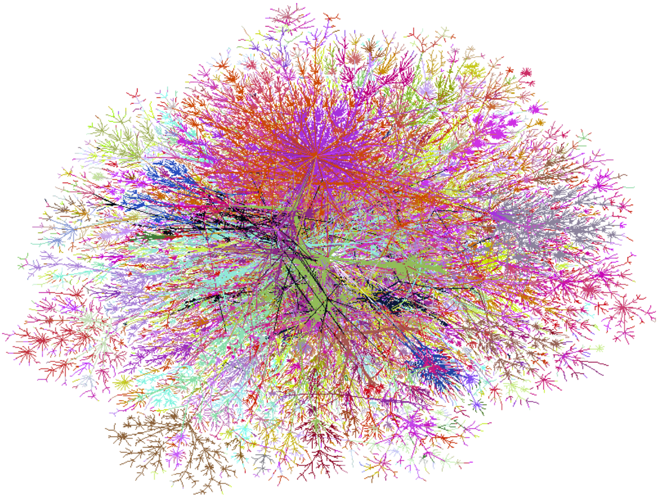
# Internet Measurement and Data Analysis (1)

Kenjiro Cho

2012-09-26

# introduction

how does the entire Internet look like?



lumeta internet mapping http://www.lumeta.com

http://www.cheswick.com/ches/map/

# introduction (cont'd)

how does the entire Internet look like?

- no one knows
- but, everyone is interested

the theme of the class

- looking at the Internet from different views
  - how to measure what is difficult to measure
  - how to extract useful information from huge data sets

this kind of approach will be increasingly important in the future information society

# Internet measurement and data analysis

- Faculty: Kenjiro Cho ⟨kjc@sfc.keio.ac.jp⟩
- TA: Yohei Kuga ⟨sora@sfc.wide.ad.jp⟩
- SA: Yukito Ueno ⟨eden@sfc.wide.ad.jp⟩
- URL: http://web.sfc.keio.ac.jp/~kjc/classes/sfc2012f-measurement/
- support email (facaulty, TA, SA): ⟨imda@sfc.wide.ad.jp⟩
- textbooks, references: the lecture slide materials will be provided online.
- programming: data processing exercises by Ruby
- evaluation: 2 assignments and a final report

# what you will learn in the class

- how to understand statistical aspects of data, and how to process and visualize data
  - which should be useful for writing thesis and other reports
- programming skills to process a large amount of data
  - beyound what the existing package software provide
- ability to suspect statistical results
  - the world is full of dubious statistical results and infomation manipulations
  - (improving literacy on online privacy)

# Big Data everywhere

the WHITE HOUSE PRESIDENT BARACK OBAMA

THE WHITE HOUSE WASHINGTON

BLOG | PHOTOS & VIDEO | BRIEFING ROOM | ISSUES | the ADMINISTRATI

*Home • The Administration • Office of Science and Technology Policy*

Office of Science and Technology Policy

About OSTP | OSTP Blog | Pressroom | Divisions | R&D Budgets | Resource Library | NST

## Big Data is a Big Deal

Subscribe

Posted by Tom Kalil on March 29, 2012 at 09:23 AM EDT

E-Mail | Tweet

*[Editor's Note: Watch*
*http://live.science360*

Today, the Obama Admi
our ability to extract kno
promises to help accele
transform teaching and l

To launch the initiative,
commitments that, toget
glean discoveries from l
address the challenges

We also want to challen
most of the opportunitie
President calls an "all ha

Some companies are al
research. Universities a
generation of "data scie
bono data collection, an
forum to highlight new p

*Tom Kalil is Deputy Di*

The New York Times

SundayReview | The Opinion Pages

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH

NEWS ANALYSIS

## The Age of Big Data

By STEVE LOHR
Published: February 11, 2012

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Enlarge This Image

Chad Hagen

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers,"

The Economist

Log in | Register | Subscribe

World politics | Business & finance | Economics | Science & technology | Culture

Current issue | Previous issues | Special reports | Politics this week | Business this

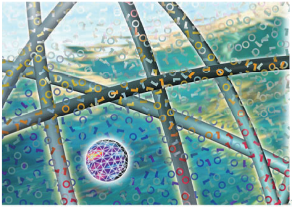
**Special report: Managing information**

## Data, data everywhere

**Information has gone from scarce to superabundant. That b benefits, says Kenneth Cukier (interviewed here)—but also**

Feb 25th 2010 | from the print edition

Like | 30



McKinsey Global Institute

Research | People | In the news | Contact us

Report

## Big data: The next frontier for innovation, competition, and productivity

May 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh

| Download | Executive Summary PDF–922KB | Full Report PDF–6MB | Kindle MOBI–4MB | eBook EPUB–3MB |
|---|---|---|---|---|

The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.

# big data by cloud computing

- “big data” becomes a trendy word, especially for marketing
- most technologies are not new
  - have been used in search ranking, online recommender systems, etc.
- big data processing used to be limited to big organizations that could collect, manage, and analyze data in-house
- now, anyone can easily use big data with cloud services
- package tools are available for collecting and analyzing online customer behaviors
- customer information can be easily used for marketing with minimal initial investment

# the age of data

- big data is not just for marketing
- technological innovations known as the data revolution are occurring in every field
- previously difficult applications become possible
  - access to huge amount of data, analysis of data constantly being updated, and applications to non-linear models
- big data analysis becomes an indispensable research method in all areas of science and technology

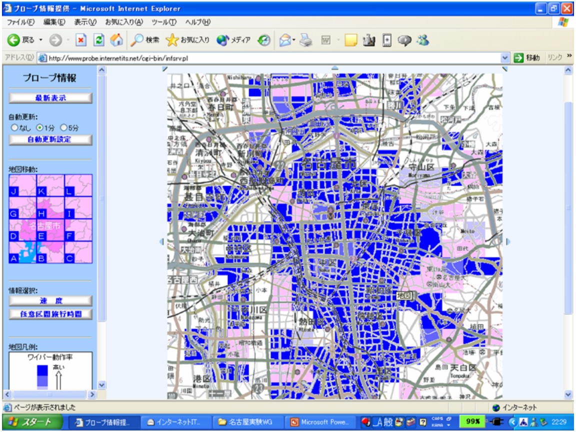
# example: impact to science

e-science: paradaigm shift?

- theory
- experiment
- simulations (enabled by computer)
- data-driven discovery (enabled by big data)

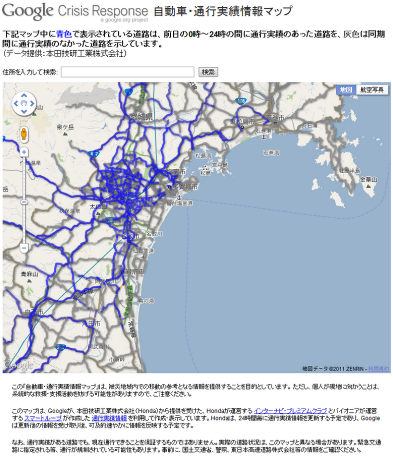
# example: Internet vehicle experiments

- by WIDE Project In Nagoya in 2001
  - location, speed and wiper usage data from 1,570 taxis
  - blue areas indicate high ratio of wiper usage, showing rainfall in detail



# Japan Earthquake

- the system is now part of ITS
- usable roads info released 3 days after the quake
  - data provided by HONDA (TOYOTA, NISSAN)



# Google's Chief Economist Hal Varian on Statistics

## The McKinsey Quarterly, January 2009

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

# self-introduction

Kenjiro Cho

- positions
  - Research Director, IIJ Research Lab
  - Guest Professor, Keio SFC
  - Adjunct Professor, JAIST
  - Board member, WIDE Project
- bio
  - BE in electronics from Kobe University in 1984.
    - started as a hardware engineer at Canon, Inc, then became interested in operating systems
  - M.Eng in computer science from Cornell University in 1993
    - studied computer science and distributed systems
  - Researcher at Sony Computer Science Labs from 1996
    - research on the Internet
  - Ph.D. (Media and Governance) from Keio University in 2001
  - Researcher at IIJ from 2004
- research topics
  - Internet measurement and management
  - networking support in operating systems

# class overview

It becomes possible to access a huge amount of diverse data through the Internet. It allows us to obtain new knowledge and create new services, leading to an innovation called "Big Data" or "Collective Intelligence". In order to understand such data and use it as a tool, one needs to have a good understanding of the technical background in statistics, machine learning, and computer network systems.

In this class, you will learn about the overview of large-scale data analysis on the Internet, and basic skills to obtain new knowledge from massive information for the forthcoming information society.

# class overview (cont'd)

## Theme, Goals, Methods

In this class, you will learn about data collection and data analysis methods on the Internet, to obtain knowledge and understanding of networking technologies and large-scale data analysis.

Each class will provide specific topics where you will learn the technologies and the theories behind the technologies. In addition to the lectures, each class includes programming exercises to obtain data analysis skills through the exercises.

## Prerequisites

The prerequisites for the class are basic programming skills and basic knowledge about statistics.

In the exercises and assignments, you will need to write programs to process large data sets, using the Ruby scripting language and the Gnuplot plotting tool. To understand the theoretical aspects, you will need basic knowledge about algebra and statistics. However, the focus of the class is to understand how mathematics is used for engineering applications.

# class schedule (1/5)

- Class 1 Introduction (9/26)
  - Big Data and Collective Intelligence
  - Internet measurement
  - Large-scale data analysis
  - exercise: introduction of Ruby scripting language
- Class 2 Data and variability (10/3)
  - Summary statistics
  - Sampling
  - How to make good graphs
  - exercise: graph plotting by Gnuplot
- NO CLASS on 10/10
- Class 3 Data recording and log analysis (10/17)
  - Network management tools
  - Data format
  - Log analysis methods
  - exercise: log data and regular expression

# class schedule (2/5)

- Class 4 Distribution and confidence intervals (10/24)
  - Normal distribution
  - Confidence intervals and statistical tests
  - Distribution generation
  - exercise: confidence intervals
  - **assignment 1**
- Class 5 Diversity and complexity (10/31)
  - Long tail
  - Web access and content distribution
  - Power-law and complex systems
  - exercise: power-law analysis
- Class 6 Correlation (11/7)
  - Online recommendation systems
  - Distance
  - Correlation coefficient
  - exercise: correlation analysis

# class schedule (3/5)

- Class 7 Multivariate analysis (11/14)
  - Data sensing
  - Linear regression
  - Principal Component Analysis
  - exercise: linear regression
- Class 8 Time-series analysis (11/22?) ***makeup class
  - Internet and time
  - Network Time Protocol
  - Time series analysis
  - exercise: time-series analysis
  - **assignment 2**
- Class 9 Topology and graph (11/28)
  - Routing protocols
  - Graph theory
  - exercise: shortest-path algorithm
- Class 10 Anomaly detection and machine learning (12/5)
  - Anomaly detection
  - Machine Learning
  - SPAM filtering and Bayes theorem
  - exercise: naive Bayesian filter

# class schedule (4/5)

- Class 11 Data Mining (12/12)
  - Pattern extraction
  - Classification
  - Clustering
  - exercise: clustering
- Class 12 Search and Ranking (12/19)
  - Search systems
  - PageRank
  - exercise: PageRank algorithm
- Class 13 Scalable measurement and analysis (12/26)
  - Distributed parallel processing
  - Cloud computing technology
  - MapReduce
  - exercise: MapReduce algorithm
- Class 14 Privacy Issues (1/9)
  - Internet data analysis and privacy issues
  - Summary of the class

# network measurement and Internet measurement

- network measurement
  - measurement in limited environment
  - snapshot at a time
- Internet measurement
  - measurement of the Internet as a large-scale open system
    - large-scale distributed system
    - open system (continuously changing)

# Internet measurement – measuring unmeasurable Internet

- need for generic measurement data for the Internet
  - example: typical packet size distribution
- the Internet is an open system continuously changing, evolving, and expanding
  - no central point, representative locations, different behaviors are observed depending on observing location and time
  - seeking for generality of the Internet: measuring unmeasurables
- for operation of the Internet, for development of protocols, equipment and services
  - seeking for the best estimates, predicting the future, and revisiting the existing knowledge
- need to consider not only from technical aspects but also from social, political and economical aspects

# importance of measurement

measurement is a basis of all technologies

- for networking, it is an attempt to observe invisible networks
- needed for operation, design, implementation, and research

- however, it has become difficult by commercialization of the Internet and widespread use
  - traffic data is confidential for providers and will not be disclosed
  - risks of leaking private information

# goals of measurement and data analysis

- operational goals
  - trouble-shooting
  - tuning for performance and reliability
  - monitoring the usage, usage reports
  - long-term planning, cost evaluation of network capacity and equipment
- engineering goals (software, hardware, protocol design and implementations)
  - design trade-offs (e.g., buffer size and its cost)
  - testing and evaluation
  - observing unexpected behaviors (in complex systems)
- research goals (theory, modeling, new findings)
  - characteristics of network behaviors
  - modeling (e.g., behavior of web services)
  - behaviors of complex systems
    - abundant data and tools
- inputs for policy or investment plans

# characteristics of network data and behavior

- skewed distributions with large variance
  - inherent mechanism to make burst transfer
  - skewed utilization: e.g., a handful users generate most traffic
- anomalies everywhere
  - bugs, mis-configurations, spec mismatches, accidents, maintenance's
- interferences among various mechanisms
  - e.g., congestion control: Ethernet's collision avoidance, packet queueing, TCP's congestion control, capacity provisioning
- traffic aggregation
  - complex behavior as a whole (more than the sum of the individual components)
- limitations of network measurement
  - many practical issues and limitations exist
  - measurement affects the observed behavior

# measurement needs combined skills

- goals could be operational, engineering, scientific
  - all inseparable, all skills required
    - knowledge of operational environment
    - engineering of measurement tools
- output can be facts, findings, new ideas
  - new ideas are not always necessary
  - facts, especially long-term measurement, are valuable
- but you should have clear goals
  - better to start with real problems to solve
    - there are many issues and problems but some are more important than others

# why traffic measurement of Internet is so hard?

- massive, diverse and changing traffic
- mechanisms at different layers in different time scale
  - interact with each other
- dynamics
  - Internet mechanisms are adaptive and resilient
  - traditional measurement techniques are often not applicable
- pathological traffic is not unusual
  - by bugs, misconfigurations, errors, mismatches, accidents
- we still don't have good understanding

# massive volume of traffic

- unprecedented scale with unprecedented growth
  - e.g., traffic volume: 1Gbps traffic
    - 120MB/sec 7GB/minute 420GB/hour 9.8TB/day
- far more data than we can analyze
  - techniques needed to reduce data size
    - filtering: e.g., record only TCP SYN packets
    - aggregation: e.g., flow-based accounting
    - sampling: e.g., record 1 in n packets
  - also, techniques needed to reduce dimensionality
- still, details matter
  - a big impact often comes
    - from small fraction
    - from minor differences

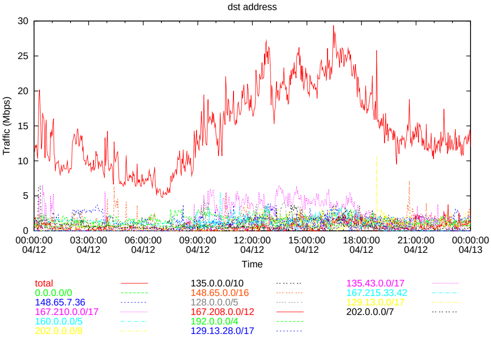
# diverse traffic

- large variation in traffic mix between sites
- backbone vs. access links
  - access line types: fiber, ADSL, modem, wireless, satellite
    - differences in bandwidth, delay, loss

typical traffic doesn't exist!

# constant change of traffic pattern

- daily, weekly traffic pattern
- trend changes over time
  - web in 90s and p2p in 2000s completely changed traffic pattern
- hard to predict future!



# limitations of Internet measurement

- problems often occur at boundaries of different networks
  - cooperation needed but not easy
- measurement affects the behavior of the observed network
- need understanding and help from operators
  - need to understand operational requirements and find suitable methods for measurement
- cost: measurement doesn't come free
  - limitations to measure high-end routers with a PC
- privacy and confidential information in data
  - barriers for researchers to access commercial data

# possible topics to be studied in the class

- search ranking (PageRank), online recommender systems(collaborative filtering)
- connections among SNS users, popular keyword extraction, shortest path search, online privacy
- SPAM filtering, MapReduce, Geolocation services, Web server log analysis
- Internet traffic, Internet topology

## summary

Internet measurement and data analysis

- measurement is basis for all technologies
- for networking, it is an attempt to observe invisible networks
- need to consider not only from technical aspects but also from social, political and economical aspects

theme of the class

- Internet measurement and data analysis as case studies
- learn how to measure what is difficult to measure
- learn how to extract useful information from huge data sets

# Introduction to Ruby

# Ruby

- a scripting language for object-oriented programming
- supports wide range of functions for text processing and system management
- free software started in 1993
- original author: Yukihiro Matsumoto
- became popular for Ruby on Rails (a web application framework)

Ruby information

Ruby official site: `http://www.ruby-lang.org/`
Ruby reference manual: `http://www.ruby-lang.org/en/documentation/`
Ruby の歩き方: `http://jp.rubyist.net/magazine/?FirstStepRuby`

# Ruby characteristics

- interpreter language: no need to compile for execution
- highly portable: runs on most platforms
- simple syntax
  - no predefined data type for variables, variables can store any data and are dynamically typed
  - no need to declare variables, variable types (local variables, global variables, instance variables) can be inferred from variable names
- garbage collection: users do not need to manage memory
- object-oriented
  - everything is an object
  - class, inheritance, methods
  - iterator and closure
    - control structures and procedures can be written in object-oriented manner
- powerful string operations/regular expressions
- built-in support for large integers
- Ruby's shortcomings: a bit slower than its competitors

# Ruby commands

- irb: Ruby's interactive interface

  ```
  $ irb --simple-prompt
  >> puts "Hello"
  Hello
  ```

- ruby: Ruby main program

  ```
  $ ruby test.rb
  ```

  or,

  ```
  $ ruby -e 'puts "Hello".reverse'
  olleH
  ```

# exercise: a program to count text lines

count the number of text lines in a file given by the argument

```
filename = ARGV[0]
count = 0
file = open(filename)
while text = file.gets
  count += 1
end
file.close
puts count
```

write to “count.rb” and then run it

```
$ ruby count.rb foo.txt
```

rewrite it in a more rubyish way

```
#!/usr/bin/env ruby
count = 0
ARGF.each_line do |line|
  count += 1
end
puts count
```

# next class

Class 2 Data and variability (10/3)

- Summary statistics
- Sampling
- How to make good graphs
- exercise: graph plotting by Gnuplot

NO CLASS on 10/10

# references


[1] Ruby official site. `http://www.ruby-lang.org/`

[2] gnuplot official site. `http://gnuplot.info/`

[3] Mark Crovella and Balachander Krishnamurthy. *Internet measurement: infrastructure, traffic, and applications*. Wiley, 2006.

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[5] Raj Jain. *The art of computer systems performance analysis*. Wiley, 1991.

[6] Toby Segaran. *Programming Collective Intelligence*. O'Reilly Media. 2007.

[7] Allen B. Downey. *Think Stats: Probability and Statistics for Programmers*. O'Reilly Media. 2011. `http://thinkstats.com/`

[8] あきみち、空閑洋平. **インターネットのカタチ**. オーム社, 2011.

[9] 井上洋, 野澤昌弘. **例題で学ぶ統計的方法**. 創成社, 2010.

[10] 平岡和幸, 掘玄. **プログラミングのための確率統計**. オーム社, 2009.