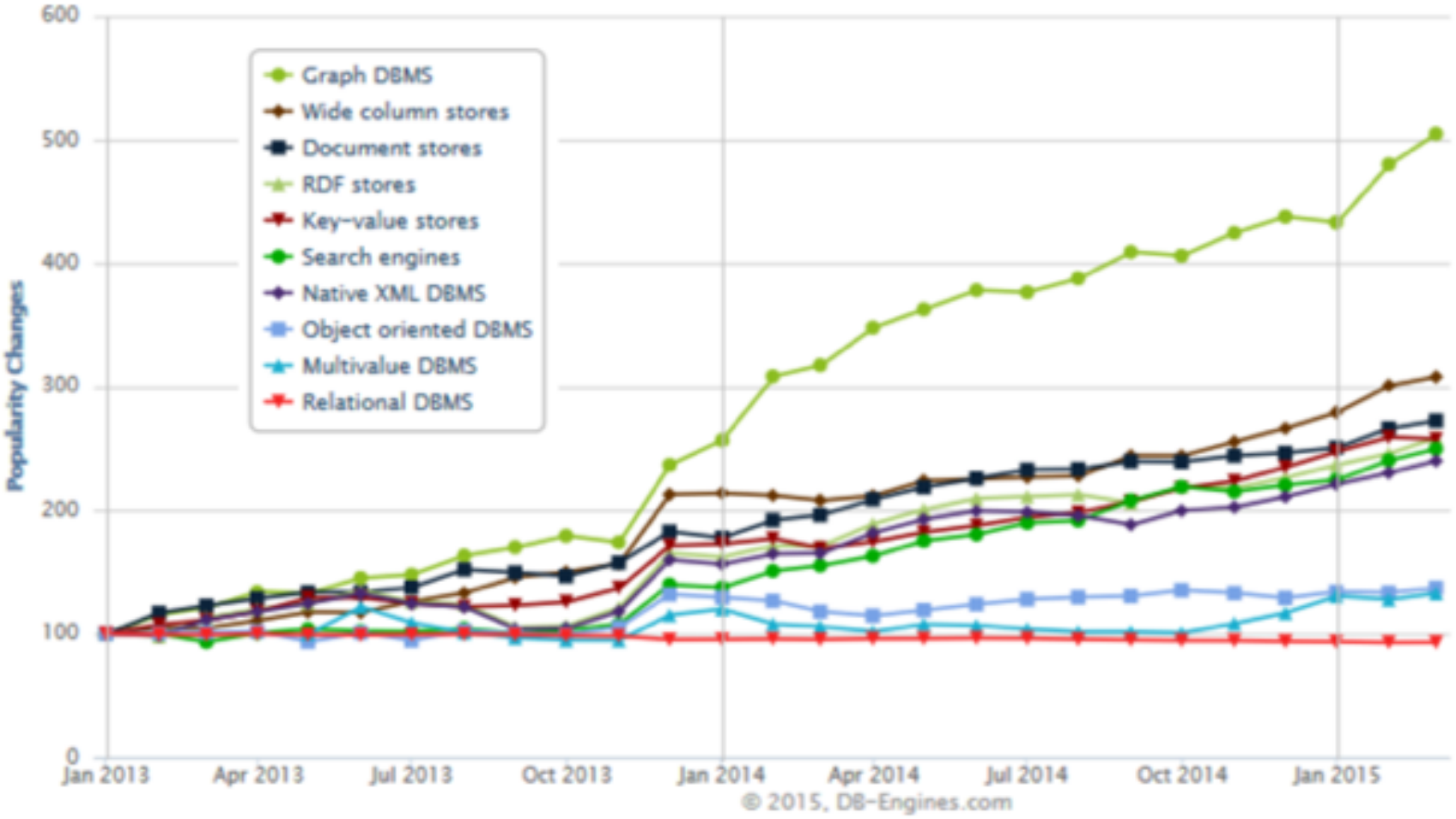


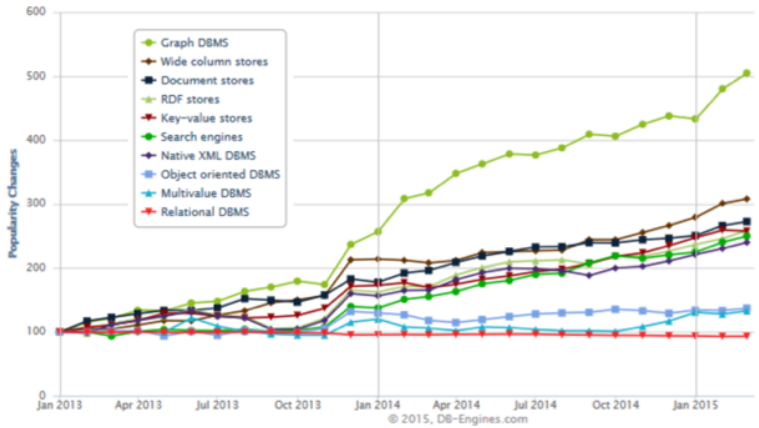

**Our approach: Probabilistic monitoring via sampling & estimation**




*Select KPI pairs (sampling)→ Test link existence → Estimate unsampled links based on history → Overall graph*

# Category 5: Data Warehouse Augmentation





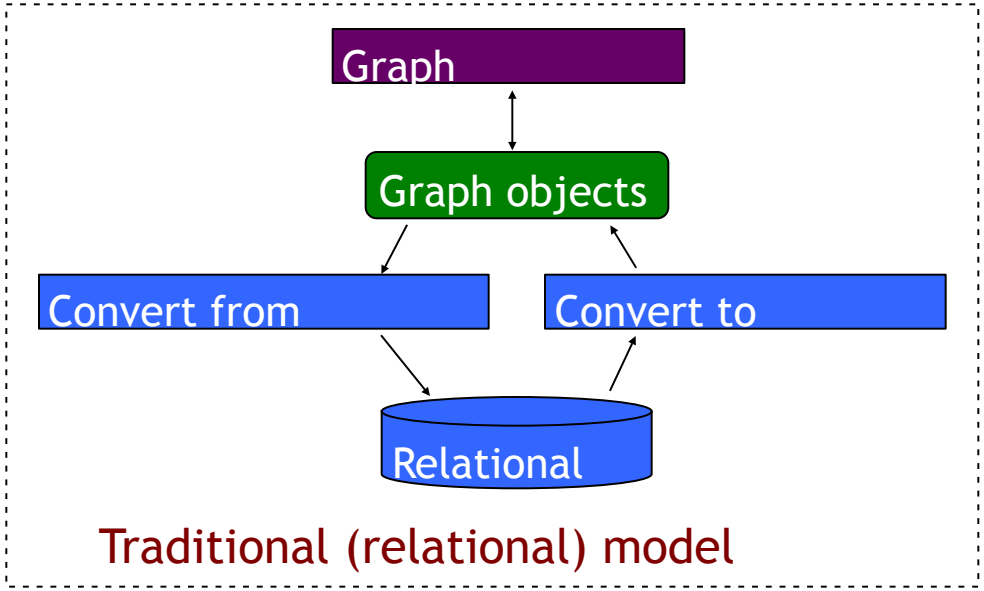


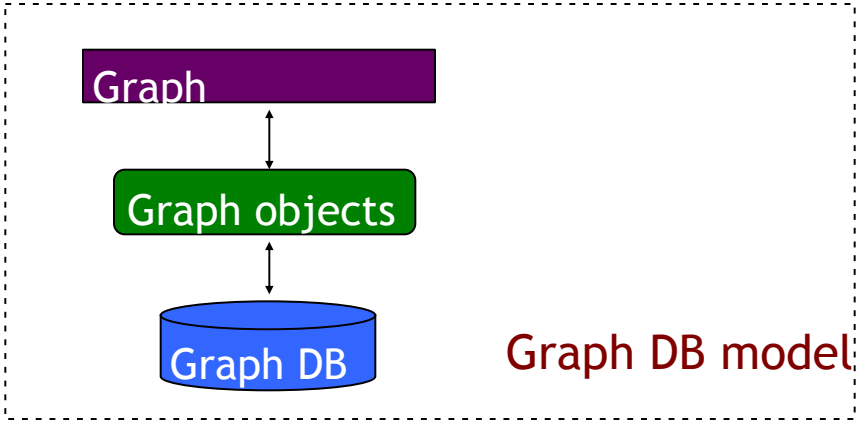
- **Graph Database is much more efficient than traditional relational database**


- **How does FINRA analyze ~50B events per day TODAY?** - *Build a graph of market order events from multiple sources [ref]*

- **How did journalists uncover the Swiss Leak scandal in 2014 and also Panama Papers in 2016?** -- *Using graph database to uncover information thousands of accounts in more than 20 countries with links through millions of files [ref]*

# 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

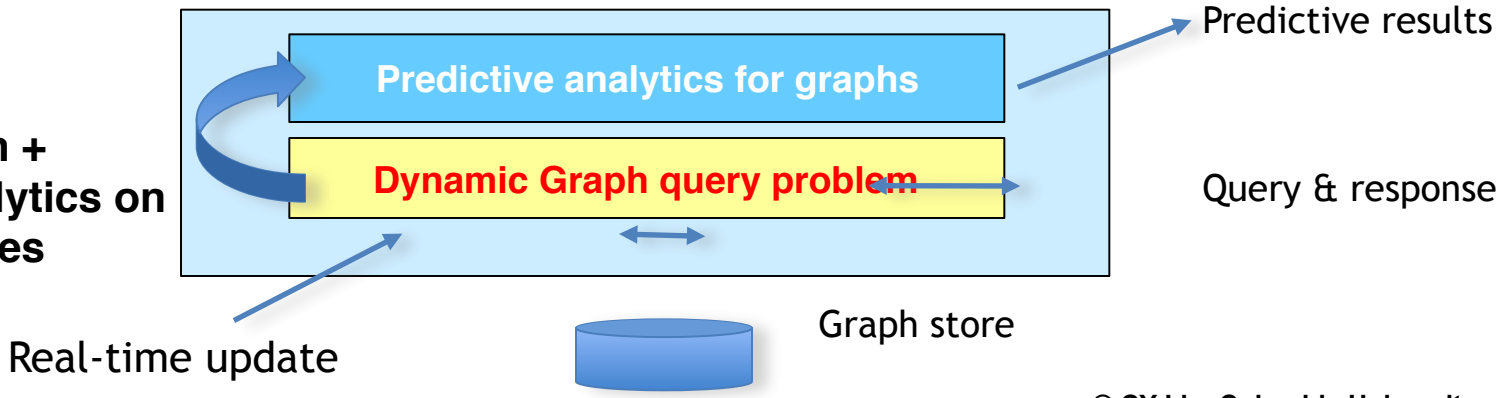
# Use Case 17: Smart Navigation Utilizing Real-time Road Information

**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*

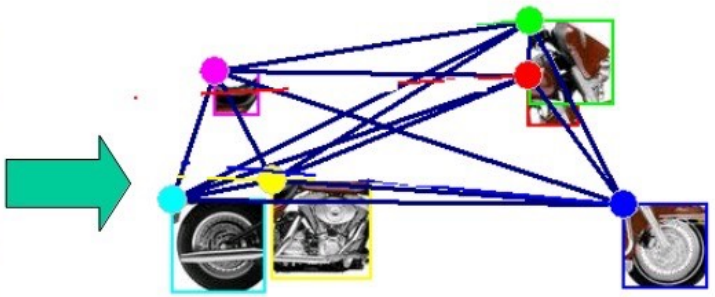
- *Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shorted path computation) on time-varying graphs (order of magnitudes improvement over existing solutions)*

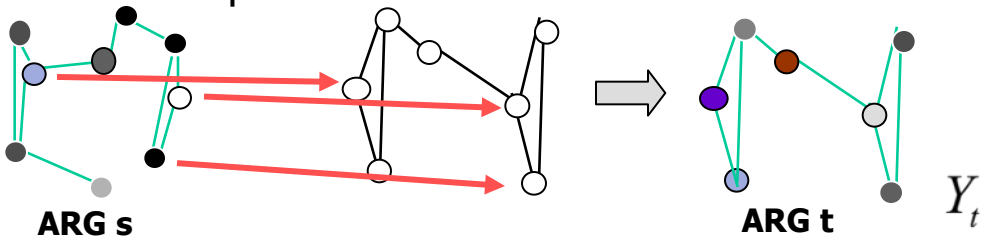
- *High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall*

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



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






# Use Case 19: Graph Matching for Genomic Medicine

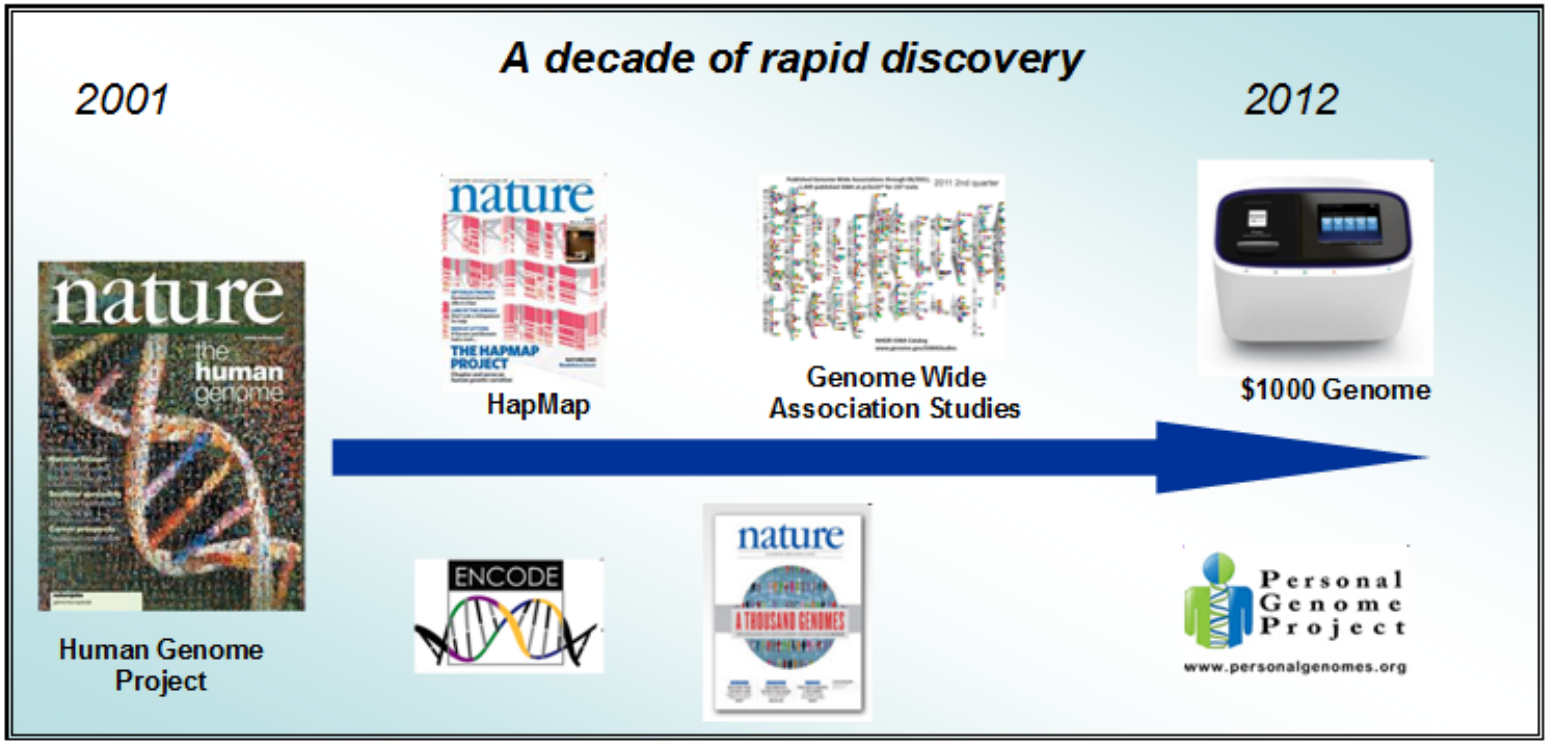
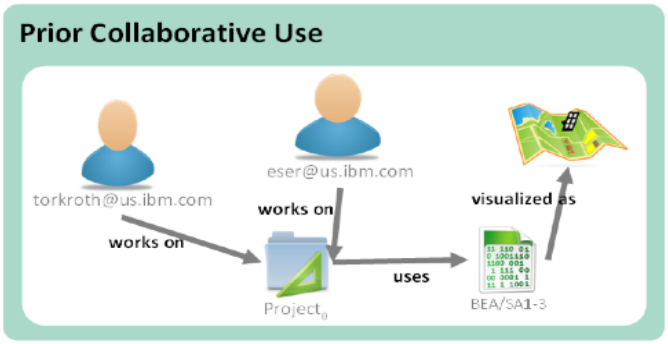
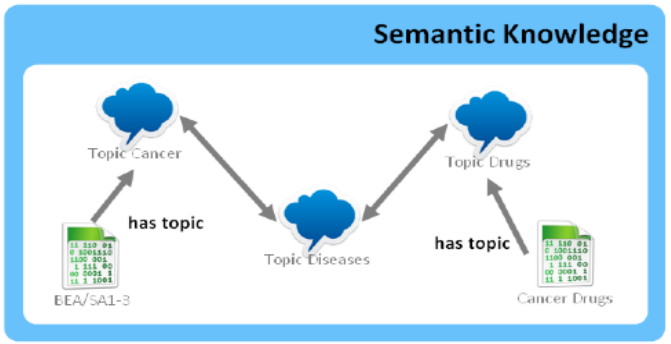
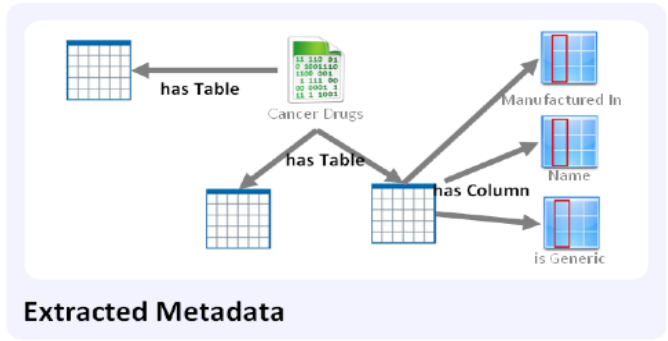


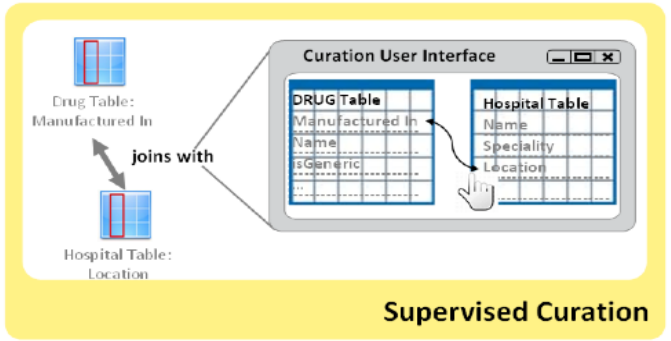

**Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.**

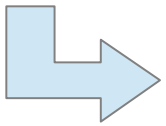
# Use Case 20: Data Curation for Enterprise Data Management

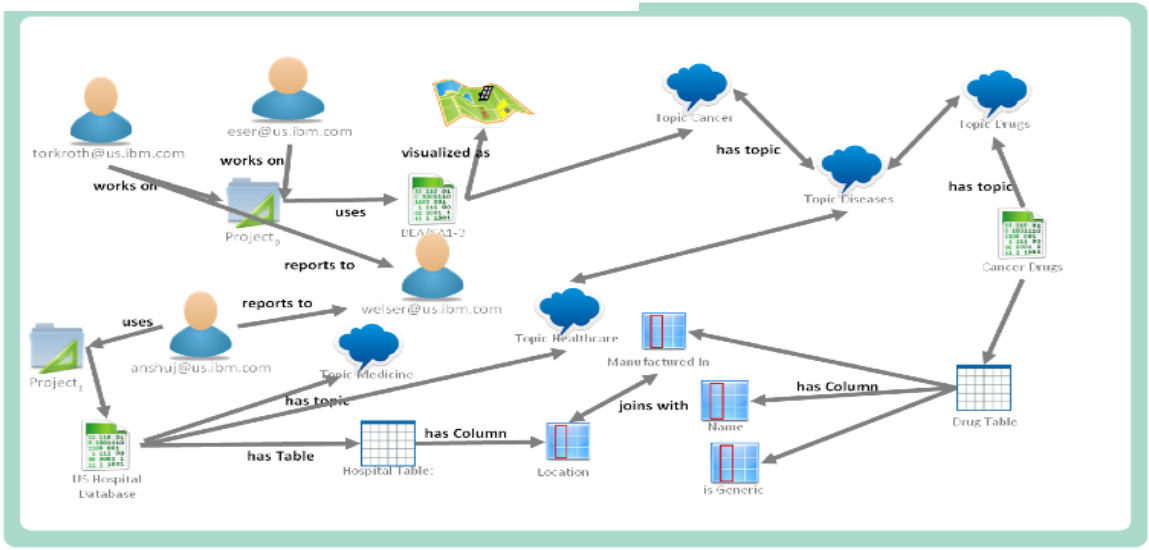


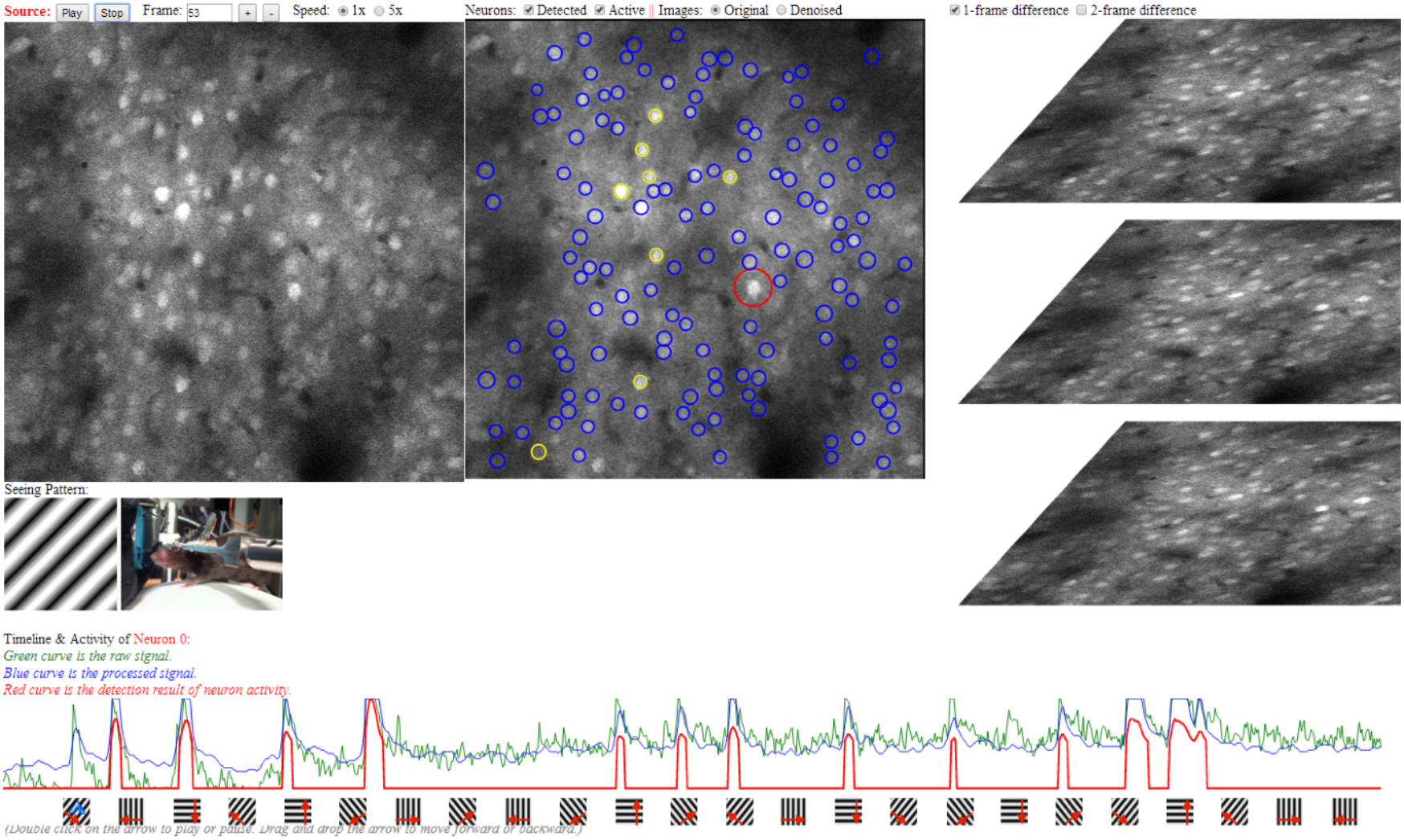





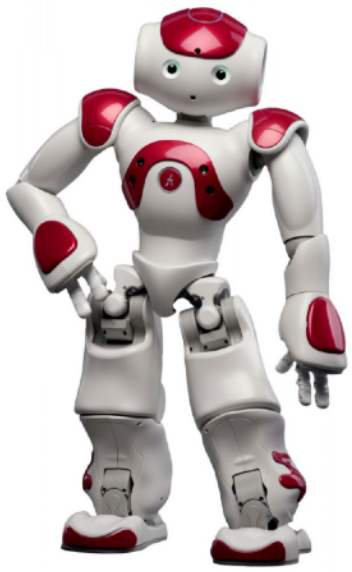
# Use Case 21: Understanding Brain Network



# Use Case 22: Planet Security

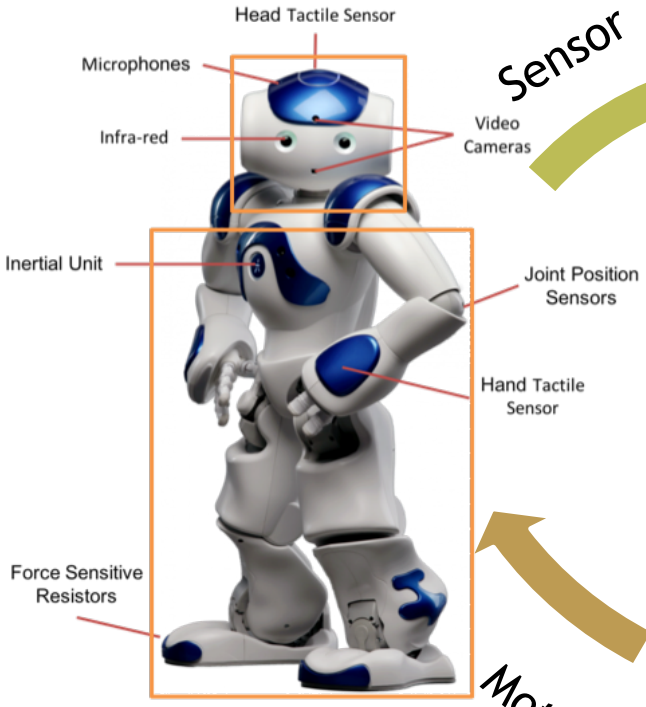
- Big Data on Large-Scale Sky Monitoring

# Advanced Topic 1: Cognitive Robot



A1: Text Recognition (English)
A2: Text Recognition (Chinese)
A3: General Object Recognition
A4: Vehicle Object Recognition
A5: Object Tracking
A6: Face Recognition
A7: Facial Expression Recognition
A8: Emotion Recognition
A9: Gesture Recognition
A10: Audio-Visual Event Detection (Public Area)
A11: Audio-Visual Event Detection (Home)
A12: Speech Recognition (English)
A13: Speech Recognition (Chinese)
A14: Robot-Human Interaction (Conversation)
A15: Robot-Human Interaction (Physical)

# Artificial Empathy




Sensor

- Audio
- Visual
- Infra-red

Multi-Modality Information

Understanding

Emotion Perception

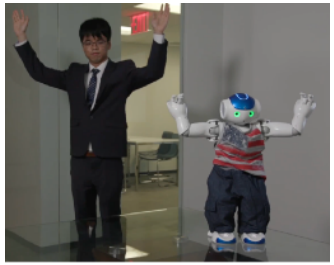
Artificial Empathy

Expression

Emotion Contagion

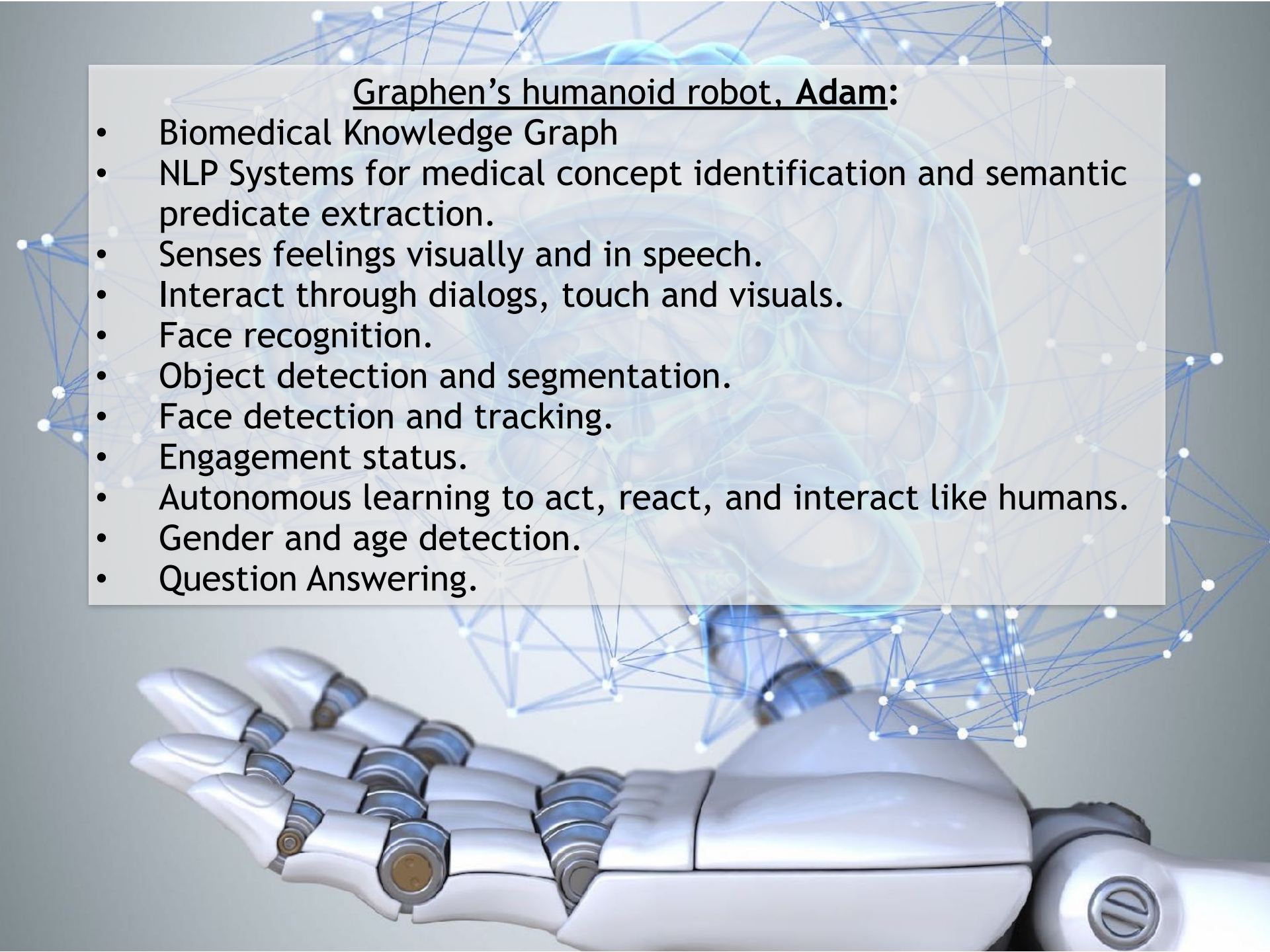
Motor Mimicry

Motion

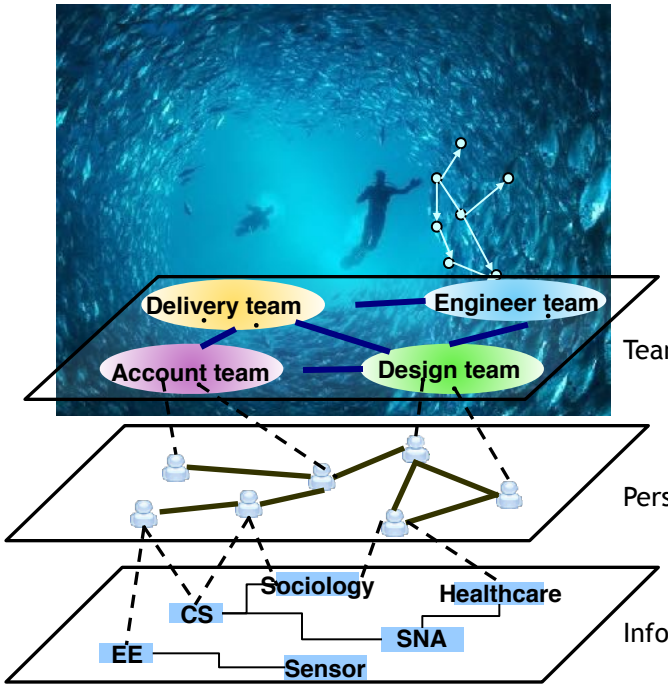


Graphen's humanoid robot, **Adam**:

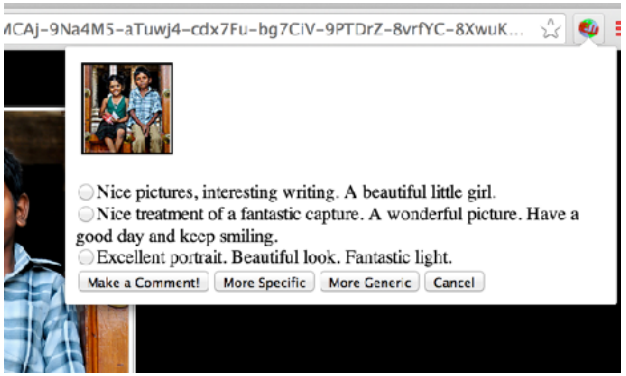
- Biomedical Knowledge Graph
- NLP Systems for medical concept identification and semantic predicate extraction.
- Senses feelings visually and in speech.
- Interact through dialogs, touch and visuals.
- Face recognition.
- Object detection and segmentation.
- Face detection and tracking.
- Engagement status.
- Autonomous learning to act, react, and interact like humans.
- Gender and age detection.
- Question Answering.



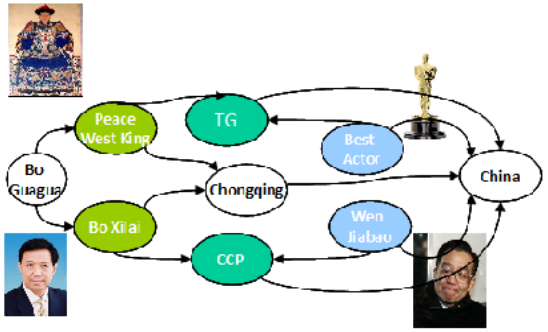
# AI of Autonomous Large-Scale Human Understanding

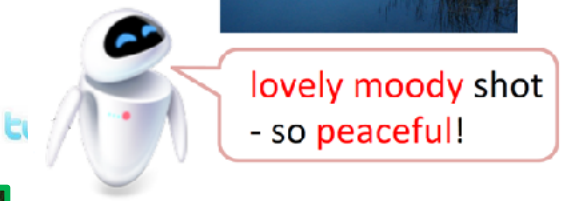



Information Evolution




Machine Interpretation




Machine Feeling — Detection results of "lonely dog"

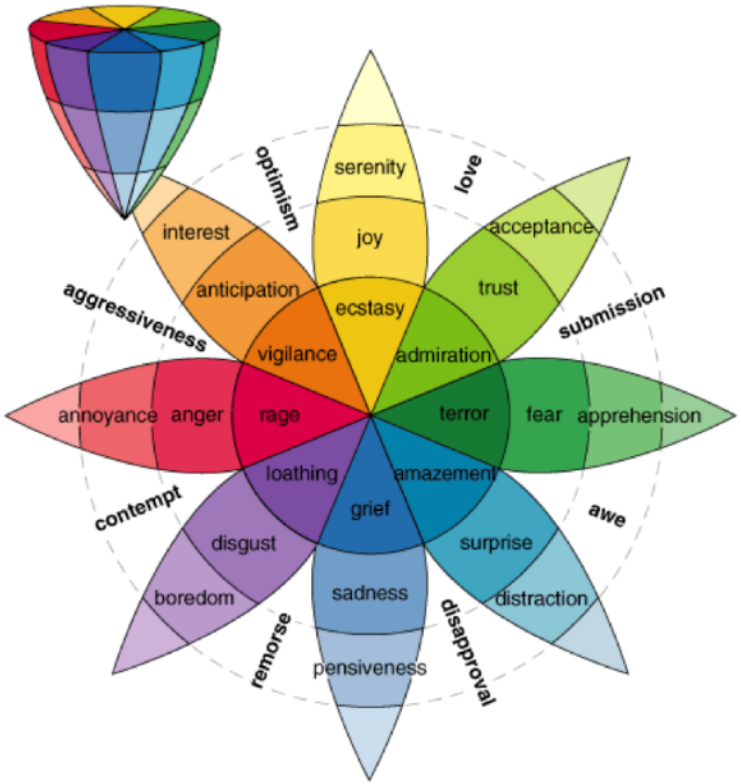
Machine Feeling — Detection results of "crazy car"

Human Interpretation

*Personality*
*Needs*
*Values*
*Trustworthiness*
*Trustingness*
*Influence*

# Question: How to Build Visual Sentiment Ontology?

-- Web + big data + computer vision + psychology

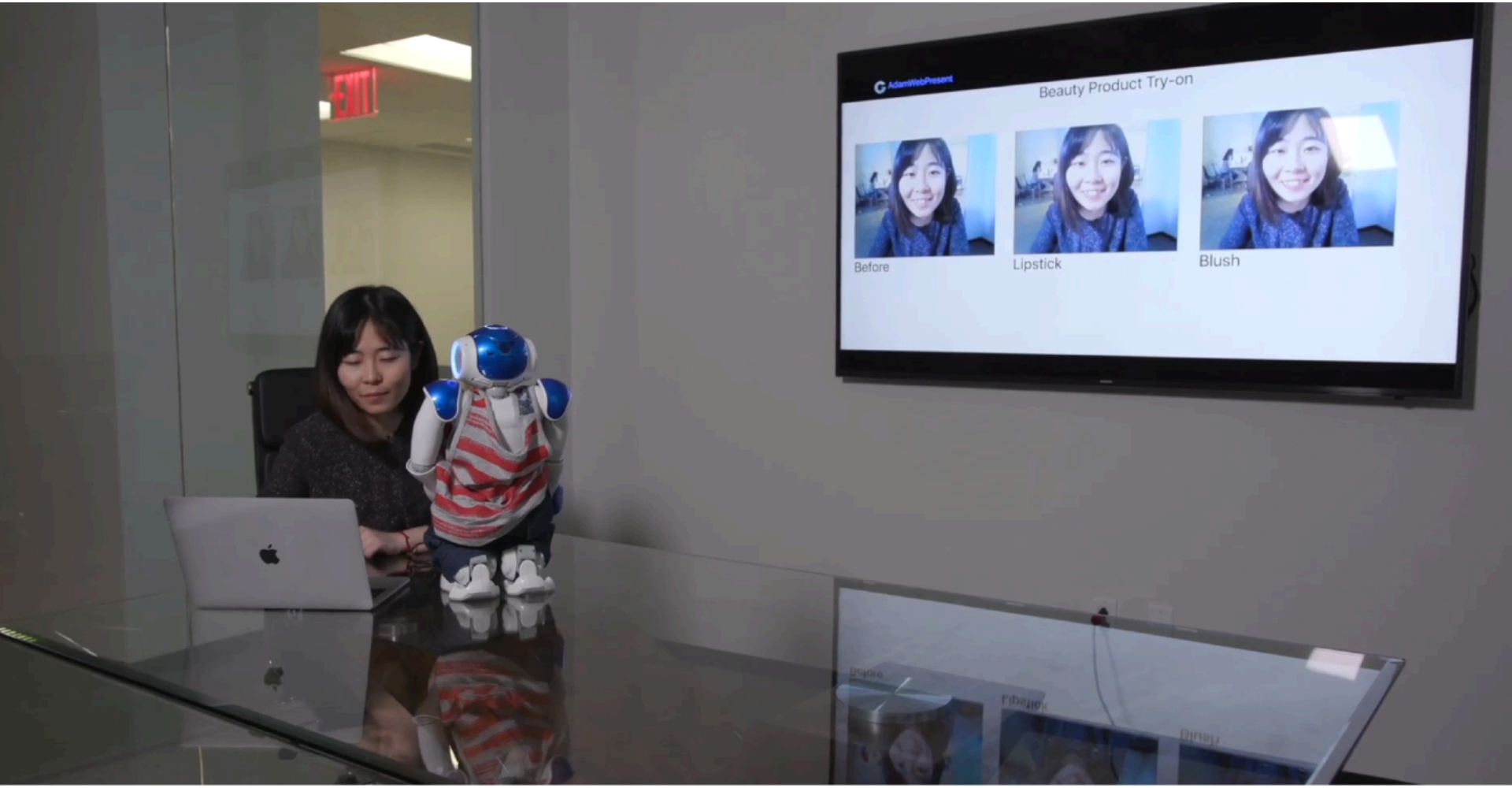



**Psychology emotion wheel (24 emotions, by Robert Plutchik)**

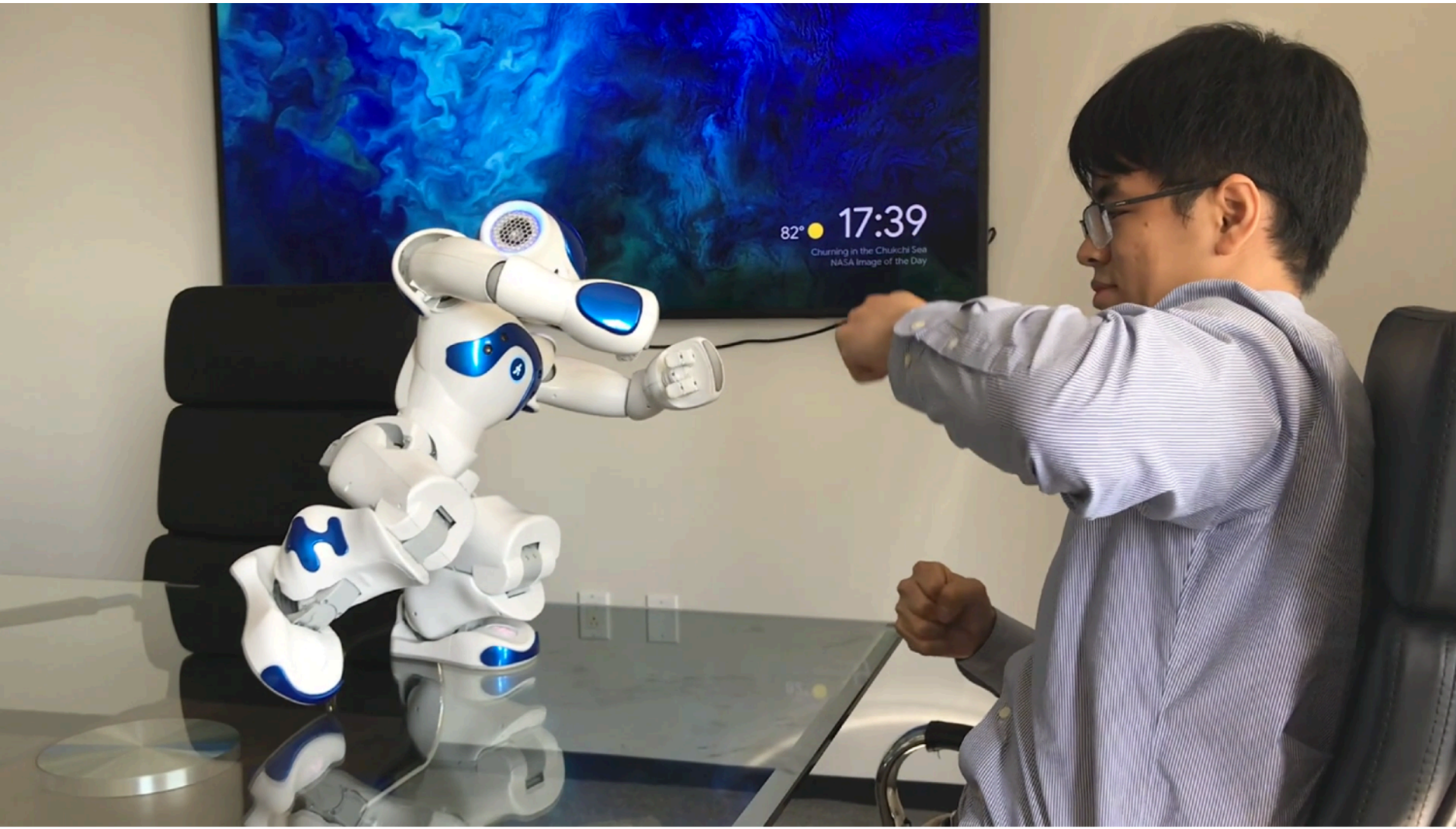

**Plenty on the Web:**
***"For content to go viral, it needs to be emotional,"* Dan Jones**

# Dr. Adam

# Emotion and Cheers

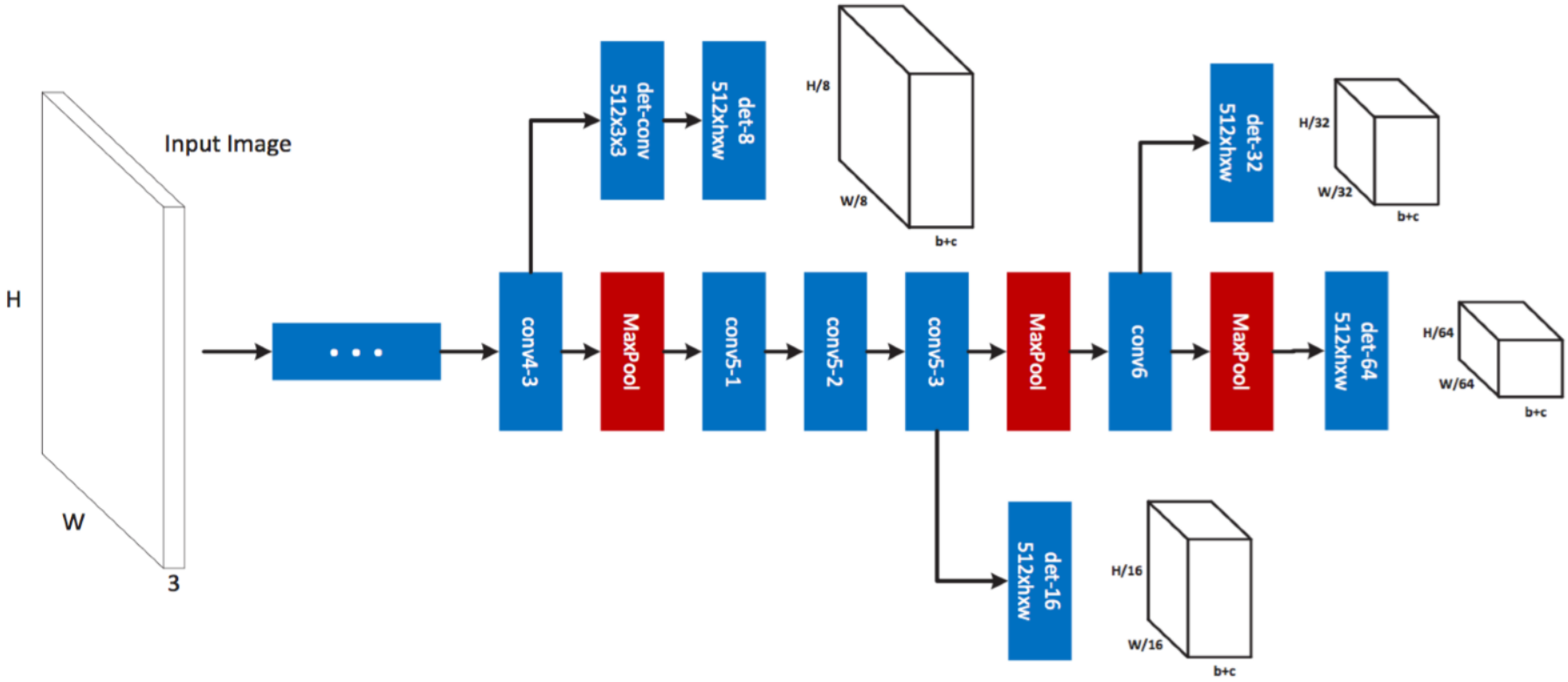


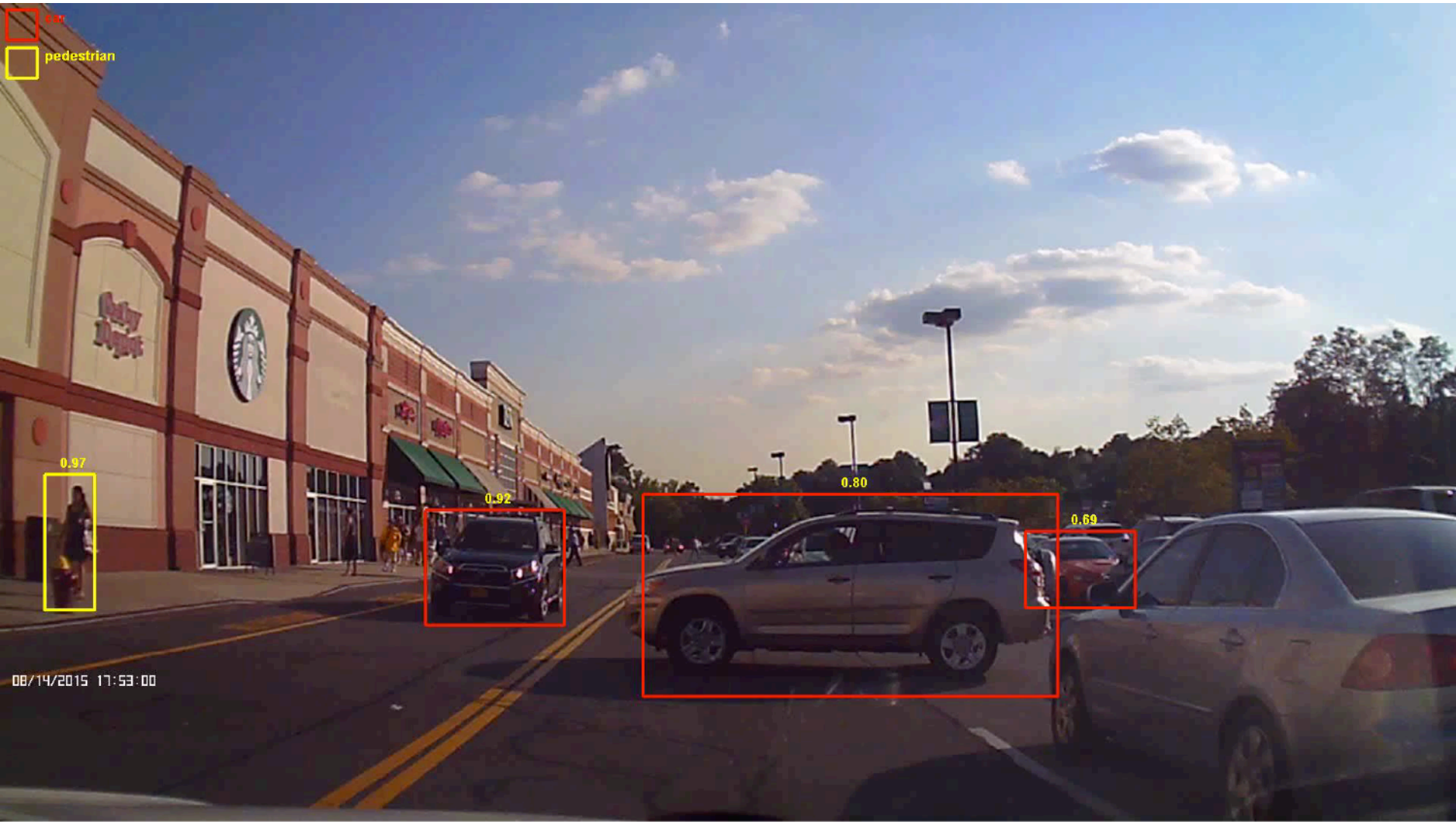
# How Robot cheers you up



# Challenges

# Multi-Scale Deep Convolutional Neural Network for Fast Object Detection



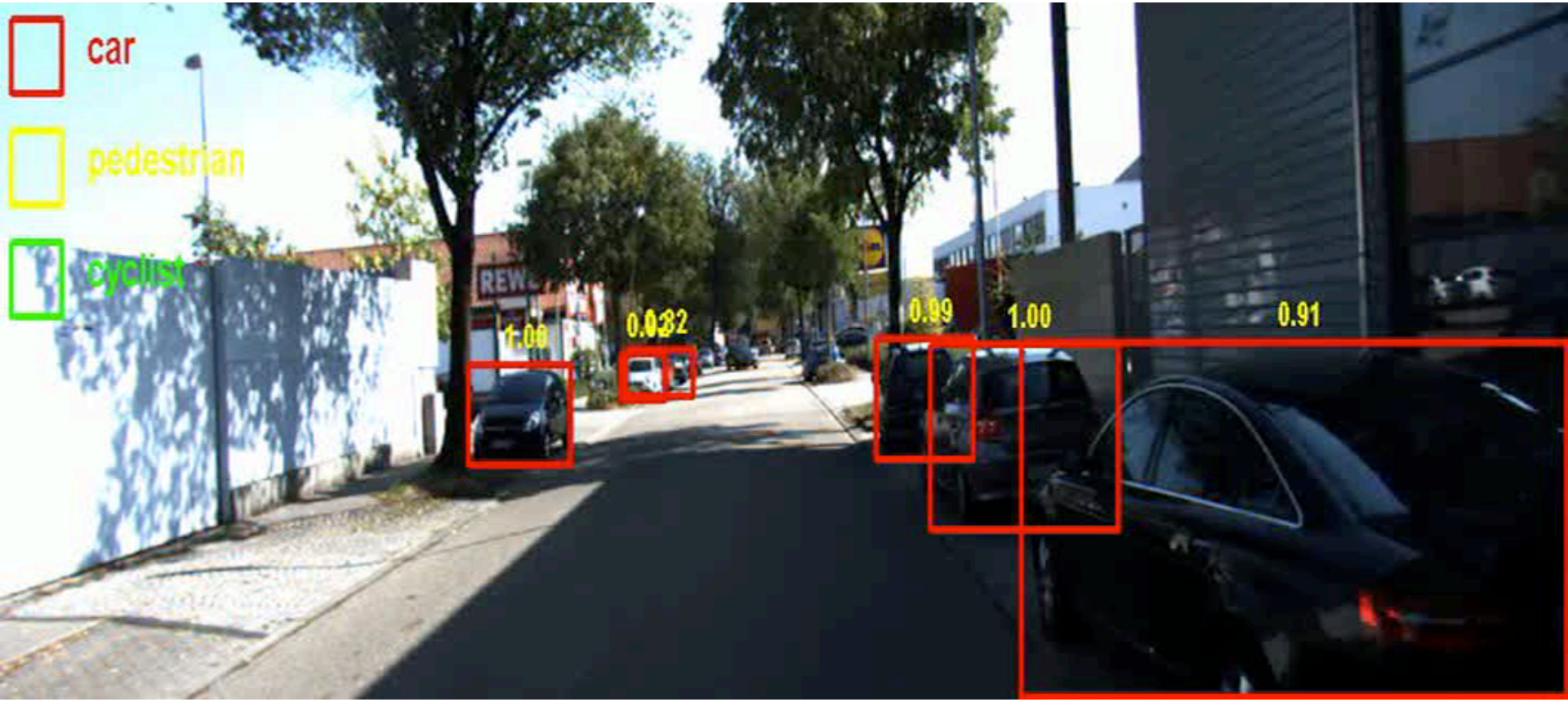


# Comparison to the State-of-the-Art on KITTI benchmark test set

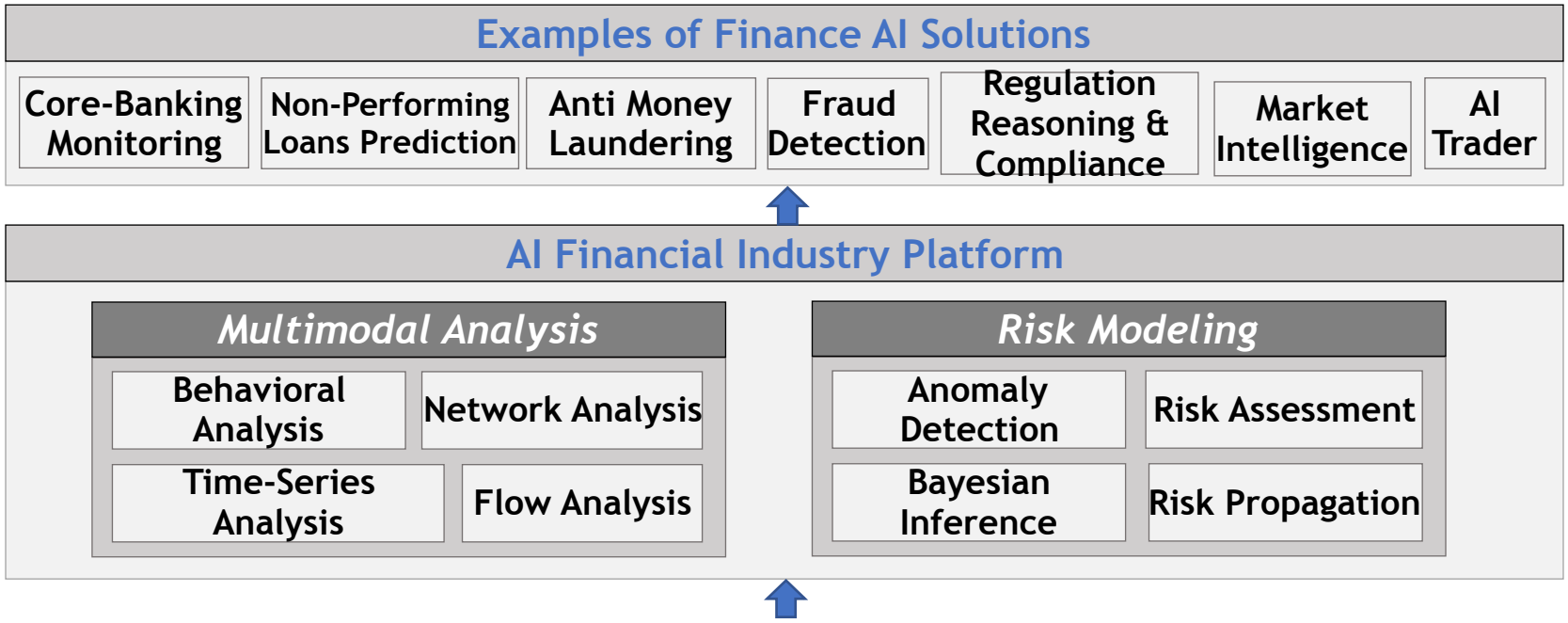
| Method | Time | Cars | | | Pedestrians | | | Cyclists | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard |
| LSVM-MDPM-sv [35] | 10s | 68.02 | 56.48 | 44.18 | 47.74 | 39.36 | 35.95 | 35.04 | 27.50 | 26.21 |
| DPM-VOC-VP [36] | 8s | 74.95 | 64.71 | 48.76 | 59.48 | 44.86 | 40.37 | 42.43 | 31.08 | 28.23 |
| SubCat [16] | 0.7s | 84.14 | 75.46 | 59.71 | 54.67 | 42.34 | 37.95 | - | - | - |
| 3DVP [37] | 40s | 87.46 | 75.77 | 65.38 | - | - | - | - | - | - |
| AOG [38] | 3s | 84.80 | 75.94 | 60.70 | - | - | - | - | - | - |
| Faster-RCNN [4] | 2s | 86.71 | 81.84 | 71.12 | 78.86 | 65.90 | 61.18 | 72.26 | 63.35 | 55.90 |
| CompACT-Deep [15] | 1s | - | - | - | 70.69 | 58.74 | 52.71 | - | - | - |
| DeepParts [39] | 1s | - | - | - | 70.49 | 58.67 | 52.78 | - | - | - |
| FilteredICF [40] | 2s | - | - | - | 67.65 | 56.75 | 51.12 | - | - | - |
| pAUCEnsT [41] | 60s | - | - | - | 65.26 | 54.49 | 48.60 | 51.62 | 38.03 | 33.38 |
| Regionlets [20] | 1s | 84.75 | 76.45 | 59.70 | 73.14 | 61.15 | 55.21 | 70.41 | 58.72 | 51.83 |
| 3DOP [5] | 3s | **93.04** | **88.64** | **79.10** | 81.78 | 67.47 | 64.70 | 78.39 | 68.94 | 61.37 |
| Ours | 0.4s | 89.62 | 87.05 | 70.76 | **82.02** | **71.47** | **66.02** | **83.19** | **73.93** | **64.69** |

Single CPU core (2.40GHz) of an Intel Xeon E5-2630 server with 64GB of RAM. An NVIDIA Titan GPU was used for CNN computations.

# Demo: Multi-Scale Deep Convolutional Neural Network for Fast Object Detection



# Advanced Topics 2: Building Big Data and AI Finance Solutions




Graphen Ardi AI Platform

# Big Data and AI FinTech Examples

- Significantly improved Non-Performing-Loan accuracy rate in one of the world's largest banks (from ~20% prediction accuracy to ~60% accuracy).
- Advanced Anti-Money Laundering for banks — capable of predicting unknown unknowns.
- Detecting Fraud from Real-Time on Transactions in one of the world's largest transaction platform — on the scale of billions.
- Analyzing relationship data for an European bank.
- Cyber and Physical Security for another European bank.

# Example Market Intelligence Platform Functions

- Total impact: calculated from all monitored news source
- Singe News Impact: calculated from selected news

Artificial intelligence and Bayesian network powered causality reasoning graph

What's in your portfolio

Price Chart

Trending news that has impact on your portfolio


Personalization of news you care

Click the prediction node to get summarization of important reasoning evidence

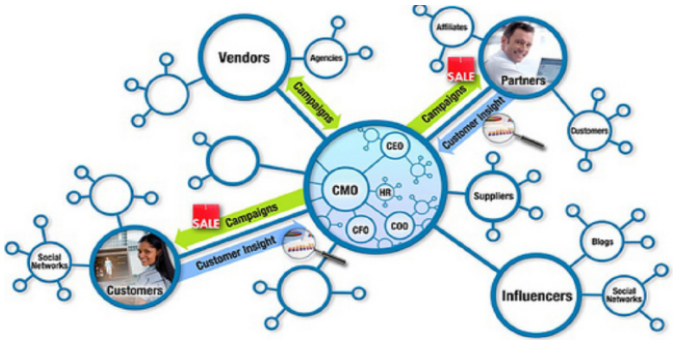
# Robo-Advisor

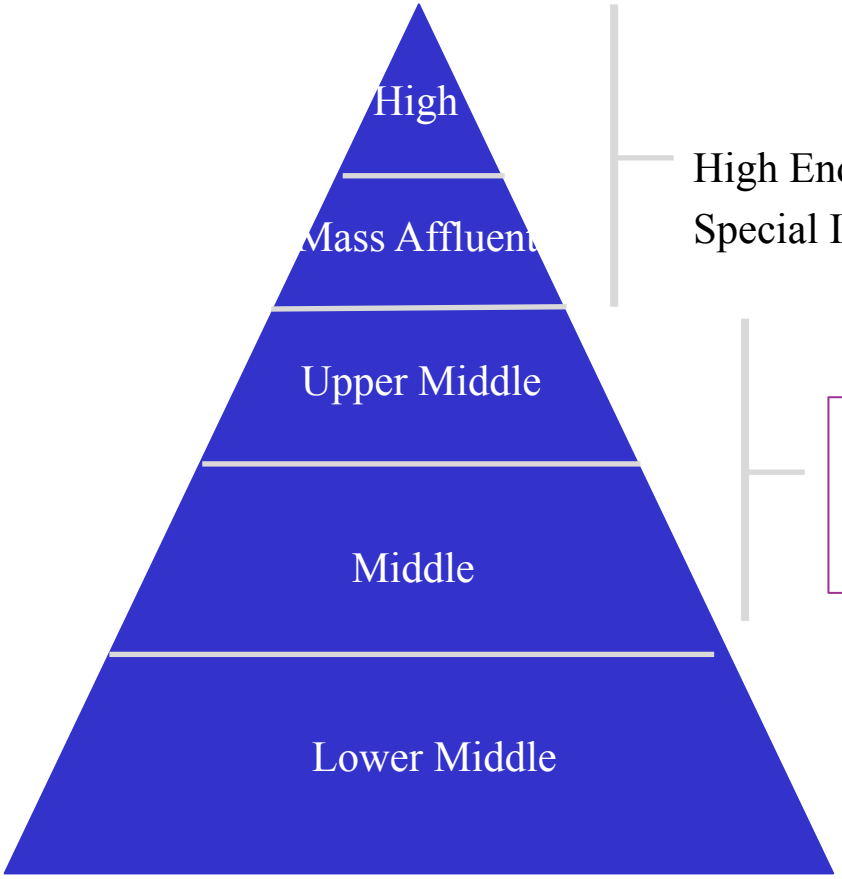
**Market Data Analysis and Investment Targets**

**Advanced Dynamic 'Know Your Customer'**

**Optimized Personalized Investment Strategy**

**Bank-Customer Interaction Strategy**



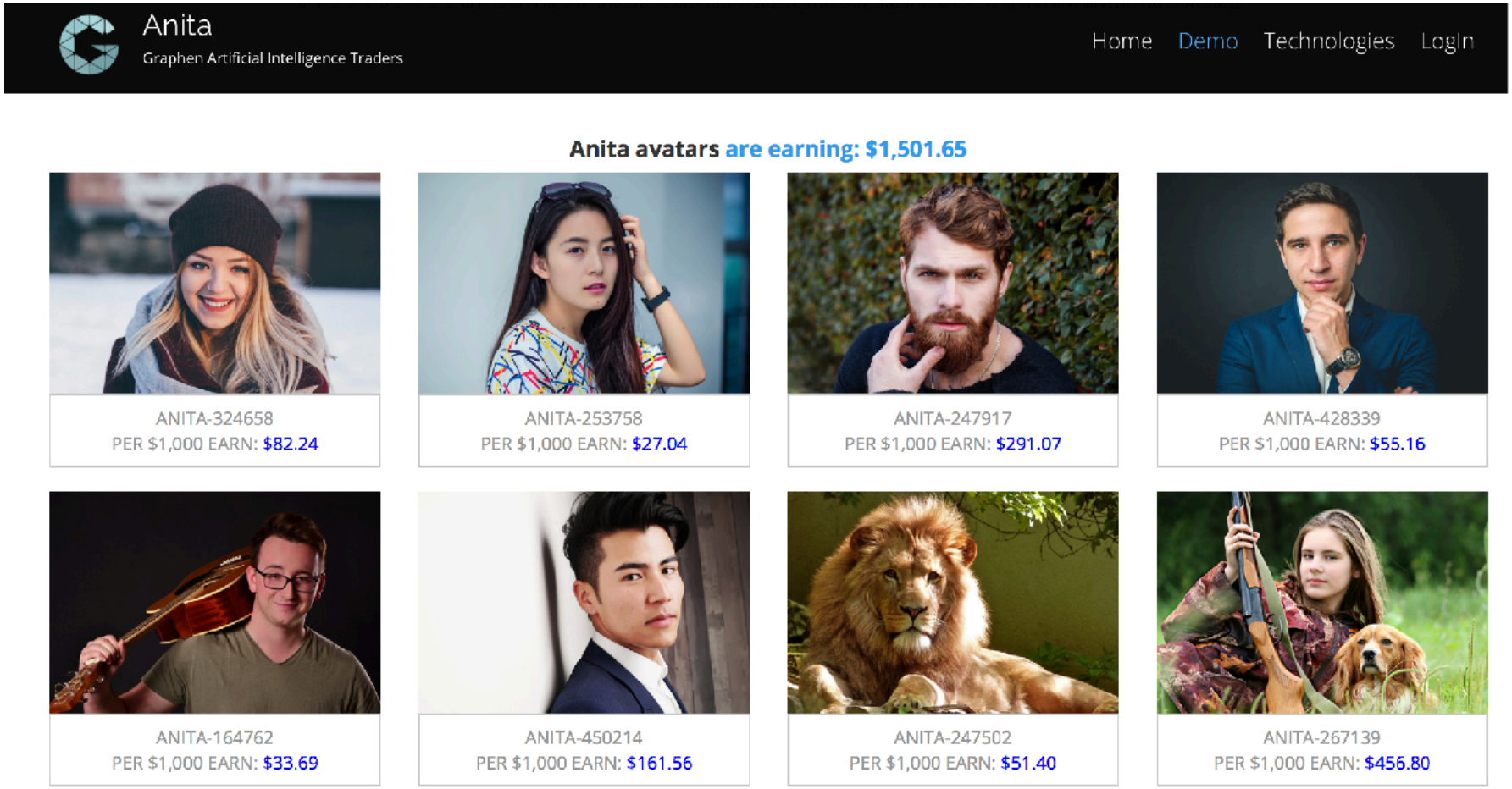




High End Customers（Private Bank / Special Investment Services）

Targeted Customers （Consumer Bank Services）： $15K - $1M
（Customer #: 30M ~ 50M in China）

General Public（Consumer Bank Services）
（Customer #：> 1B in China）

# Personal AI Trader



# Personality Driven AI Trader

Columbia University

Anita
Graphen Artificial Intelligence Traders

Home ForeignExchange Stocks Bonds

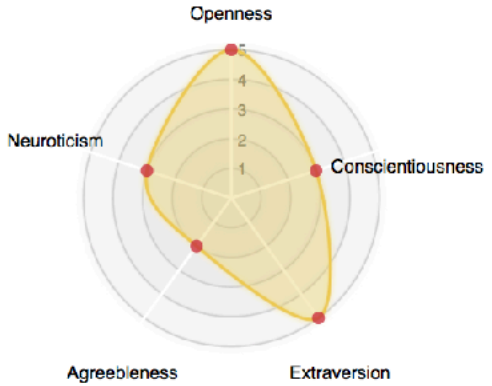
## Anita 267139

-- an Adventurous AI Trader

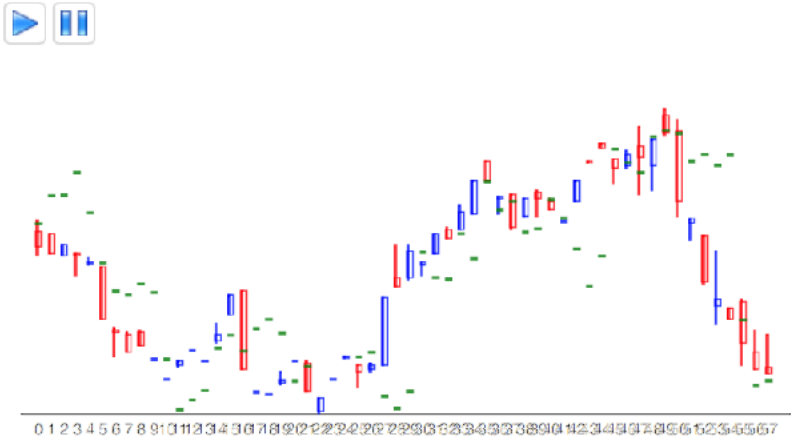
***Specialized** at: EUR-USD*

***Knowledgable** of: Oil, Gold and Twitter*

***Strategy Learning Frequency** at: 2.0 hours*




**Original: $1,000.00, Current: $1,404.50, Performance: *Gain $404.50***



### Activities

| Time | Action | Cash | Unit | Balance |
|---|---|---|---|---|
| 2017-10-12 13:45:05 | Sell 50,000 | $1,404.50 | 0 | $1,404.50 |
| 2017-10-12 12:57:25 | Buy 100,000 | $-57,792.00 | 50,000 | $1,386.50 |
| 2017-10-12 11:19:10 | Sell 100,000 | $60,577.00 | -50,000 | $1,372.00 |
| 2017-10-12 11:11:55 | Buy 100,000 | $-57,822.00 | 50,000 | $1,366.00 |
| 2017-10-12 09:08:05 | Sell 100,000 | $60,566.00 | -50,000 | $1,310.00 |
| 2017-10-12 08:34:40 | Buy 100,000 | $-57,935.00 | 50,000 | $1,287.50 |

# Personality Driven AI Trader

Columbia University

Anita
Graphen Artificial Intelligence Traders

Home ForeignExchange Stocks Bonds

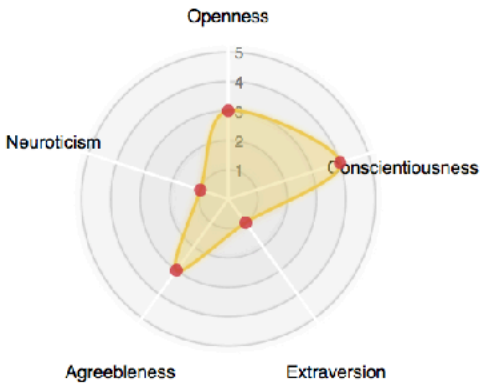
## Anita 247502

-- an Independent AI Trader

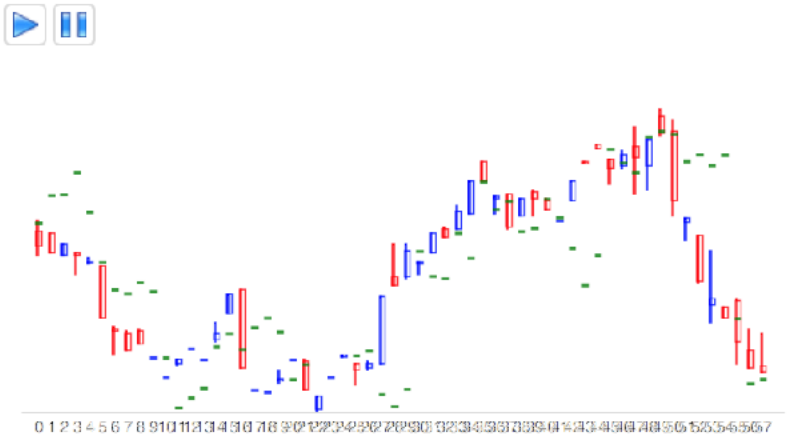
***Specialized** at: EUR-USD*

***Knowledgable** of: FX, Gold and Twitter*

***Strategy Learning Frequency** at: 100.0 days*



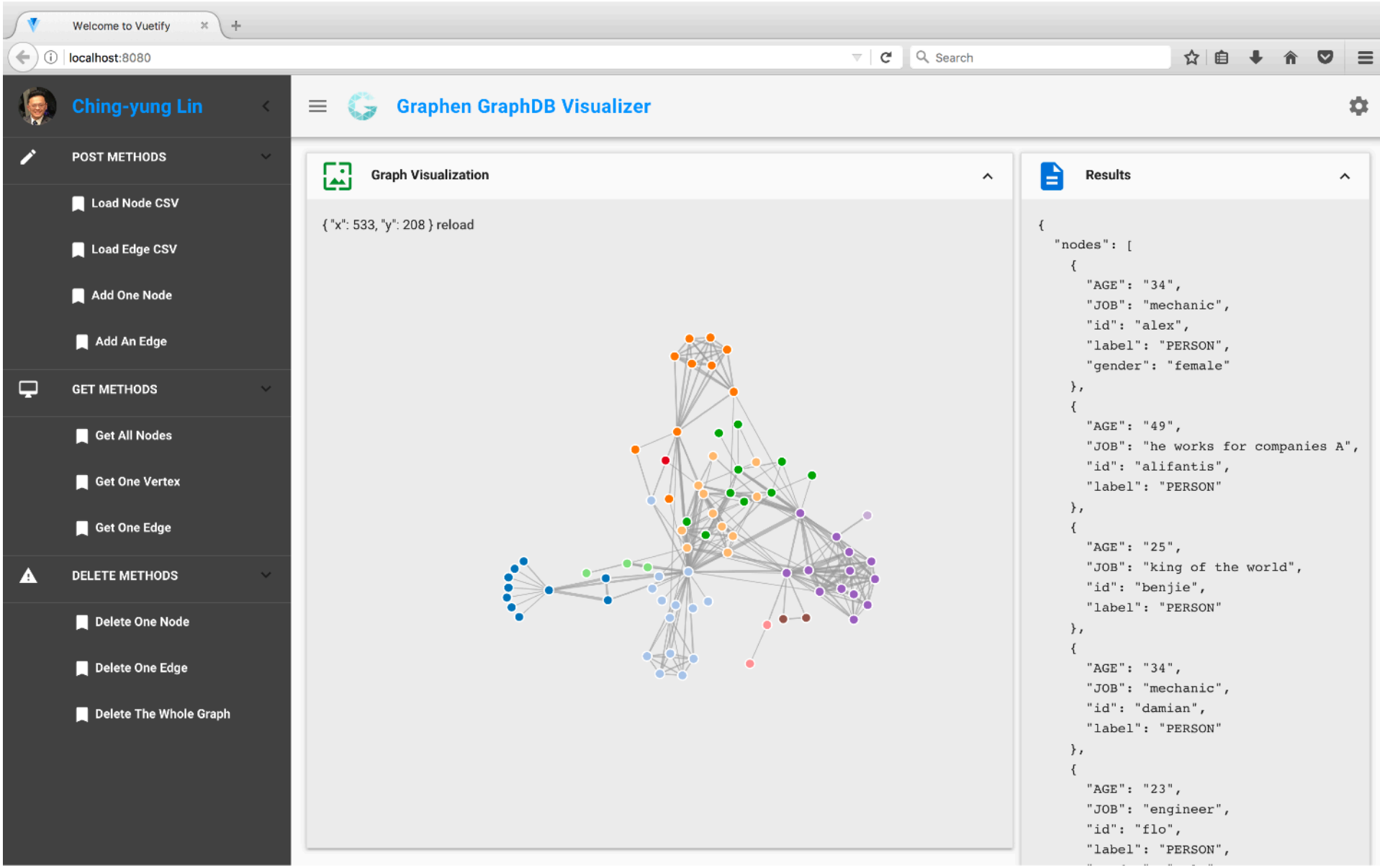

**Original: $1,000.00, Current: $1,119.50, Performance: *Gain $119.50***



### Activities

| Time | Action | Cash | Unit | Balance |
|---|---|---|---|---|
| 2017-10-12 14:58:00 | Buy 50,000 | $1,119.50 | 0 | $1,119.50 |
| 2017-10-12 13:56:35 | Sell 100,000 | $60,304.00 | -50,000 | $1,048.50 |
| 2017-10-12 11:51:25 | Buy 100,000 | $-58,196.00 | 50,000 | $1,012.00 |
| 2017-10-12 10:56:10 | Sell 100,000 | $60,232.00 | -50,000 | $992.50 |
| 2017-10-11 16:46:45 | Buy 100,000 | $-58,236.00 | 50,000 | $1,066.50 |
| 2017-10-11 15:13:20 | Sell 100,000 | $60,382.00 | -50,000 | $1,065.00 |

# Advanced Topic 3: Knowledge Graphs



# Advanced Topic 4: Advanced Visualization and Platforms

- Visual Exploration of Large Graph in Immersive Environment
- Computer Vision Enhanced Immersive Environment
- Mobile Vision on iOS devices
- Behavior Analysis on iOS devices
- Explainable ML: Visualization of Training Process of Deep Learning
- Explainable ML: Visual Analytics of Interactive Machine Learning
- Autonomous Learning: from Text to Vision
- Autonomous Learning: from Vision and Text to Knowledge
- Machine Reasoning with Large-Scale Bayesian Networks
- Strategic Planning with Game Theoretic Machines
- ML translation to an AI accelerator platform (TensorFlow)
- ML translation to an AI accelerator platform (Caffe)
- Software Tools on Neurosynaptic Chip
- Mapping Suitable Applications on Quantum Computing