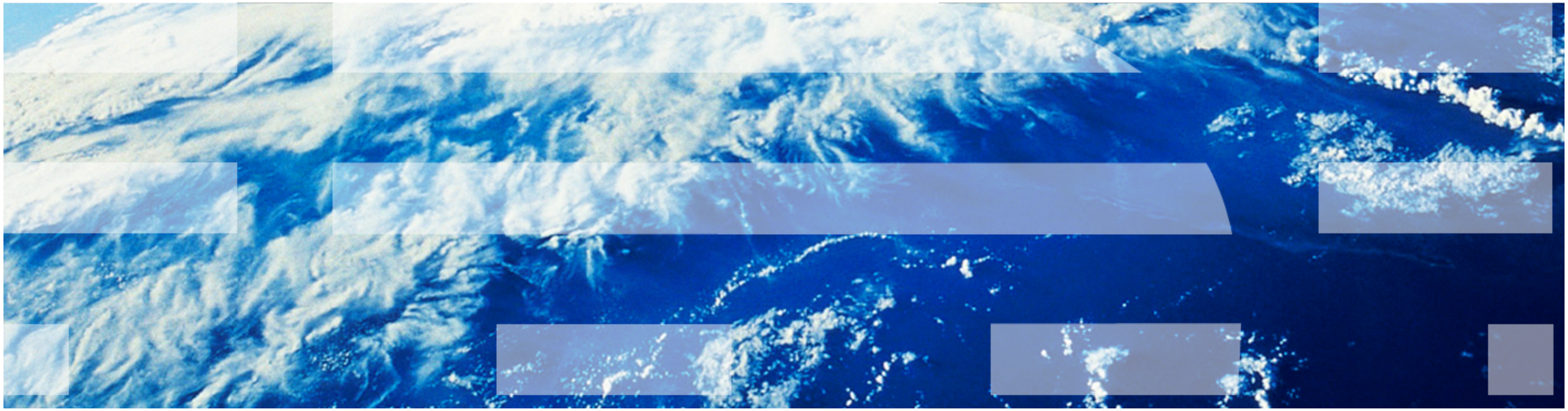
EECS E6893 Big Data Analytics Lecture 1:

Overview of Big Data Analytics

Ching-Yung Lin, Ph.D.

Adjunct Professor, Depts. of Electrical Engineering and Computer Science

IEEE Fellow



September 6th, 2024

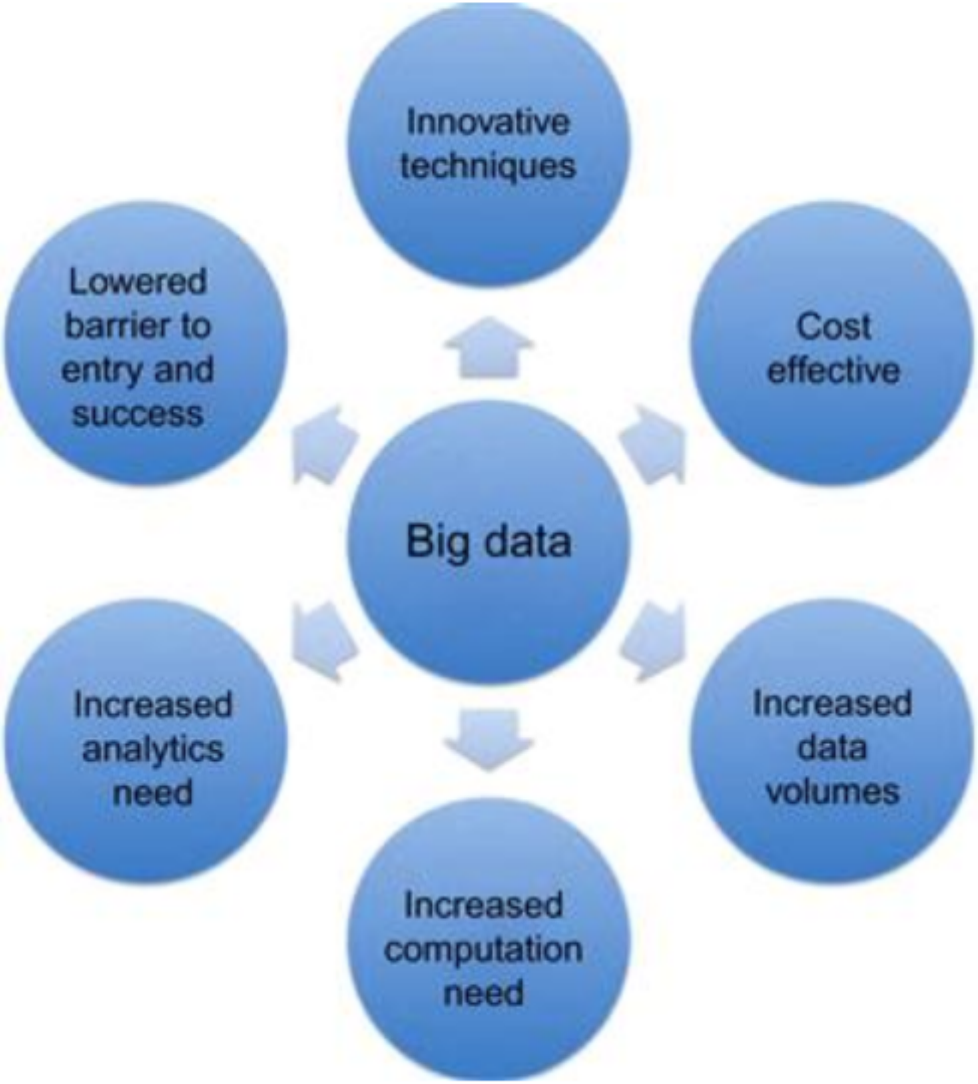
Definition and Characteristics of Big Data

"*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

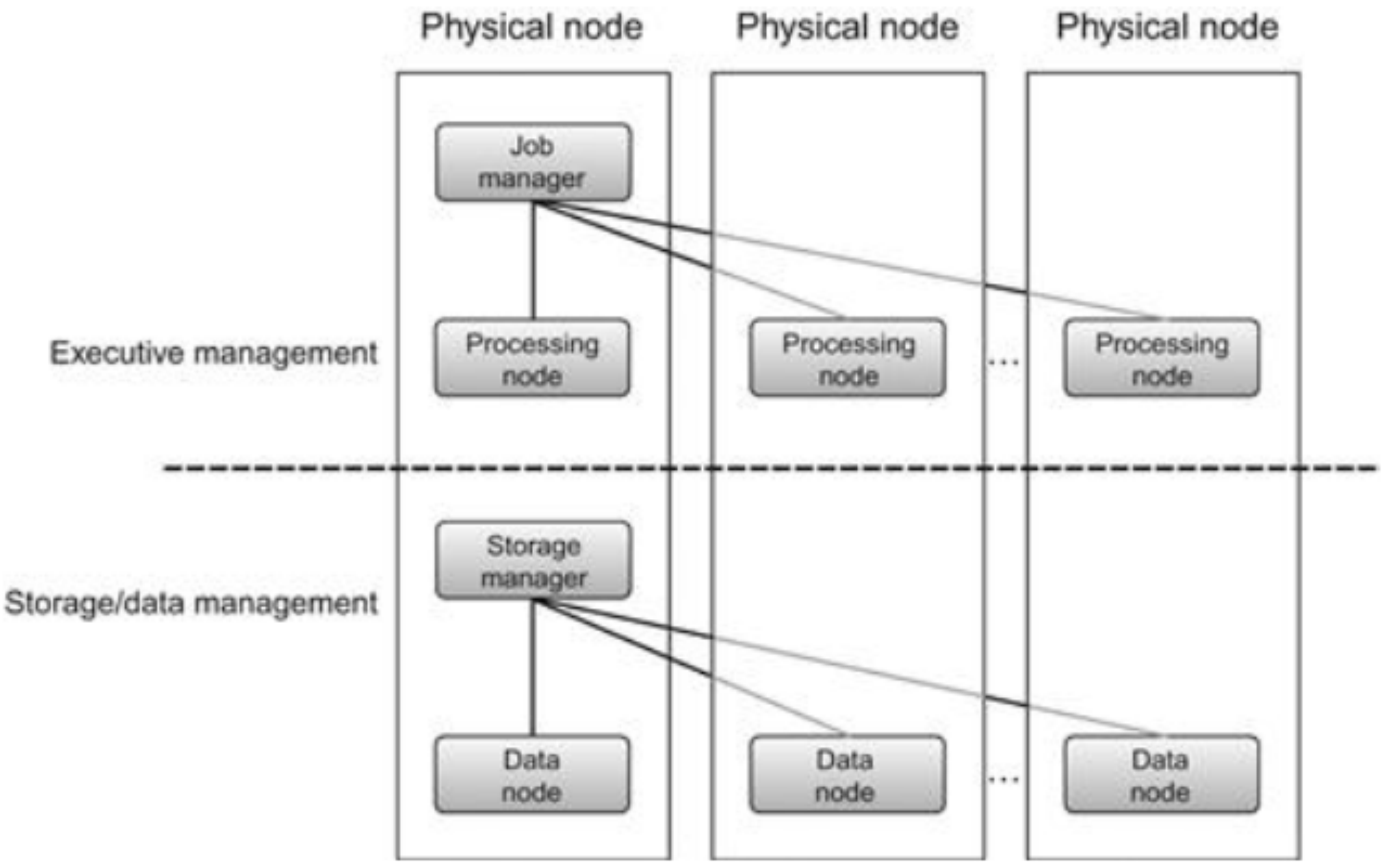
What made Big Data needed?



"Big Data Analytics", David Loshin

Key Computing Resources for Big Data

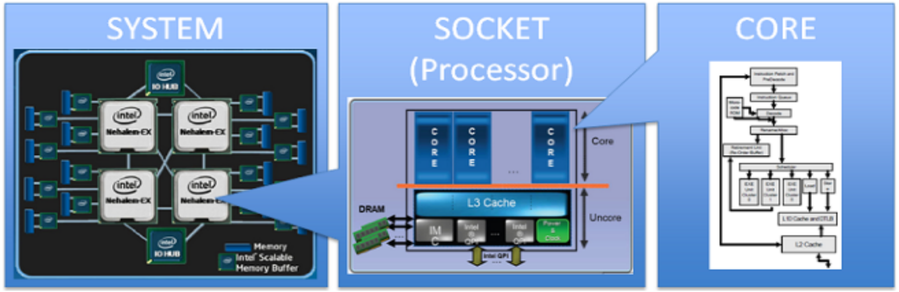
- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network

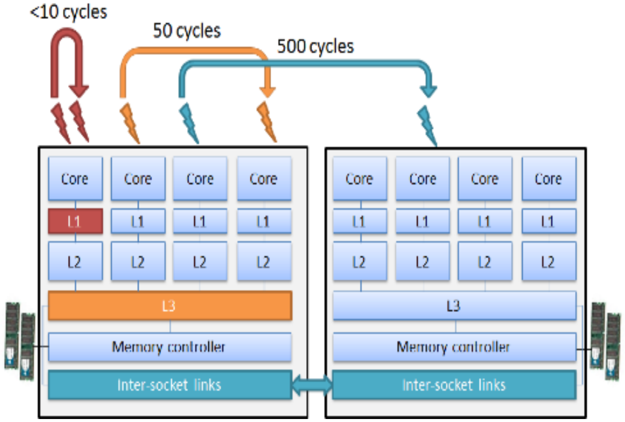


“Big Data Analytics”, David Loshin

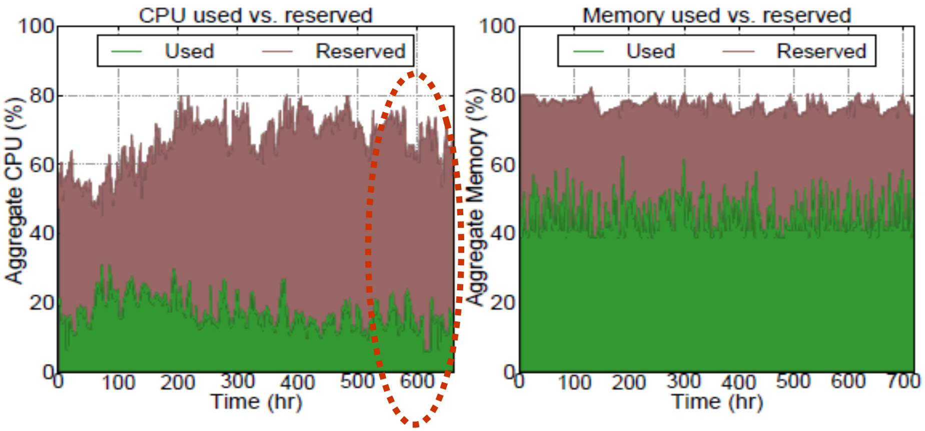
Scalability — Scale Up & Scale Out

- 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





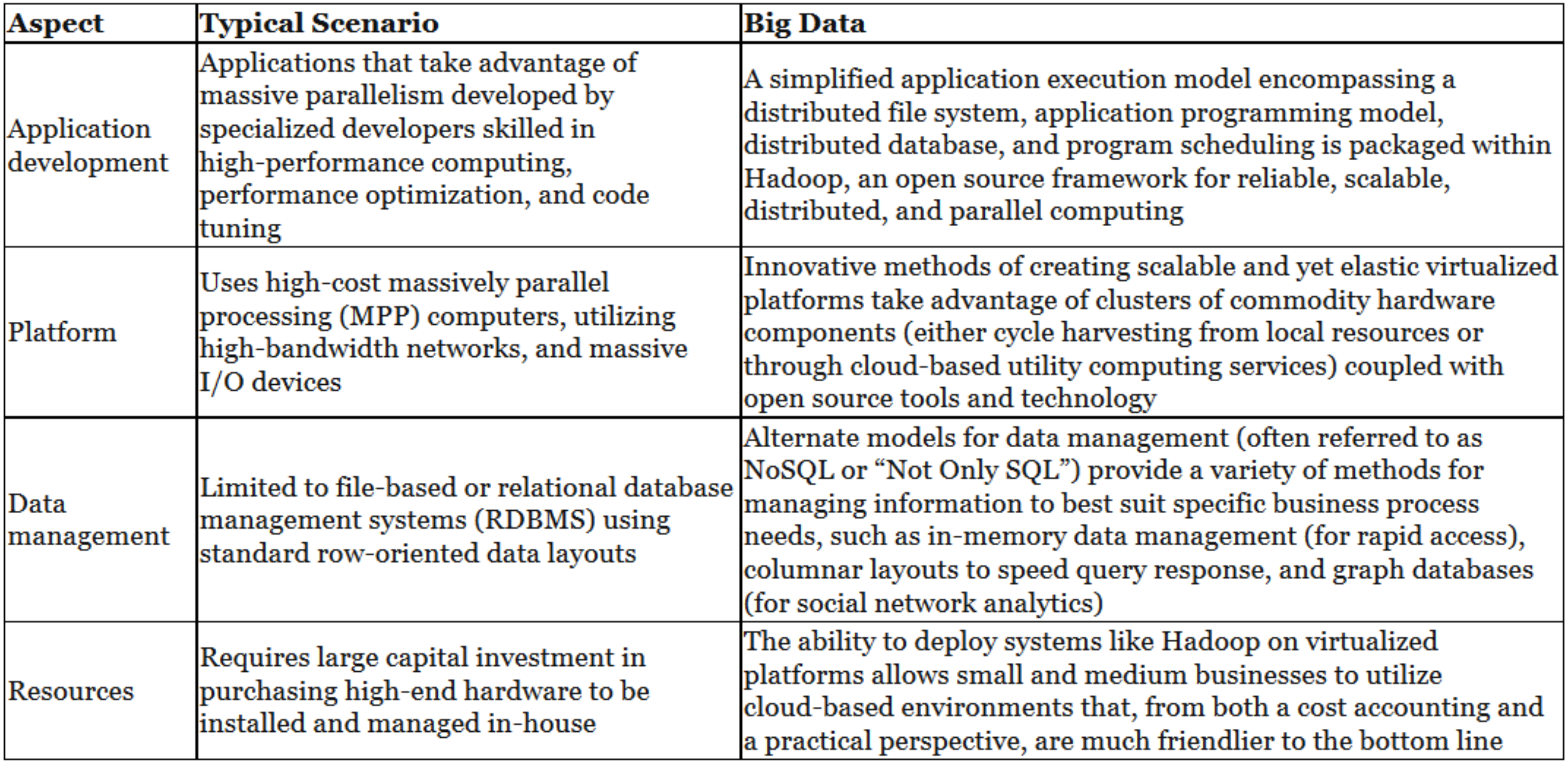
Example: Resource utilization for a large production cluster at Twitter data center



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

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?

Why Big Data now?

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

Why Big Data now?

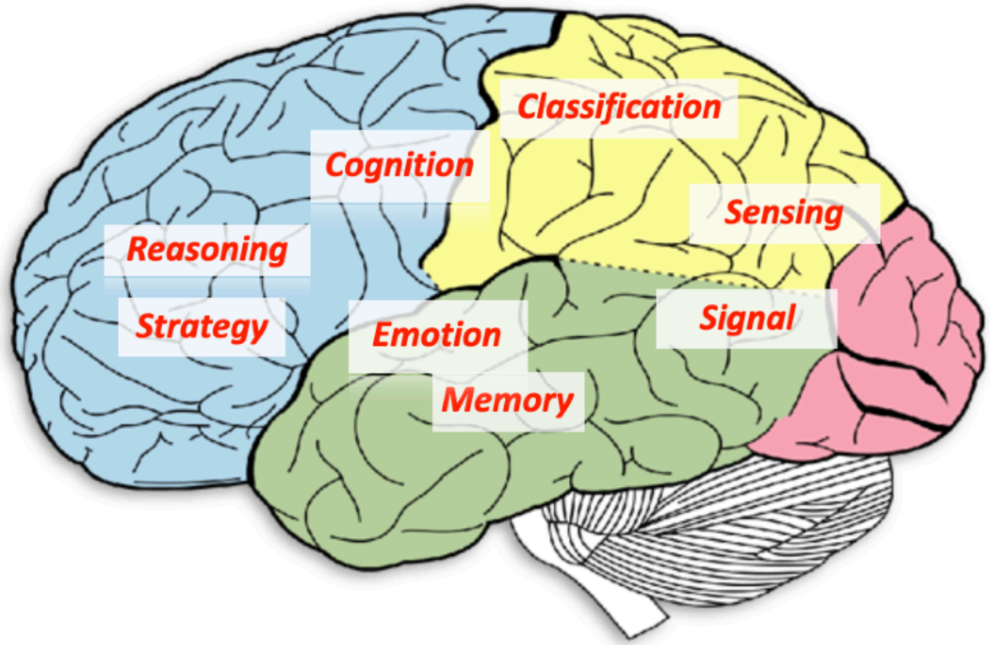
- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

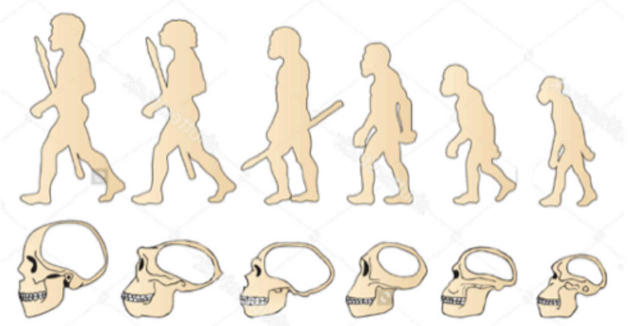
→
- High-Volume
- High-Velocity
- High-Variety

→ Artificial Intelligence

Evolution of Intelligence

Direction of the Evolution of Intelligence





Evolution of Artificial Intelligence is similar, but much faster

Timeline of the Evolution of Artificial Intelligence

2020s 2010s 2000s 1990s 1960s - 1980s

Ardi AI Platform

Reasoning
- Bayesian Networks
- Causality Inference
- Behavior Prediction

Strategy
- Game Theory
- Optimal Solution

Personality
- Humanization
- Ethics

Emotion
- Feeling
- Affection

Expression
- Big Data Visualization
- Explainability
- Interaction

Learning
- Machine Learning and Deep Learning
- Autonomous Learning

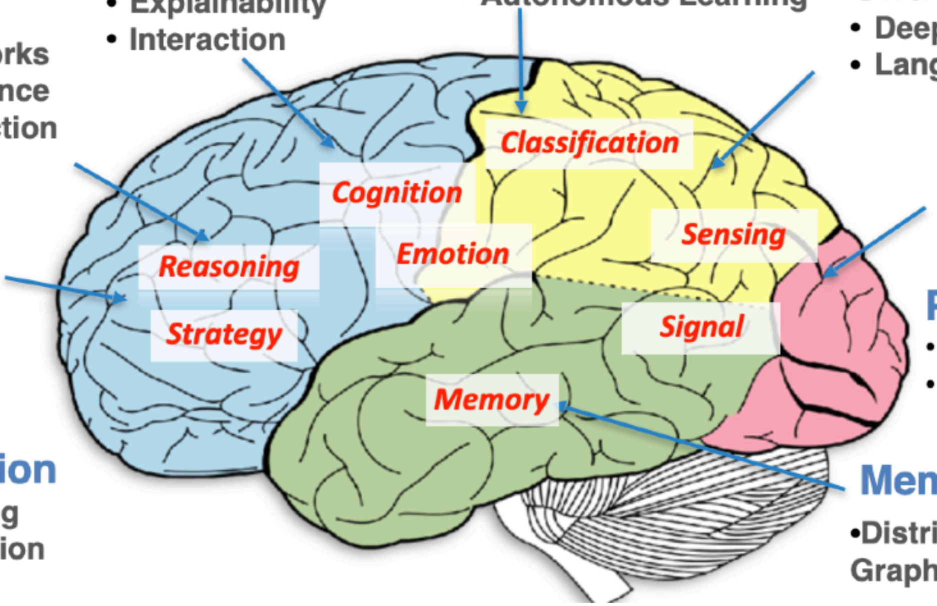
Understanding
- Deep Vision Understanding
- Language Understanding

Perception
- Network Analysis
- Feature Analysis

Action
- Control
- Behavior

Pipeline
- Coordination
- Flow

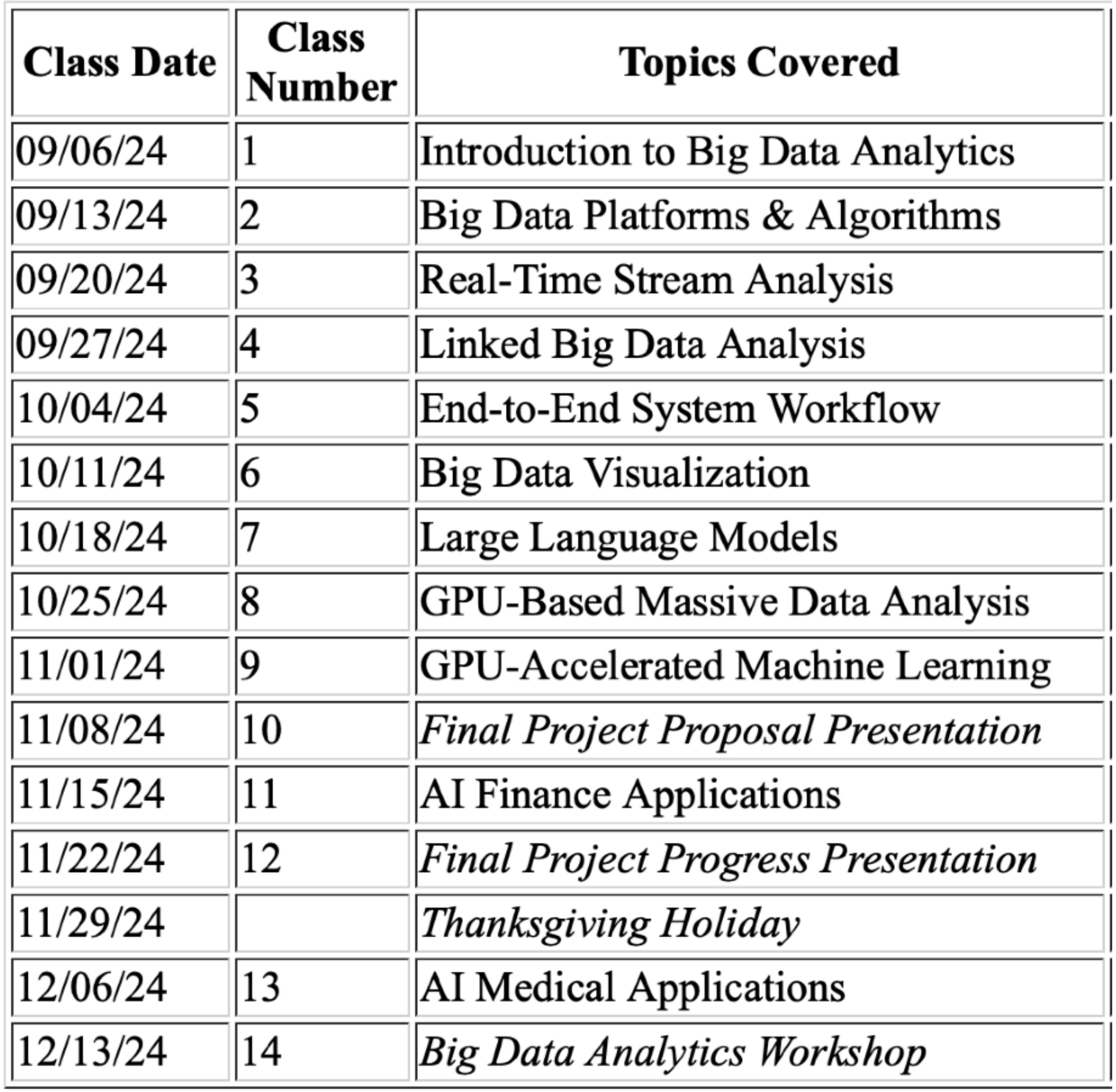
Memory
- Distributed Native Graph Database



https://www.graphen.ai/products/ardi.html

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

Course Outline



Class Date	Class Number	Topics Covered
09/06/24	1	Introduction to Big Data Analytics
09/13/24	2	Big Data Platforms & Algorithms
09/20/24	3	Real-Time Stream Analysis
09/27/24	4	Linked Big Data Analysis
10/04/24	5	End-to-End System Workflow
10/11/24	6	Big Data Visualization
10/18/24	7	Large Language Models
10/25/24	8	GPU-Based Massive Data Analysis
11/01/24	9	GPU-Accelerated Machine Learning
11/08/24	10	*Final Project Proposal Presentation*
11/15/24	11	AI Finance Applications
11/22/24	12	*Final Project Progress Presentation*
11/29/24		*Thanksgiving Holiday*
12/06/24	13	AI Medical Applications
12/13/24	14	*Big Data Analytics Workshop*

Course Main Thrust 1: Apache Hadoop and Big Data

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

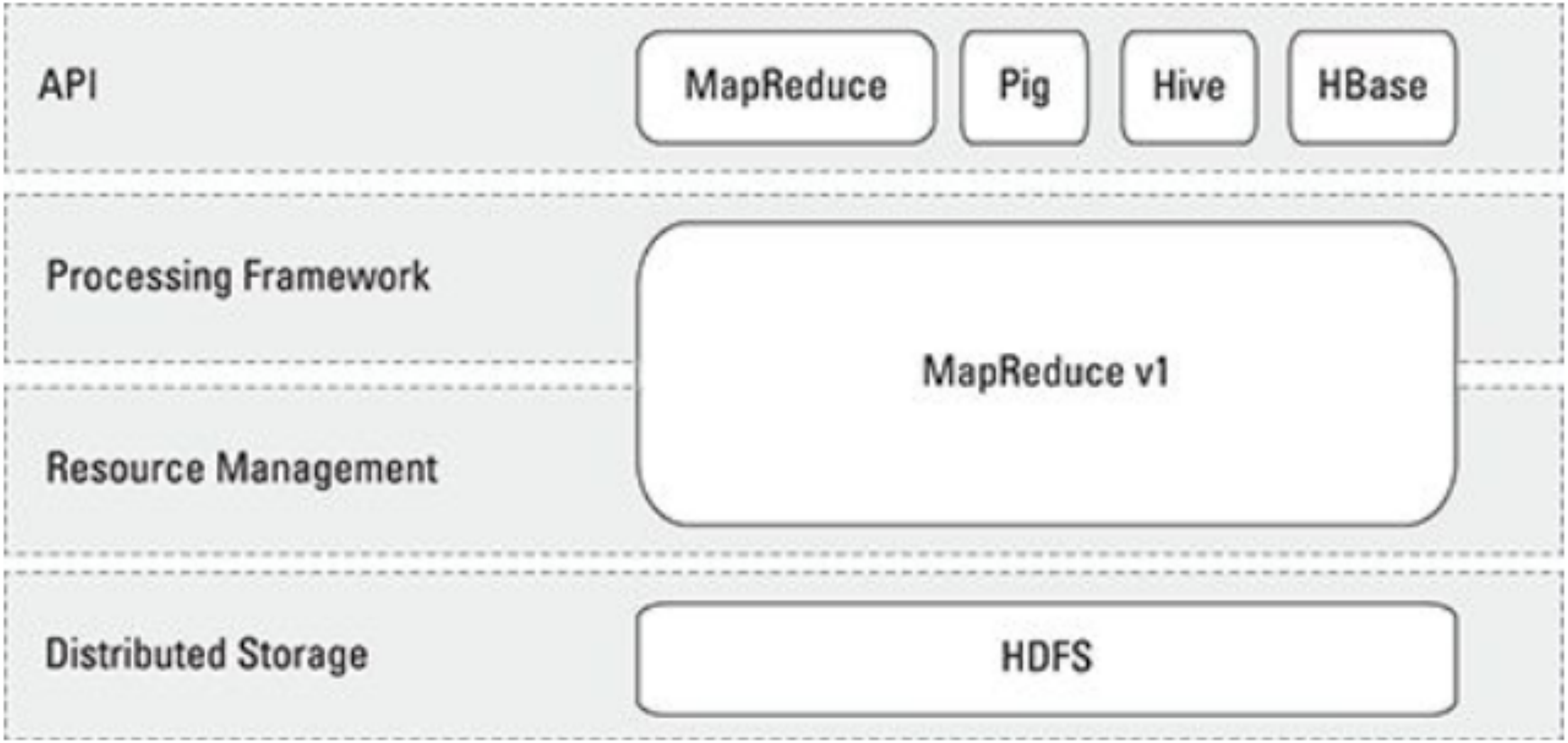
The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than 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:

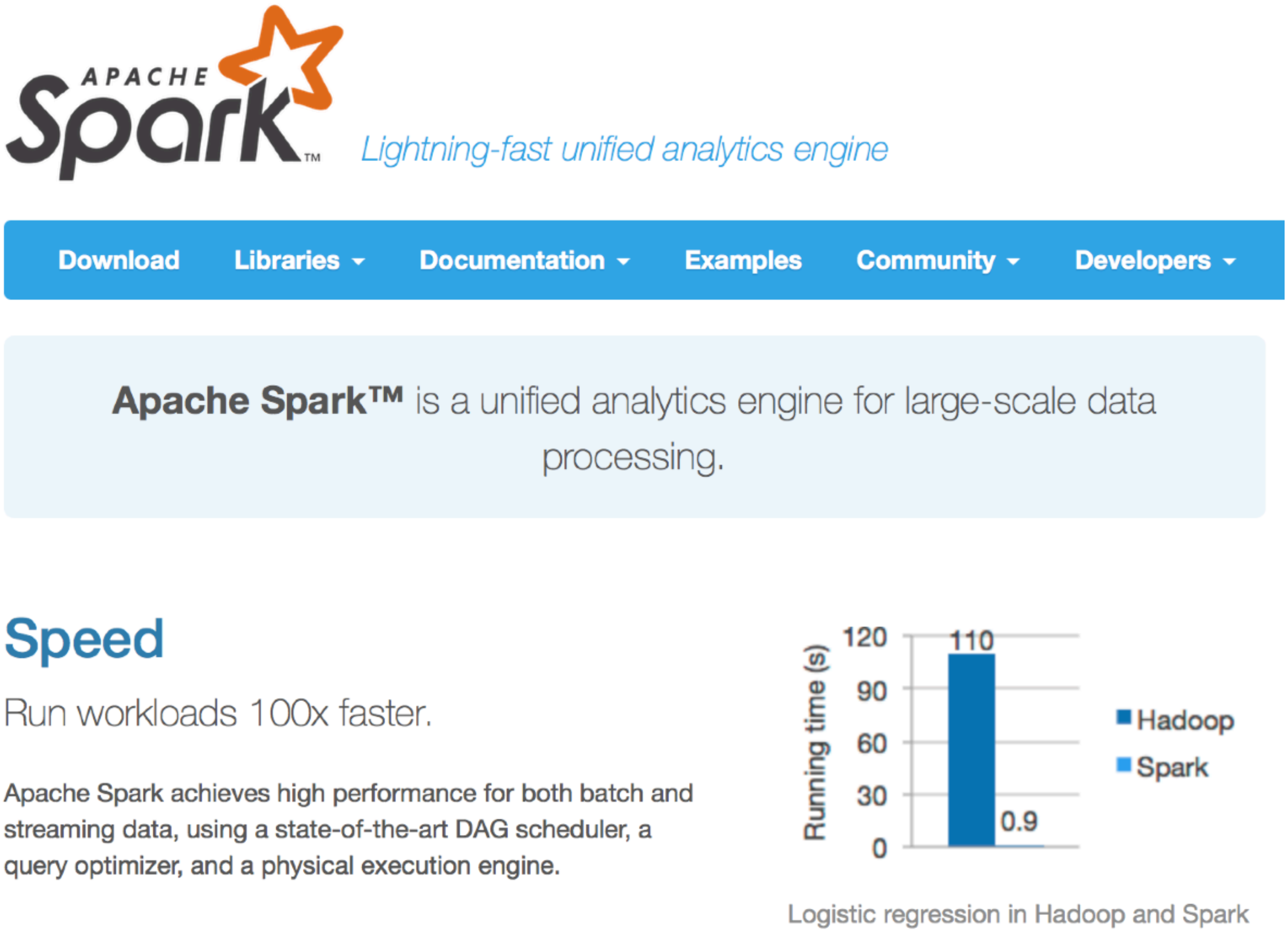
- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop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

http://hadoop.apache.org

Four distinctive layers of Hadoop



Course Main Thrust 2: Apache Spark and ML Library



Lightning-fast unified analytics engine

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

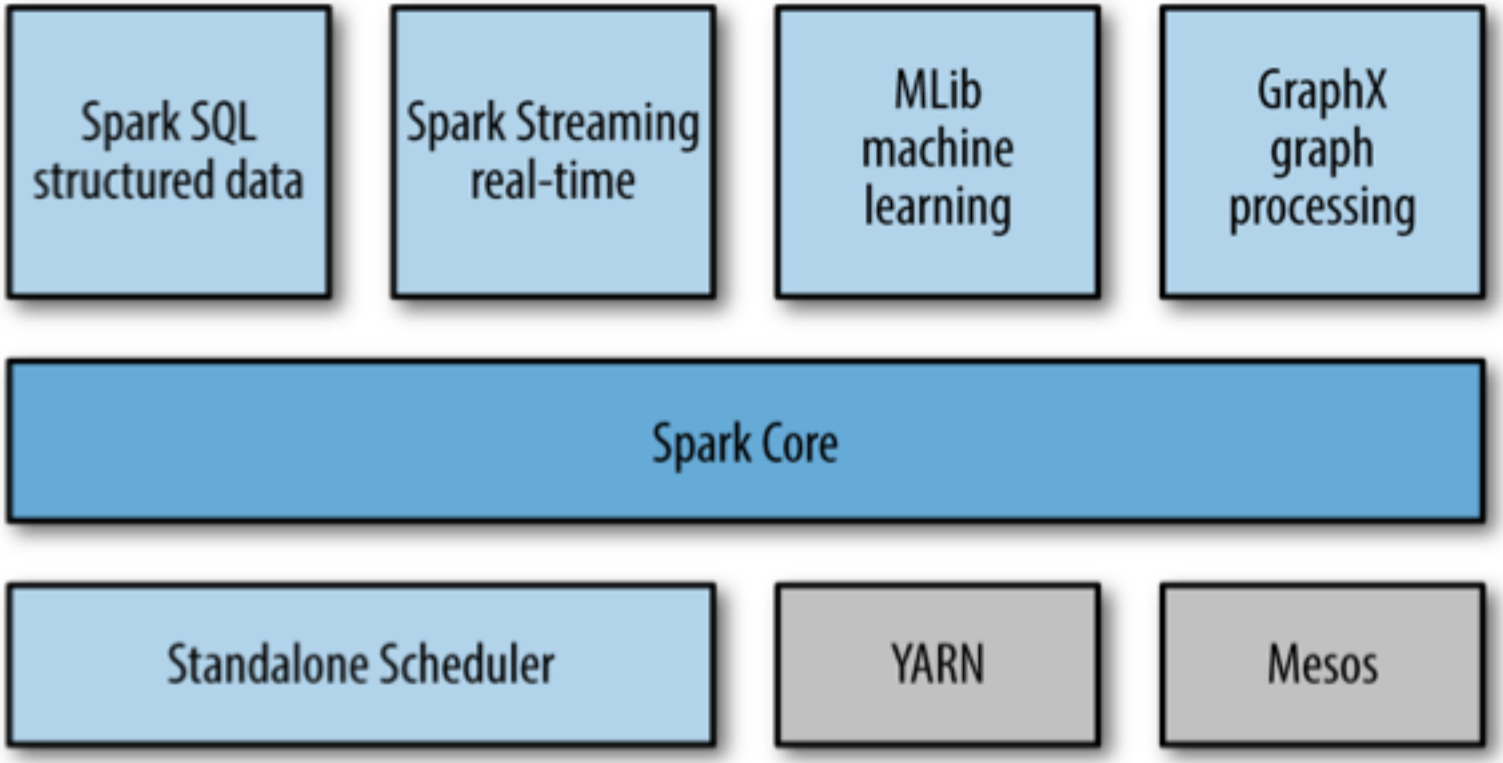
Speed

Run workloads 100x faster.

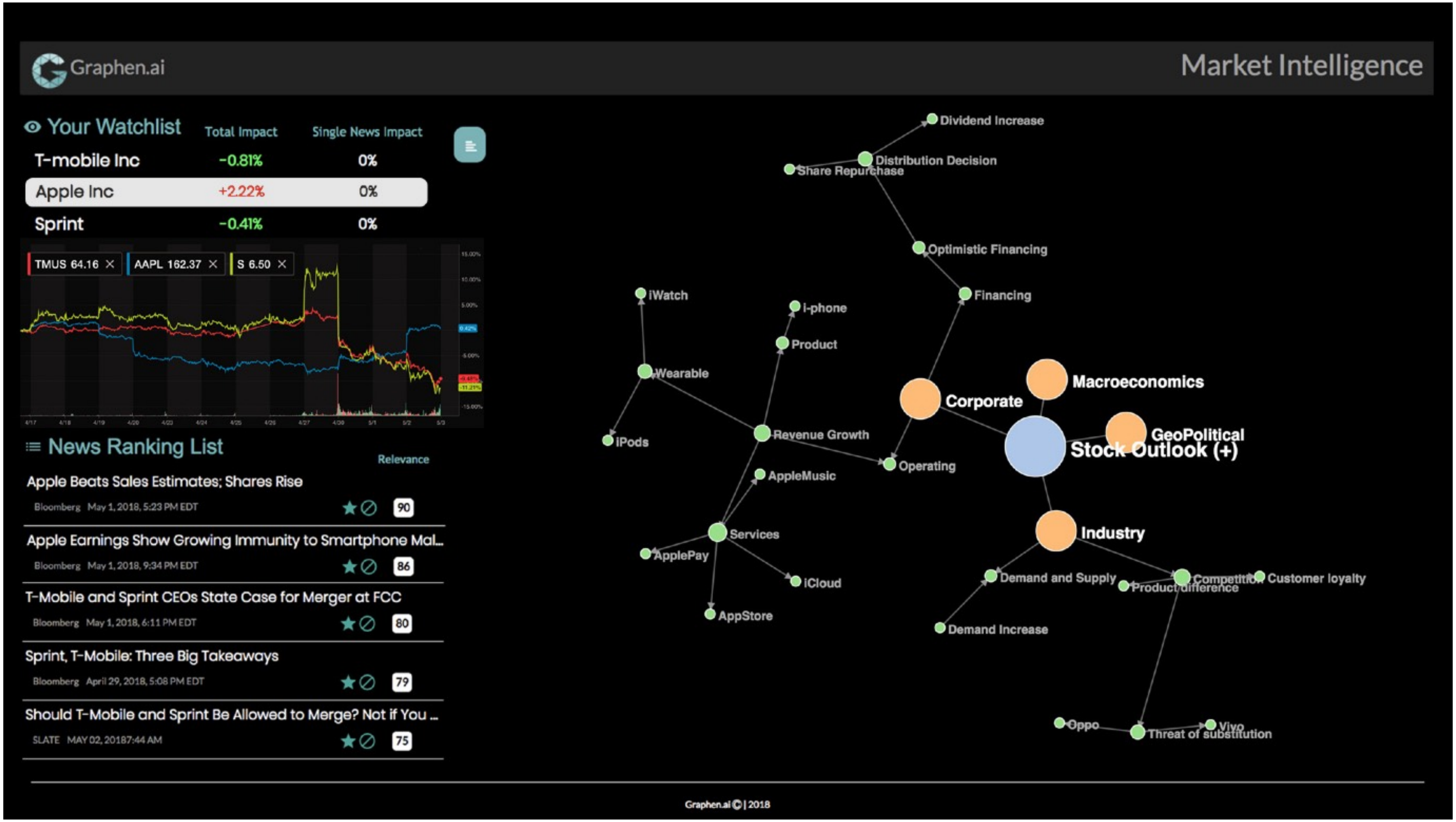
Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Logistic regression in Hadoop and Spark

Main Spark Stack



Course Main Thrust 3: Streaming and Linked Big Data Analytics



Graphen.ai

Market Intelligence

Your Watchlist

	Total Impact	Single News Impact
T-mobile Inc	-0.81%	0%
Apple Inc	+2.22%	0%
Sprint	-0.41%	0%

TMUS 64.16 × AAPL 162.37 × S 6.50 ×

News Ranking List

Relevance

Apple Beats Sales Estimates; Shares Rise
Bloomberg May 1, 2018, 5:23 PM EDT 90

Apple Earnings Show Growing Immunity to Smartphone Mal...
Bloomberg May 1, 2018, 9:34 PM EDT 86

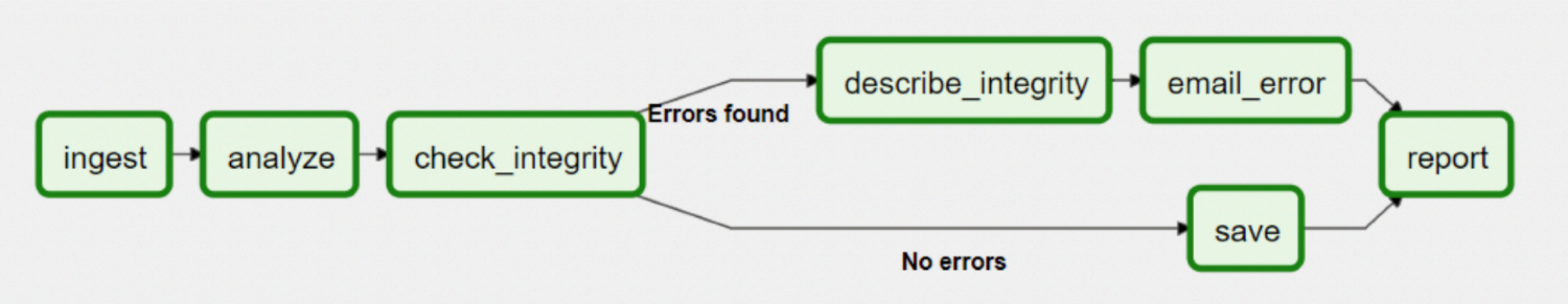
T-Mobile and Sprint CEOs State Case for Merger at FCC
Bloomberg May 1, 2018, 6:11 PM EDT 80

Sprint, T-Mobile: Three Big Takeaways
Bloomberg April 29, 2018, 5:08 PM EDT 79

Should T-Mobile and Sprint Be Allowed to Merge? Not if You ...
SLATE MAY 02, 20187:44 AM 75

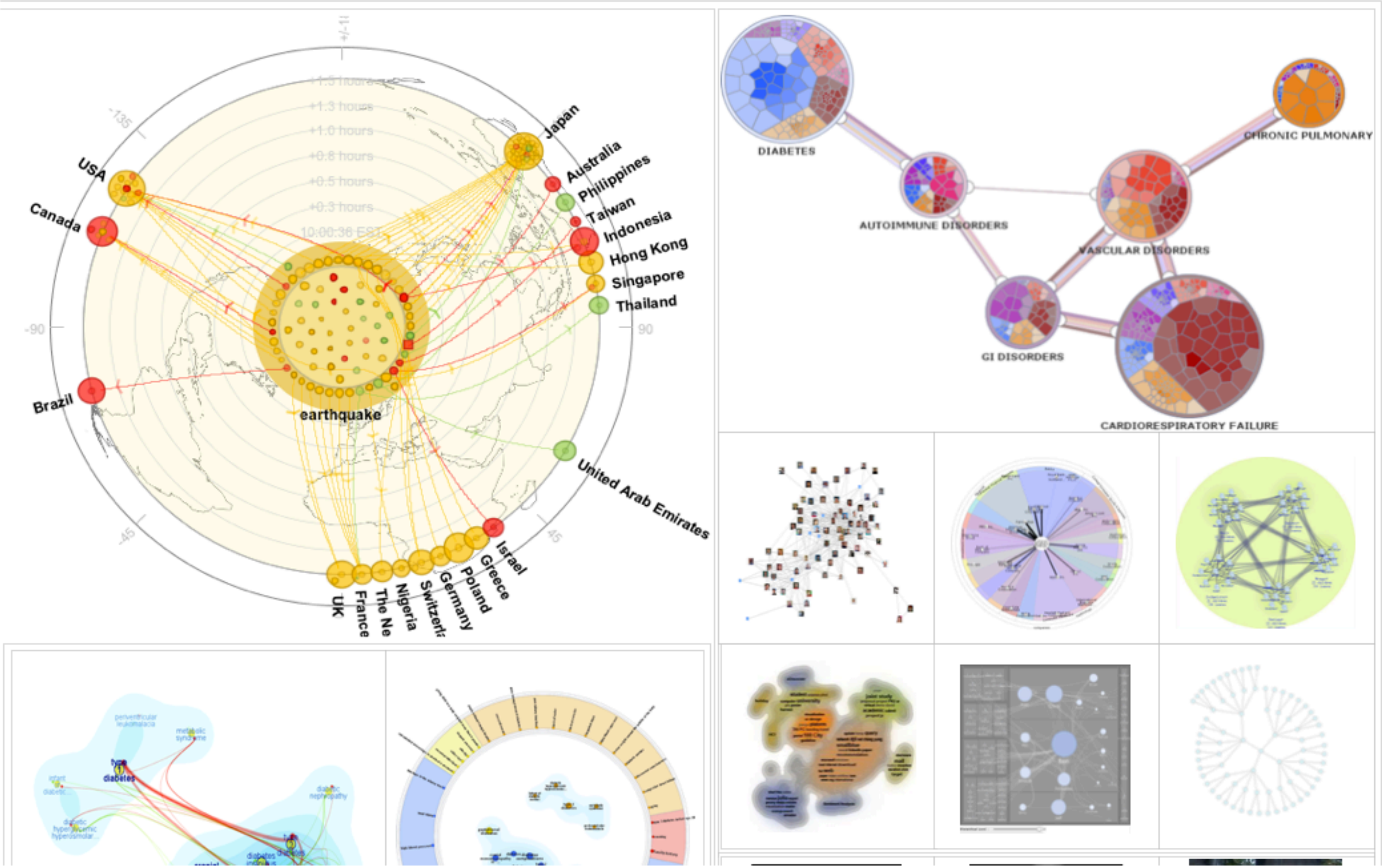
Graphen.ai © | 2018

Course Main Thrust 4: Workflow and Analytics Pipeline

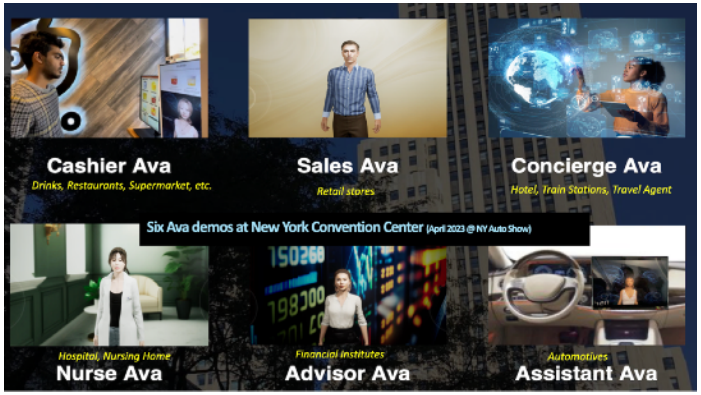


- A scheduler, which handles both triggering scheduled workflows, and submitting Tasks to the executor to run.
- An executor, which handles running tasks. In the default Airflow installation, this runs everything *inside* the scheduler, but most production-suitable executors actually push task execution out to *workers*.
- A *webserver*, which presents a handy user interface to inspect, trigger and debug the behaviour of DAGs and tasks.
- A folder of *DAG files*, read by the scheduler and executor (and any workers the executor has)
- A *metadata database*, used by the scheduler, executor and webserver to store state.

Course Main Thrust 5: Big Data Visualization

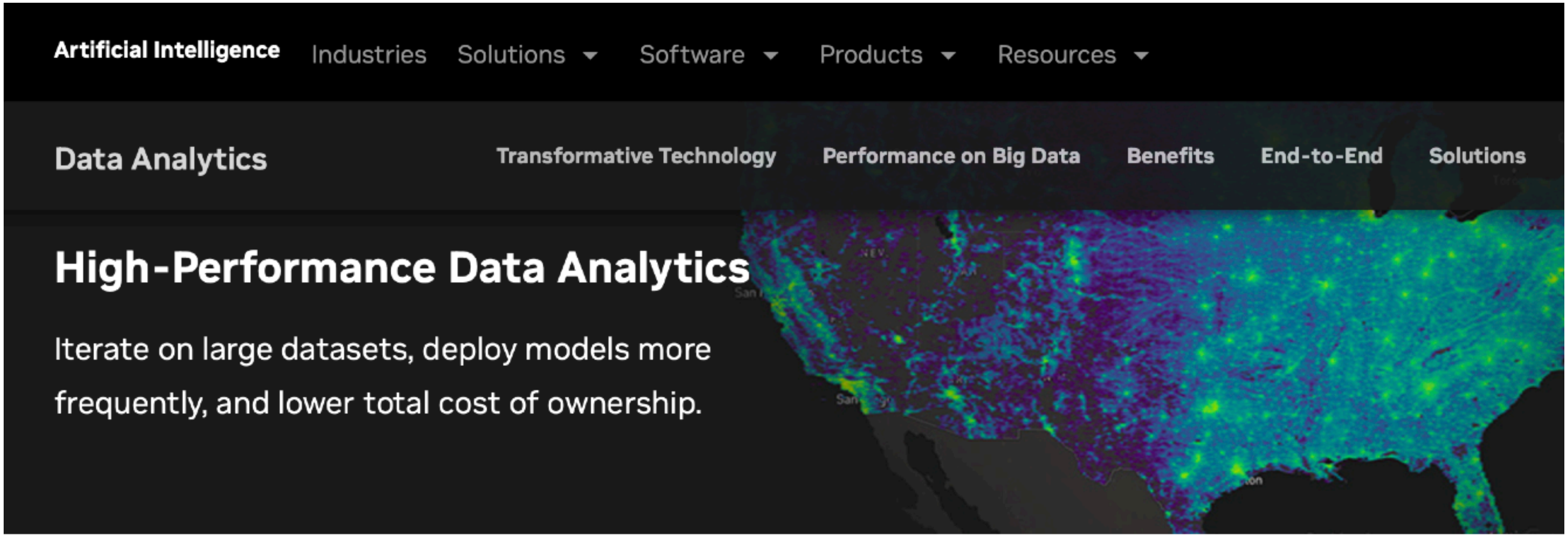


Course Main Thrust 6: Generative AI and Large Language Model


Cashier Ava
Drinks, Restaurants, Supermarket, etc.
Sales Ava
Retail stores
Concierge Ava
Hotel, Train Stations, Travel Agent
Six Ava demos at New York Convention Center (April 2023 @ NY Auto Show)
Hospital, Nursing Home
Nurse Ava
Financial Institutes
Advisor Ava
Automotives
Assistant Ava

Graphen Aiia — World's First AI Digital Human for Daily Life

Course Main Thrust 7: GPU-Based Big Data Analysis



High-Performance Data Analytics

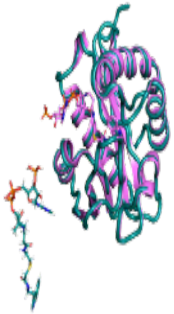
Iterate on large datasets, deploy models more frequently, and lower total cost of ownership.

Data analytics workflows have traditionally been slow and cumbersome, relying on CPU compute for data preparation, training, and deployment. Accelerated data science can dramatically boost the performance of end-to-end analytics workflows, speeding up value generation while reducing cost.

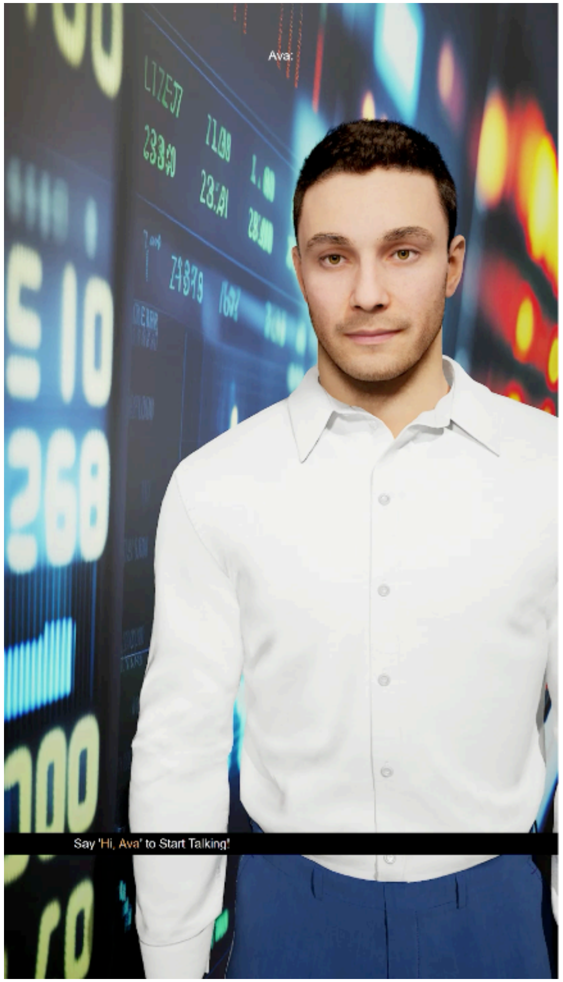
Course Main Thrust 8: Big Data AI Solutions

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

Graphen Small Mole Drug Dev ➔ 1/27 of the Time; 1/9000 of the Cost, comparing to traditional methods



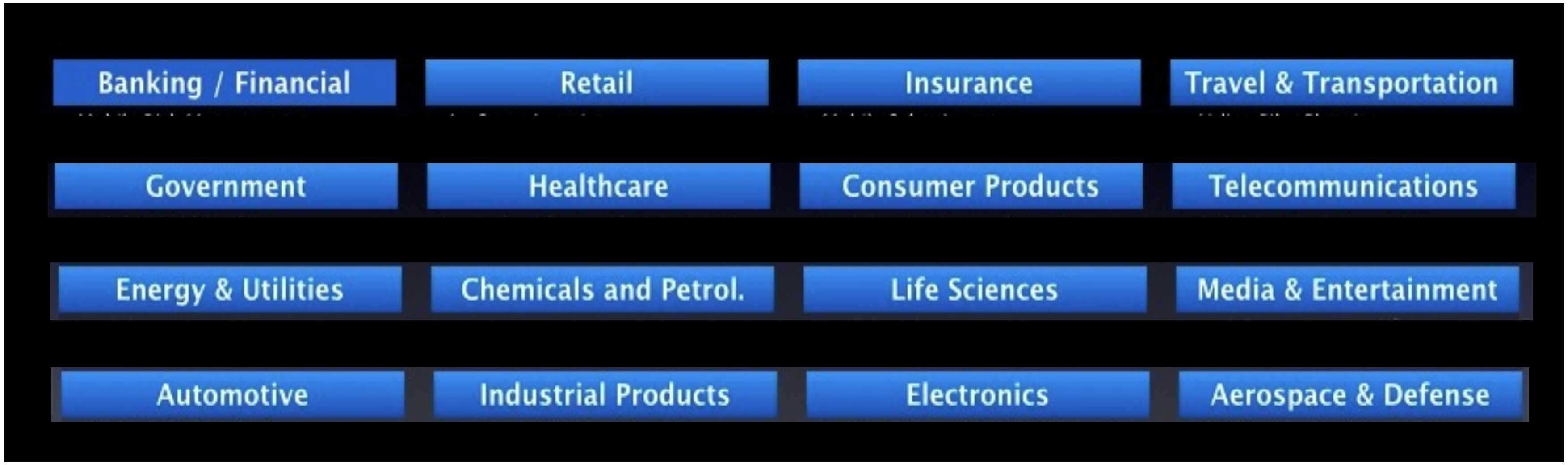
*"Tools from established companies like **Google**, startups like **Graphen**, and AI chipsets from vendors like **NVIDIA** and **Intel** will help accelerate the speed of drug discovery, development, and testing, allowing pharmaceutical companies and healthcare authorities to combat the pandemic." – ABI research, May 2020*



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.

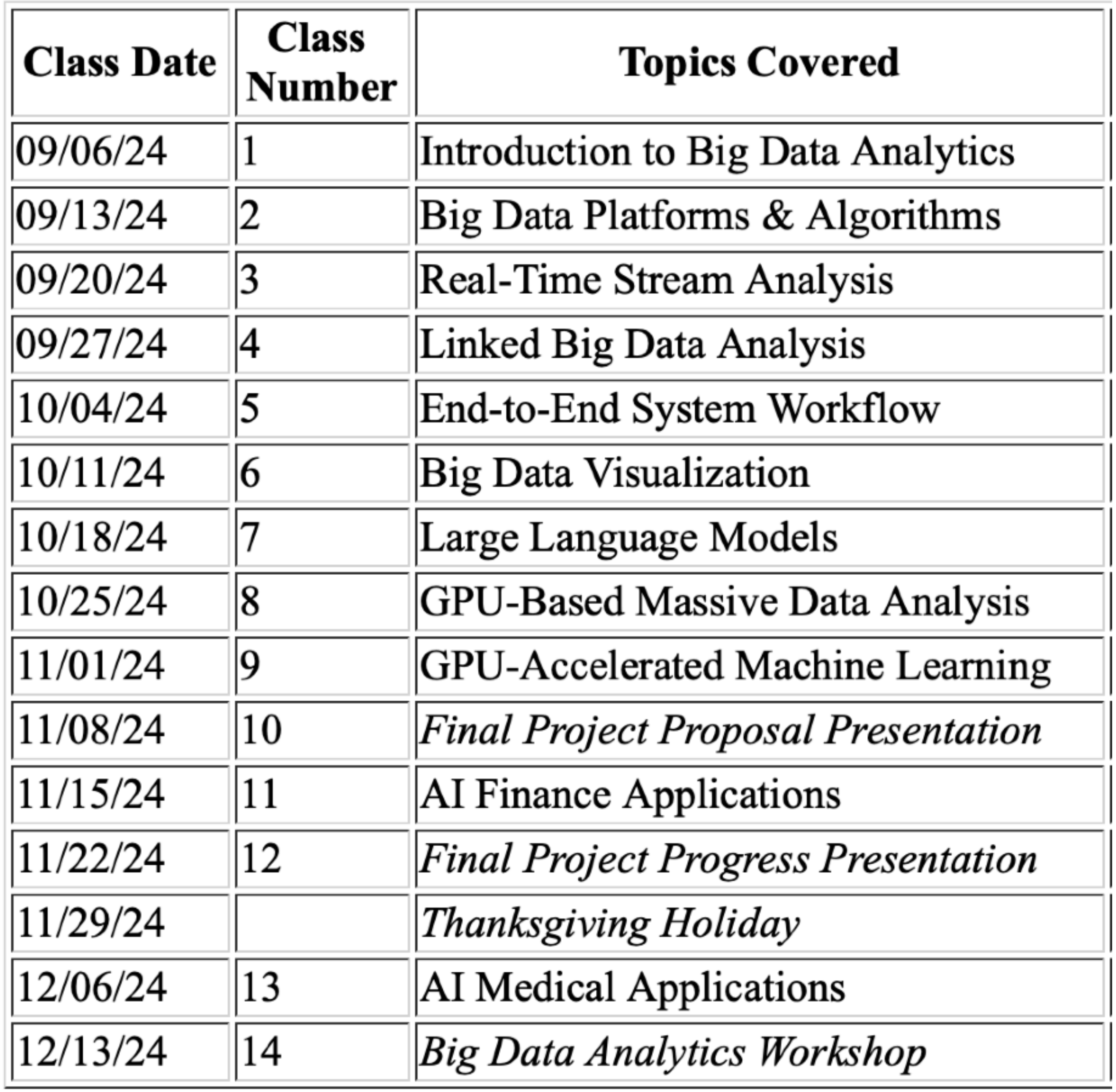
- **Sapphire Big Data Analytics Open Source Applications:** Create a Big Data open source toolsets for various industries (and disciplines)



Banking / Financial	Retail	Insurance	Travel & Transportation
Government	Healthcare	Consumer Products	Telecommunications
Energy & Utilities	Chemicals and Petrol.	Life Sciences	Media & Entertainment
Automotive	Industrial Products	Electronics	Aerospace & Defense

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

Course Outline



Class Date	Class Number	Topics Covered
09/06/24	1	Introduction to Big Data Analytics
09/13/24	2	Big Data Platforms & Algorithms
09/20/24	3	Real-Time Stream Analysis
09/27/24	4	Linked Big Data Analysis
10/04/24	5	End-to-End System Workflow
10/11/24	6	Big Data Visualization
10/18/24	7	Large Language Models
10/25/24	8	GPU-Based Massive Data Analysis
11/01/24	9	GPU-Accelerated Machine Learning
11/08/24	10	*Final Project Proposal Presentation*
11/15/24	11	AI Finance Applications
11/22/24	12	*Final Project Progress Presentation*
11/29/24		*Thanksgiving Holiday*
12/06/24	13	AI Medical Applications
12/13/24	14	*Big Data Analytics Workshop*

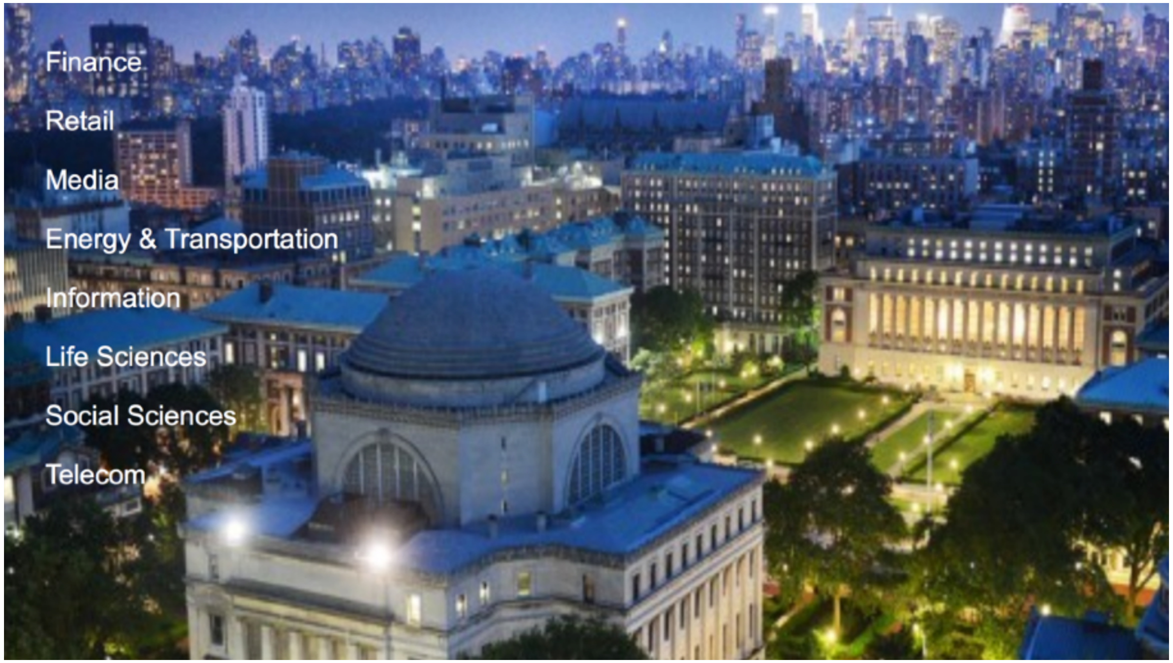
Course Information

- Website:

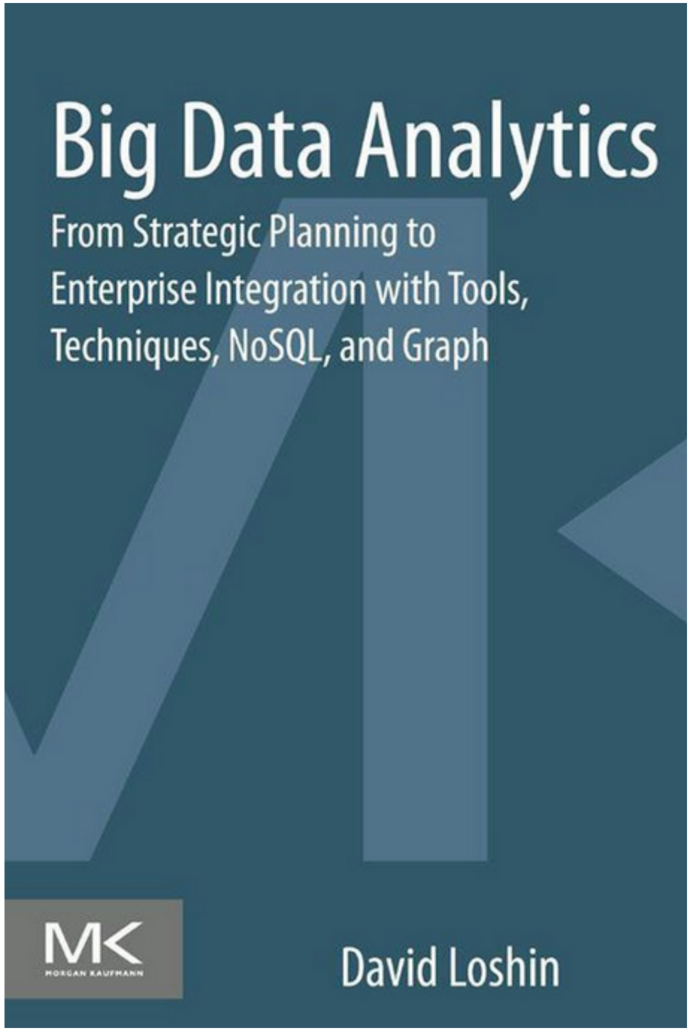
  http://www.ee.columbia.edu/~cylin/course/bigdata/

- Textbook:

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



Reading Reference for Lecture 1


Big Data Analytics
From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph
MK
MORGAN KAUFMANN
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

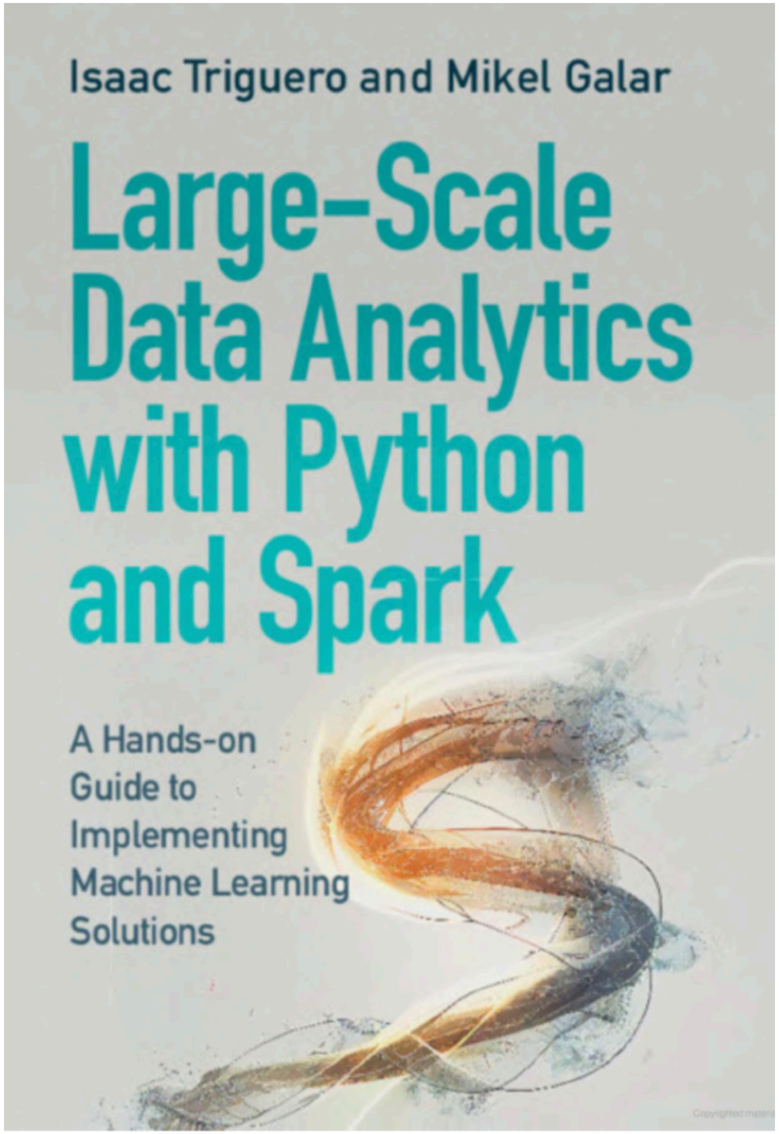
Chapter 8: Developing Big Data Applications

Chapter 9: NoSQL Data Management for Big Data

Chapter 10: Using Graph Analytics for Big Data

Chapter 11: Developing the Big Data Roadmap

Highly Recommended Reference Book for Lectures 1-4


Isaac Triguero and Mikel Galar
Large-Scale Data Analytics with Python and Spark
A Hands-on Guide to Implementing Machine Learning Solutions

Part I: Understanding and Dealing with Big Data

Chapter 1: Introduction

Chapter 2: MapReduce

Part II: Big Data Frameworks

Chapter 3: Hadoop

Chapter 4: Spark

Chapter 5: Spark SQL and DataFrames

Part III: Machine Learning for Big Data

Chapter 6: Machine Learning with Spark

Chapter 7: Machine Learning for Big Data

Chapter 8: Implementing Classic Methods: k-Means and Linear Regression

Chapter 9: Advanced Examples – Semi-Supervised, Ensembles, Deep Learning Model Deployment

Course Grading

- **5 Homeworks: 50%**

  -- **Individual work**; Language Requirement: Python, JavaScript; Get familiar with Linux

  -- **Report (including description of the work, discussions, experiments, etc) and source code**

  - **HW #0: Big Data Environment Setup and Testing**
  - **HW #1: Analytic Algorithms and System Monitoring**
  - **HW #2: Graph Analysis and Analytics Pipeline**
  - **HW #3: Big Data Visualization and LLM**
  - **HW #4: GPU-Based Big Data Analysis**

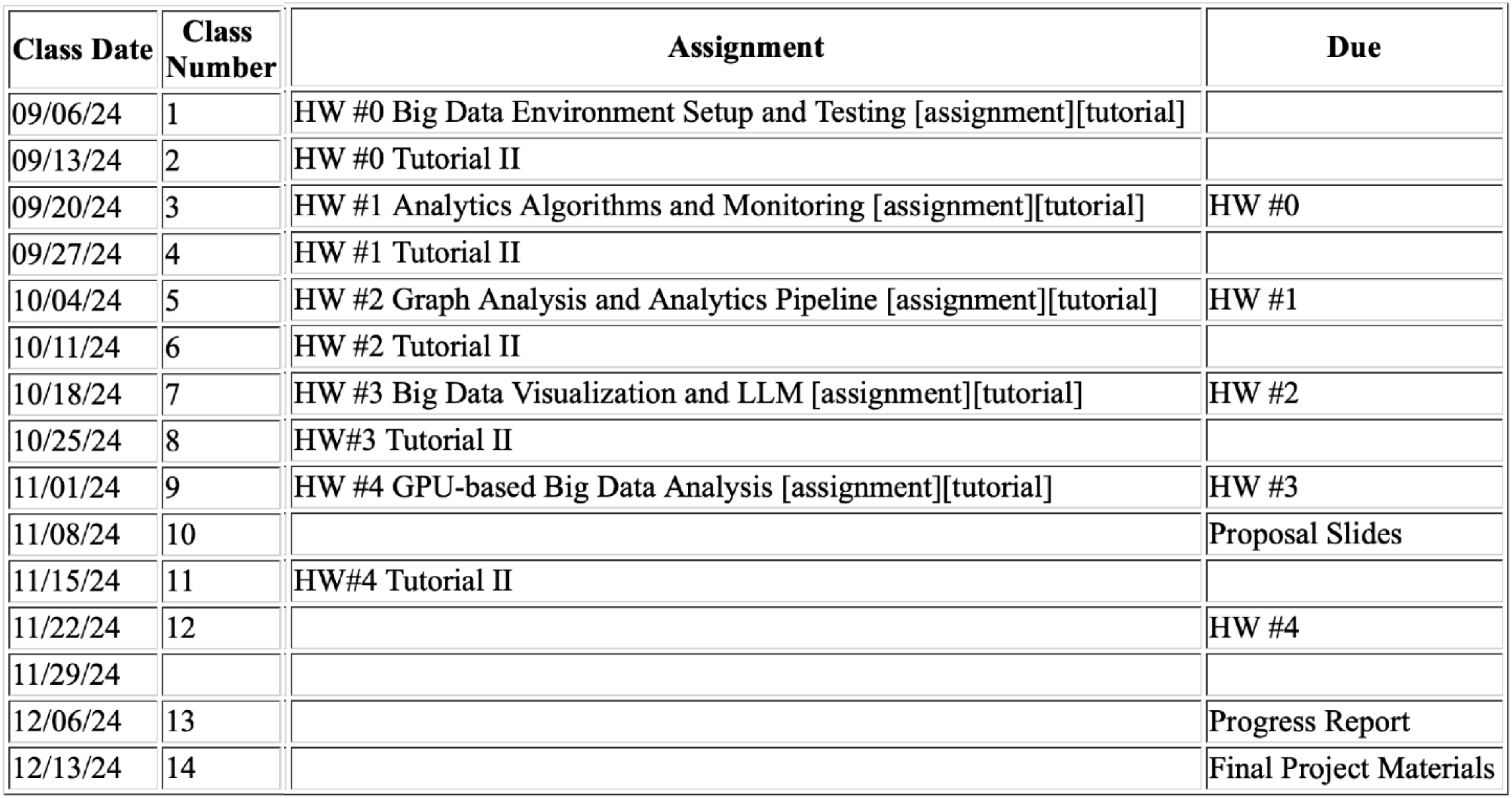
Course Grading

- **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)
  - **Progress Presentation** (slides — short presentation in the class)
  - **Progress Report** (report)
  - **Final Report** (paper, up to 10 pages)
  - **Workshop Presentation** (Oral and Demo)
  - **Open Source Codes**
  - **Video Presentation** (on YouTube)

Assignments and Submissions



Class Date	Class Number	Assignment	Due
09/06/24	1	HW #0 Big Data Environment Setup and Testing [assignment][tutorial]	
09/13/24	2	HW #0 Tutorial II	
09/20/24	3	HW #1 Analytics Algorithms and Monitoring [assignment][tutorial]	HW #0
09/27/24	4	HW #1 Tutorial II	
10/04/24	5	HW #2 Graph Analysis and Analytics Pipeline [assignment][tutorial]	HW #1
10/11/24	6	HW #2 Tutorial II	
10/18/24	7	HW #3 Big Data Visualization and LLM [assignment][tutorial]	HW #2
10/25/24	8	HW#3 Tutorial II	
11/01/24	9	HW #4 GPU-based Big Data Analysis [assignment][tutorial]	HW #3
11/08/24	10		Proposal Slides
11/15/24	11	HW#4 Tutorial II	
11/22/24	12		HW #4
11/29/24			
12/06/24	13		Progress Report
12/13/24	14		Final Project Materials

Other Issues

- Professor Lin:
  - Office Hours:

    By appointment

  - Contact: c.lin@columbia.edu

- TA:
  - Apurva Patel (amp2365): Monday 10am-12pm (onsite) and Wednesday 12pm-2pm (online)
  - Linyang He (lh3288): Tuesday 3pm-5pm (onsite) and Thursday 1pm-3pm (online)

  - Location: EE Student Lounge (next to the EE office, Mudd 13th Floor)

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

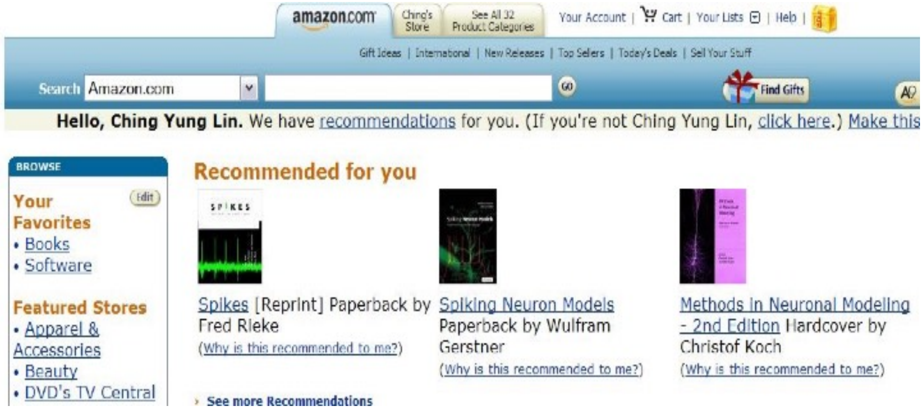
Data Warehouse Augmentation
Integrate big data and data warehouse capabilities to increase operational efficiency

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



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

Gift Ideas | International | New Releases | Top Sellers | Today's Deals | Sell Your Stuff

Search Amazon.com GO 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 edit
- Books
- Software

Featured Stores
- Apparel & Accessories
- Beauty
- DVD's TV Central

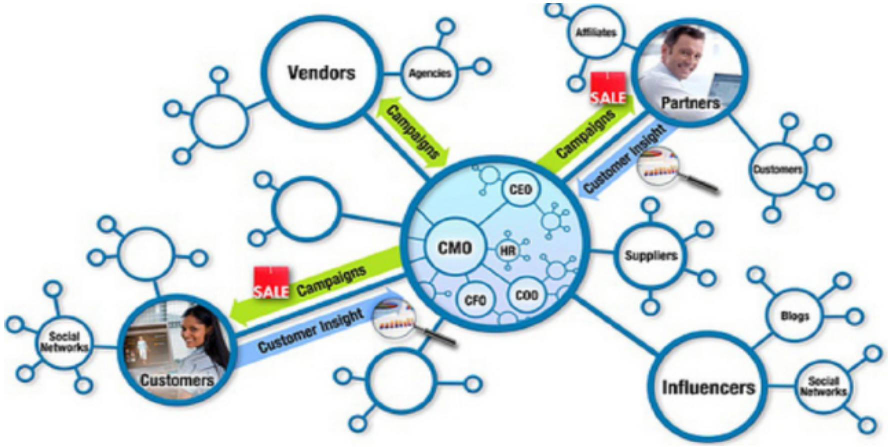
Recommended for you

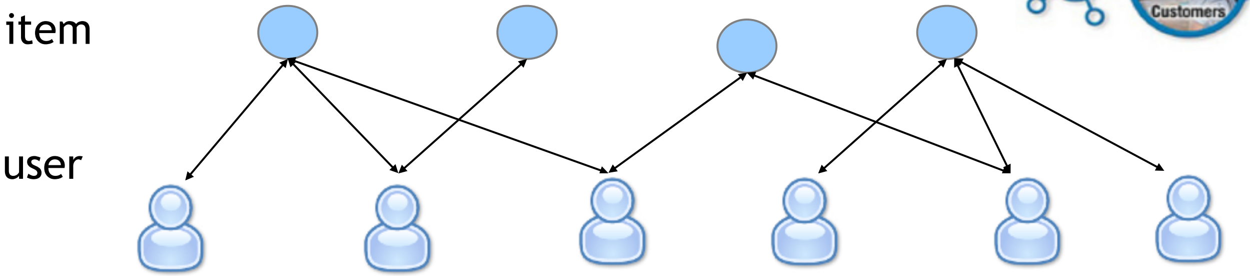
Spikes [Reprint] Paperback by Fred Rieke (Why is this recommended to me?)

Spiking Neuron Models Paperback by Wulfram Gerstner (Why is this recommended to me?)

Methods in Neuronal Modeling - 2nd Edition Hardcover by Christof Koch (Why is this recommended to me?)

See more Recommendations

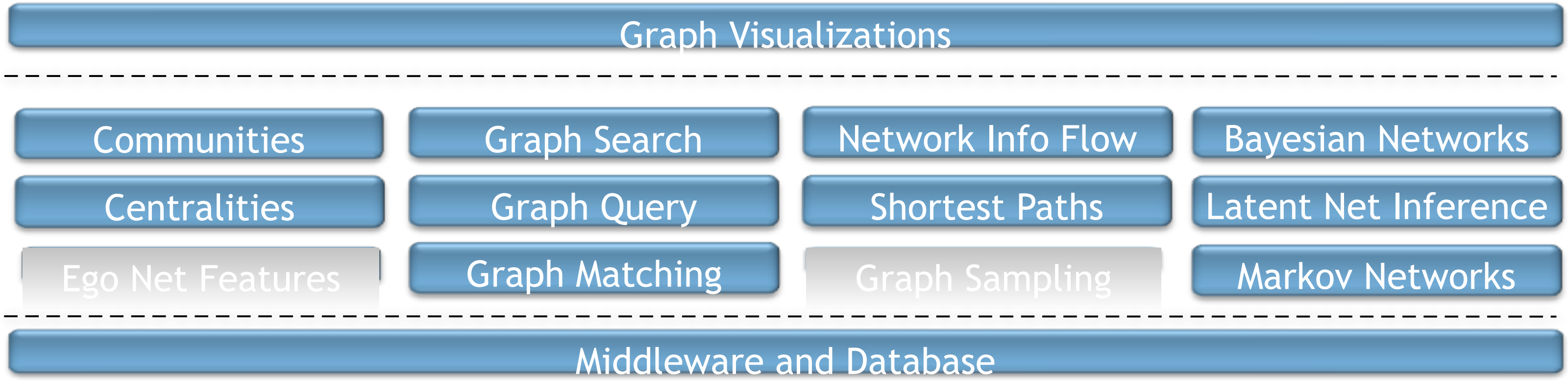




Enhancing:

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

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

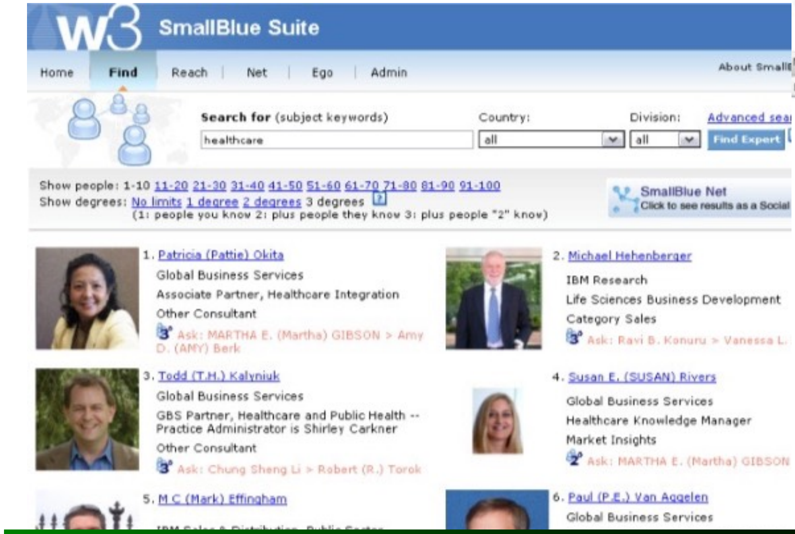
1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

200,000 people's consulting project & earning d

Shortest Paths

Centralities

Graph Search



W3 SmallBlue Suite

Home | Find | Reach | Net | Ego | Admin — About SmallB

Search for (subject keywords): healthcare — Country: all — Division: all — Advanced sea — Find Expert

Show people: 1-10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81-90 91-100
Show degrees: No limits 1 degree 2 degrees 3 degrees
(1: people you know 2: plus people they know 3: plus people "2" know)

SmallBlue Net
Click to see results as a Social

1. Patricia (Pattie) Okita
Global Business Services
Associate Partner, Healthcare Integration
Other Consultant
Ask: MARTHA E. (Martha) GIBSON > Amy D. (AMY) Berk

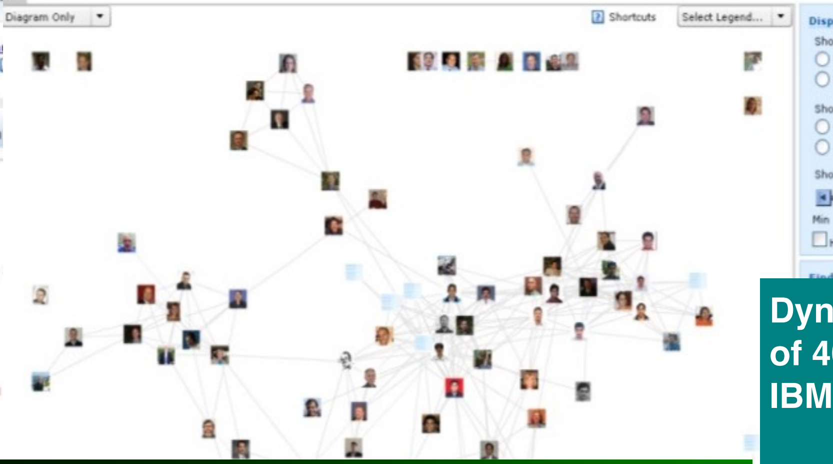
2. Michael Hehenberger
IBM Research
Life Sciences Business Development
Category Sales
Ask: Ravi B. Konuru > Vanessa L.

3. Todd (T.H.) Kalyniuk
Global Business Services
GBS Partner, Healthcare and Public Health -- Practice Administrator is Shirley Carkner
Other Consultant
Ask: Chung Sheng Li > Robert (R.) Torok

4. Susan E. (SUSAN) Rivers
Global Business Services
Healthcare Knowledge Manager
Market Insights
Ask: MARTHA E. (Martha) GIBSON

5. M C (Mark) Effingham

6. Paul (P.E.) Van Aggelen
Global Business Services



W3 SmallBlue Suite

Home | Find | Reach | Net | Ego | Admin

Show network for (subject keywords): healthcare

Diagram Only — Shortcuts — Select Legend...

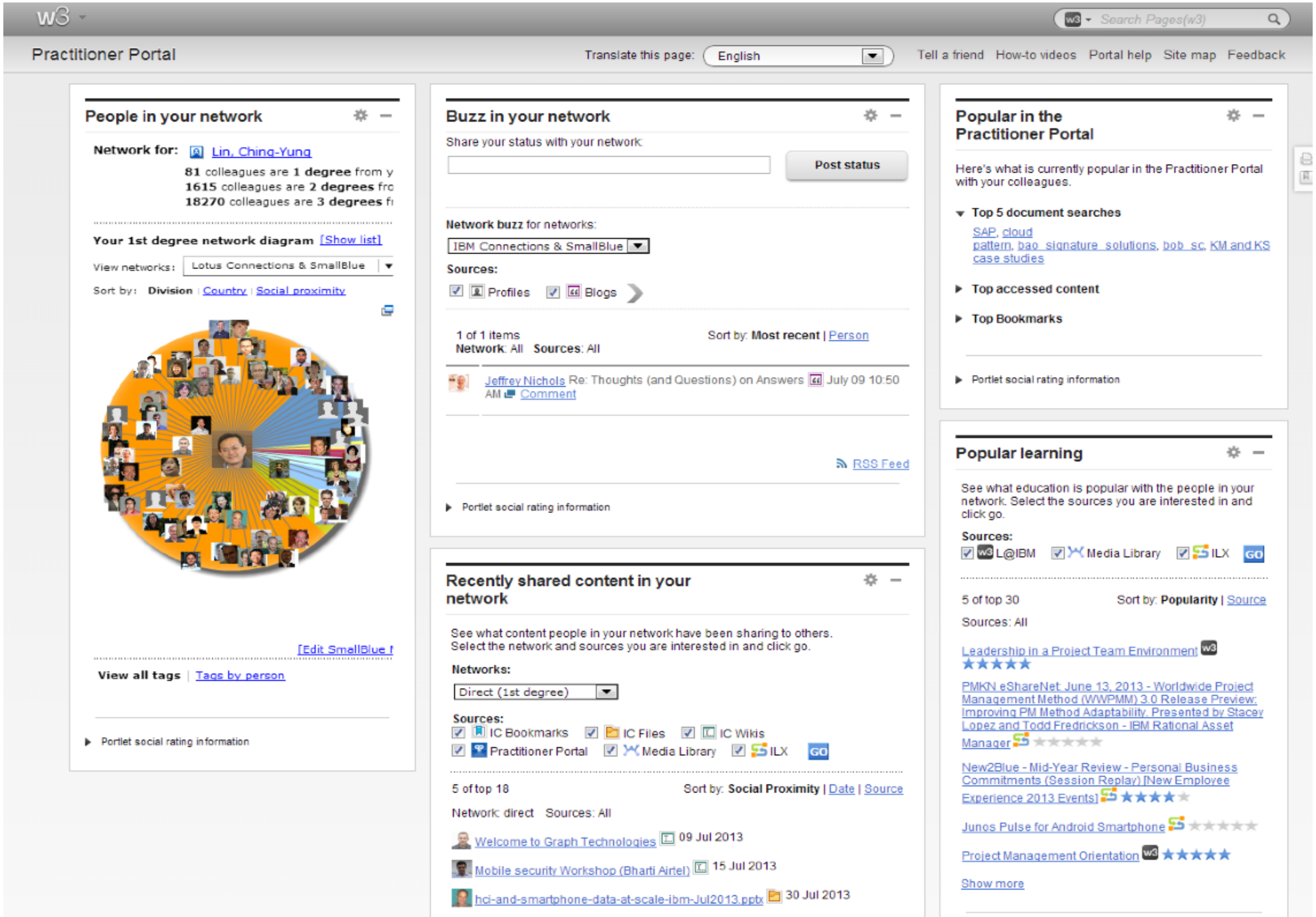
Display Settings
Show node information: ○ Names ○ Ranking ○ Statistics ● None
Show node icon: ○ Business Unit ○ Country ○ None ● Picture
Show people by rank: Min 100 Max
☐ Hide Isolates — Redraw

- 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.

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 y
1615 colleagues are **2 degrees** fro
18270 colleagues are **3 degrees** fr

Your 1st degree network diagram [Show list]

View networks: Lotus Connections & SmallBlue

Sort by: **Division** | Country | Social proximity

[Edit SmallBlue I

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

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

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 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

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 ★★★★★

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 ☆☆☆☆☆

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

Junos Pulse for Android Smartphone ☆☆☆☆☆

Project Management Orientation ★★★★★

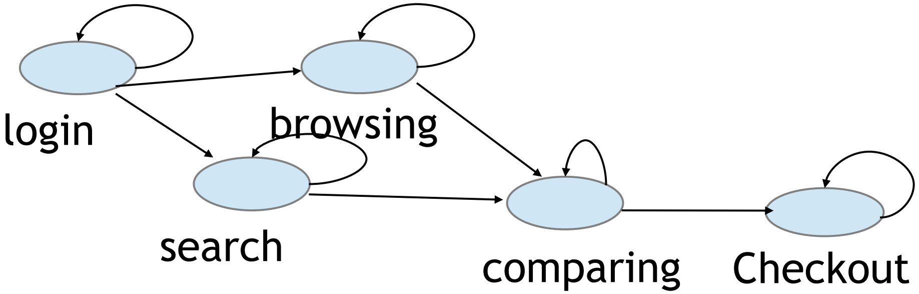
Show more

Use Case 3: Customer Behavior Sequence Analytics

Markov Network

Latent Network

Bayesian Network

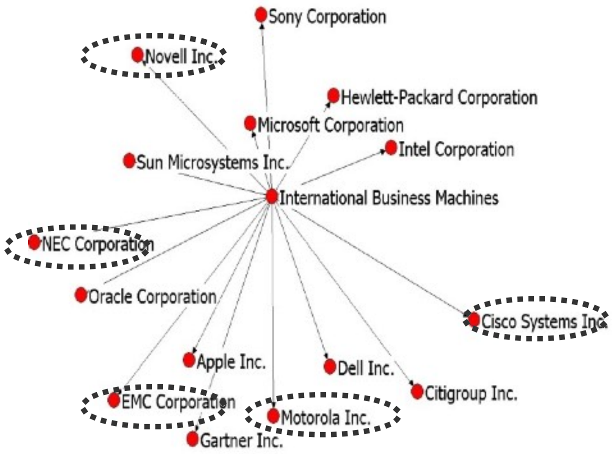


- Behavior Pattern Detection
- Help Needed Detection

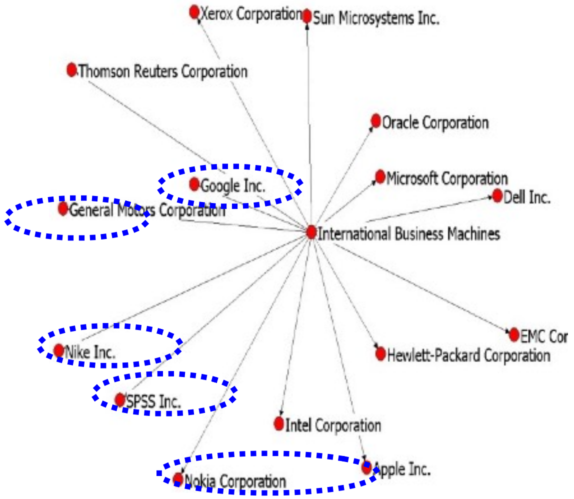
Use Case 4: Graph Analytics for Financial Analysis

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

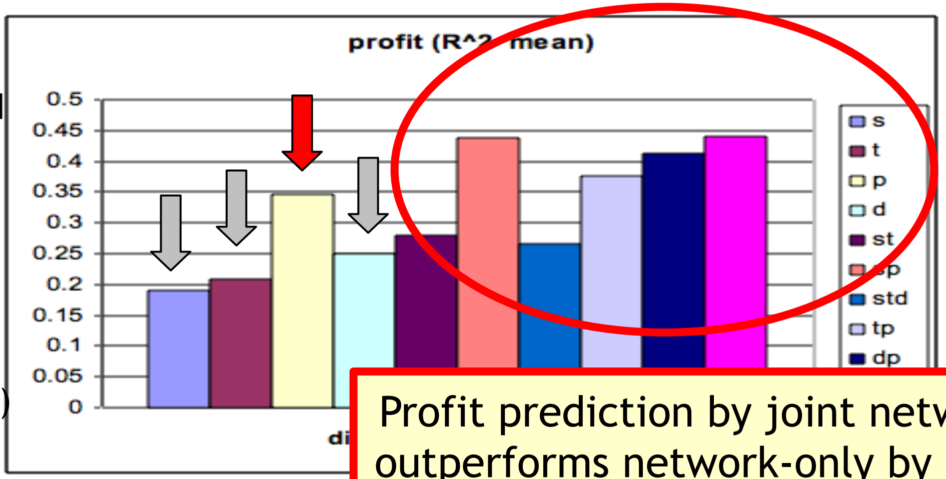


- Data Source:
  - Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

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

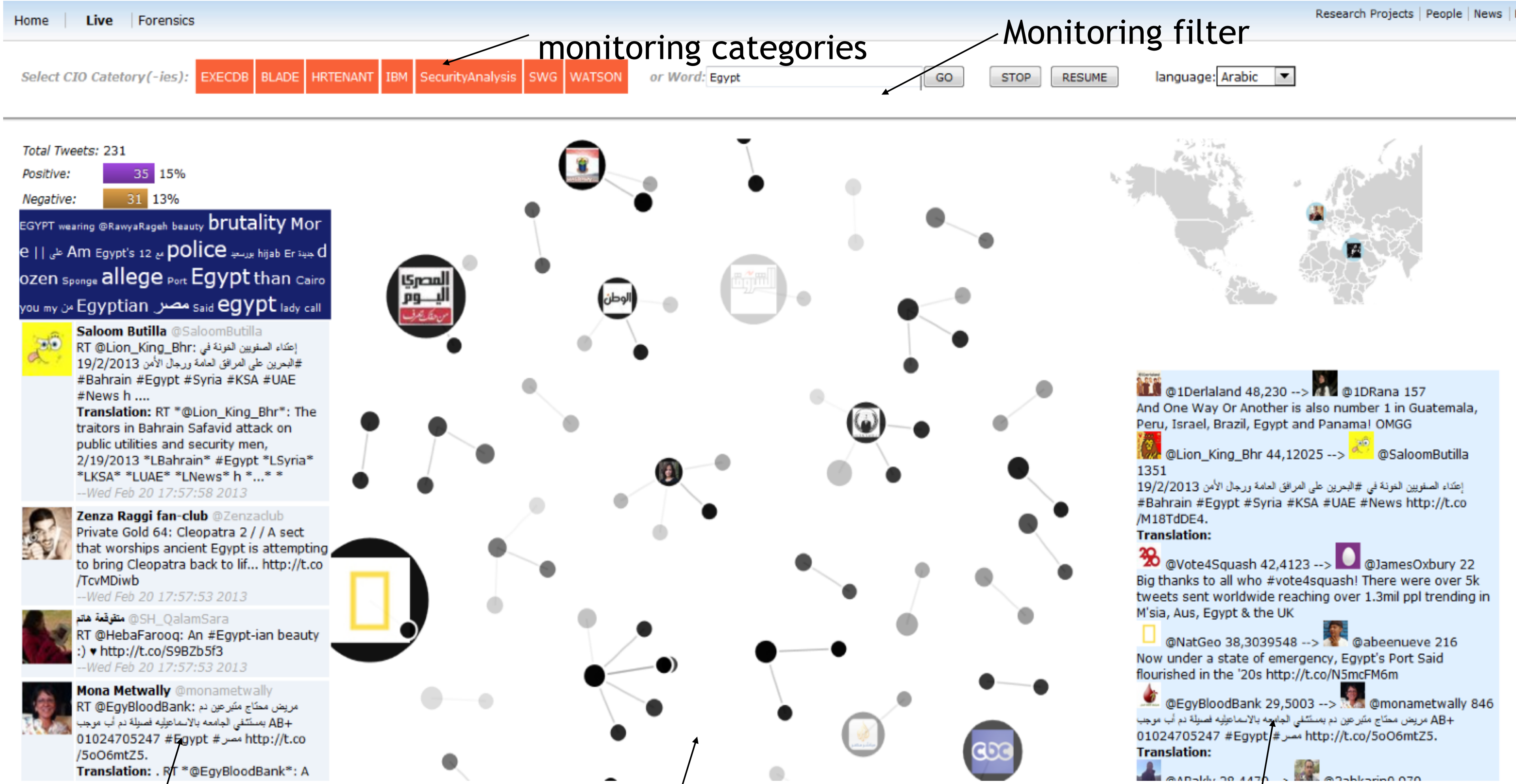
Model: Support Vector Regression (RBF kernel)



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
News

Select CIO Catetory(-ies): EXECDB BLADE HRTENANT 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 || على Am Egypt's 12 مع police يرسم hijab Er دبية d
ozen Sponge allege Port Egypt than Cairo
you my من Egyptian مصر Said egypt lady call

Saloom Butilla @SaloomButilla
RT @Lion_King_Bhr: إعتداء الصفويين الخونة في #البحرين على المرافق العامة ورجال الأمن 19/2/2013 #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 @Zenzaclub
Private Gold 64: Cleopatra 2 / / A sect that worships ancient Egypt is attempting to bring Cleopatra back to lif... http://t.co/TcvMDiwb
--Wed Feb 20 17:57:53 2013

متقوقعة هاتم @SH_QalamSara
RT @HebaFarooq: 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+ بمستشفي الجامعه بالاسماعيليه فصيلة دم أب موجب 01024705247 #Egypt #مصر http://t.co/5oO6mtZ5.
Translation: . RT *@EgyBloodBank*: A

@1Derlaland 48,230 --> @1DRana 157
And One Way Or Another is also number 1 in Guatemala, Peru, Israel, Brazil, Egypt and Panama! OMGG

@Lion_King_Bhr 44,12025 --> @SaloomButilla 1351
إعتداء الصفويين الخونة في #البحرين على المرافق العامة ورجال الأمن 19/2/2013 #Bahrain #Egypt #Syria #KSA #UAE #News http://t.co/M18TdDE4.
Translation:

@Vote4Squash 42,4123 --> @JamesOxbury 22
Big thanks to all who #vote4squash! There were over 5k tweets sent worldwide reaching over 1.3mil ppl trending in M'sia, Aus, Egypt & the UK

@NatGeo 38,3039548 --> @abeenueve 216
Now under a state of emergency, Egypt's Port Said flourished in the '20s http://t.co/N5mcFM6m

@EgyBloodBank 29,5003 --> @monametwally 846
مريض محتاج متبرعين دم بمستشفي الجامعه بالاسماعيليه فصيلة دم أب موجب AB+ 01024705247 #Egypt #مصر http://t.co/5oO6mtZ5.
Translation:

@ABaldy 28,4470 --> @2abkarin0 070

monitoring categories

Monitoring filter

Live Tweets, Sentiment, Keywords

Dynamic Graph

Zooming / Panning

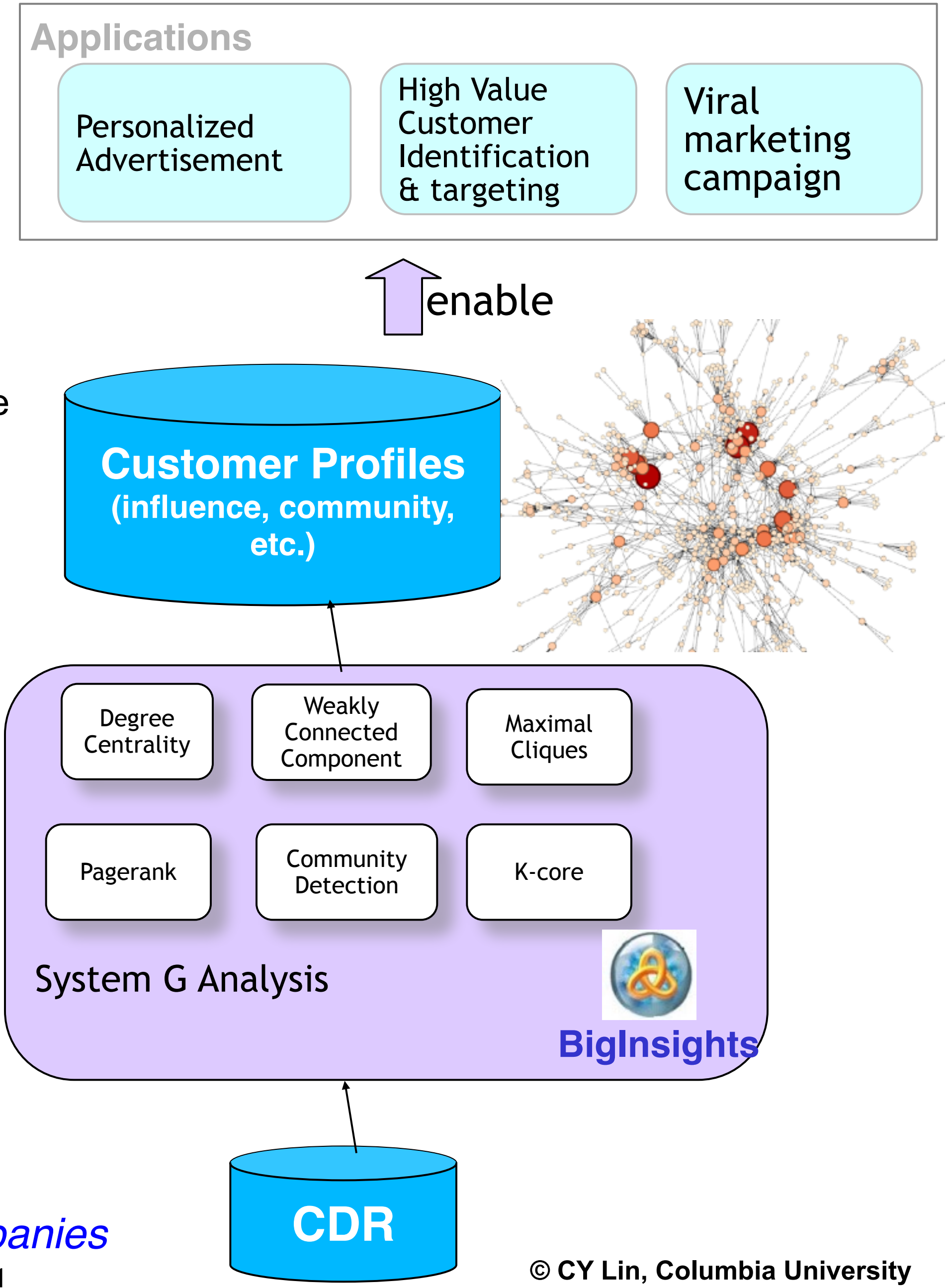
Real-Time Translation, Loca

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



Applications

Personalized Advertisement

High Value Customer Identification & targeting

Viral marketing campaign

enable

Customer Profiles
(influence, community, etc.)

Degree Centrality

Weakly Connected Component

Maximal Cliques

Pagerank

Community Detection

K-core

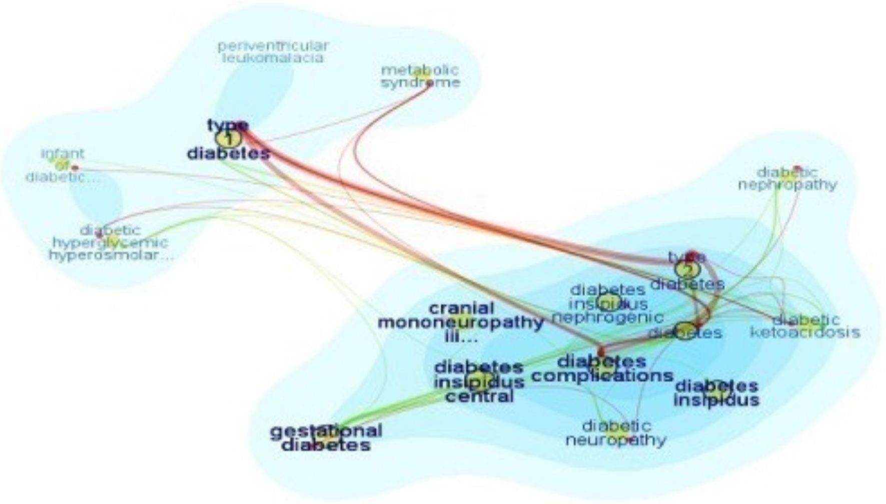
System G Analysis

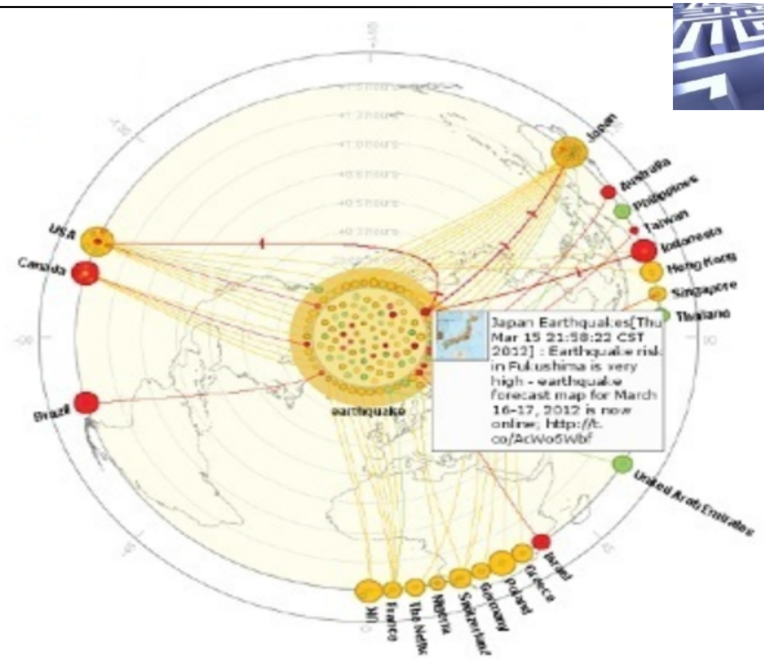
BigInsights

CDR

PoCs with Chinese and Indian Telecomm companies

Category 2: Data Exploration





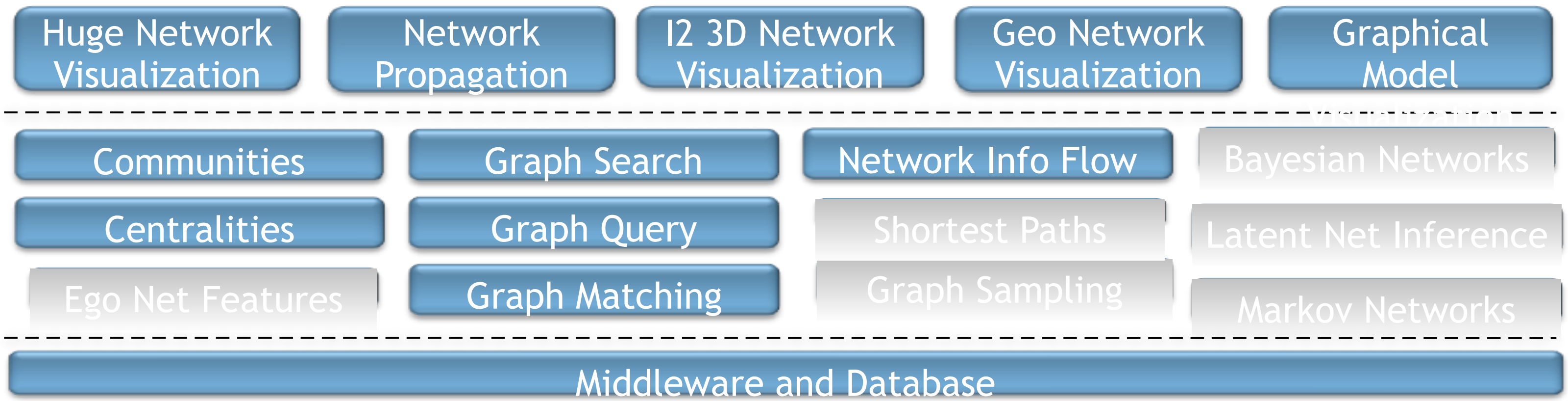
Japan Earthquakes[Thu Mar 15 21:58:22 CST 2012] : Earthquake risk in Fukushima is very high - earthquake forecast map for March 16-17, 2012 is now online; http://t.co/AcWo5Wbf

Enhancing:

Master Data Management

Vivísimo

CÚRAM SOFTWARE


Huge Network Visualization

Network Propagation

I2 3D Network Visualization

Geo Network Visualization

Graphical Model

Communities

Centralities

Ego Net Features

Graph Search

Graph Query

Graph Matching

Network Info Flow

Shortest Paths

Graph Sampling

Bayesian Networks

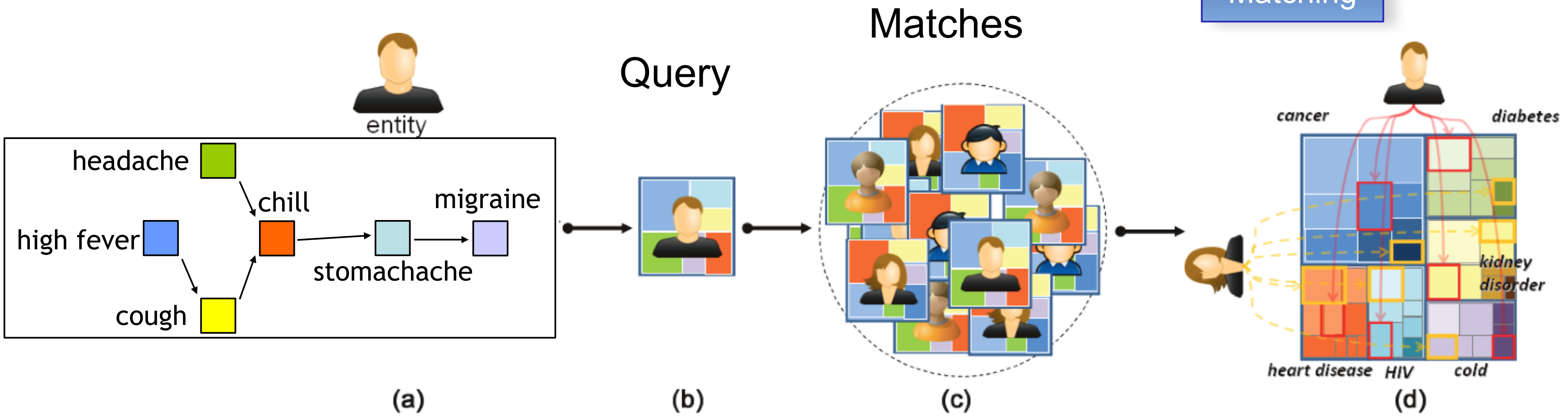
Latent Net Inference

Markov Networks

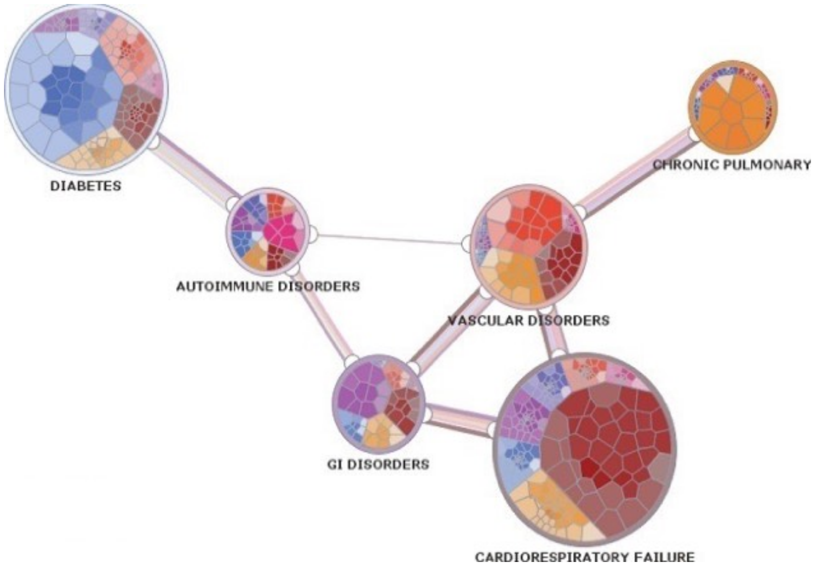
Middleware and Database

Use Case 7: Graph Analytics and Visualization

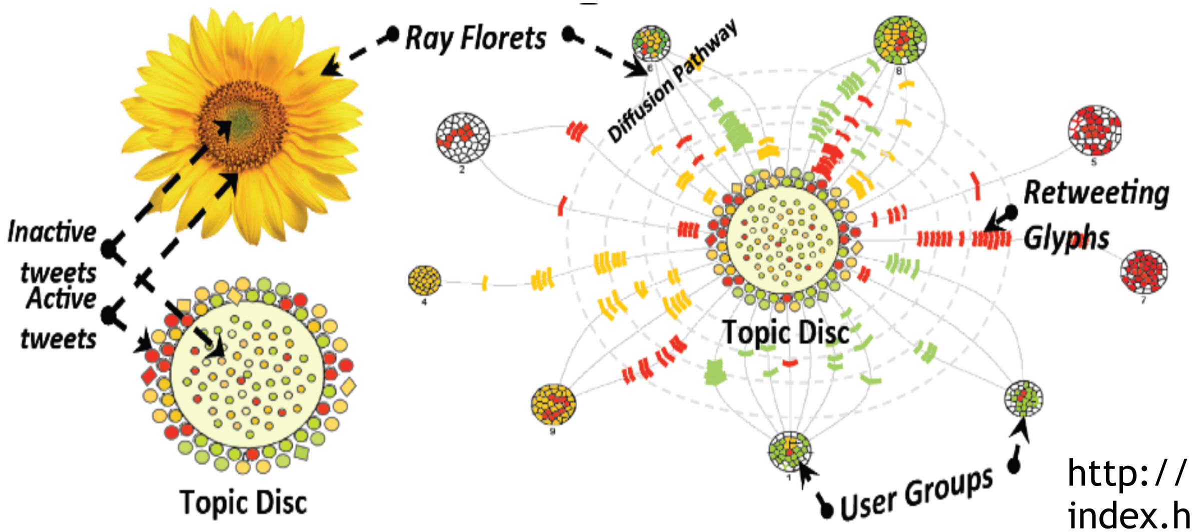
Graph Matching



Graph Communities



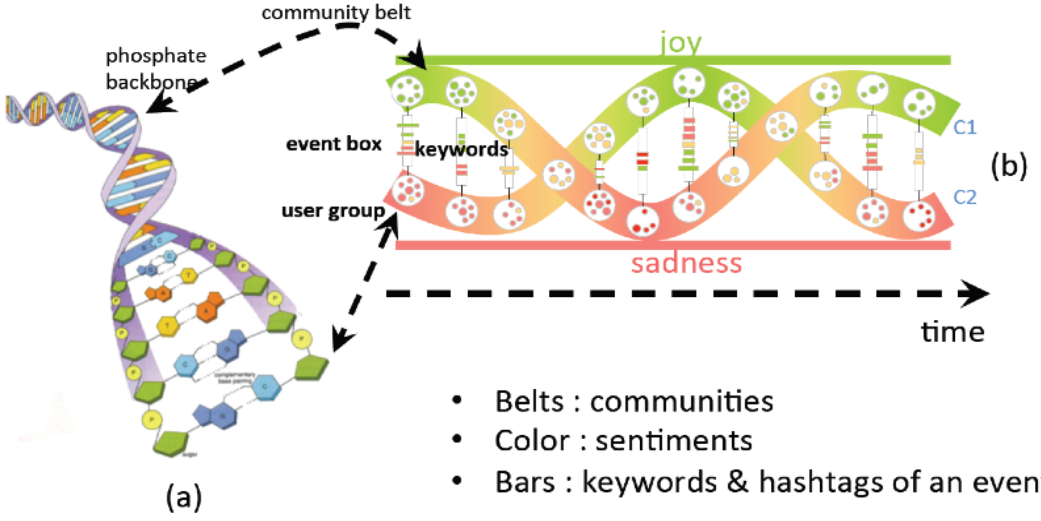
User Case 8: Visualization for Navigation and Exploration



Whisper : Tracing the information diffusion in Social Media

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

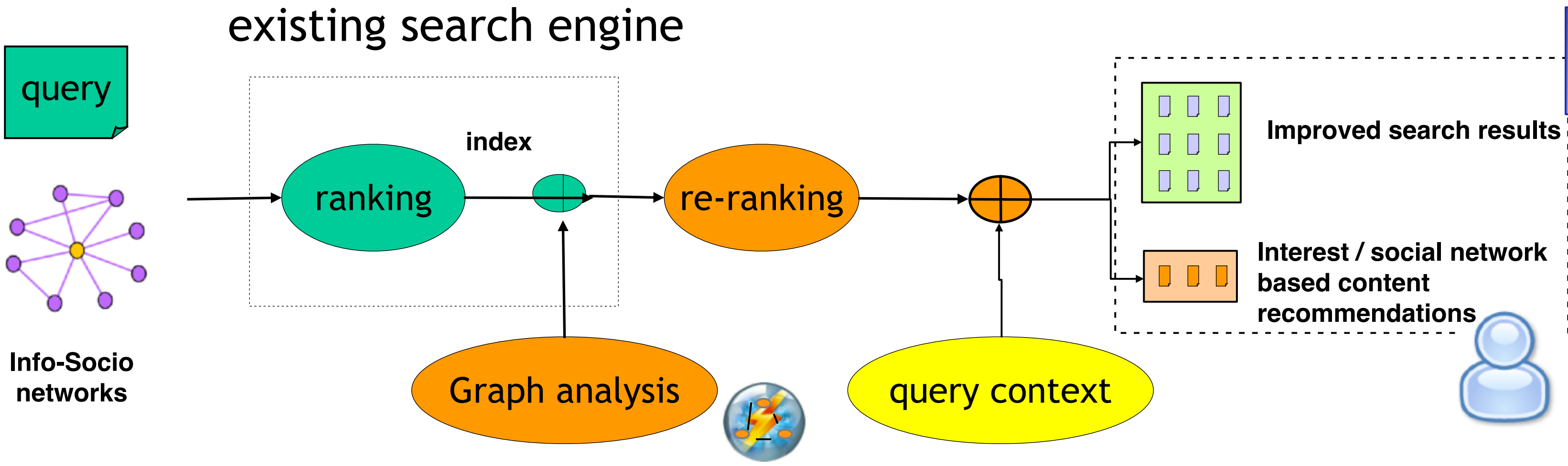
SocialHelix: Visualizaiton of Sentiment Divergence in Social Media

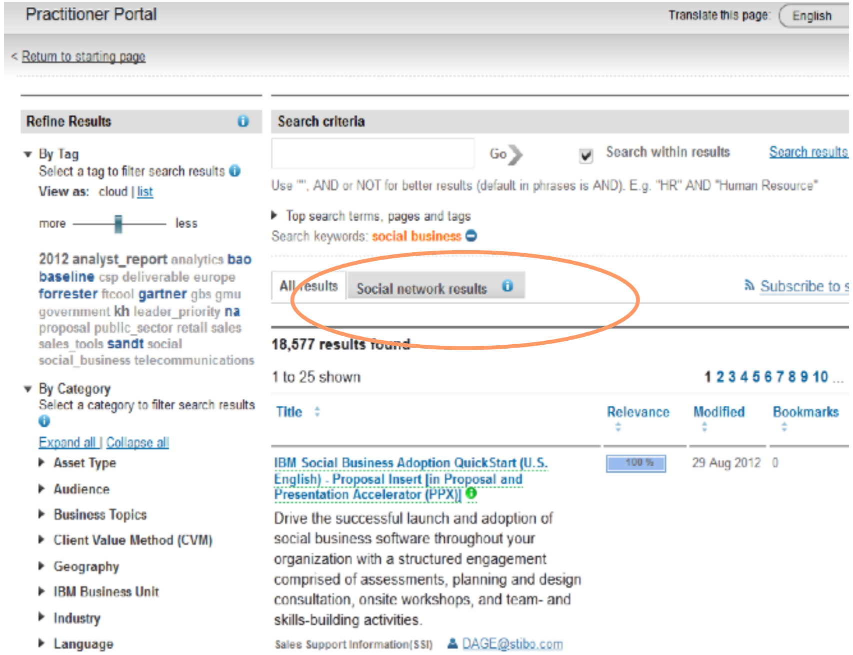


- Belts : communities
- Color : sentiments
- Bars : keywords & hashtags of an event

Use Case 9: Graph Search

Graph Search





Practitioner Portal

Translate this page: English

< 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 ftcool gartner gbs gmu government kh leader_priority na proposal public_sector retail sales sales_tools sandt 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

Go ☑ Search within results Search results

Use "", AND or NOT for better results (default in phrases is AND). E.g. "HR" AND "Human Resource"

- Top search terms, pages and tags

Search keywords: social business

All results | Social network results

Subscribe to s

18,577 results found

1 to 25 shown

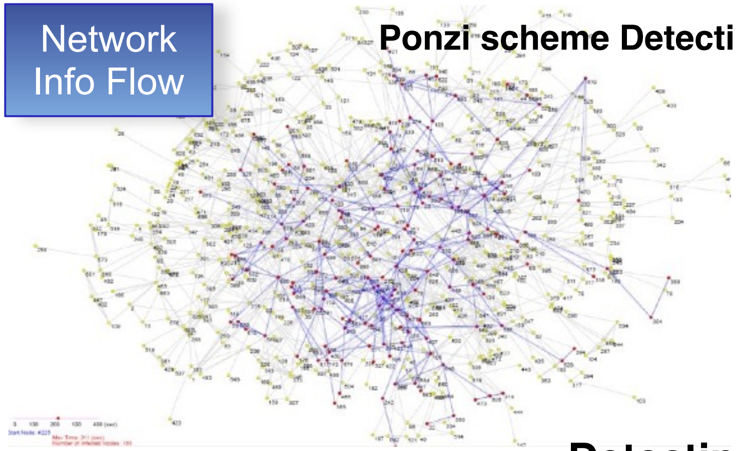
1 2 3 4 5 6 7 8 9 10 ...

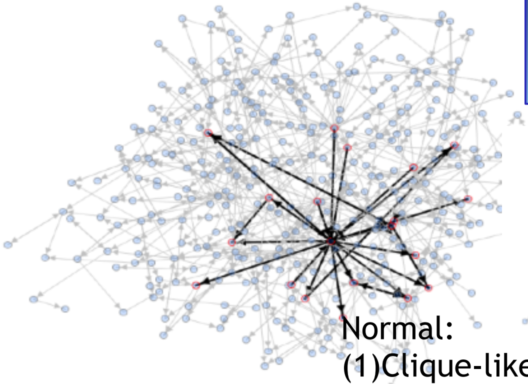
Title	Relevance	Modified	Bookmarks
IBM Social Business Adoption QuickStart (U.S. English) - Proposal Insert [in Proposal and Presentation Accelerator (PPX)] Drive the successful launch and adoption of social business software throughout your organization with a structured engagement comprised of assessments,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

Category 3: Security

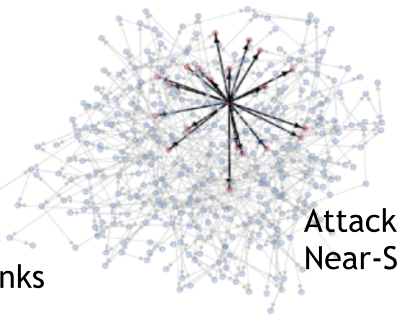
Network Info Flow

Ponzi scheme Detection





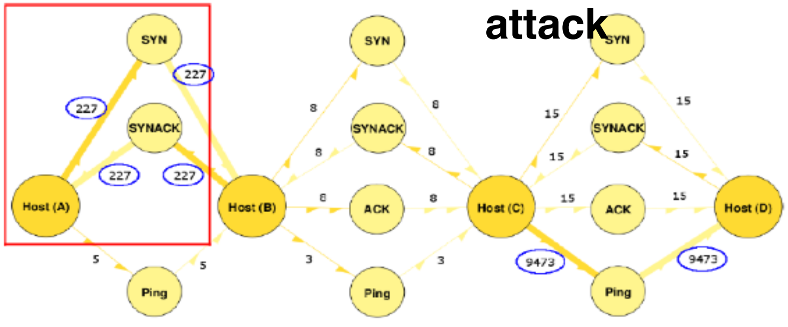
Ego Net Features



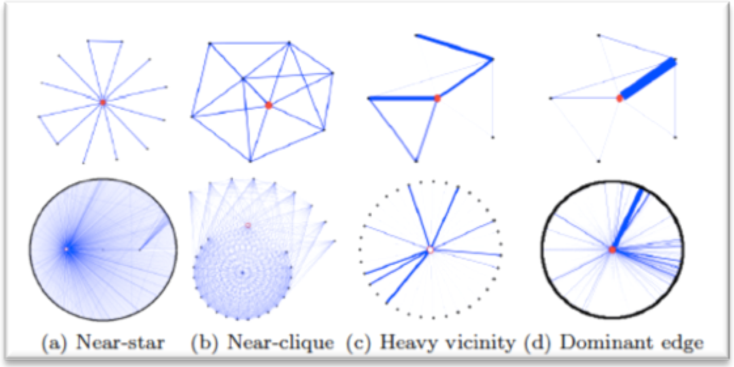
Normal:
(1)Clique-like
(2)Two-way links

Attacker:
Near-Star

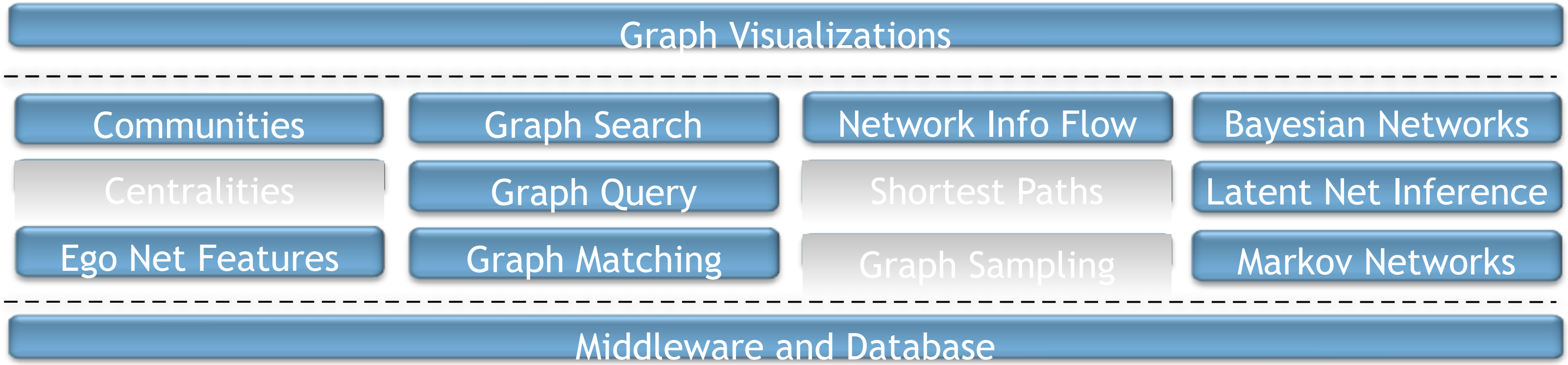
Detecting DoS attack



(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



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.

THE WALL STREET JOURNAL.

Many Past Espionage Cases Had Links to

Since 2009, Justice Department lawyers have pursued at least 19 cases of corporate espionage. Most had connections to China, according to case summaries presented in a new U.S. strategy report on trade secrets Wednesday.

Among high-profile cases:

◆ In November 2012, Shanshan Du, a former General Motors engineer, and her husband Yu Qin

help trading efficiency in China.

◆ In August 2012, Hanjuan Jin, a former Motorola engineer, was se
tenced in Illinois to four years in
prison for stealing Motorola's pro
etary IDEN technology for hersel
and for a company that works fo
the Chinese military.

◆ In March 2012, former DuP
scientist Tze Chao pleaded guilty
California to conspiracy to comm
economic espionage, admitting he

agent posin
gence offic

U.S. Ups Ante for Spying

"Enterprise Information Leakage Impacted economy and jobs" Feb 2013

npr

news ▸ business

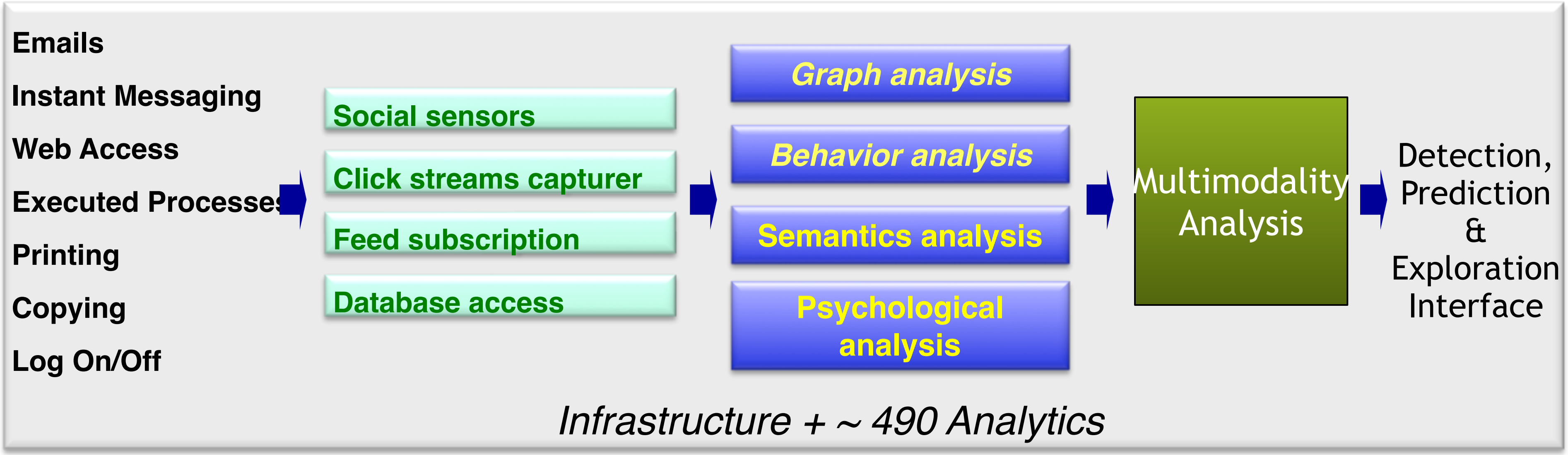
To Catch Worker Misconduct, Companies Hire Corporate Detectives

by AILSA CHANG

January 10, 2013 6:25 PM

"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

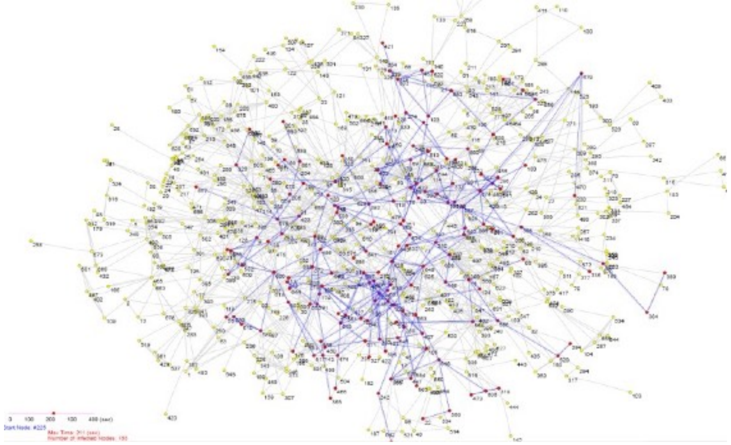
Infrastructure + ~ 490 Analytics

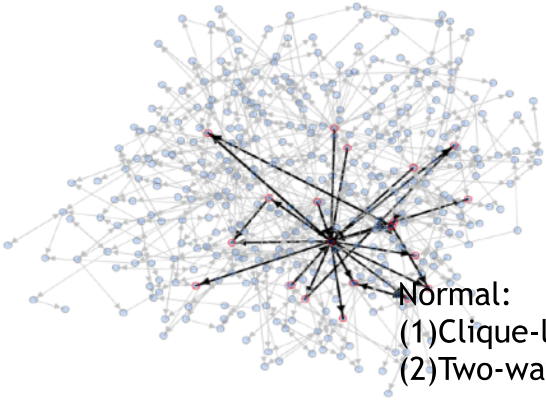
Use Case 11: Fraud Detection for Bank

Network Info Flow

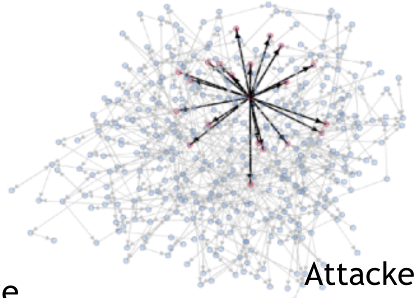
Ego Net Features

Ponzi scheme Detection

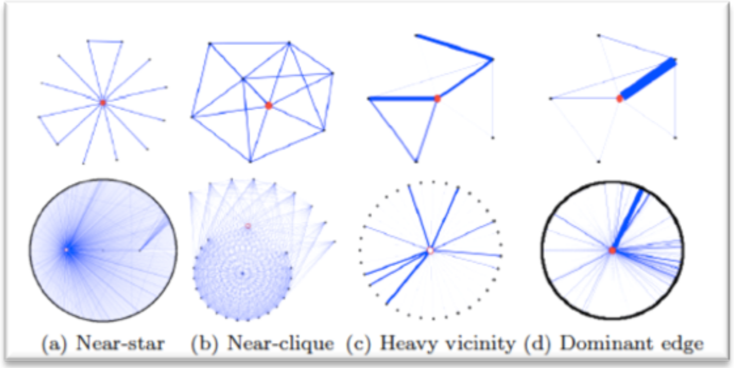




Normal:
(1)Clique-like
(2)Two-way links



Attacker:
Near-Star



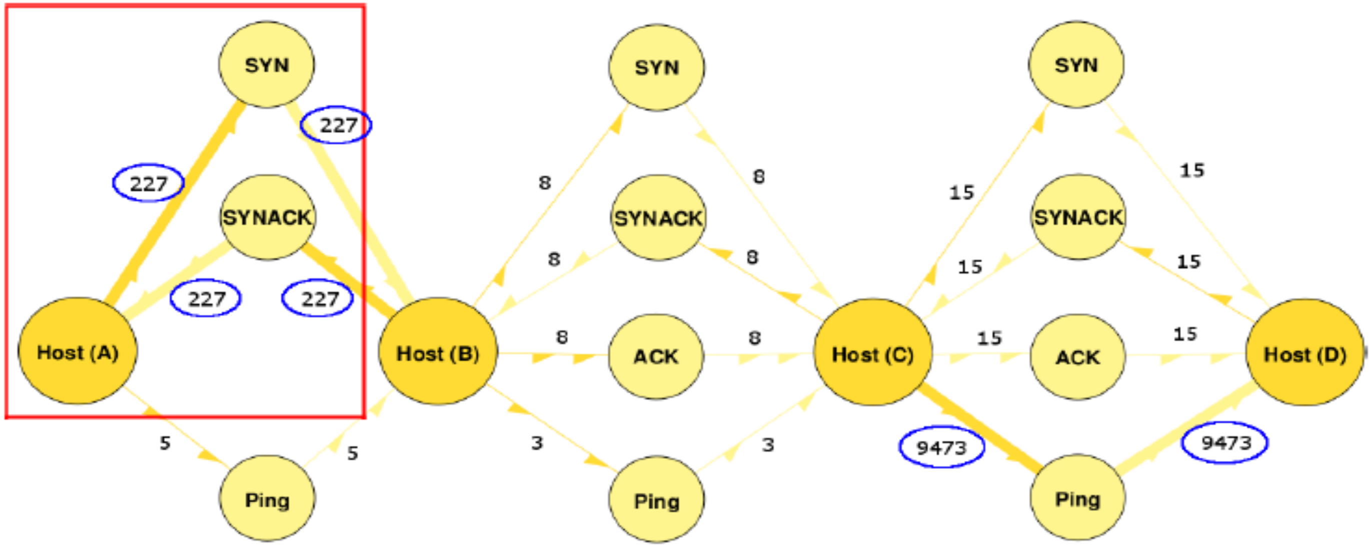
(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

Use Case 12: Detecting Cyber Attacks

Network Info Flow

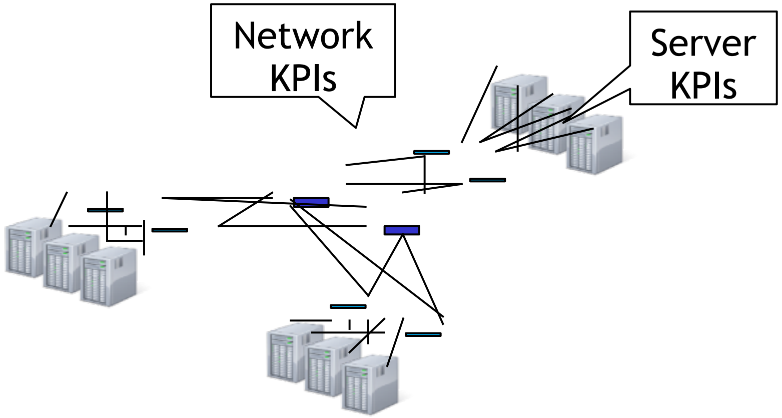
Ego Net Features

Detecting DoS attack

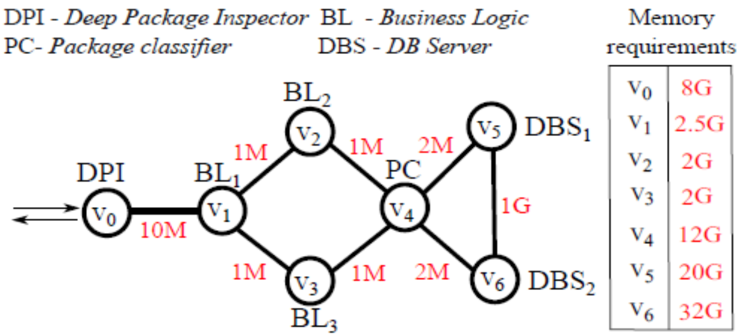


(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

Category 4: Operations Analysis



Cloud Service Placement

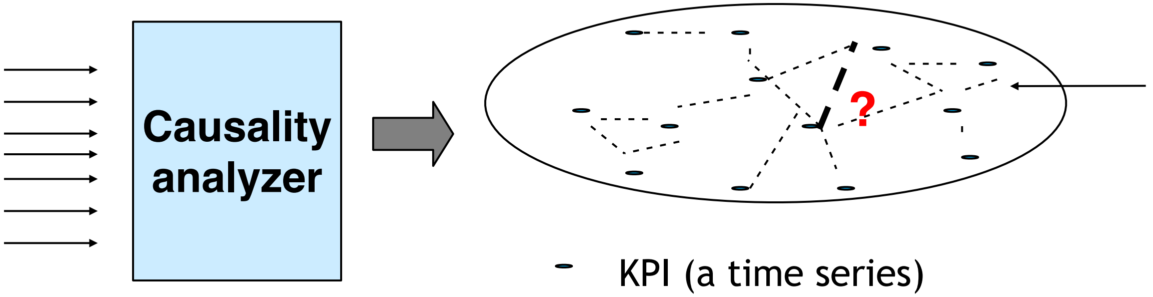


Memory requirements	
V$_0$	8G
V$_1$	2.5G
V$_2$	2G
V$_3$	2G
V$_4$	12G
V$_5$	20G
V$_6$	32G

Graph Matching

Bayesian Network

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



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

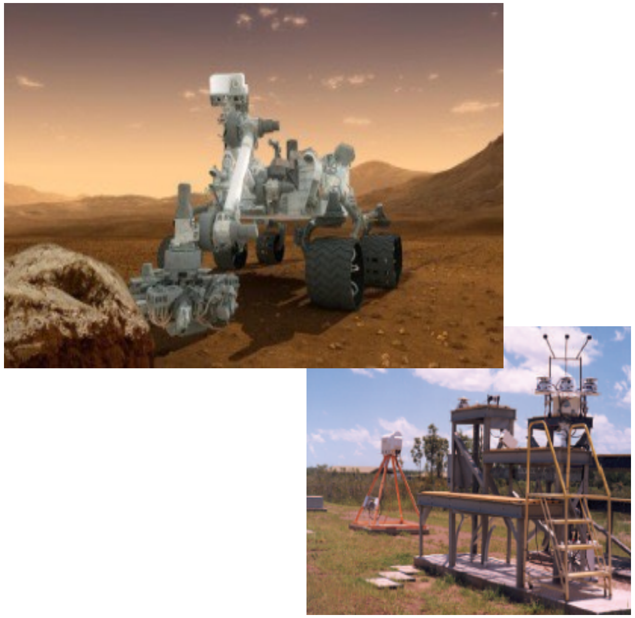
Middleware and Database

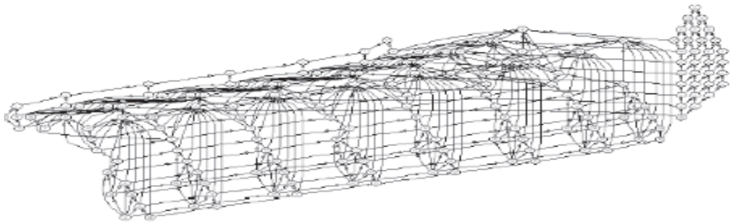
Use Case 13: Smarter *another* Planet

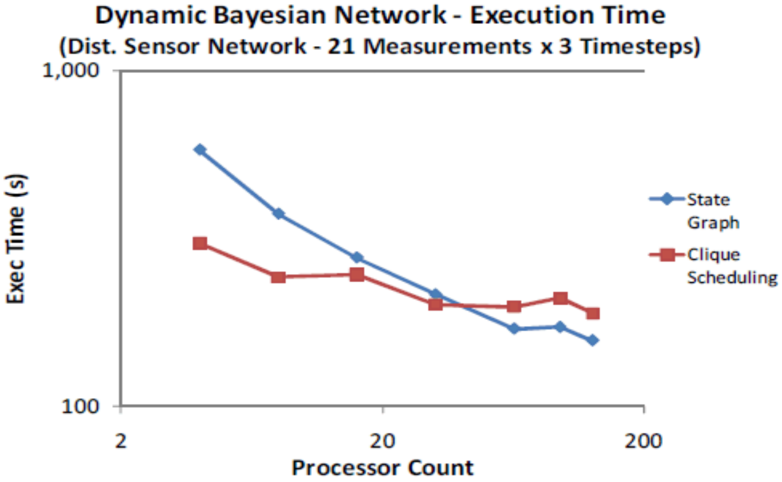
Bayesian Network

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.





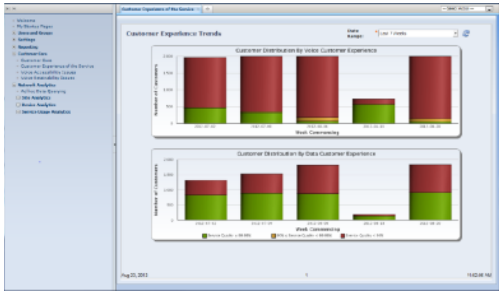

Dynamic Bayesian Network - Execution Time
(Dist. Sensor Network - 21 Measurements x 3 Timesteps)
1,000
100
Exec Time (s)
2
20
200
Processor Count
State Graph
Clique Scheduling

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

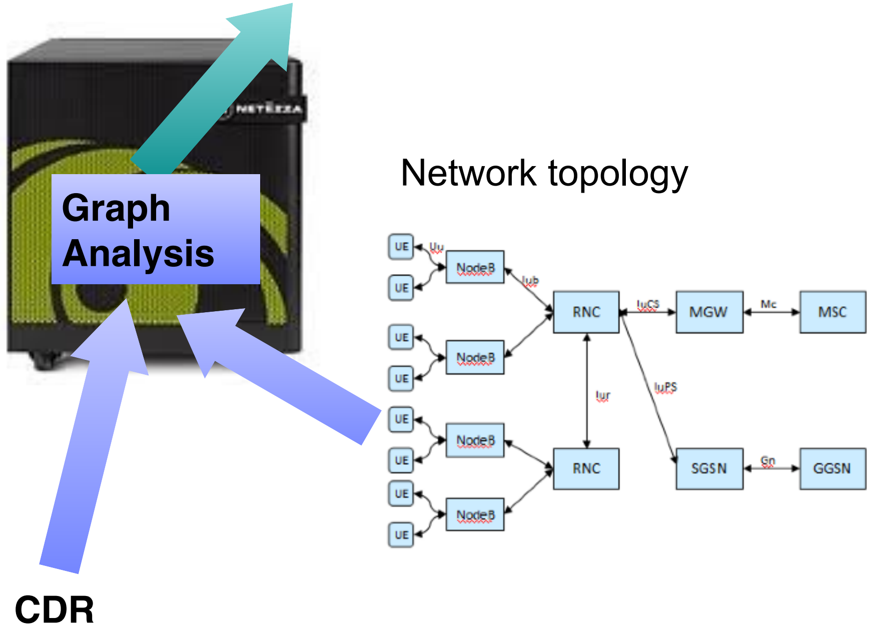
Use Case 14: Cellular Network Analytics in Telco Operation

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

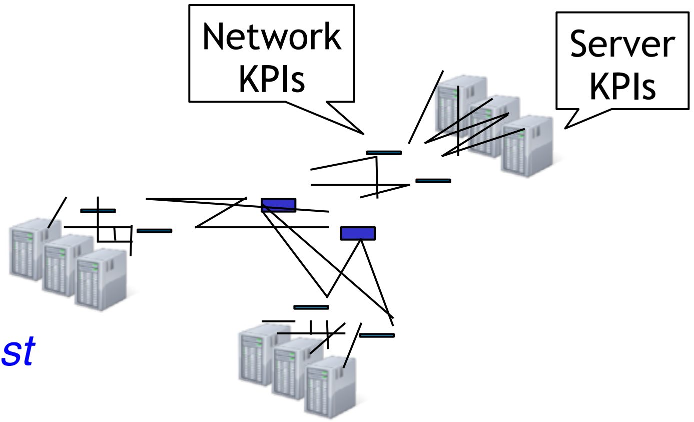
- Applied Graph Analytics to telco network analytics based on CDRs (call detail records): estimate traffic load on CSP network with low monitoring overhead
  - (1)CDRs, already collected for billing purposes, contain information about voice/data calls
  - (2)Traditional NMS* and EMS** typically lack 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



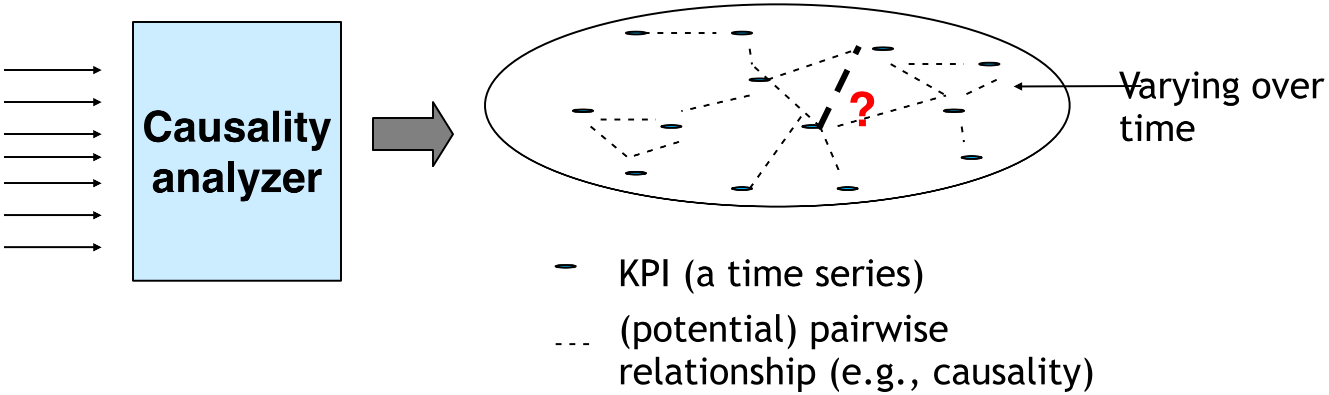
Use Case 15: Monitoring Large Cloud

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)*



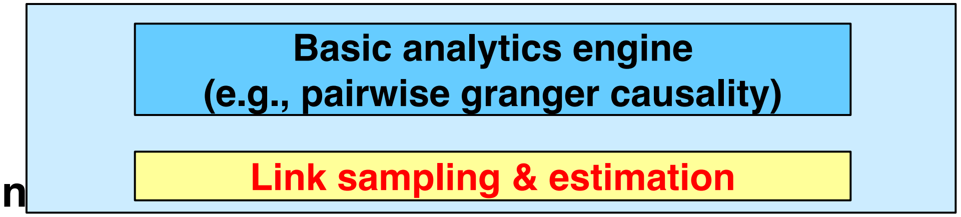
KPI time series (e.g., server performance/load, network performance/load)



- KPI (a time series)

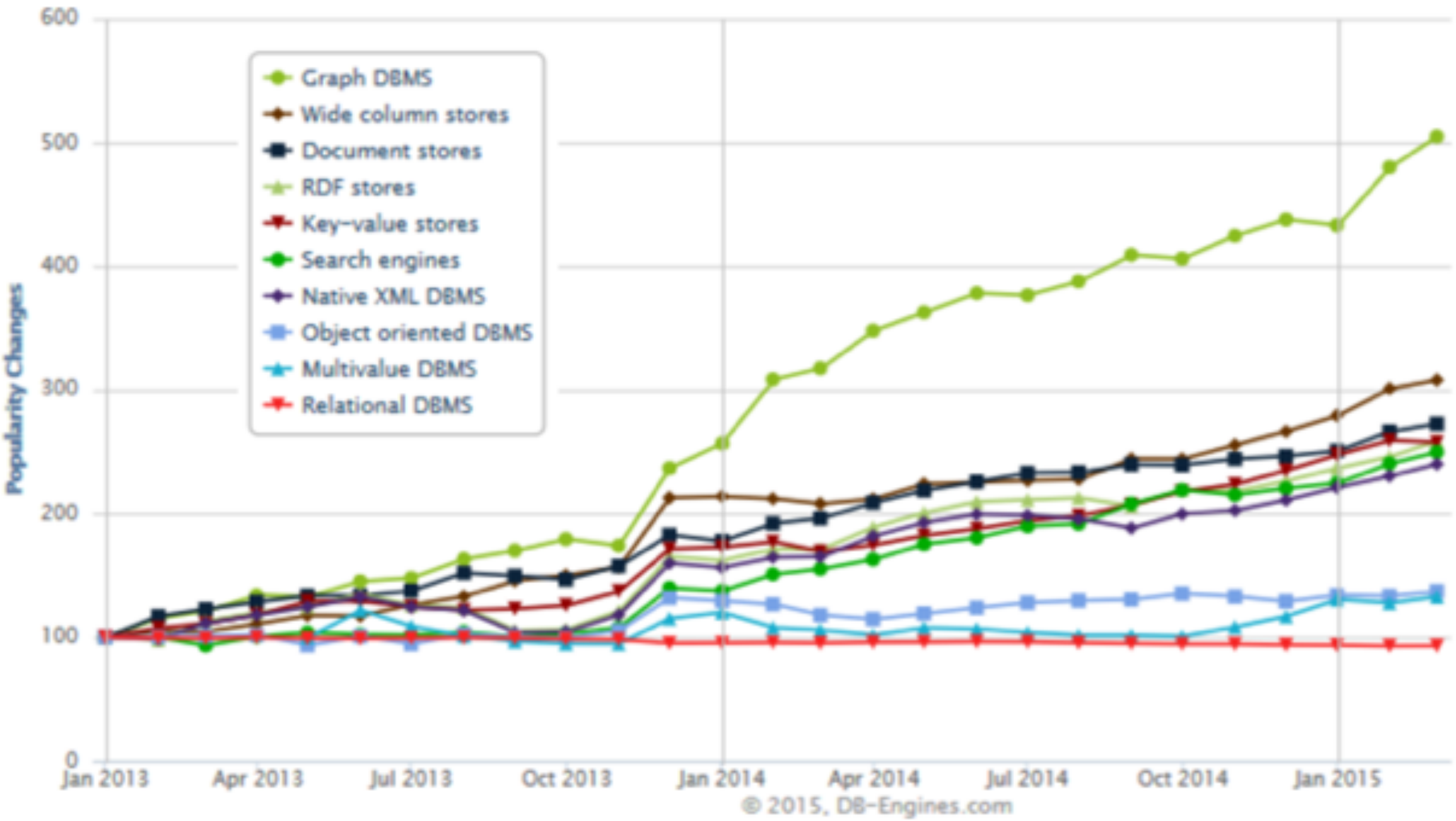
- - - (potential) pairwise relationship (e.g., causality)

Our approach: Probabilistic monitoring via sampling & estimation

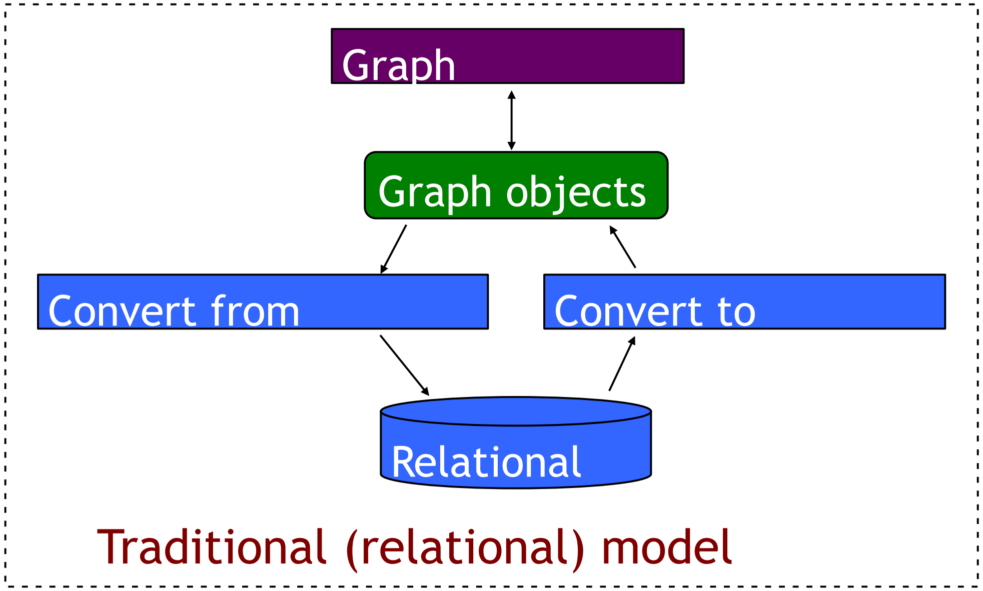

Basic analytics engine (e.g., pairwise granger causality)
Link sampling & estimation

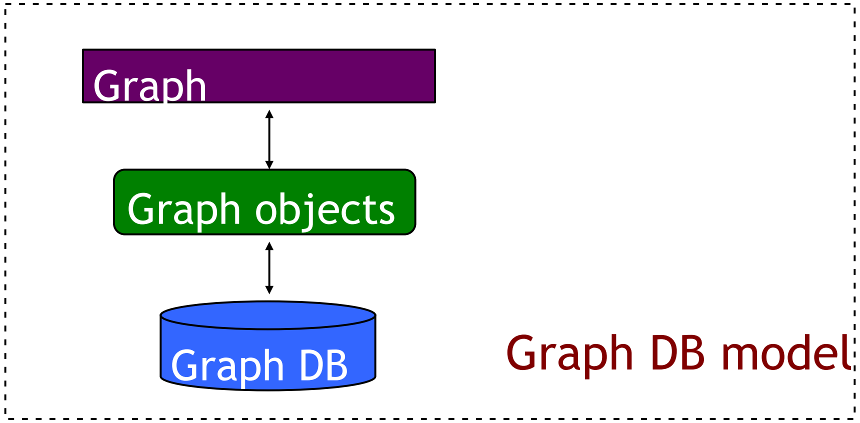
Select KPI pairs (sampling)→ Test link existence → Estimate unsampled links based on history → Overall graph

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

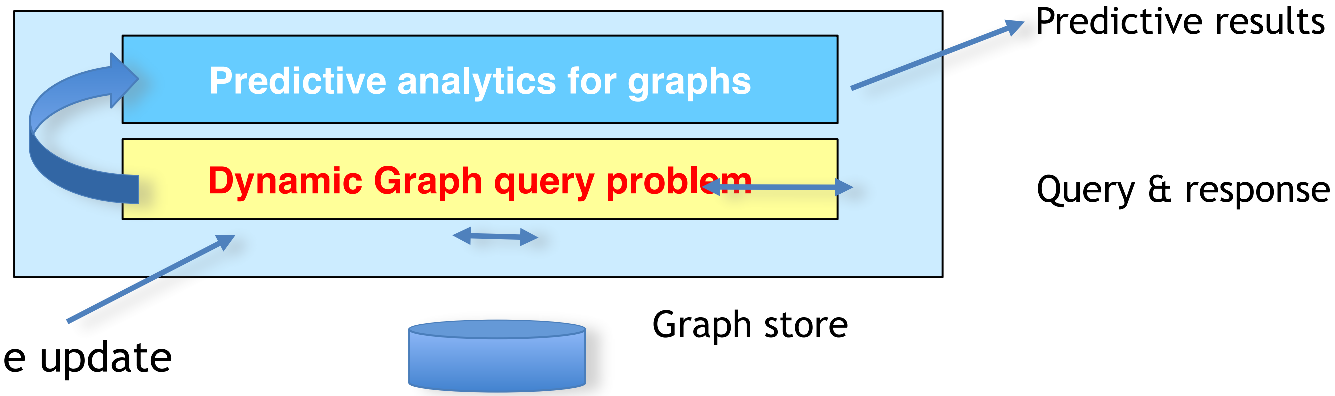
Use Case 17: Smart Navigation Utilizing Real-time Road Information

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*

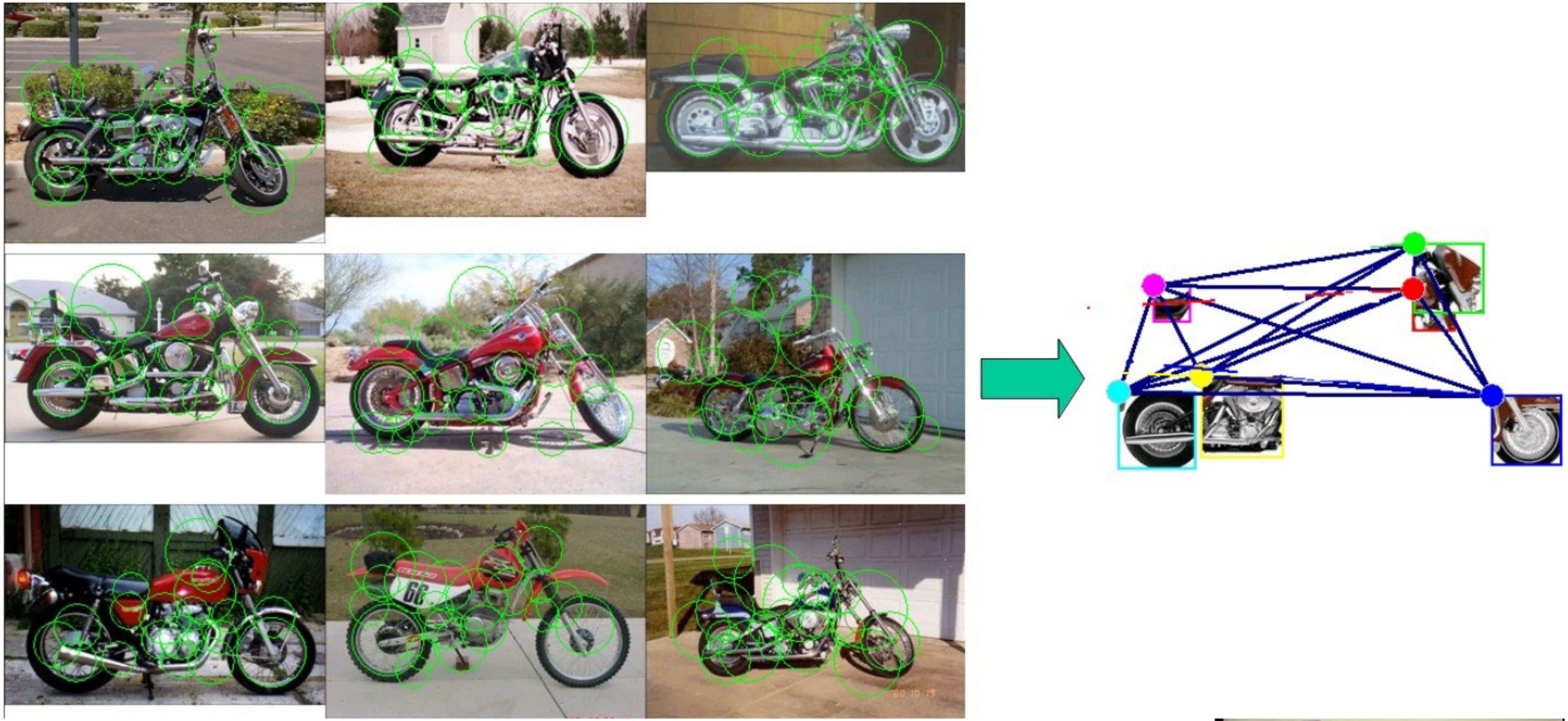
- *Dynamic graph algorithms implemented in System G provide* ***highly efficient graph query computation*** *(e.g. shorted path computation) on time-varying graphs (order of magnitudes improvement over existing solutions)*

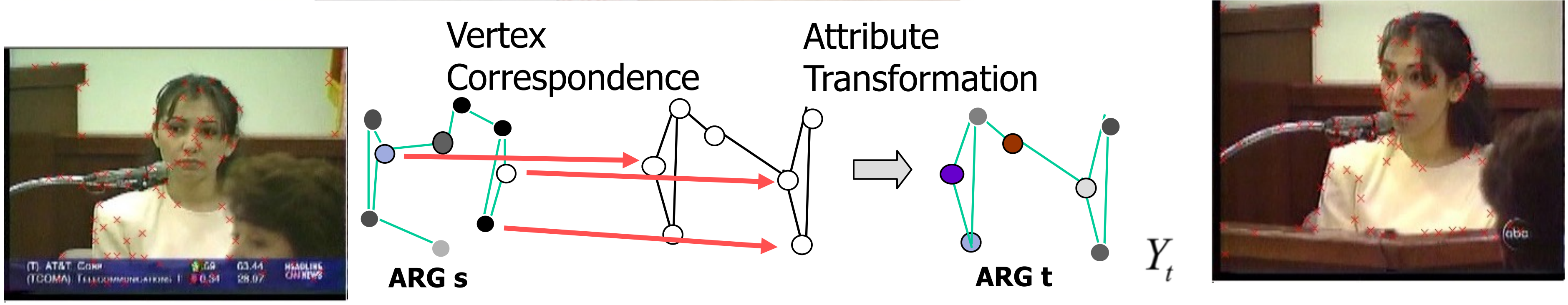
- *High-throughput* ***real-time predictive analytics*** *on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall*

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

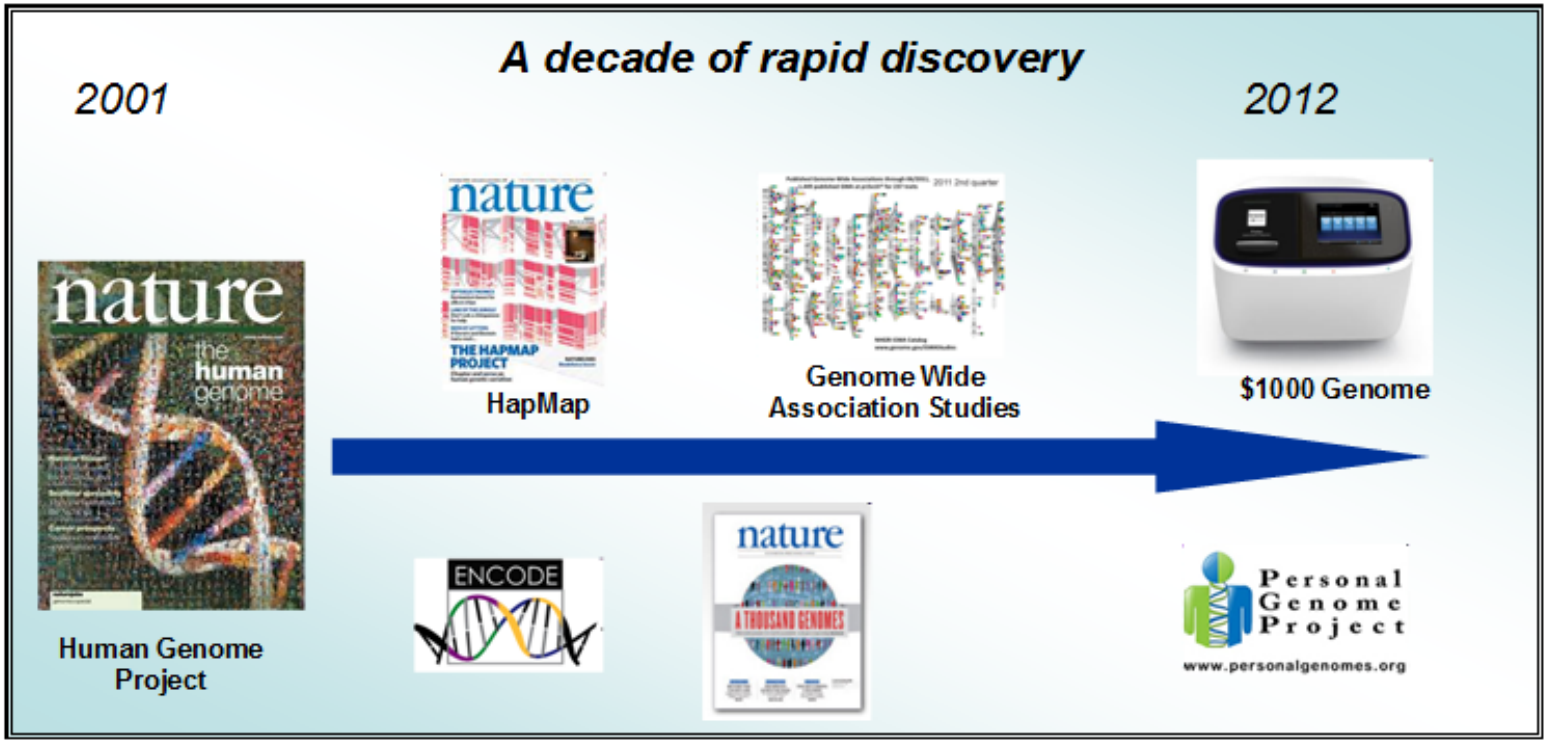
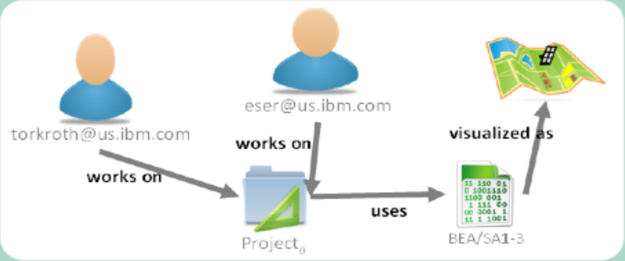


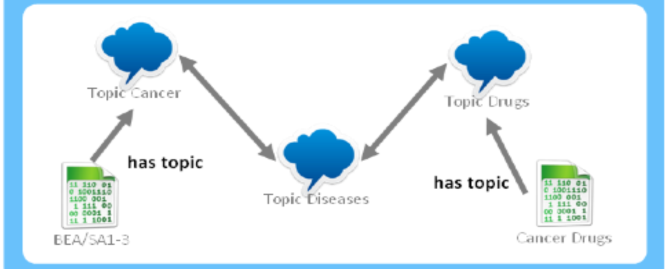
Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.

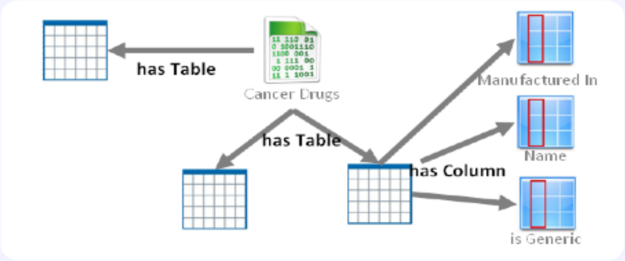
Use Case 20: Data Curation for Enterprise Data Management

Prior Collaborative Use

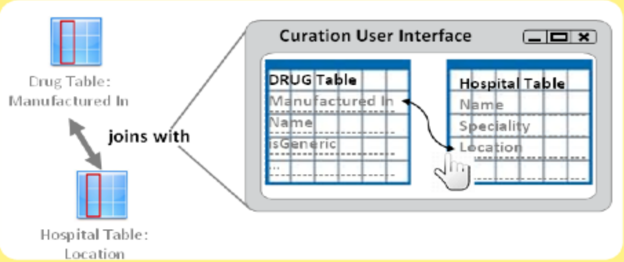


Semantic Knowledge

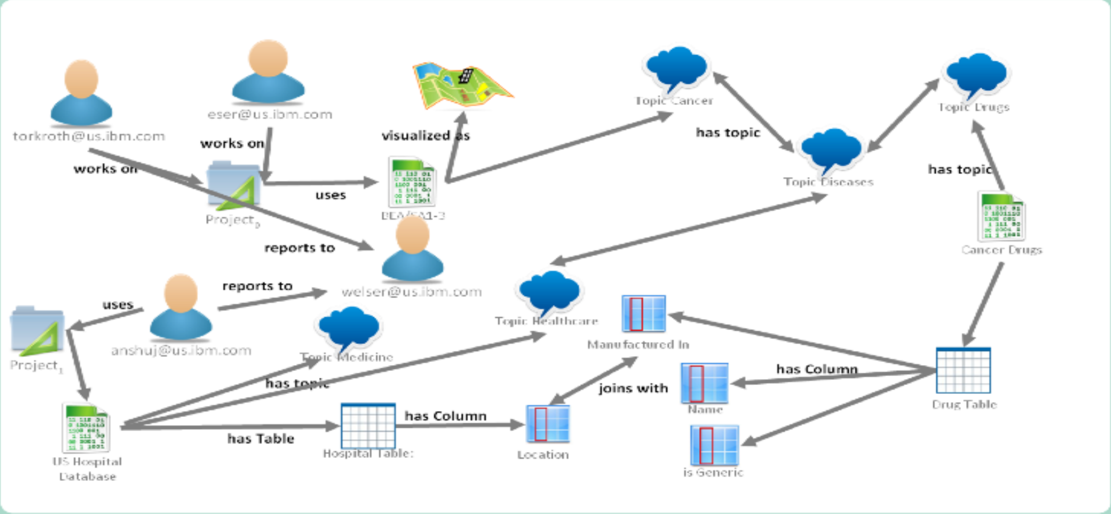




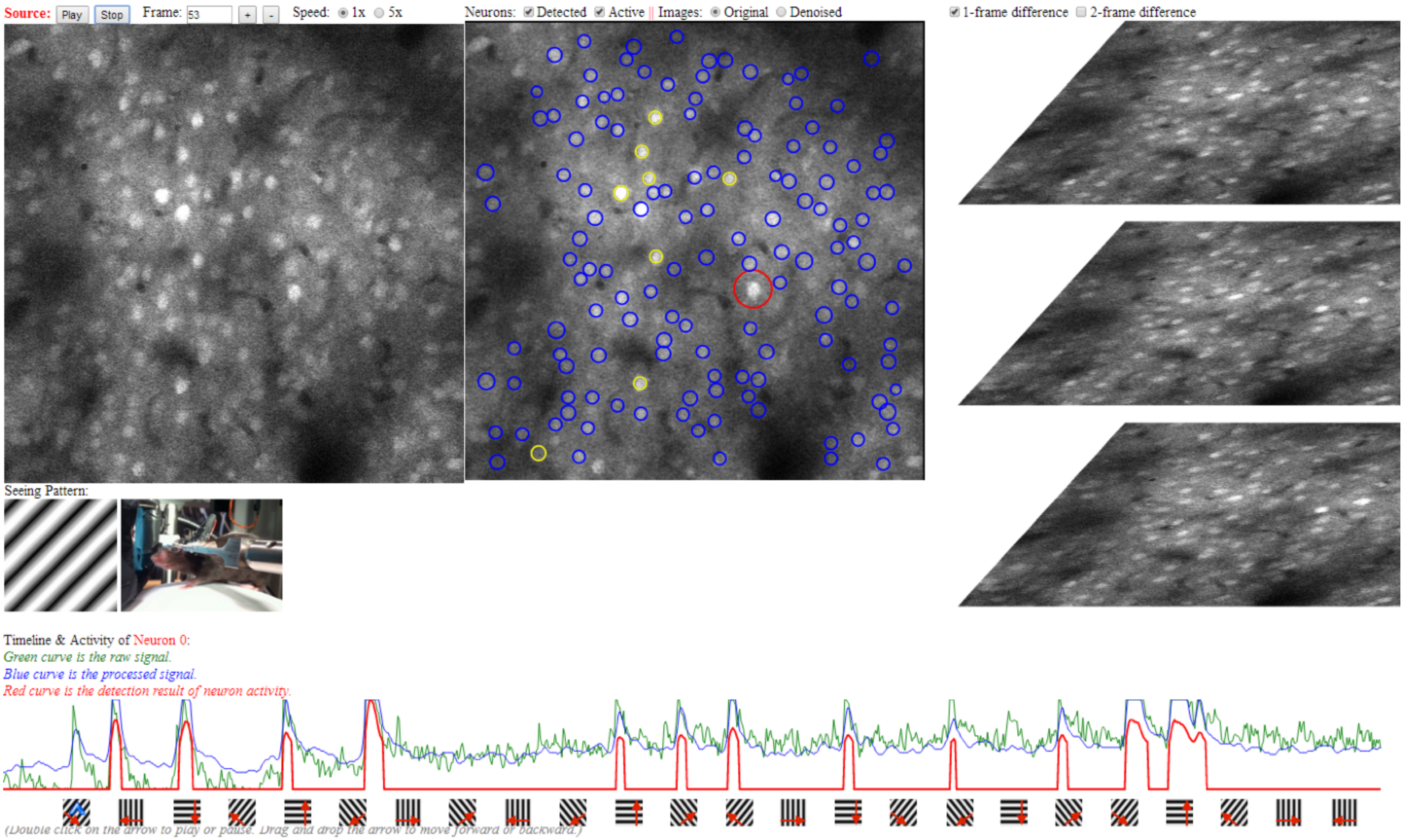
Extracted Metadata



Supervised Curation



Use Case 21: Understanding Brain Network



Source: Play Stop Frame: 53 + - Speed: ● 1x ○ 5x

Neurons: ☑ Detected ☑ Active || Images: ● Original ○ Denoised

☑ 1-frame difference □ 2-frame difference

Seeing Pattern:

Timeline & Activity of Neuron 0:

Green curve is the raw signal.

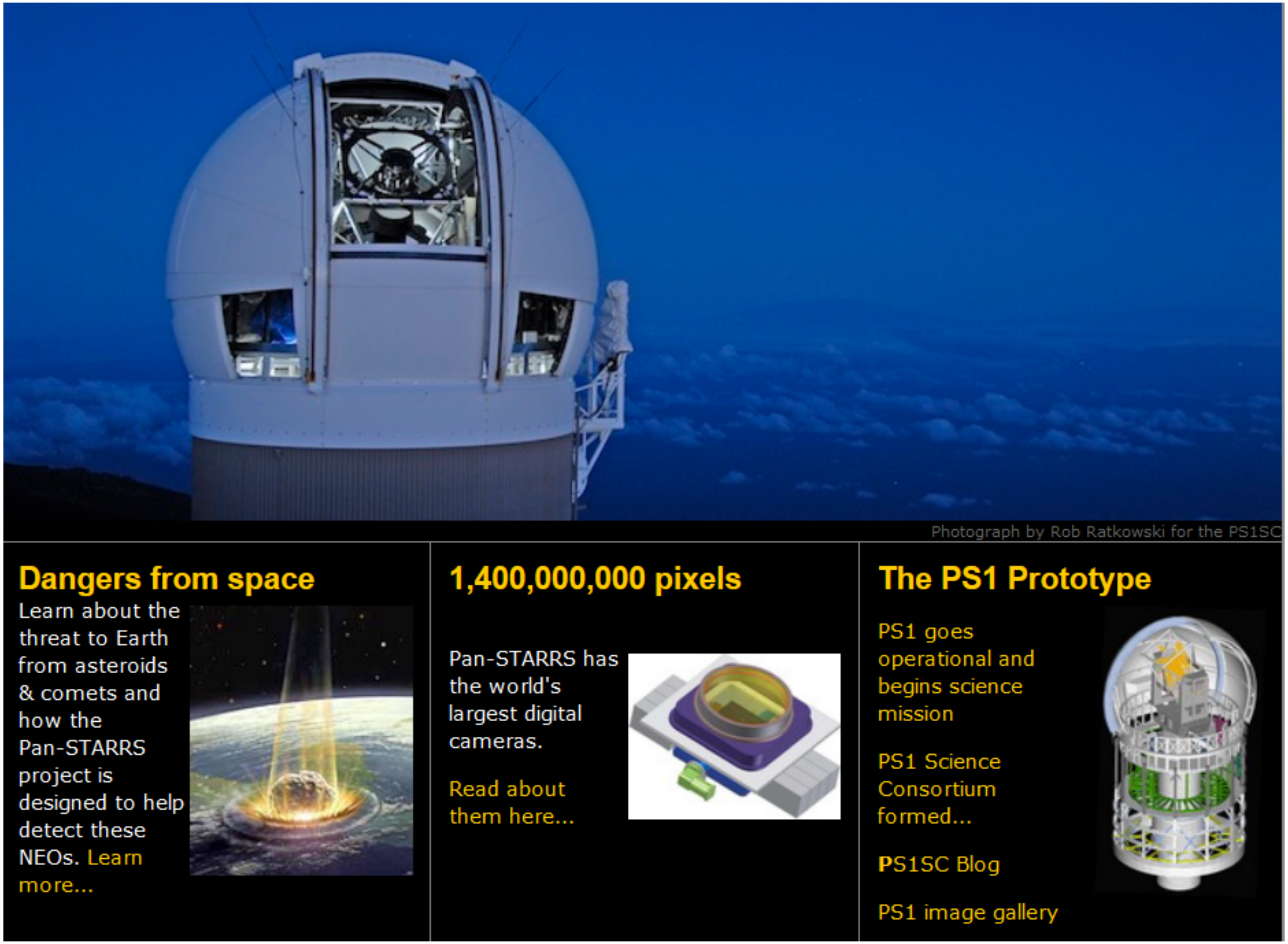
Blue curve is the processed signal.

Red curve is the detection result of neuron activity.

(Double click on the arrow to play or pause. Drag and drop the arrow to move forward or backward.)

Use Case 22: Planet Security

- Big Data on Large-Scale Sky Monitoring



Photograph by Rob Ratkowski for the PS1SC

Dangers from space

Learn about the threat to Earth from asteroids & comets and how the Pan-STARRS project is designed to help detect these NEOs. Learn more...

1,400,000,000 pixels

Pan-STARRS has the world's largest digital cameras.

Read about them here...

The PS1 Prototype

PS1 goes operational and begins science mission

PS1 Science Consortium formed...

PS1SC Blog

PS1 image gallery

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

Homework Late Submission Policy

5pm: submission deadline
Next Day midnight: 10% penalty
Two Days late midnight: 20% penalty
Three Days late midnight: 30% penalty
Any late submission more than 3 days will not be accepted.

Please do your each homework as early as possible!! They are all quite ‘heavy’.