Big Data Systems and Analytics: Opportunities and Challenges

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

Theme of this Course

Big Data Systems Big Data Analytics

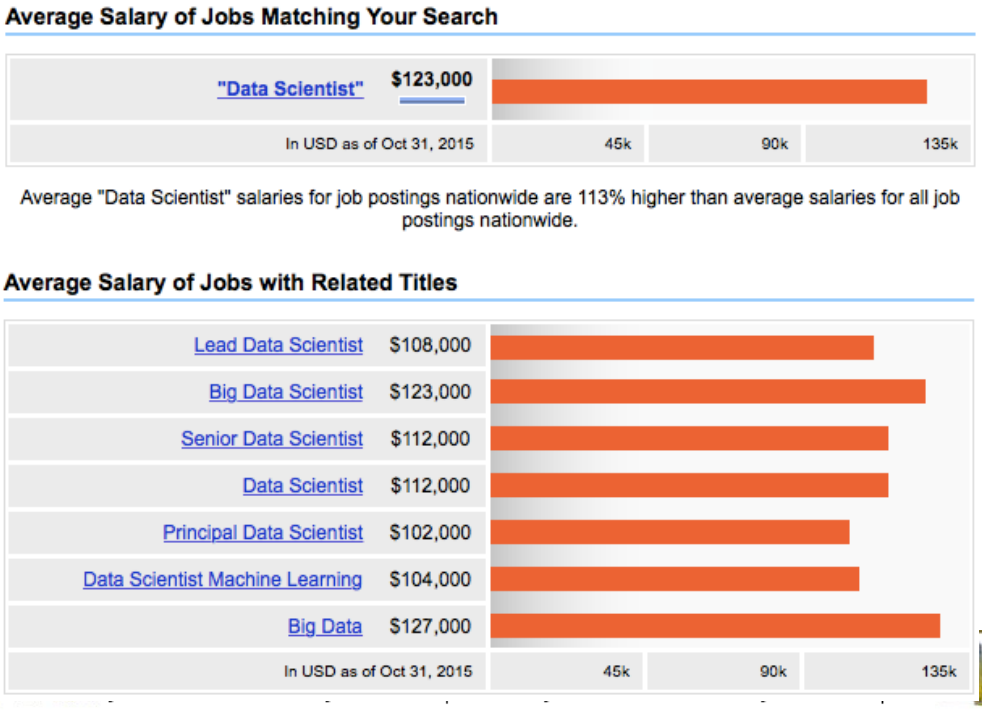
Scope of the Course

- How various parts of a big data computer system (hardware, software and applications) are put together?
- What are the appropriate approaches to realize high-performance, scalability, reliability, and security in practical big data systems?
- How to effectively manage very large amounts of data and extract value and knowledge from them

Administravia

- Course Website:
http://www.cc.gatech.edu/~lingliu/courses/8803/2016Fall/cs8803-BDS.html
- Office Hours
  - Thursdays 11am -12noon, or by appointment
- Grading
  - Class Participation 15%
  - Homework Assignments (4) 20%
  - Project 50%
  - Final (Technology Review) 15 %
- Announcements
  - In class and on T-Square

From Computer Programing to Data Scientists



College of Computing

Outline

- Big Data and Big Data Computing
- Big Data Challenges
- Innovation Opportunities in Big Data Computing
  - Programmable algorithm abstractions
  - Parallel processing abstractions
  - Leveraging Multiple Learning Models
- Challenges of Big Data
  - Data Privacy
  - Data Quality

Georgia Institute of Technology

Characteristics of Big Data Sets

- Huge
- Distributed
  - Dispersed over many servers
- Dynamic
  - Items add/deleted/modified continuously
- Heterogeneous
  - Many agents access/update data
- Noisy
  - Inherent
  - Unintentional
  - Malicious
- Unstructured / semi-structured
  - No database schema

Big Data Puzzles

What is Big Data?

What makes data, "Big" Data?

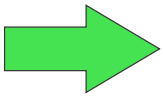
Why Big Data now?

What is Big Data: A CS Perspective

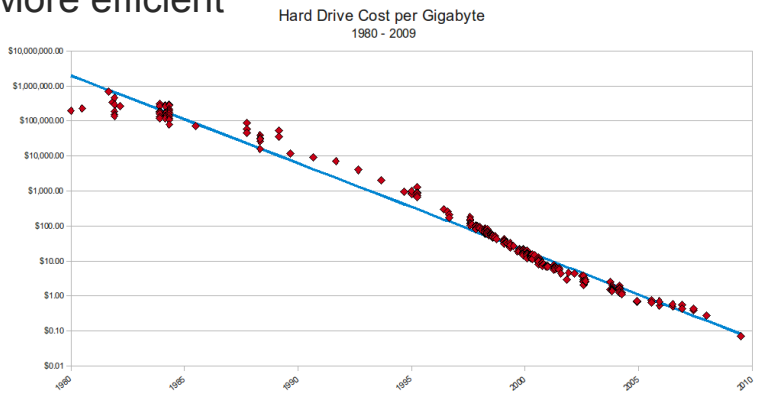
- Big data refers to datasets
  - that are *beyond the ability of legacy approaches* to manage at an acceptable level of quality and / or
  - that *exceed the capacity of conventional systems* (hardware and/or software) to process within an acceptable elapsed time.
- Definition of big data: Subjective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

Why Big Data now:
Better storage technology

- Storage & disks
  - Cheaper
  - More volume
  - Physically smaller
  - More efficient



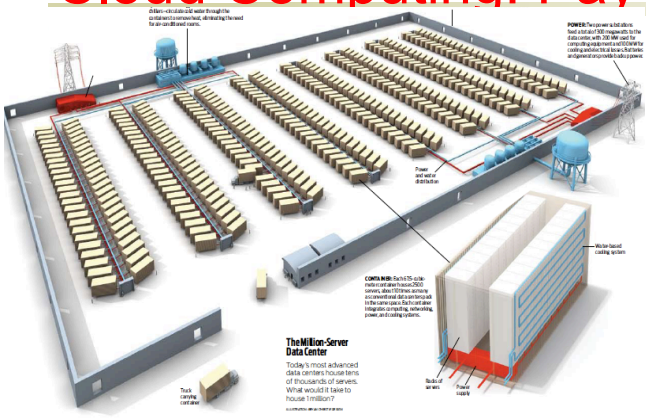
Large data sets are affordable



How many songs a TB disk can hold?

256KB/sec
4min/song
190K songs

Why Big Data now:
Cloud Computing: Pay per use

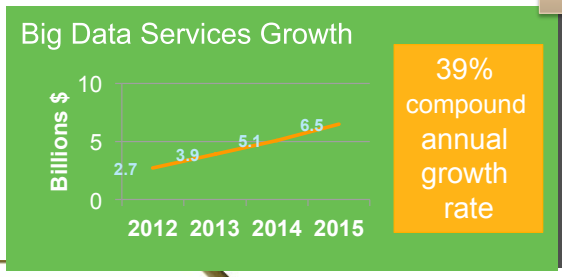


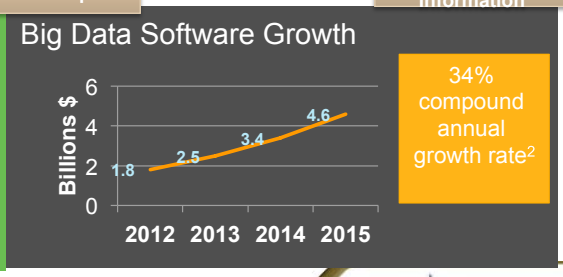
- Pay-as-you-go
- Elasticity
- Multi-tenancy
- Economics of Scale

More afforable to perform big data analytics

Georgia Tech College of Computing

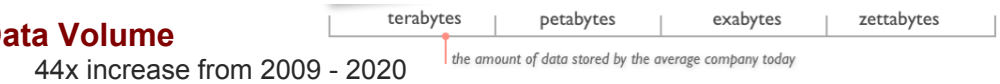
Why now → Data grows faster than intelligence

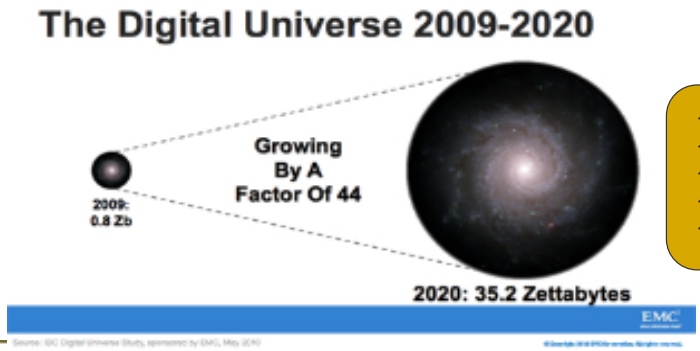




Characteristics of Big Data (1): Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 - 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





1 terabyte (TB) =1,024GB ~10^3GB
1 petabyte (PB) =1,024TB ~ 10^6GB
1 exabyte (EB) =1,024PB~10^9GB
1 zettabyte (ZB) =1,024EB~10^{12}GB.
1 yottabyte (YB) =1,024ZB~10^{15}GB.

Big Data: The Volume Challenge

- Challenges for sensing, collection, generation of more data, more meaningful data, more useful data
- Challenges for computation on data bigger in volume
- Challenges for communication on data bigger in volume
- Challenges for analysis (algorithms) on data bigger in volume
- How would we design **volume adaptive** big data processing **framework, systems** and **services**?

Characteristics of Big Data (2): Speed (Velocity)

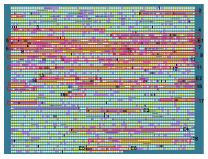
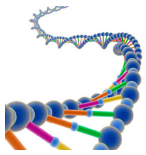
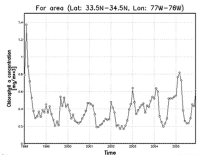
- Data are generated fast and need to be processed fast
- Online Data Analytics + Real time Analytics
- Delayed decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

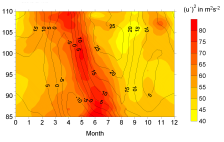
Big Data: The Velocity Challenge

- Challenges for collecting sensor data in real time (fast rates of data generations and consumption, regarding data input sources, data output rate, varying rates to sense and collect data, the varying rates of data output
- Challenges for analyzing sensor data in real time: how fast to analyze, and its implication/ impact.
- What would be **velocity adaptive** big data processing **framework, systems** and **services**?

Characteristics of Big Data (3): Complexity (Varity)

- Input data related complexity
  - Various formats, types, and structures
  - Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
  - Static data vs. streaming data
  - A single application may be generating/collecting many types of data



To extract knowledge➔ all these types of data need to linked together

Characteristics of Big Data (3): Complexity (Varity)

- Algorithms related complexity
  - Different algorithms may derive different insights/ outputs from the same collection of datasets
  - What analysis does make sense?
- Metrics (Quality and Performance measurement)
  - Different measures can be used to measure performance, effectiveness, and utility of big data
  - **What metrics do make sense?**

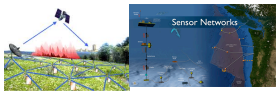
To leverage different algorithms and metrics ➔ derive high quality value and data utility

Many Sources Generate Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation are no longer hindered by the ability to collect data; but by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

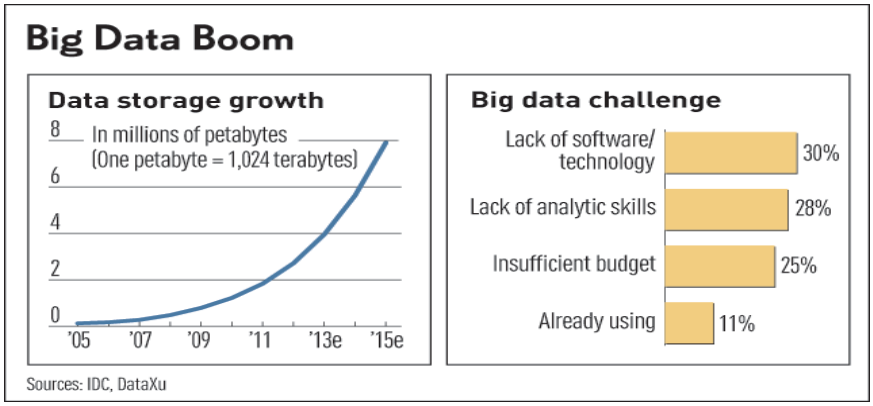
The Evolution of Data Generation Models

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: everyone is generating data and consuming data

Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed in data collection, storage, processing, data center management, data security and privacy …
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

The 'Big Data' Phenomenon

Big Data Drivers:

- The proliferation of data capture and creation technologies
- Increased "interconnectedness" drives consumption (creating more data)
- Inexpensive storage makes it possible to keep more, longer
- Innovative software and analysis tools turn data into information

Big Data encompasses not only the **content itself**, but **how it's consumed.**

- Every **gigabyte** of stored content can generate a **petabyte** or more of transient data*
- The information about you **is much greater** than the information you create

*Source: IDC 2011, Cloudera 2011

Big Data Challenges

It's not just about "big"

- Cost-effectively managing the **volume, velocity and variety** of data
- Deriving value across **structured and unstructured** data
  - Flexible and effective inference/analytics/correlation analysis techniques
  - Time series, histograms, graphs
- Adapting to **context changes** and integrating **new data sources and types**
  - Seamlessly adaptive, highly available
  - Failure tolerant, highly reliable

New challenges
The Scalability Dilemma

- State-of-the Art Machine Learning techniques do not scale to large data sets.
  - Memory constraints
  - Network bandwidth constraints
- Data Analytics frameworks can't handle lots of incomplete, heterogeneous, dirty data.
- Processing architectures struggle with increasing diversity of programming models and job types.
- Goal driven feature extractions demand a lot more innovations on learning and modeling techniques
- …

Abstract model of computing

$f : X^n \rightarrow Y$ (n is very large)

$x \in X^n$ **Data**

Approximation of $f(x)$

- **Approximation of f(x) is sufficient**
- **Program can be randomized**

Examples

Mean

$X = [0, 1]$,
$f(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$

Parity

$X = \{0, 1\}$,
$f(x) = \sum_{i=1}^{n} x_i \bmod 2$

Models for Computing over Big Data

- Optimization with
  - Random sampling
  - Data Streams
  - Sketching
  - Discrete Optimization with submodular function
  - Correlation Analysis (if time permits)
- Parallel Data/Computation Partition
  - Horizontal partition (range/row partition)
  - Vertical partition (column partition)
  - Hash partition
  - Graph partition

Random Sampling

$f : X^n \rightarrow Y$ (n is very large)

$x \in X^n$ 

Examples

Mean

O(1) queries

Parity

n queries

Query a few data items

Approximation of $f(x)$

Random Sampling

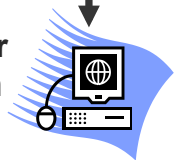
- Advantages
  - Ultra-efficient
    - Sub-linear running time & space (could even be independent of data set size)
- Disadvantages
  - May require random access
  - Doesn't fit many problems
  - Hard to sample from disorganized data sets

Data Streams

$f : X^n \rightarrow Y$ (n is very large)

$x \in X^n$ **Data**

Computer Program



Examples

Mean

O(1) memory

Parity

1 bit of memory

Stream through the data; Use limited memory

Approximation of $f(x)$

Data Streams

- Advantages
  - Sequential access
  - Limited memory
- Disadvantages
  - Running time is at least linear
  - Too restricted for some problems

Sketching

$f : X^n \times X^n \rightarrow Y$ (n is very large)

Sketch1 Data1

$x \in X^n$

Sketch2 Data2

$y \in X^n$

Compress each data segment into a small "sketch"

Compute over the sketches

Approximation of $f(x, y)$

Examples

Equality

O(1) size sketch

Hamming distance

O(1) size sketch

L_p distance (p > 2)

$\Omega(n^{1-2/p})$ size sketch

Sketching

- Advantages
  - Appropriate for distributed data sets
  - Useful for "dimension reduction"
- Disadvantages
  - Too restricted for some problems
  - Usually, at least linear running time

Discrete Optimization

Many discrete optimization problems can be formulated as follows:

Given finite set V, we want to select a subset A (subject to some constraints) maximizing utility F(A).

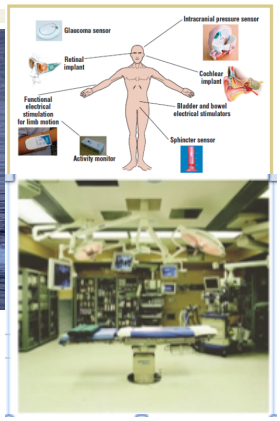
$$\max_{A \subseteq V} F(A)$$

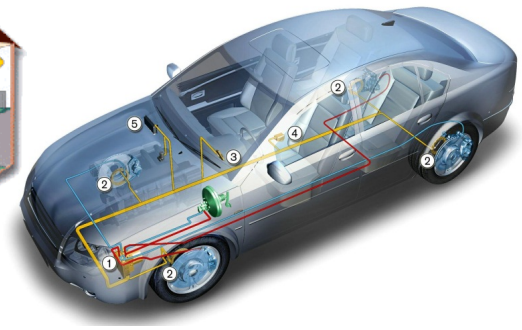
Opportunities for Big Data Analytics

reliance on non-renewable resources

pollution

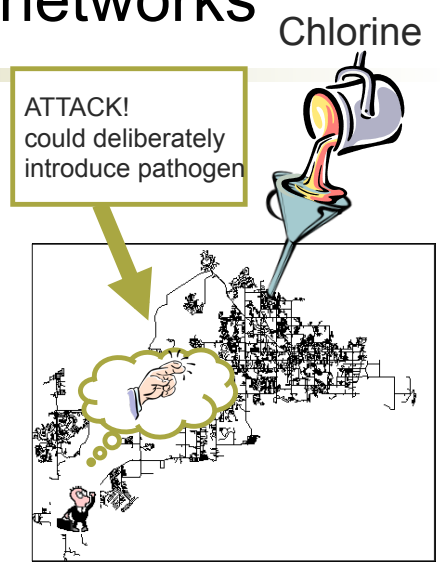
global warming





Water distribution networks

- Water distribution in a city → very complex system
- Pathogens in water can affect thousands (or millions) of people
- Currently: Add chlorine to the source and hope for the best



Simulator from EPA

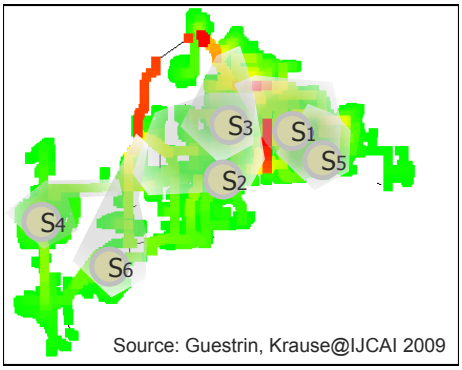
Source: Guestrin, Krause@IJCAI 2009

Sensor placement

Given: finite set V of locations, sensing quality F

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

NP-hard!



Source: Guestrin, Krause@IJCAI 2009

Greedy algorithm:

Start with A = {}

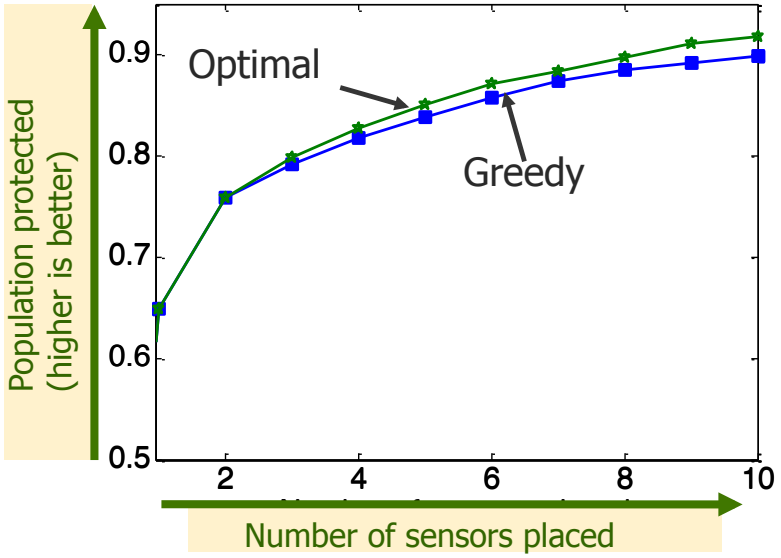
For i = 1 to k

s* := argmax_s F(A U {s})

A := A U {s*}

How well can this simple heuristic do?

Performance of greedy algorithm



Small subset of Water networks data

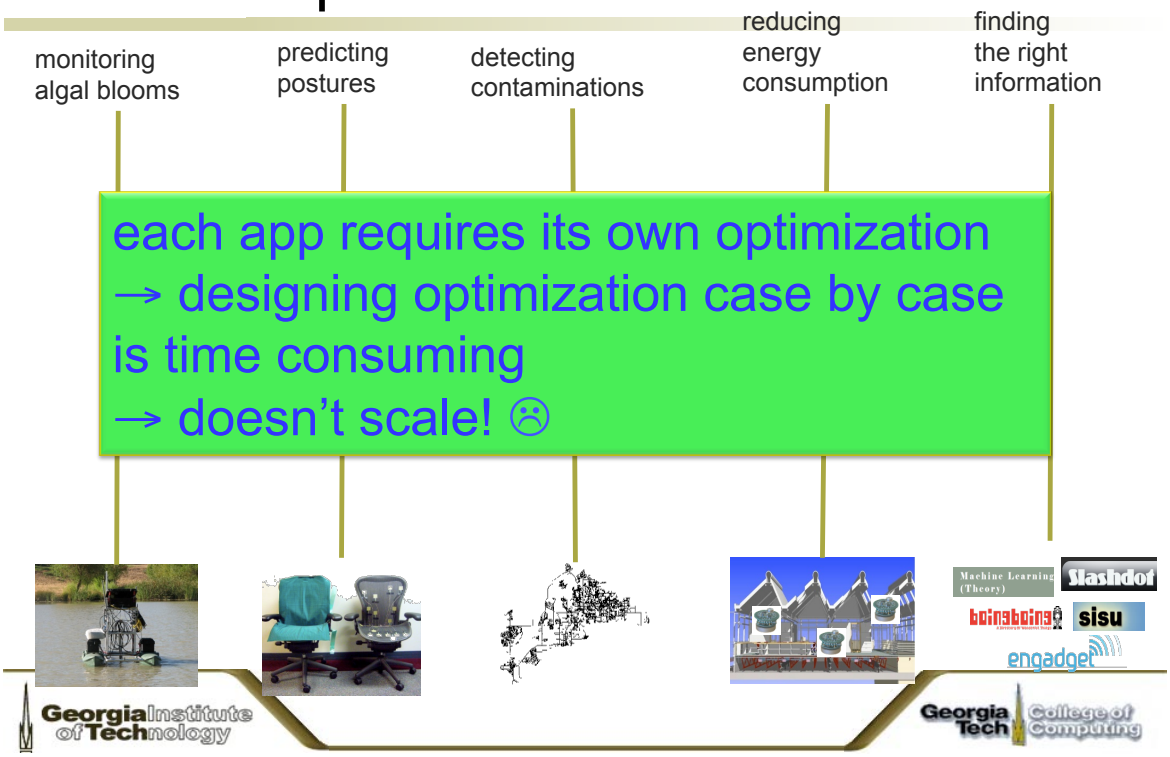
Greedy score empirically close to optimal. Why?

One reason submodularity is useful

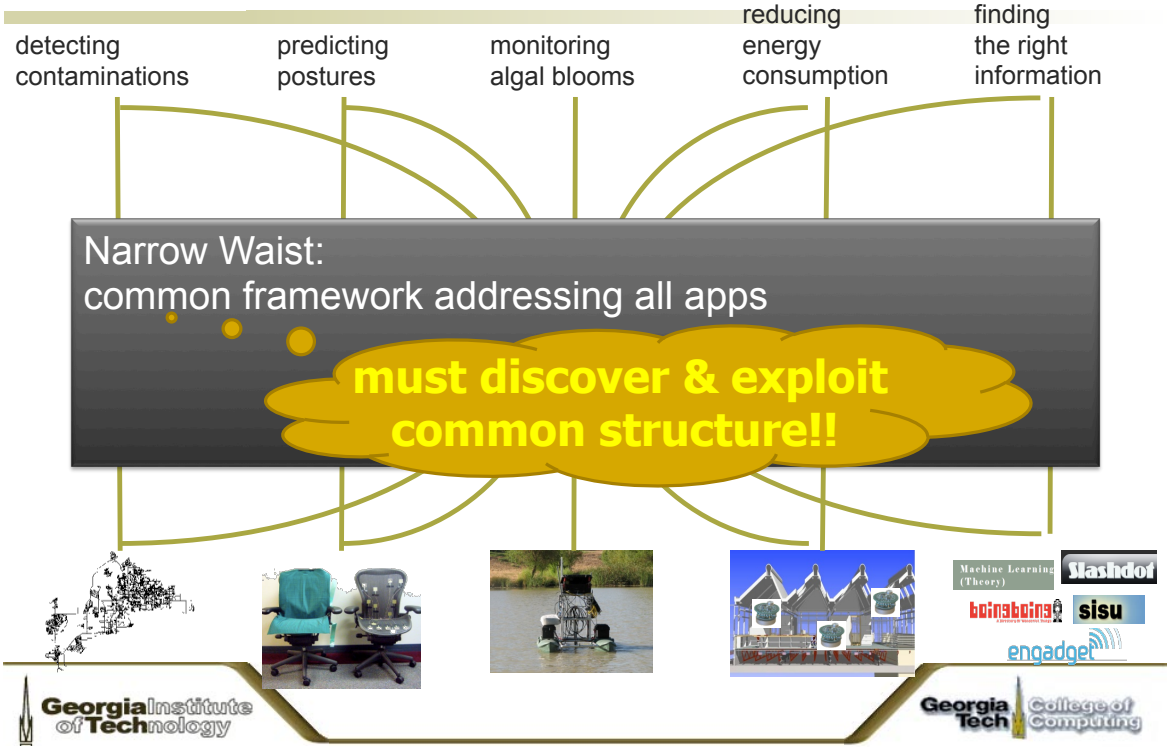
Theorem [Nemhauser et al '78]
Suppose F is *monotonic* and *submodular*. Then greedy algorithm gives constant factor approximation: $F(A_{greedy}) \geq \underbrace{(1-1/e)}_{\sim 63\%} \max_{|A| \leq k} F(A)$

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

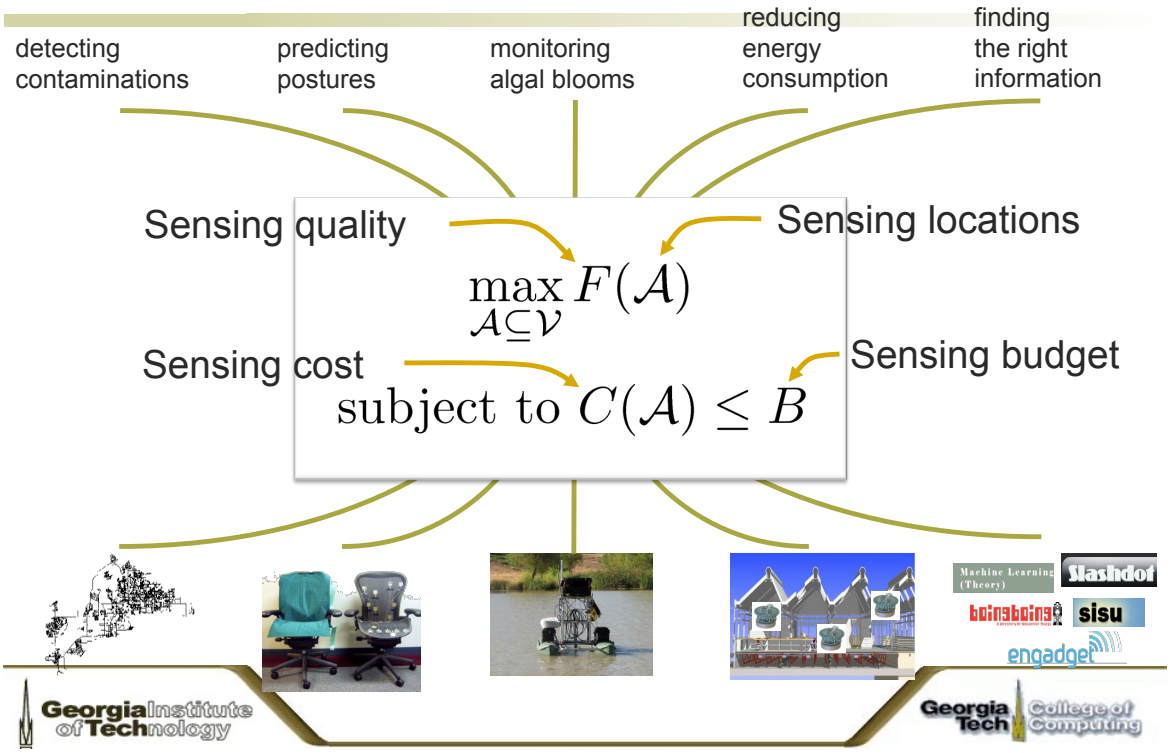
Many big data applications use discrete optimization…



The quest for the optimization narrow waist



The quest for the optimization narrow waist



More Examples

- ❑ **Finding top k most influential people in a large social network**
- ❑ **Finding top k most interesting blogs from sea of blogs**
- ❑ **Finding top k best matches of keyword search**

Social Influence: Problem Setting

- Given
  - a limited budget B for initial advertising (e.g. give away free samples of product)
  - estimates for influence between individuals
- Goal
  - trigger a large cascade of influence (e.g. further adoptions of a product)
- Question
  - Which set of individuals should B target at?
- Application besides product marketing
  - spread an innovation
  - detect stories in blogs

Keyword search is not enough

still too much information

top responses ecessarily need

doesn't highlight any structure in the space

Approximation Optimization: Analytics of the Web Data

77% read blogs
184M blogs
[Universal McCann, '08]

Slashdot engadget boingboing

TIME
IN PARTNERSHIP WITH CNN

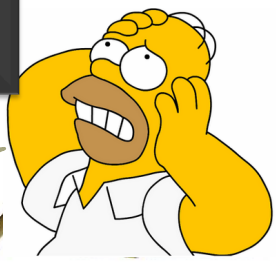
Thursday, Nov. 20, 2008

How Many Blogs Does the World Need?

By Michael Kinsley

you have 10 minutes a day to read blogs / news

which of the million blogs should you read?



information overload!!! ☹

Georgia Institute of Technology

Information Cascades

[Leskovec, Krause, G., Faloutsos, VanBriesen, Glance '07]

Slashdot | Machine Learning (Theory) | boingboing | sisu | engadget

time

learn about story after us!

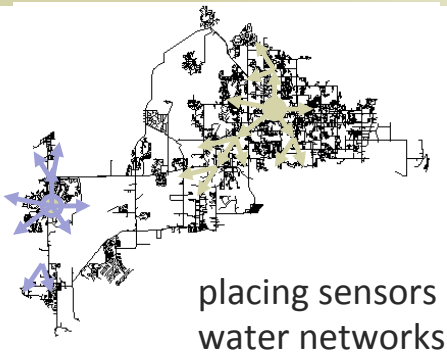
information cascade

a good blog learns about big cascades early on

Georgia Institute of Technology

Georgia Tech College of Computing

Water vs. Web



placing sensors in water networks

vs.

selecting informative blogs

want to pick nodes to **detect big cascades early**

in both apps, utility functions submodular ☺

Georgia Institute of Technology

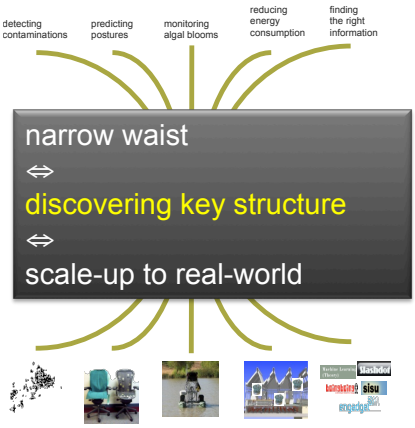
The power of the efficient narrow waist

constrained maximization of submodular functions

placing *k* sensors	robust sensing	complex constraints	sequential sensing
finding needles in haystack	better generalization	better use of "attention"	personalization

same algorithms new insights

Structural insights: challenges of next decade



past 30 years

independence	convexity
• Bayes nets, Markov nets, CRFs, PRMs…	• Kernel machines, SVMs, GPs, MLE…

next 30 years?

Big Data Analytics as a Service ☺

The basic foundations of data analytics are changing

new computing architectures

new notions of "sensing"

unique new challenges → new structure
new intelligence

- distributed algorithms for learning over massive datasets on huge computer clusters
- parallel probabilistic inference
- Correlation analysis and inference of multiple time series
- …

Data Analysis as a Service: Open Issues

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

- **Privacy Protection** at both Individual and Corporate/Enterprise Level (handling the risk of unwanted privacy intrusion)
- **Noise / Quality Control** on the input, during the computation and on the output phases of data analytics (handling the risk of misuse of BIG DATA)

Questions