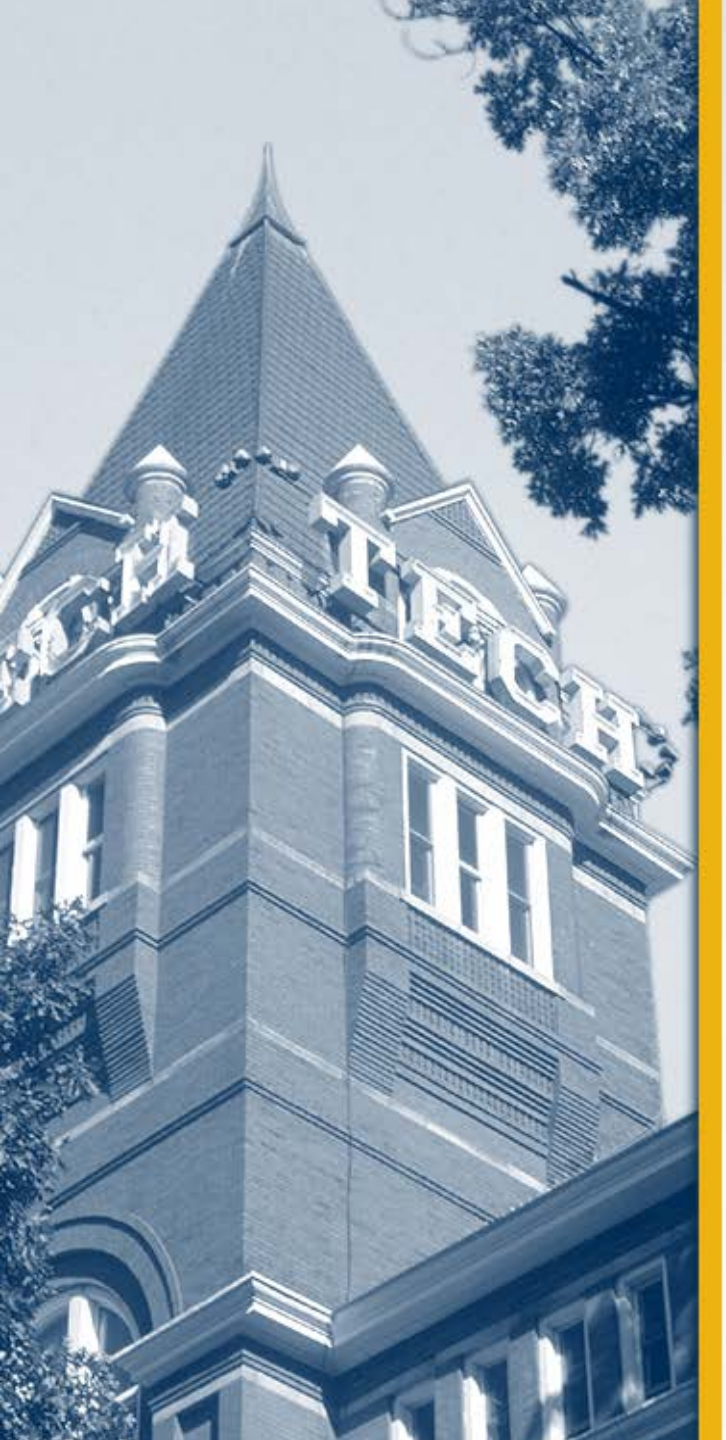


Connecting Big Data with Data Analytics: A CS perspective

Ling Liu
Professor
Distributed Data Intensive Systems Lab
School of Computer Science
Georgia Institute of Technology

Two Questions from Day 1's Presentations

- How can multiple IEEE societies work together under the IEEE big data initiative
- How can we increase the value of the IEEE big data ecosystem
  - Members
  - IEEEXplore
  - Journals/Magazines/Conferences
  - Other Products

Three Opportunities and Challenges

- (1) Data Utility driven, unified underlying big data computing infrastructure
  - Different applications/users with different big data problems may benefit from a unified underlying big data computing infrastructure for data analysis tasks
- (2) Leveraging Different Strengths of Big Algorithms
  - The Value/Utility of Big Data is in the eyes of beholders
- (3) Utility-driven big data policies and standards
  - How big data should be analyzed and used while respect privacy, trust and authorization

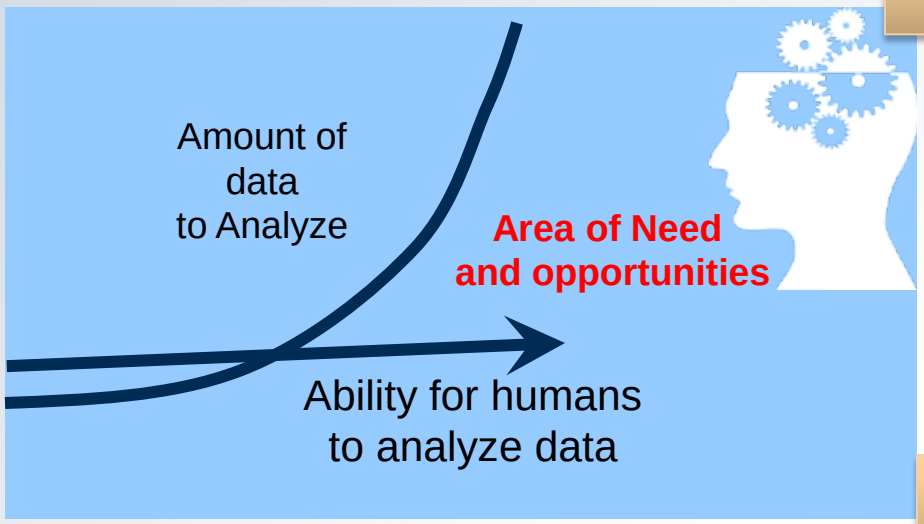
What is Big Data: A CS Perspective

- Big data refers to datasets
  - that are *beyond the ability of legacy approaches* to manage at an acceptable level of quality and / or
  - that *exceed the capacity of conventional systems* (hardware and/or software) to process within an acceptable elapsed time.

- Definition of big data: Subjective & evolving
  - As technology advances over time, the size of datasets that qualify as big data will also increase.
  - The definition is varying by sector, depending on
    - what kinds of software tools are commonly available and
    - what sizes of datasets are common in a particular industry or science domain.

4

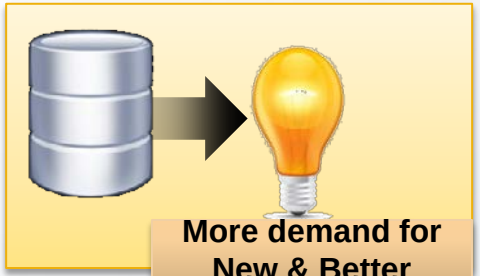
Why now → Data grows faster than intelligence



More Content

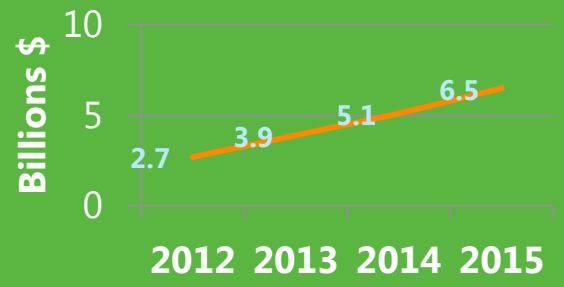
More Devices

More Consumption



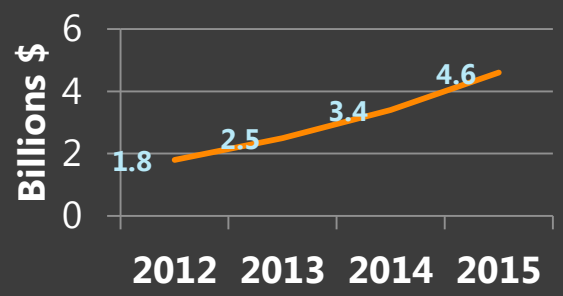
More demand for New & Better Information

Big Data Services Growth



39% compound annual growth rate

Big Data Software Growth



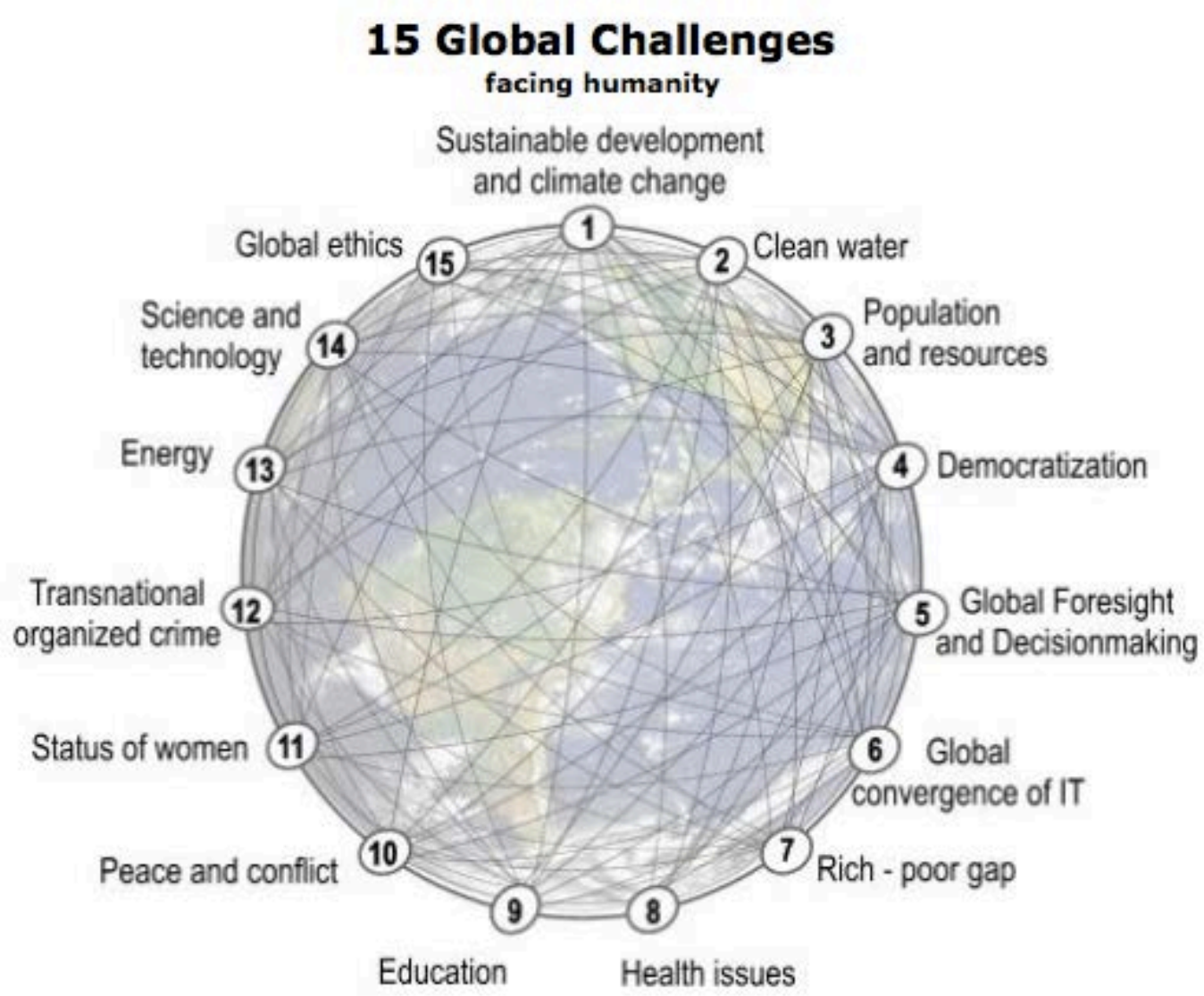
34% compound annual growth rate[2]

IDC Market Analysis, Worldwide Big Data Technology and Services 2012–2015 Forecast , 2012
[Jack Tang & Darien Nagle, Microsoft APC 2012]

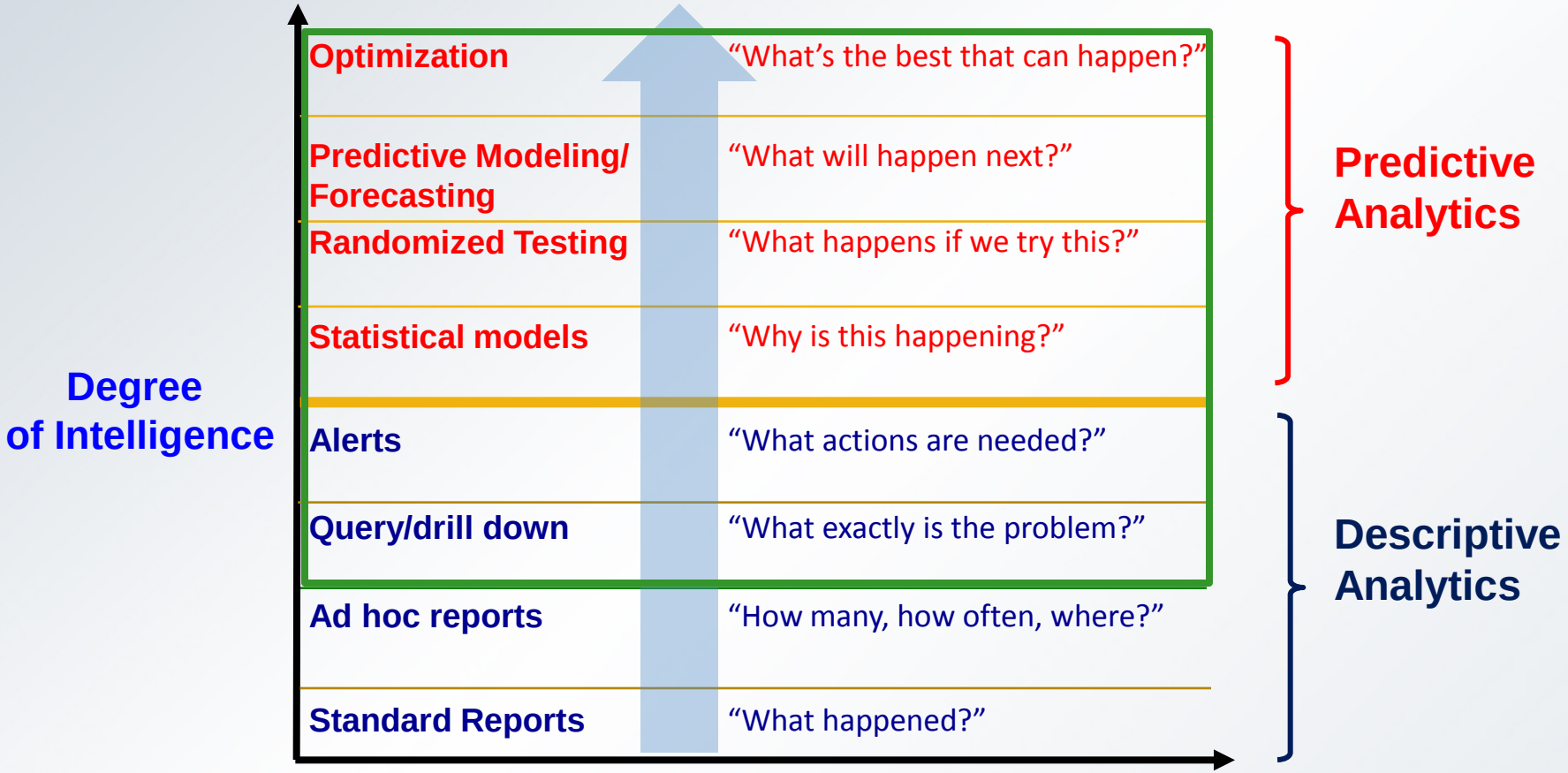
Georgia Tech

Humanity Challenges →
Connecting Big Data with Data Analytics

http://www.millennium-project.org/millennium/challenges.html



Descriptive Analytics vs. Predictive Analytics



Courtesy Tom Davenport, Charlotte Informatics 2012

Georgia Tech

Fundamental Data Analytic Problems

- **What information should I use to perform analytics ?**
- **How should I perform analytics efficiently ?**

Fundamental Data Analytic Problems

- What information should I use to perform analytics ?
- How should I perform analytics efficiently ?

Connecting Big Data with Data Analytics Opportunity 1:

Utilize programmable algorithm abstractions for many seemingly domain-dependent data analytics tasks

Discrete Optimization Problems

- **Social Network Analysis**
  - **How to discover the most influential k people for marketing of a product/an idea/an innovation on a social network?**
  - Want maximum effect of advertisement **with limited marketing budget**
- **Sensor placements**
  - with budget constraints (Krause *et al.* 2005a)
  - for maximizing information at minimum communication cost (Krause *et al.* 2006)
  - **Where should we place sensors to quickly detect fresh water contamination?**
    - Want get most useful information **at lowest cost**
    - Water sensors are expensive / limited, plus power consumption, personal time, ...

Find top K most influential people in a social network?

***Given*:** a finite set V of people in a social network graph, and the social influence utility function F

***Want*:** find a subset A of V such that

$$\mathcal{A}^* \subseteq \mathcal{V}$$
$$\mathcal{A}^* = \operatorname*{argmax}_{|\mathcal{A}| \leq k} F(\mathcal{A})$$

NP-hard!

We can use Greedy algorithm to get an approximation:

Greedy algorithm:

Start with A = {}

For i = 1 to k

s* := argmax$_s$ F(A U {s})

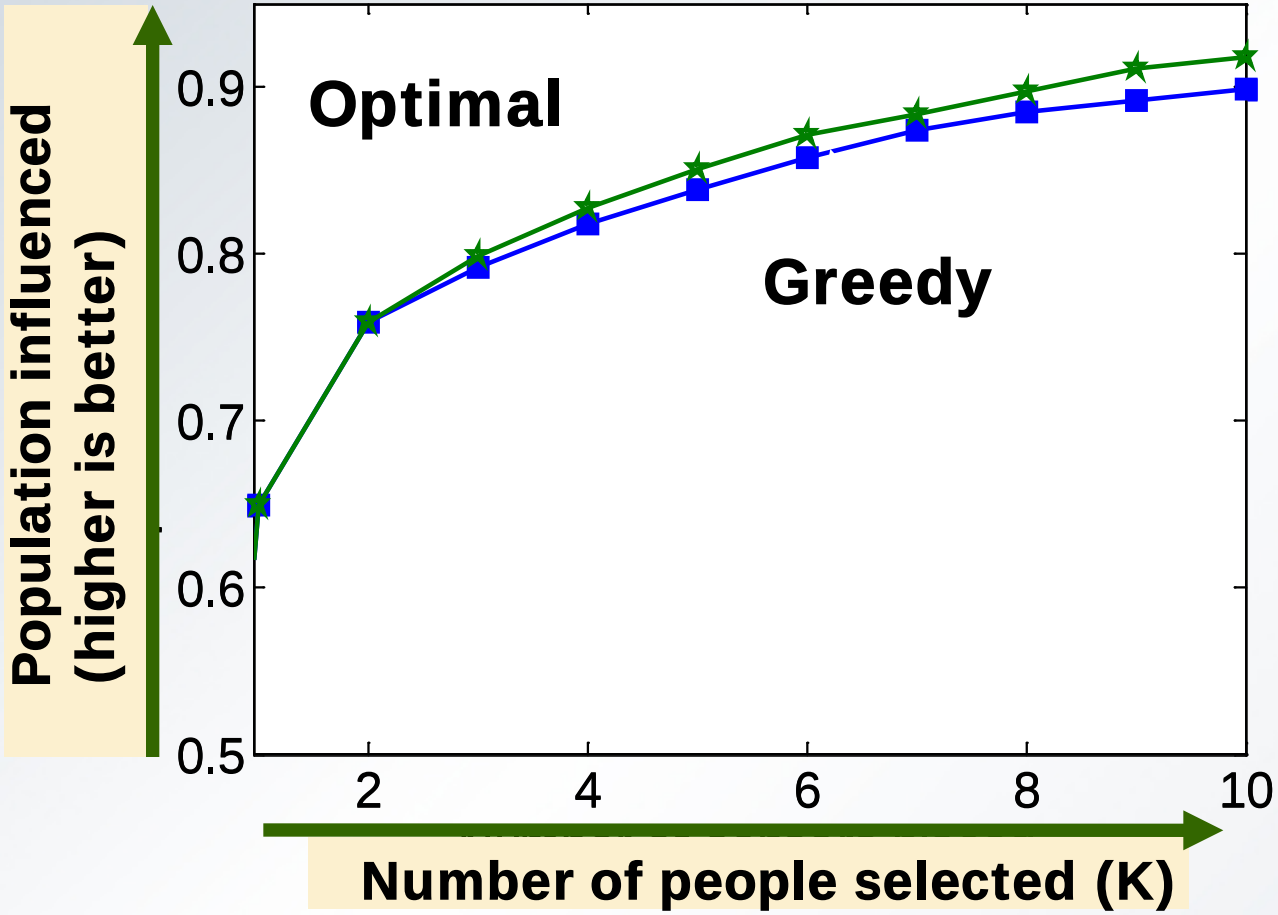
A := A U {s*}

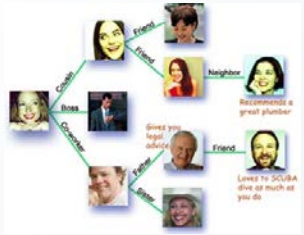
→ **how well can this simple heuristic do?**

11

Georgia Tech

Performance of greedy algorithm





Small subset of
a social network

Greedy score empirically close to optimal
if F is monotonic and submodular.

Why?

http://www.select.cs.cmu.edu/tutorials/icml08submodularity.html

Georgia Tech

Theorem on submodularity

Theorem [Nemhauser et al '78]

Suppose F is *monotonic* and *submodular*. Then greedy algorithm gives constant factor approximation:

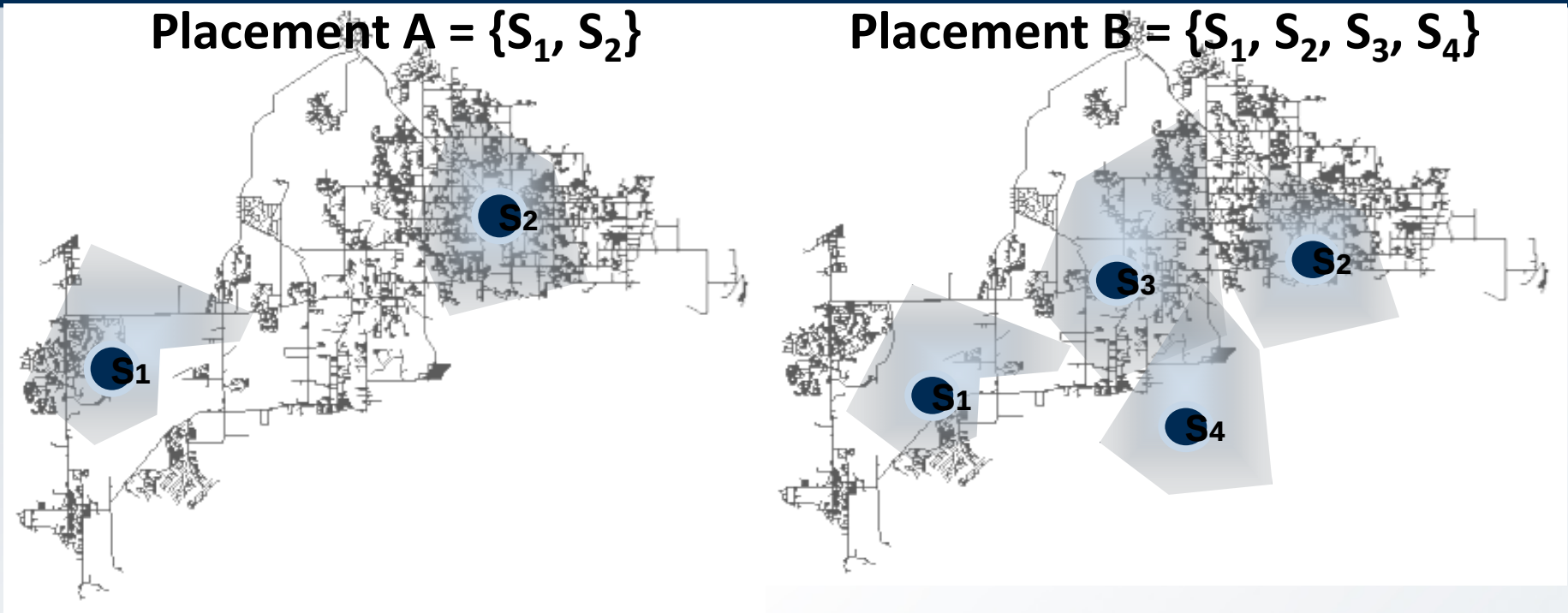
$$F(A_{greedy}) \geq \underbrace{(1 - 1/e)}_{\sim 63\%} \max_{|A| \leq k} F(A)$$

- Greedy algorithm gives near-optimal solution!
- In general, guarantees best possible unless P = NP!

G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, Maximizing submodular set functions, *Mathematical Programming*, 1978.

Georgia Tech

Key property 1: Monotonicity



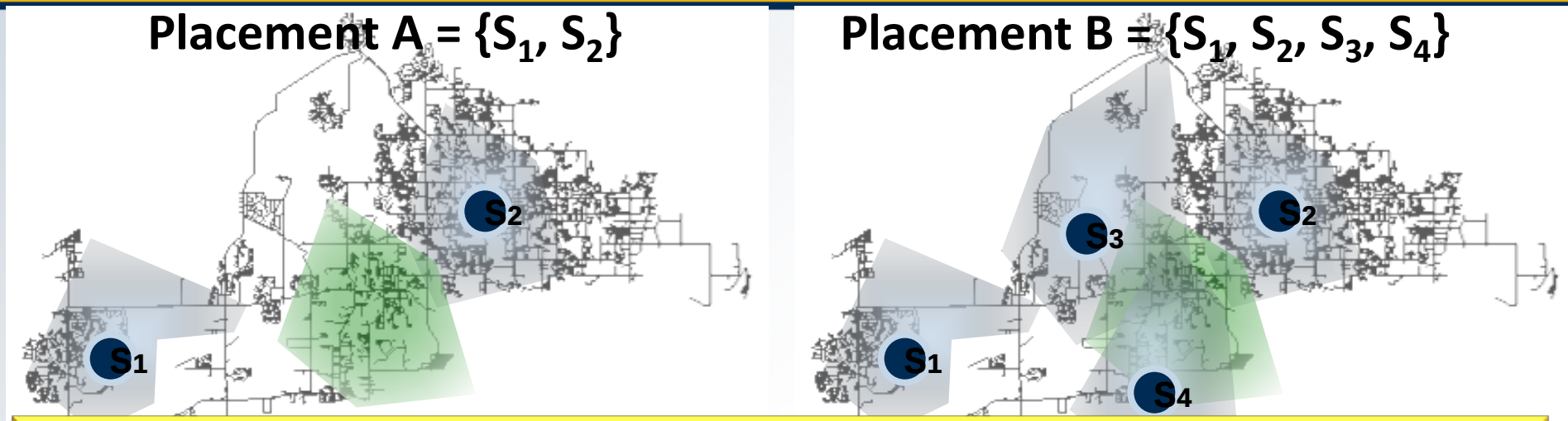
F is monotonic: $\forall A \subseteq B : F(A) \leq F(B)$

Adding sensors can only help

Source: Guestrin, Krause@IJCAI 2009, www.select.cs.cmu.edu/tutorials/icml08submodularity.html

Georgia Tech

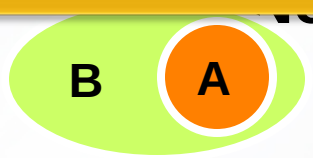
Key property 2: Diminishing returns



Theorem: [Krause, Leskovec, G., Faloutsos, VanBriesen '08]

sensing quality F(A) in water networks is submodular!

Submodularity:



\+ s' **Large improvement**

\+ s' **Small improvement**

$$\forall A \subseteq B, s' \notin B : F(A \cup \{s'\}) - F(A) \geq F(B \cup \{s'\}) - F(B)$$

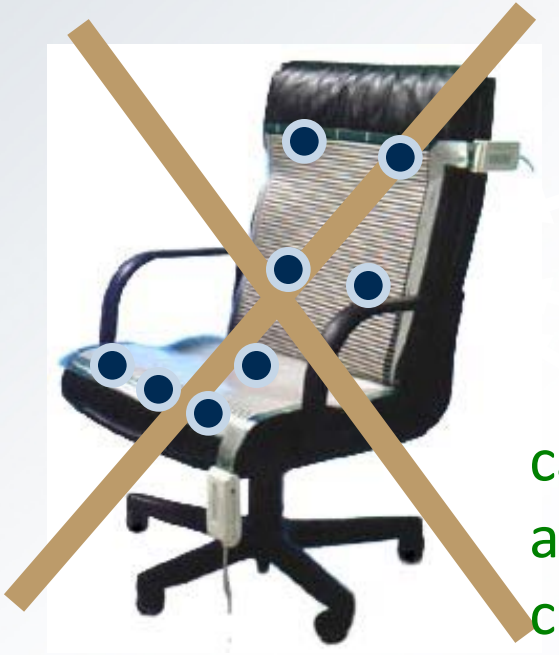
Source: Guestrin, Krause@IJCAI 2009

Georgia Tech

Building a sensing chair

[Mutlu, Krause, Forlizzi, G., Hodgins '07]

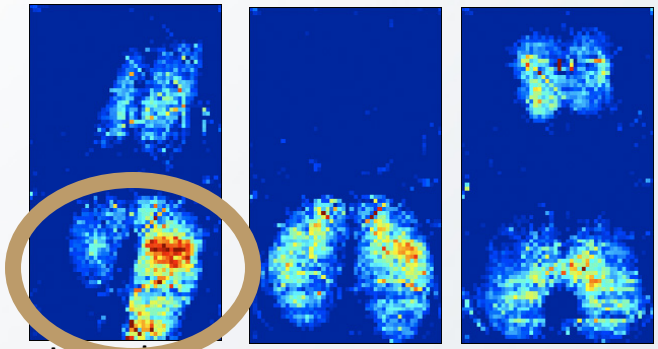
- People sit a lot
- Activity recognition in assistive technologies
- Seating pressure as user interface



equipped with
1 sensor per cm^2!

sheet costs $6,000!
☹

can we get similar accuracy with fewer, cheaper sensors?



82% accuracy on 10 postures! [Tan et al]

Georgia Tech

How to place sensors on a chair?

- Predict posture Y; Possible locations V
- Goal: **minimize uncertainty** in prediction ⇔ **maximize information gain** (IG):

$$\mathcal{A}^* = \operatorname*{argmax}_{|\mathcal{A}| \leq k} \mathrm{IG}(Y; \mathcal{A})$$

Possible locations

Theorem: information gain is submodular!* [Krause, G. '05]

Placed sensors, did a user study:

	Accuracy	Cost
Before	82%	**$6,000** ☹
After		

random placement: 53%; uniform Placement: 73%

similar accuracy at <2% of cost!

Analytics As a Service: Discrete Optimization

finding the right information

Finding the Most influential people

monitoring algal blooms

predicting postures

detecting contaminations

reducing energy consumption

each app requires its own optimization
→ designing optimization case by case
→ time consuming, doesn't scale! ☹

Machine Learning (Theory)

Slashdot

boingboing

sisu

engadget

Tech

The quest for programming model: treating the class of optimization problems as a service

Finding top K Most influential people

predicting postures

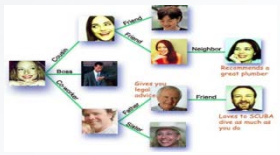
Detecting contaminations

reducing energy consumption

finding the right information

Programmable algorithm abstractions for a class of discrete optimization problems

Greedy Approximation with Theoretical Guarantee



Machine Learning (Theory)

Slashdot

boingboing

sisu

engadget

Georgia Tech

Reuse Opportunity: Programmable algorithm abstractions

detecting contaminations

predicting postures

monitoring algal blooms

reducing energy consumption

finding the right information

Sensing quality

Sensing locations

$$\max_{\mathcal{A} \subseteq \mathcal{V}} F(\mathcal{A})$$

Sensing cost

Sensing budget

$$\text{subject to } C(\mathcal{A}) \leq B$$

Machine Learning (Theory)

Slashdot

boingboing

sisu

engadget

Georgia Tech

Big Data Analytics as a service: Approximation

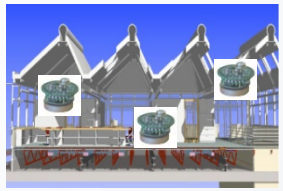
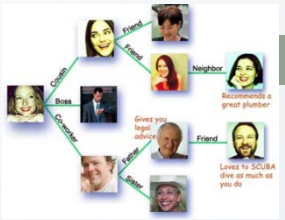
Theoretical:

Approximation algorithms that have theoretical guarantees and scale to large problems

Applied:

Unified Computing infrastructure to support real deployments of both large datasets and programmable algorithm abstractions



Machine Learning (Theory)

Slashdot

boingboing

sisu

engadget

Georgia Tech

Fundamental Data Analytic Problems

- What information should I use to perform analytics ?
- How should I perform analytics efficiently ?

Connecting Big Data with Data Analytics Opportunity 2:

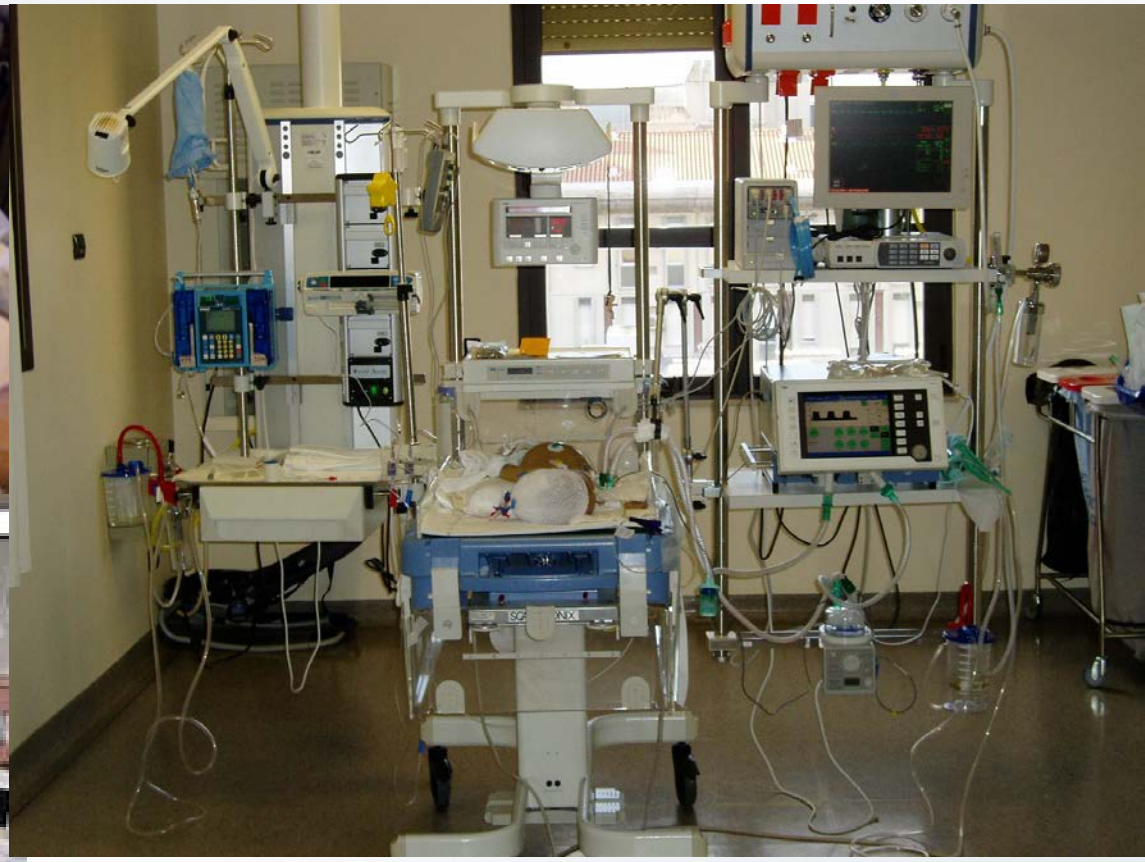
Explore data utility in domain specific data analytics tasks for cross-domain fertilization of big data analytics

IBM's Jeopardy machine

- IBM's new question-answering system → Dr. Watson,
  - won an actual Jeopardy competition on US national TV. Smarter Than You Think
- **Argument/Hypothesis**
  - No single algorithm can simulate the human ability to parse language and facts.
- **Approach/Methodology**
  - **Multi-Channel Analysis and Learning**: Watson uses more than a hundred algorithms at the same time to analyze a question in different ways, generating hundreds of possible solutions.
  - **Majority Voting and Ranking**: Another set of algorithms ranks these answers according to plausibility; for example, if dozens of algorithms working in different directions all arrive at the same answer, it's more likely to be the right one.

Time Series Correlation Analysis

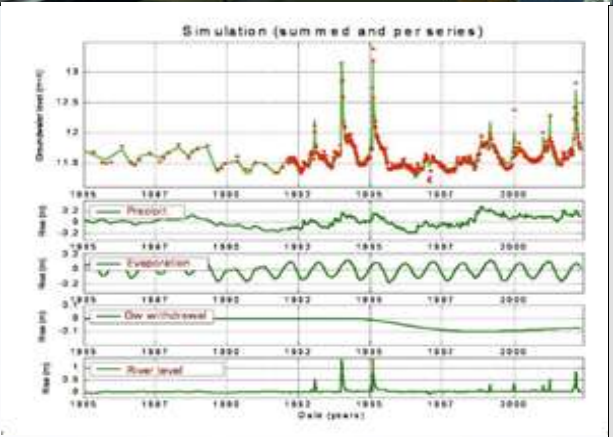
- **ICU Patients Wellness Prediction**

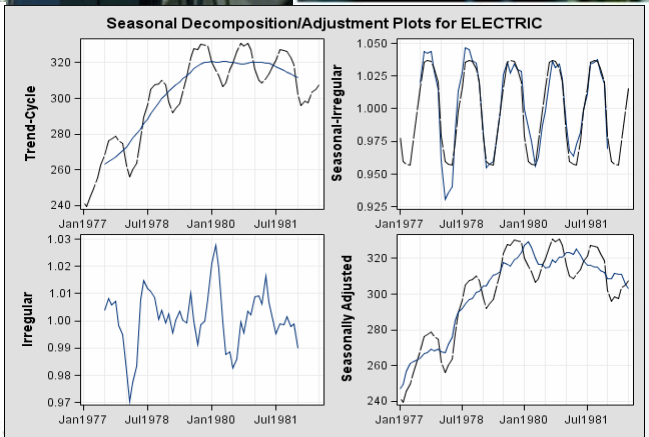


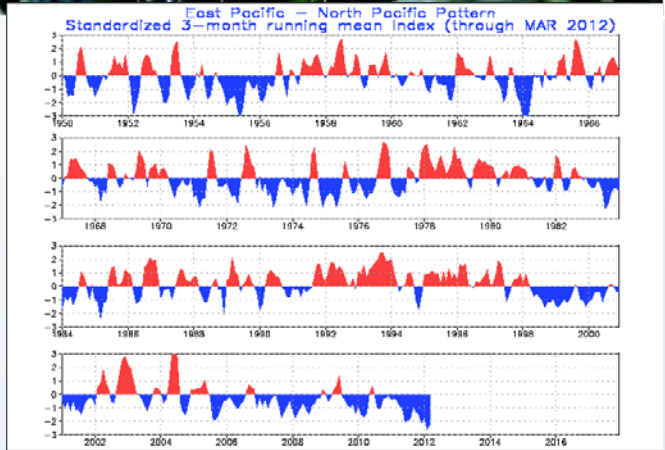
Georgia Tech

Time Series Correlation Analysis

- ICU Patients Wellness Prediction

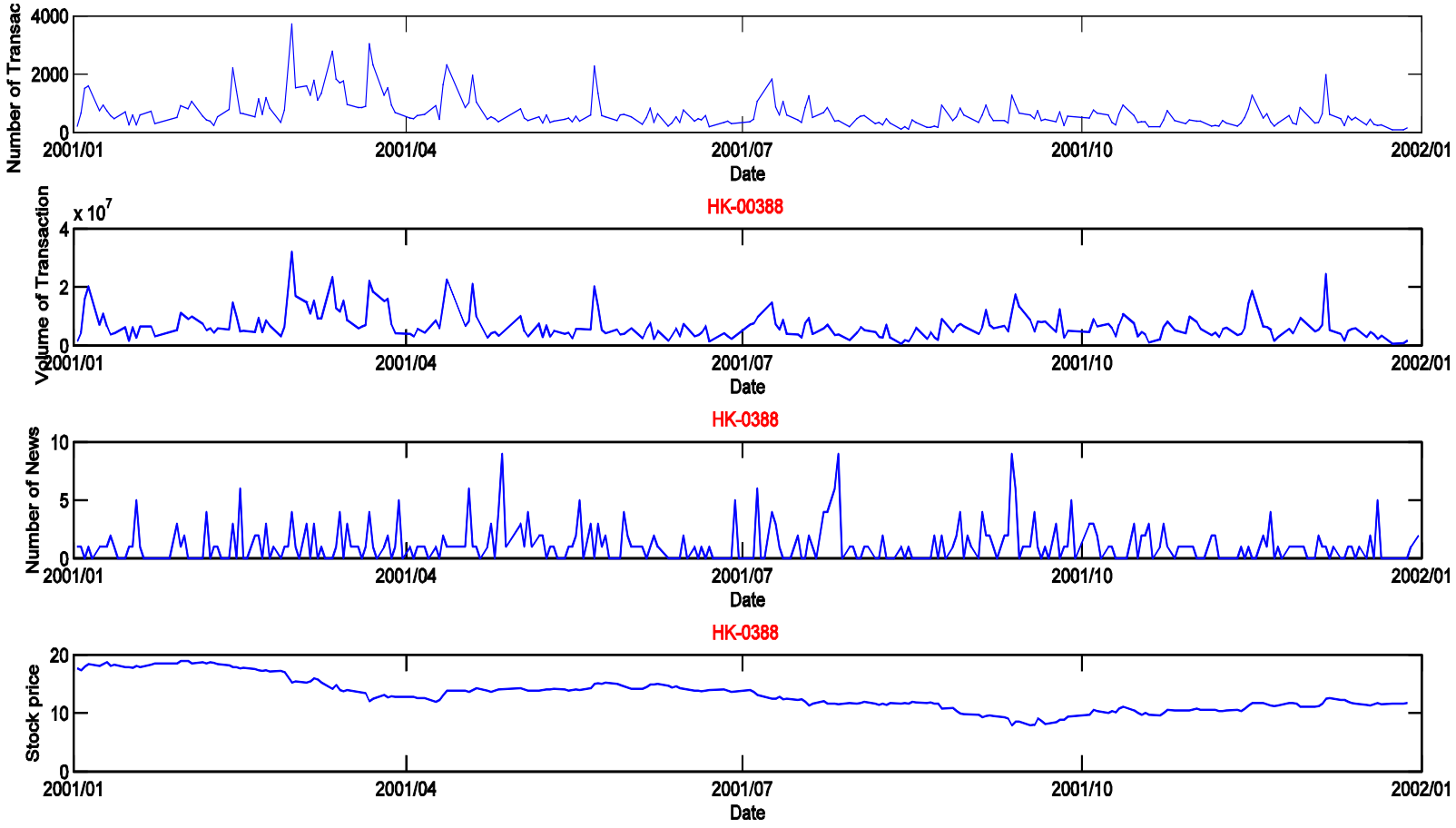






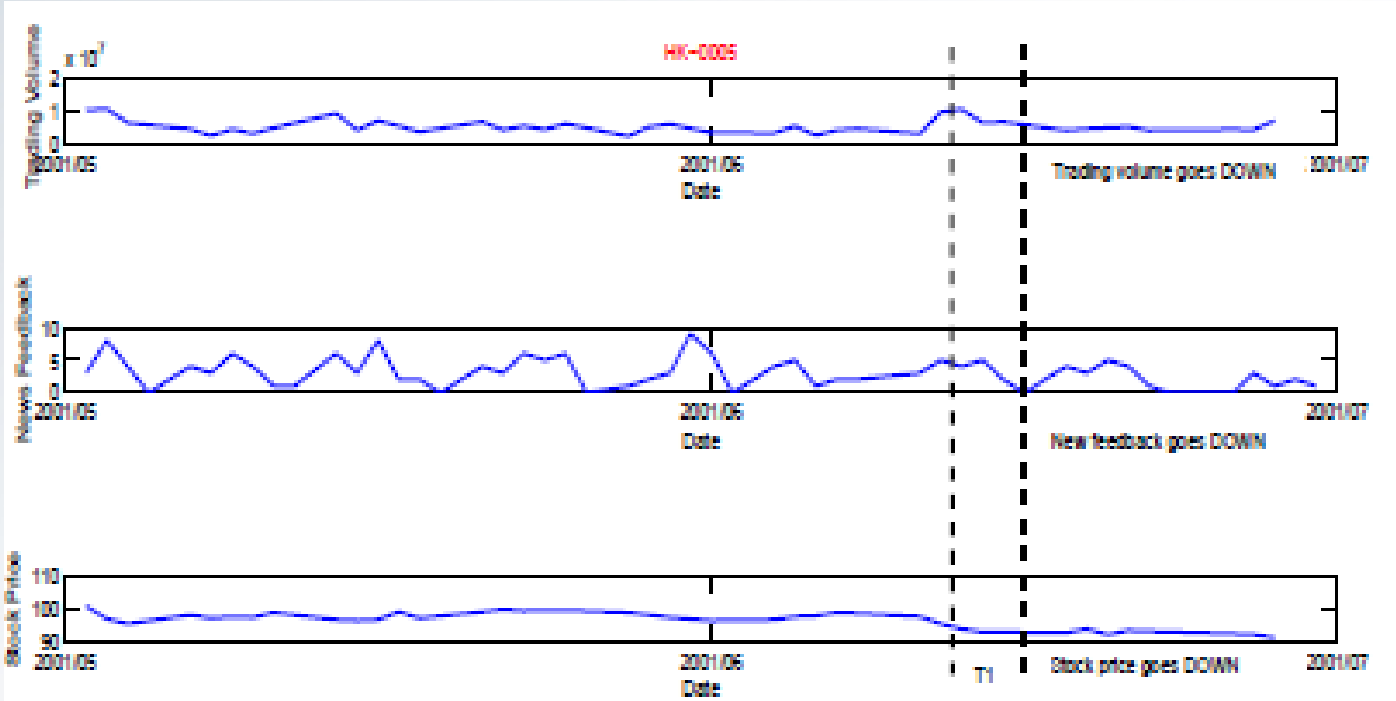
Time Series Correlation Analysis

- Stock Market Volatility Prediction
  - Multiple Learning Kernels: Historical price, News, Transaction Volume
  - Prediction of future market trend: Up, Down, Hold



Time Series Correlation Analysis

- Stock Market Volatility Prediction (**Easy case**)

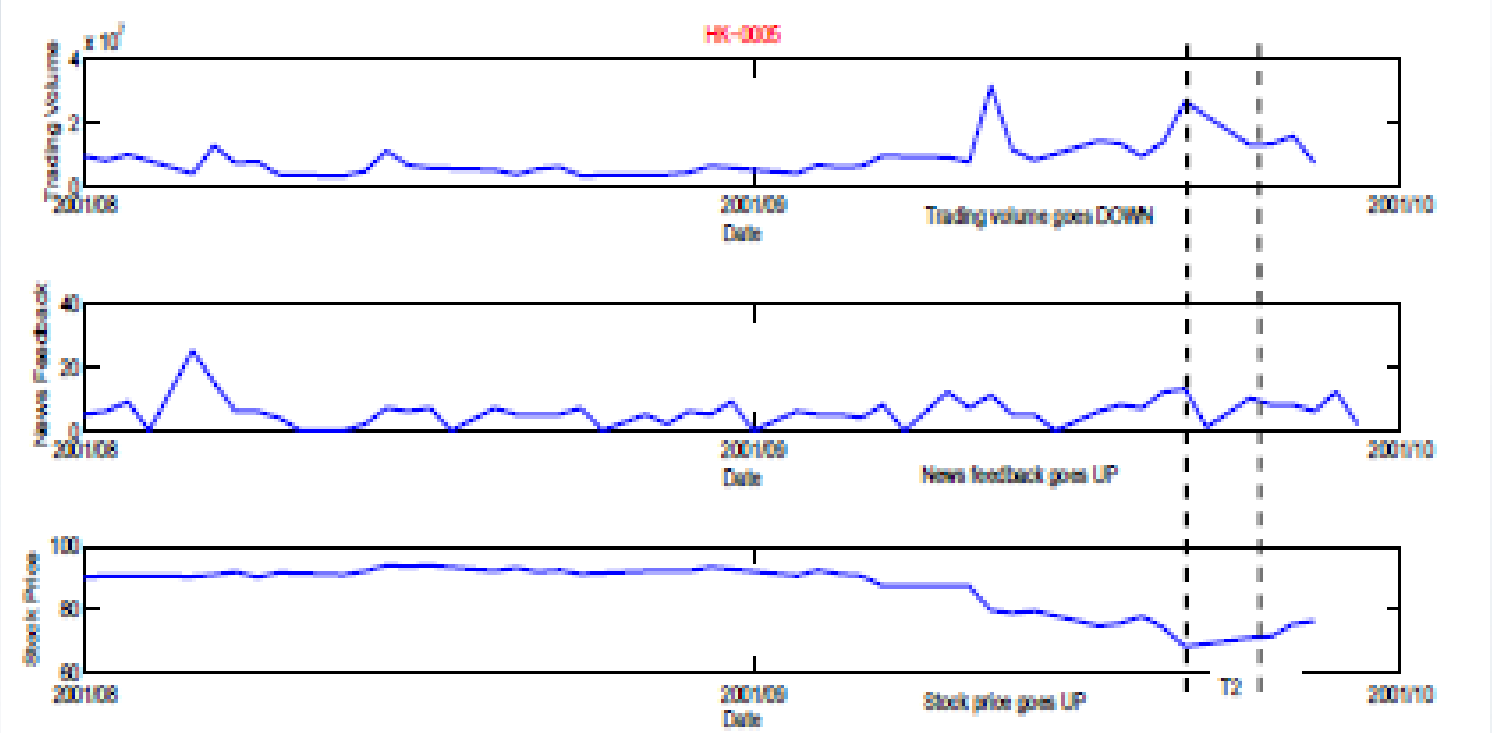


(a) Stock 0005.HK (2001/05-2001/07)

27

Time Series Correlation Analysis

- Stock Market Volatility Prediction (**Hard case**)



(b) Stock 0005.HK (2001/08-2001/10)

28

Observation

- Two different domains + Two different Datasets
- Same Category of Data: Time Series
- Same Learning Objectives: Multi-model classification-based learning
- Advances in one can stimulate advances in the other

Information Networks (Graphs) are everywhere

- Web graph
- Social networks
- Biological networks
- Internet communication networks
- Knowledge networks (e.g., RDFs)

NETFLIX

user-movie ratings graph

ENCODE

DNA interaction graph

flickr

6 Billion
Flickr Photos

facebook

Social Graph
750 Million Users
140 billion Connections

24 Million
Wikipedia Pages

YouTube

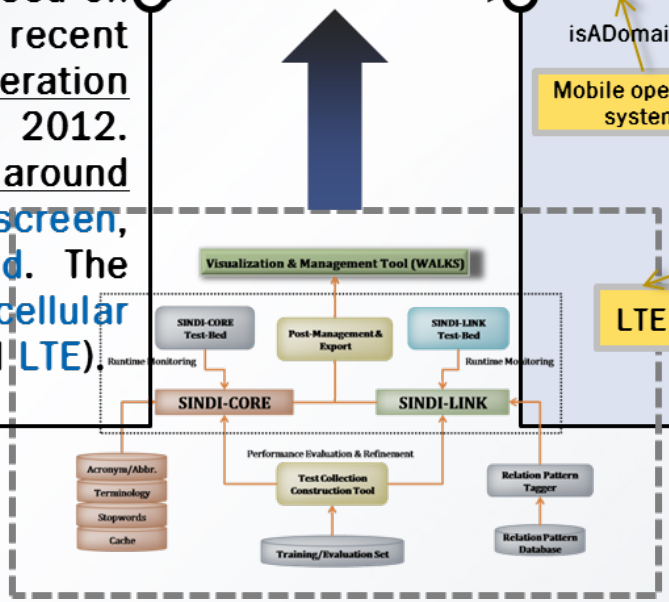
48 Hours a Minute
YouTube

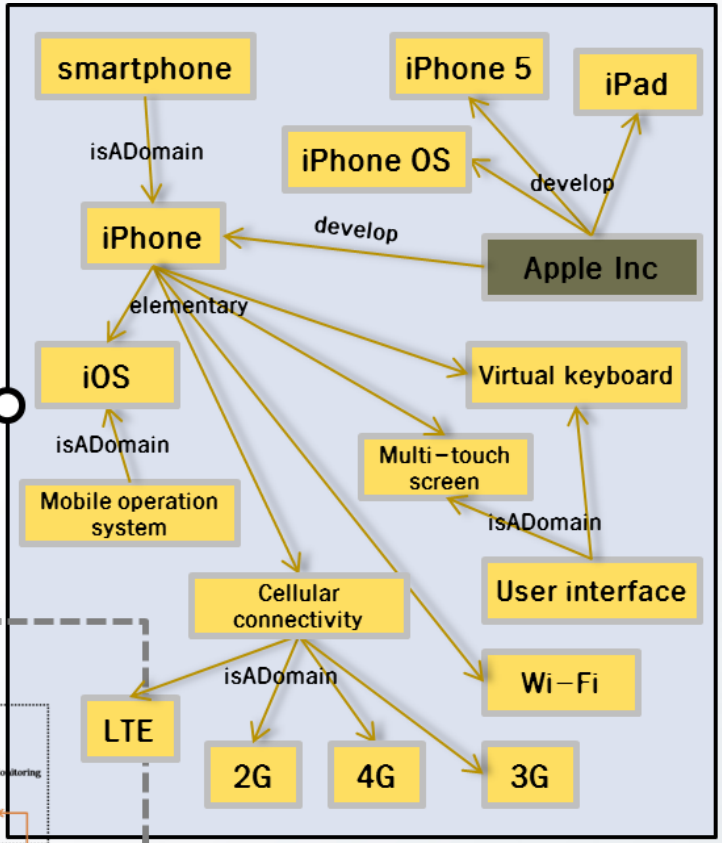
Georgia Tech

Semantic Analysis using RDF Graph

The iPhone is a line of smartphones designed and marketed by Apple Inc. It runs Apple's iOS mobile operating system, known as the "iPhone OS" until mid-2010, shortly after the release of the iPad. The first generation iPhone was released on June 29, 2007; the most recent iPhone, the sixth-generation iPhone 5, on September 21, 2012. The user interface is built around the device's multi-touch screen, including a virtual keyboard. The iPhone has Wi-Fi and cellular connectivity (2G, 3G, 4G, and LTE).

Term Extraction & Semantic Relation

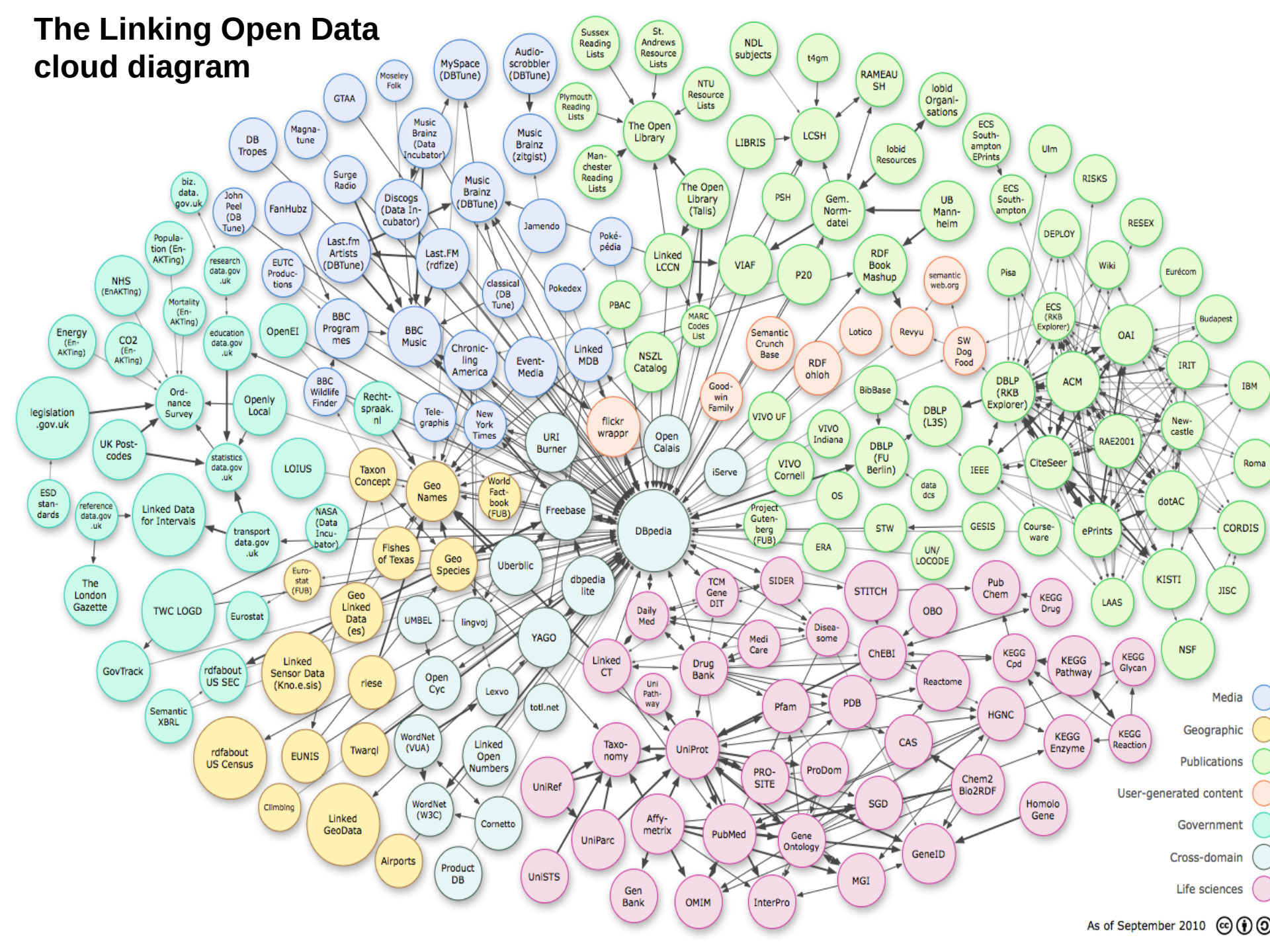




J. Gim, S. Lim, H. Jung: 2013 ISO/IEC JTC1 SC32

Georgia Tech

The Linking Open Data cloud diagram

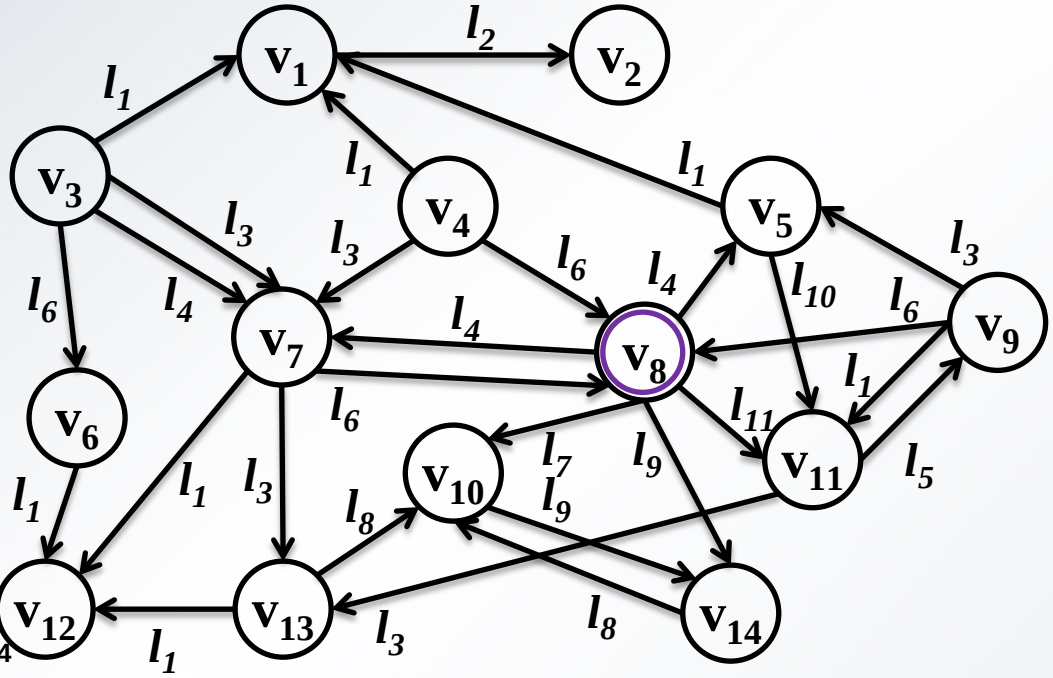


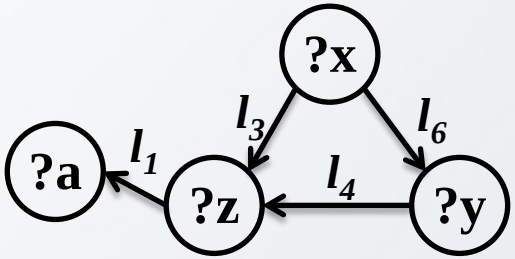
Characterizing Graph Computation

- **Two broad classes of problems:**
  - **Graph queries** to find matchings (e.g., subgraph matchings)
  - **Iterative Algorithms** to find clusters, orderings, paths, patterns, …
- **Graph Kernel**
  - traversal, shortest path algorithms, flow algorithms, spanning tree algorithms, topological sort, …
- **Many factors can influence the choices of graph analytic algorithms and performance optimization techniques**
  - graph sparsity (edge/vertex ratio), Static/dynamic, diameter, graph heterogeneity, weighted/unweighted (and weight distribution), vertex degree distribution, directed/undirected, simple/multi/hyper graph, problem size, granularity of computation (vertices/edges), problem-specific or domain specific characteristics

Graph Queries (Pattern matching)

- **Graph pattern queries** are subgraph matching problems
  - One of the most fundamental graph operations
- **Executing** a graph pattern query
  - Find a set of **subgraphs** in a given graph, which match the given graph query pattern if we can substitute the query variables with vertices and edges in the graph.
  - Variables are denoted by a prefix **"?"**





Q4: *(?x, l_3, ?z), (?x, l_6, ?y), (?y, l_4, ?z), (?z, l_1, ?a)*

Georgia Tech

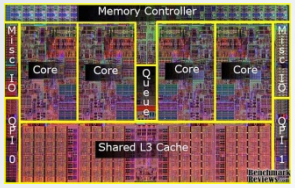
Scaling Graph Analysis

Common Techniques

- **Compression**
  - Compact storage on disk and compact data structure in memory
- **Data placement (disk, memory)**
  - Balance computation with storage
  - maximize sequential access and minimize random access to edges and/or vertices
- **Indexing (vertex, edge)**
  - utilizing sequential access to reduce unnecessary random access
- **Caching (multiple levels)**
  - Performance gain for repeated vertex/edge access
- **Parallel Computation (multiple levels)**
  - Multi-threads, Multi-cores, Disk and memory optimization, Cluster-computing
  - Minimizing parallel overhead (minimizing communications & maximize local computation)

Parallel Graph Processing: Growing Demand

- Wide array of different parallel architectures:



GPUs **Multicore** **Clusters** **Science Clouds** **Commercial Clouds**

- Different Graph Processing Algorithms
  - Subgraph matching (Graph Queries)
  - Finding patterns, rules, anomalies (Graph Mining)
  - Ranking web pages (Search Engine)
  - ‘Viral' or 'word-of-mouth' marketing (Social Influence)
  - Identifying interactions among proteins (Bioinformatics)
  - Detecting anomalies in email traffic (Computer security)

36

Processing Graph Queries: Challenges

- Graph datasets often exhibit **higher data correlations**
  - Entities (vertices) are **correlated** through both direct and indirect links (edges)
  - High **heterogeneity**
    - heterogeneous types of entities (vertices)
    - heterogeneous types of links (edges)
  - Highly **skewed distribution** (some high degree vertices, many low degree vertieces)

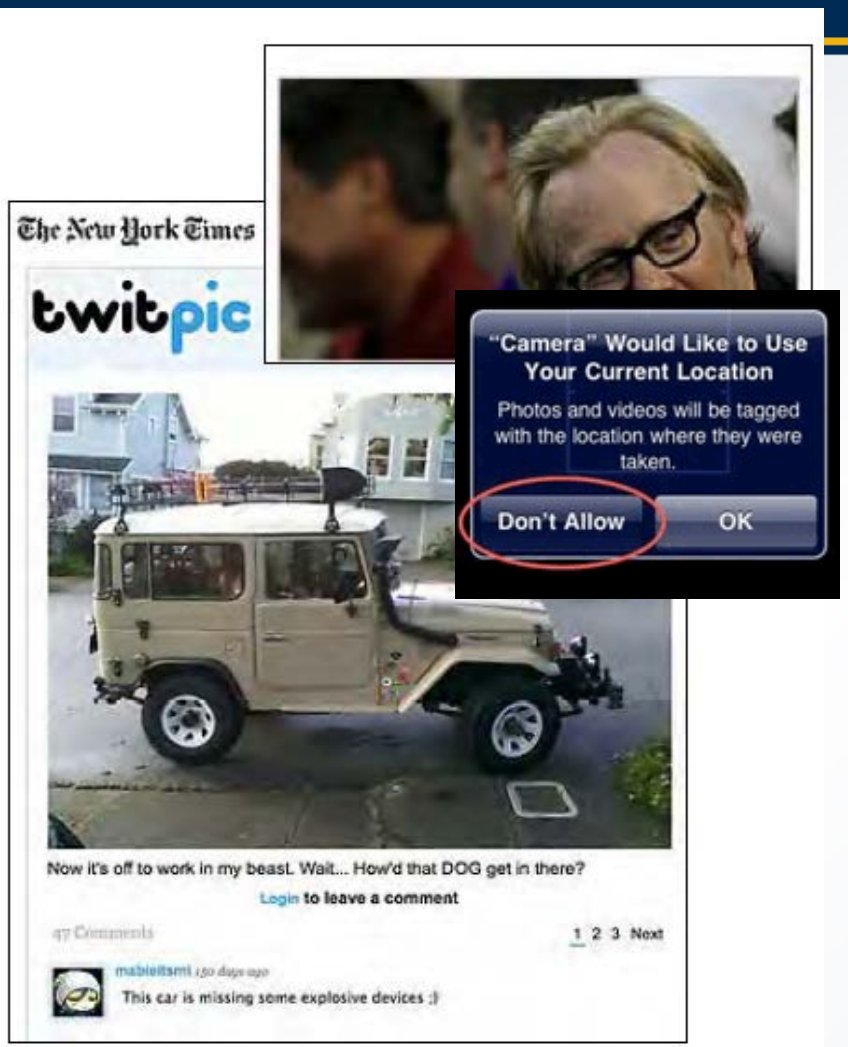
- Graph computations often **exceed the processing capacity** of conventional hardware, software systems and tools
  - Intermediate results size exceeds the available memory
  - Fail to deliver the computation within acceptable latency
    - Time complexity with respect to Disk IO, Network IO

Connecting Big Data with Data Analytics: Another Grant Challenge

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

- **Privacy at both Individual and Corporate/Enterprise Level**
- **Noise / Quality Control at input, computation and output phases of analytics**

Big Data Analytics v.s. Privacy Risks



Read the full story here: http://nyti.ms/917h

- In August of 2010, Adam Savage, of "Myth Busters," took a photo of his vehicle using his smartphone. He then posted the photo to his Twitter account including the phrase "Off to work."
- Since the photo was taken by his smartphone, the image contained metadata revealing the exact geographical location the photo was taken
- By simply taking and posting a photo, Savage revealed the exact location of his home, the vehicle he drives, and the time he leaves for work

Georgia Tech

Dean Compher, IBM 2013

Privacy: Is it a dark side of Big Data Era

- **What is the biggest threat to privacy (government, social media, advertisers, data brokers, ourselves)?**
  - Understanding Data utility based Privacy Risks
  - Having utility-driven privacy policies
  - Providing software for policy compliances
- **What are the most promising technological solutions for protecting privacy? What are the most over-hyped?**
- **Is privacy really dead? What would it take to resurrect it?**
  - The privacy defined by the amount of information shared is dead
  - The privacy should be defined by how our data should be used and protected

Questions

41