Internet Measurement and the Impact of Big Data

Kenjiro Cho (IIJ/WIDE)

Big Data everywhere

the WHITE HOUSE PRESIDENT BARACK OBAMA

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 h
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

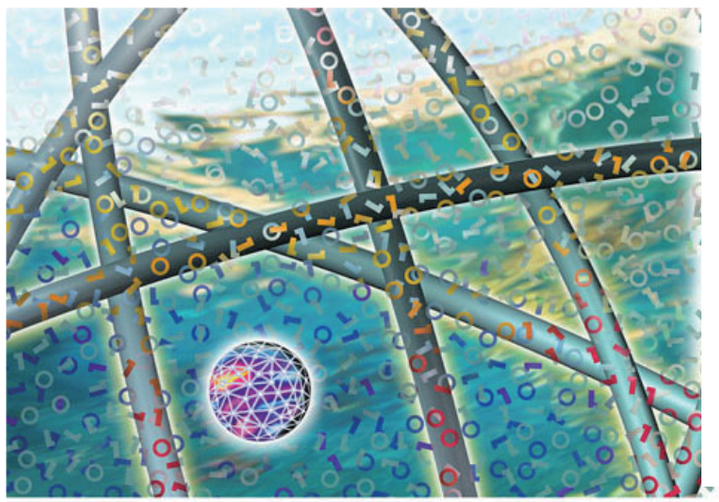
Special report: Managing information

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.

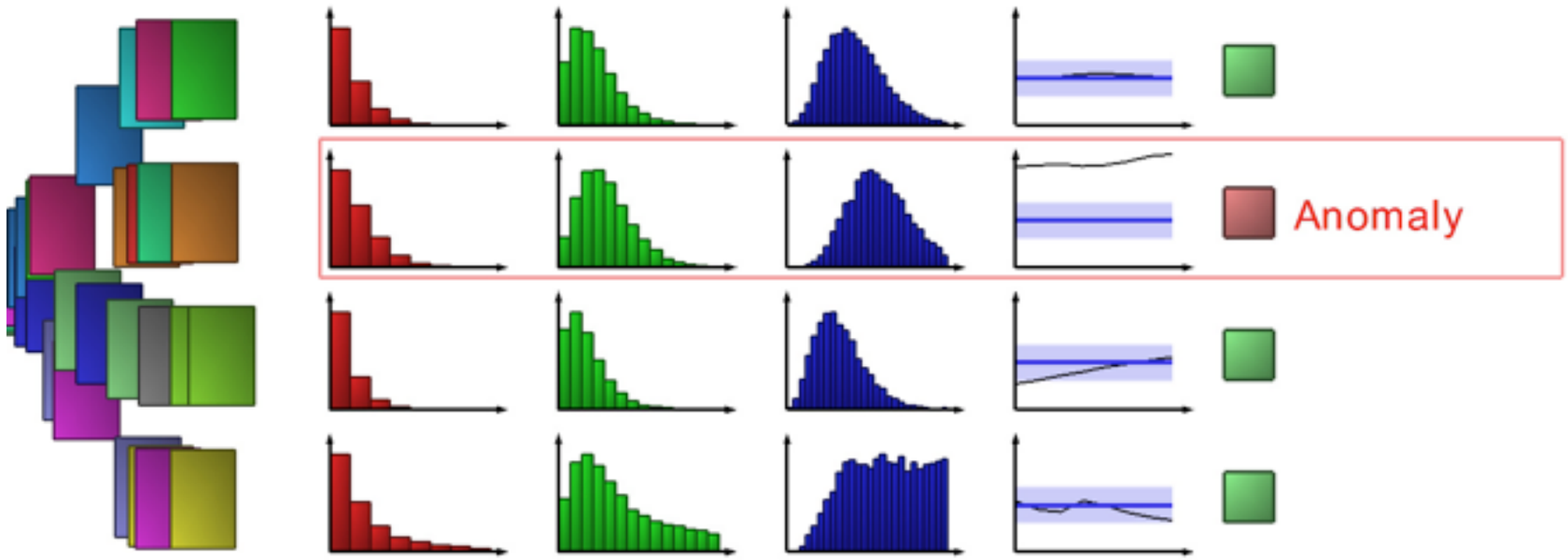
Are we on the edge of a big wave?

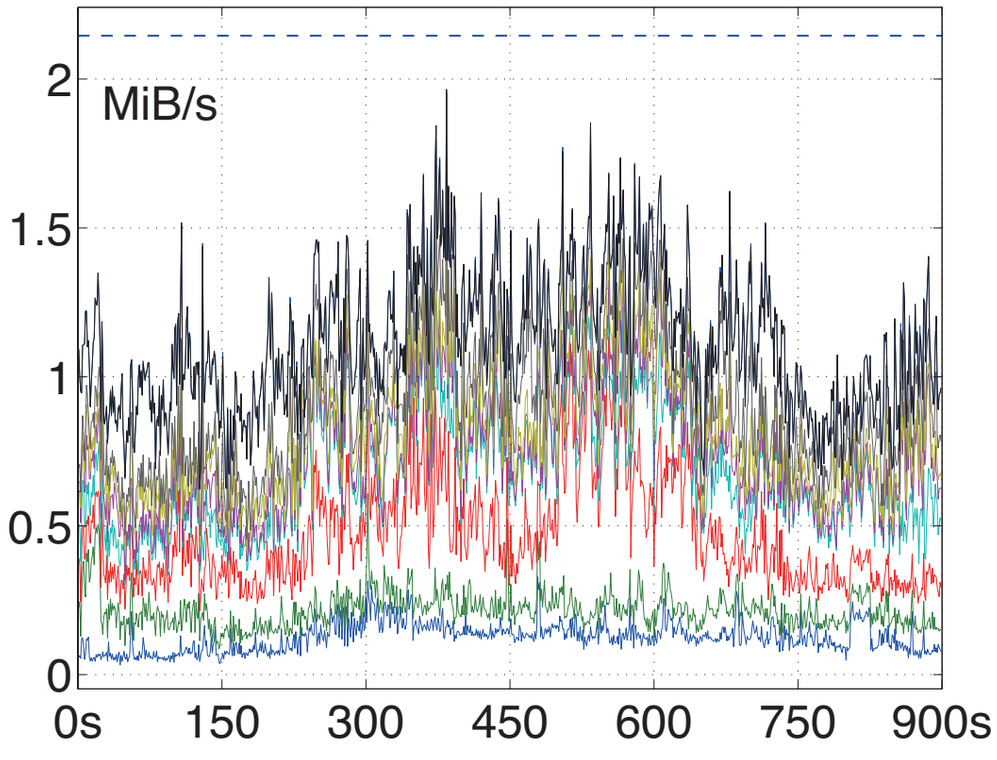
- Big data: extracting hidden useful info from huge amount of unstructured data
- we have been doing it for Internet measurement for 20 years!

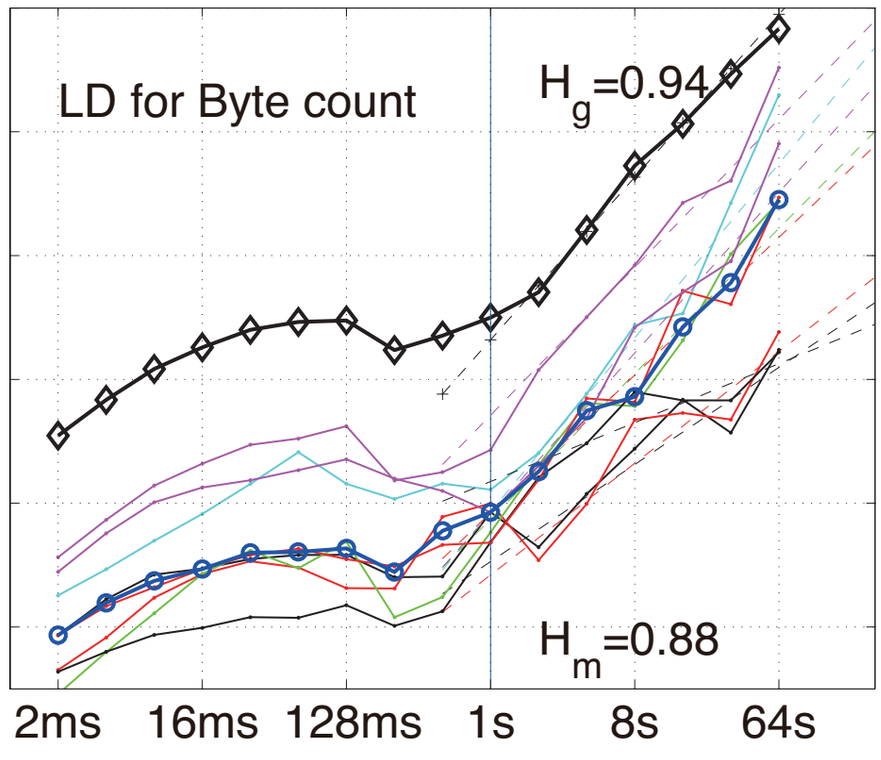
lessons learned from Internet measurement research

- data collection
  - quality and integrity of data
  - trust/relationship with data owners and users
- data sharing
  - third-party verification: fundamental to science
  - beneficial for society
- privacy in data
  - technical, legal, moral issues
  - social benefits vs. privacy risks

example: anomaly detection by sketch and statistical feature comparison

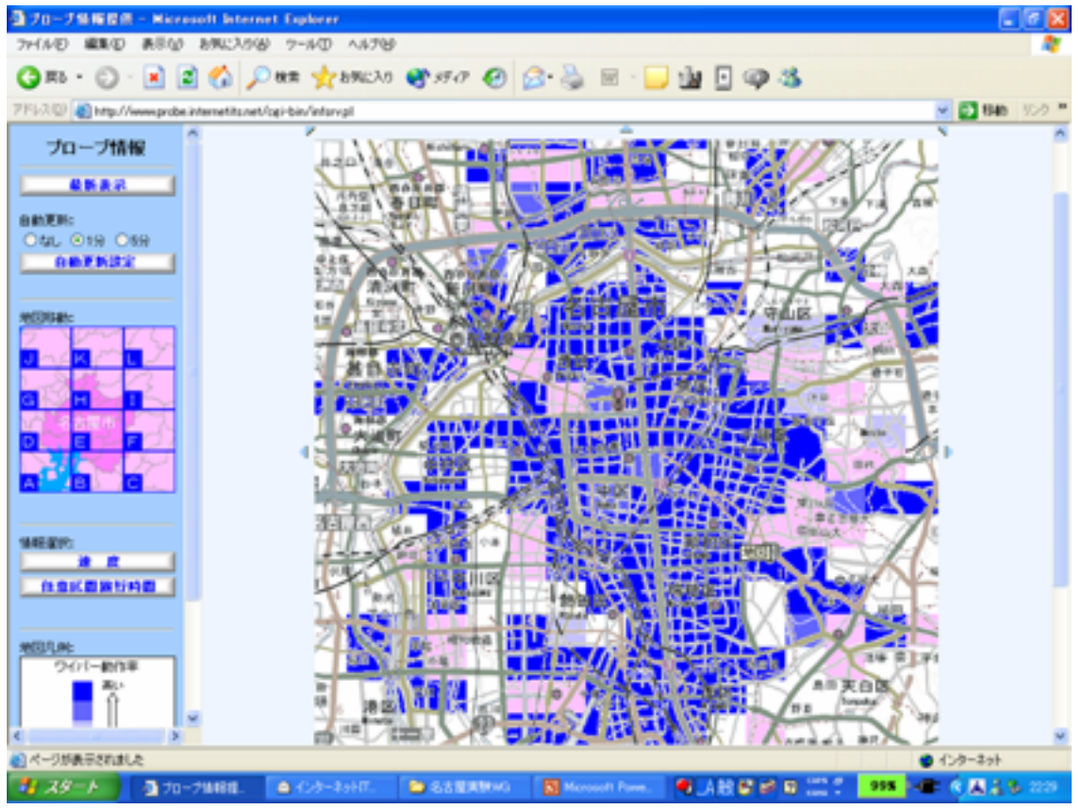






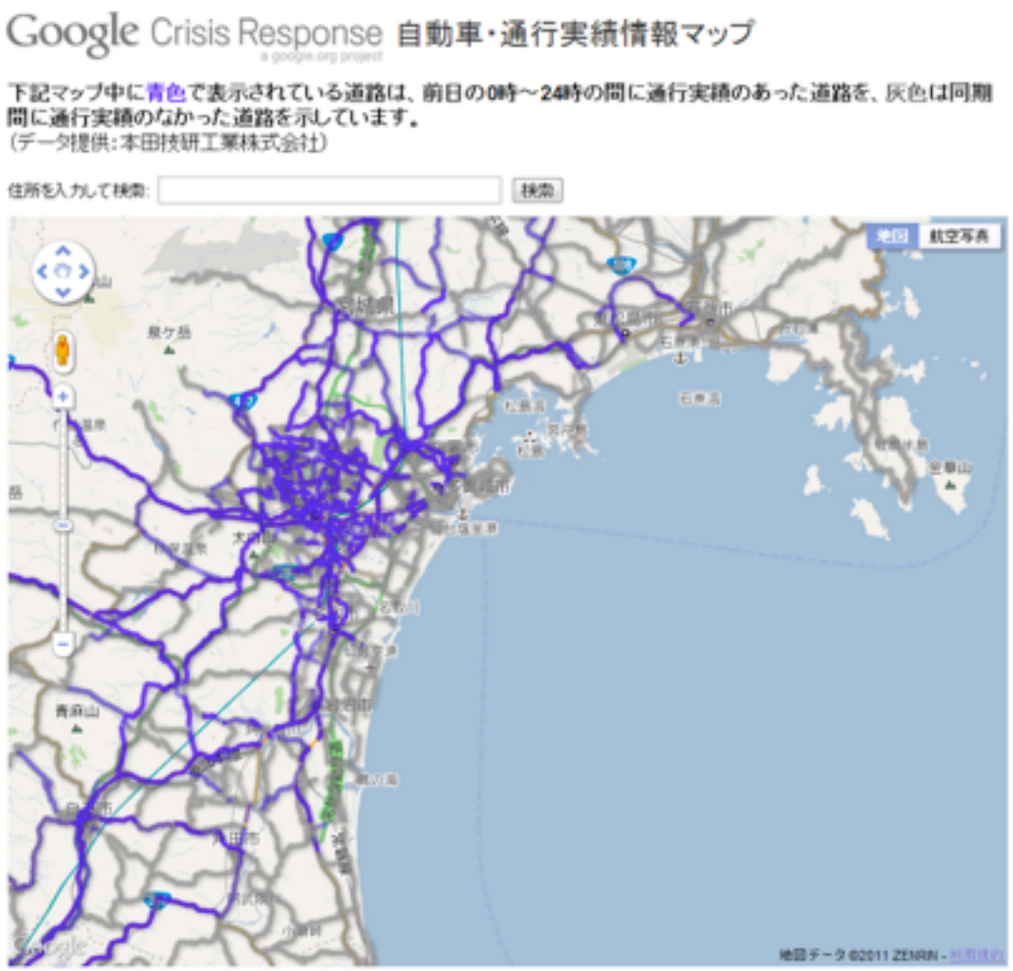
example: Internet vehicle experiments

- by WIDE Project in Nagoya in 2001
  - location, speed, and wiper usage data from 1,570 taxis
  - blue area 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)

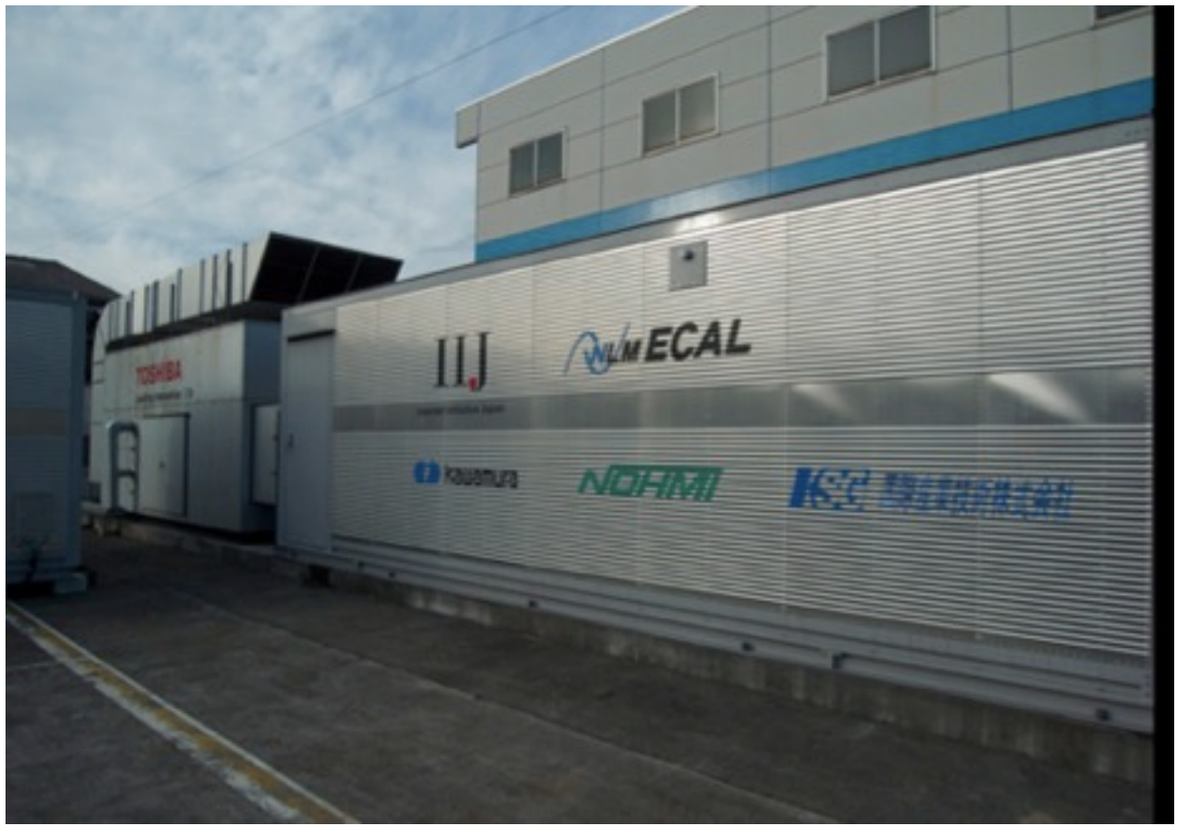


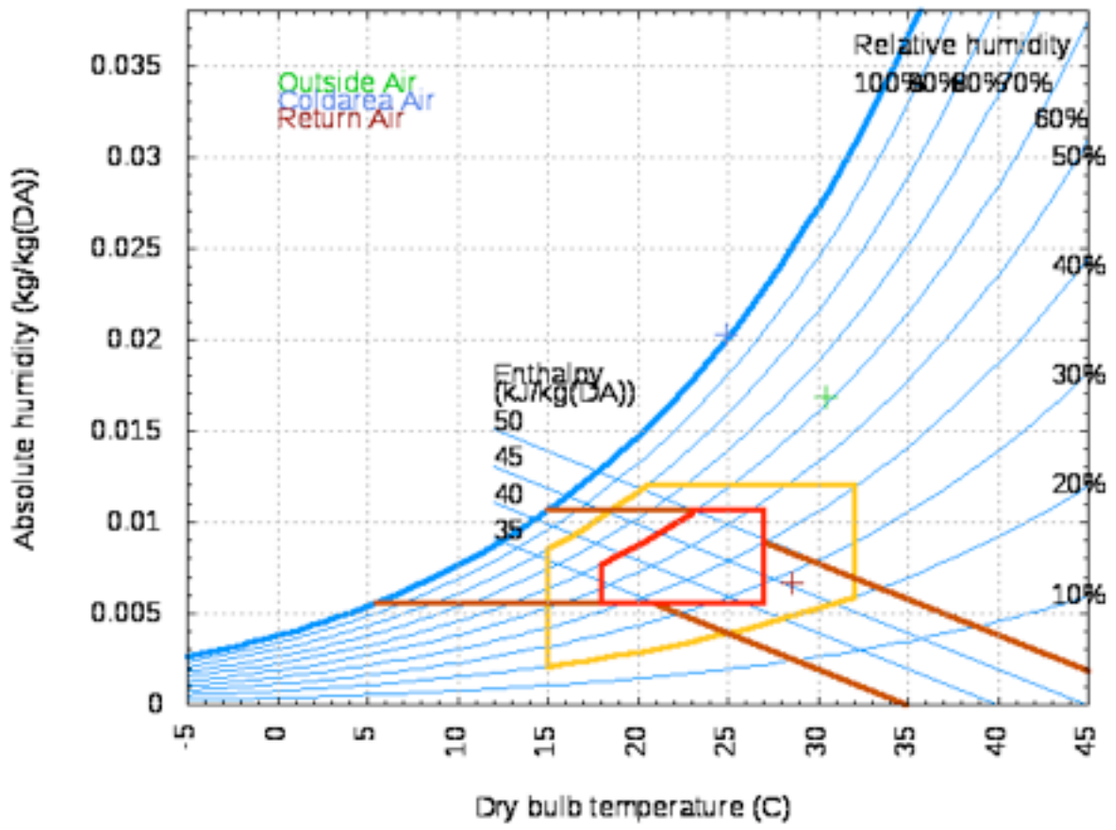
この「自動車・通行実績情報マップ」は、被災地域内での移動の参考となる情報を提供することを目的としています。ただし、個人が現地に向かうことは、系統的な救援・支援活動を妨げる可能性がありますので、ご注意ください。

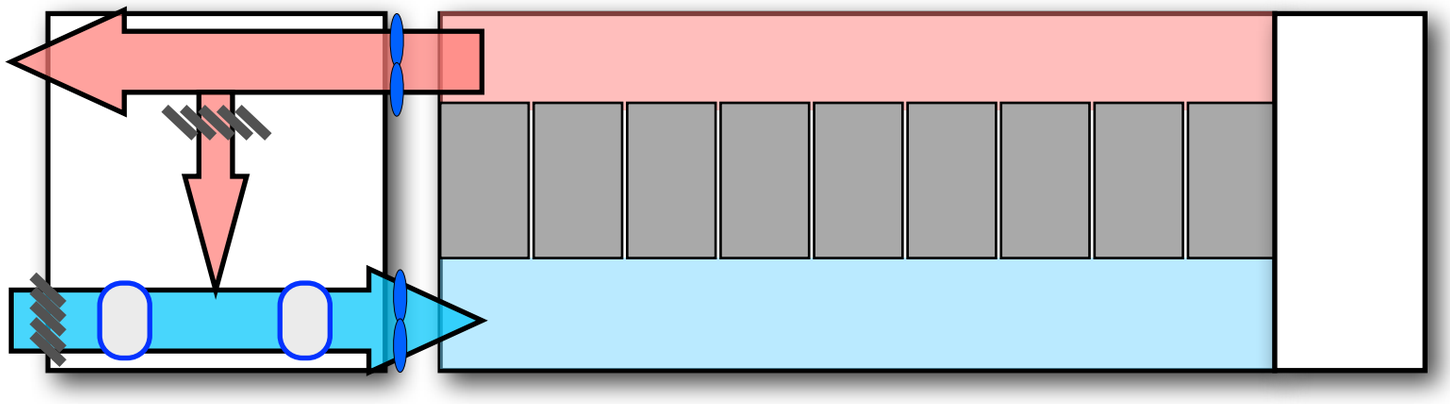
このマップは、Googleが、本田技研工業株式会社（Honda）から提供を受けた、Hondaが運営するインターナビ・プレミアムクラブとパイオニアが運営するスマートループが作成した通行実績情報を利用して作成・表示しています。Hondaは、24時間毎に通行実績情報を更新する予定であり、Googleは更新後の情報を受け取り次第、可及的速やかに情報を反映する予定です。

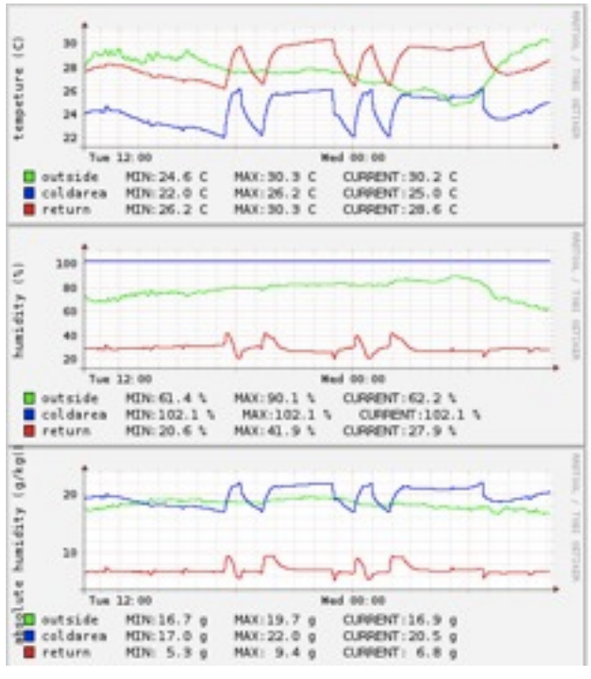
なお、通行実績がある道路でも、現在通行できることを保証するものではありません。実際の道路状況は、このマップと異なる場合があります。緊急交通路に指定される等、通行が規制されている可能性もあります。事前に、国土交通省、警察、東日本高速道路株式会社等の情報をご確認ください。

example: data center as data









Should we take advantage of big data or stay away from it?

- pros:
  - it helps to convince people for the need of data
  - it attracts researchers, students, and money
  - many useful tools have been developed
- cons:
  - it's just a hype, technically nothing new
  - dubious about those who jump on the bandwagon
- How can we make use of the big data trend?

Google's Chief Economist Hal Varian on Statistics

The McKinsey Quarterly, January 2009

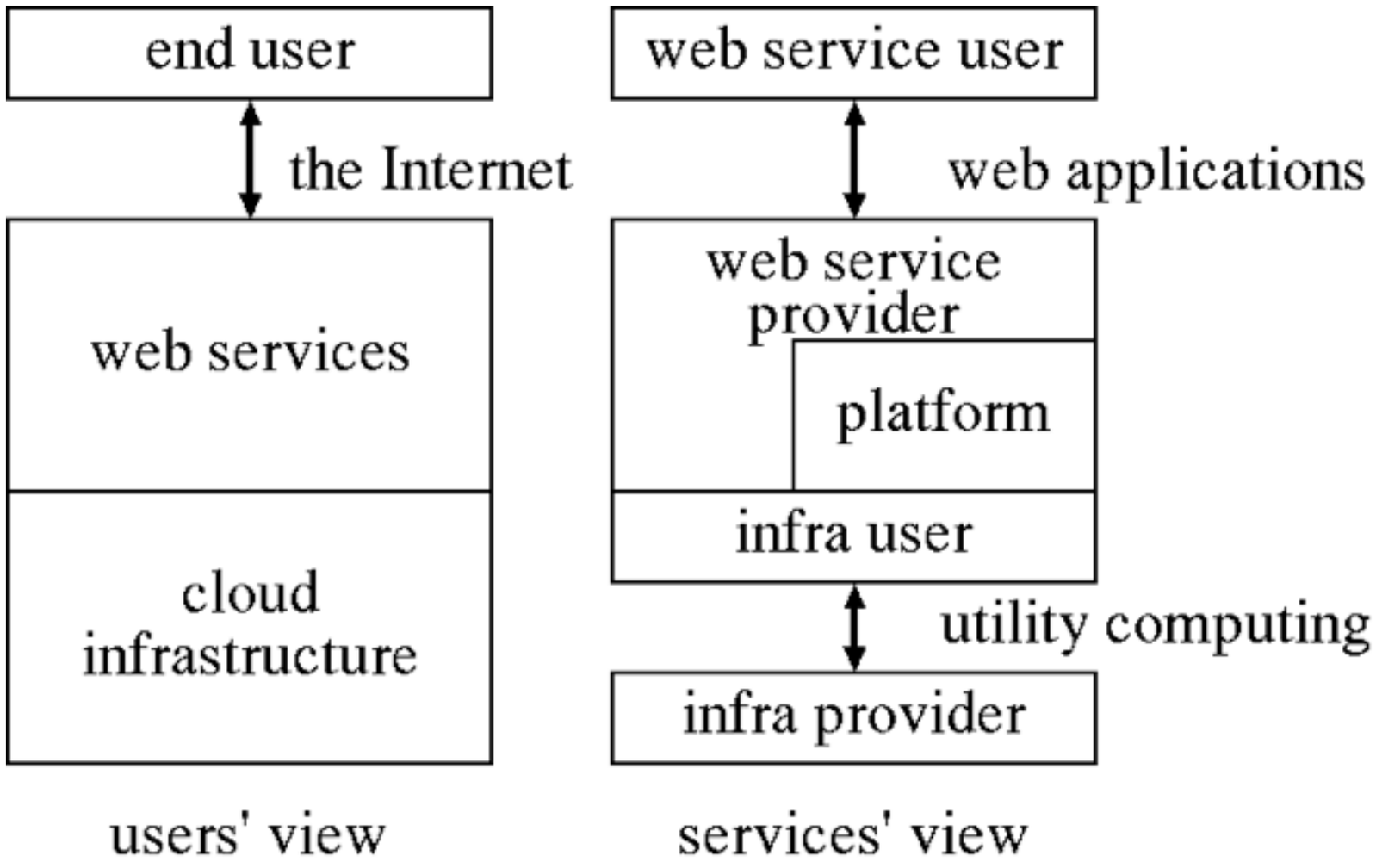
"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, 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."

big data by cloud services

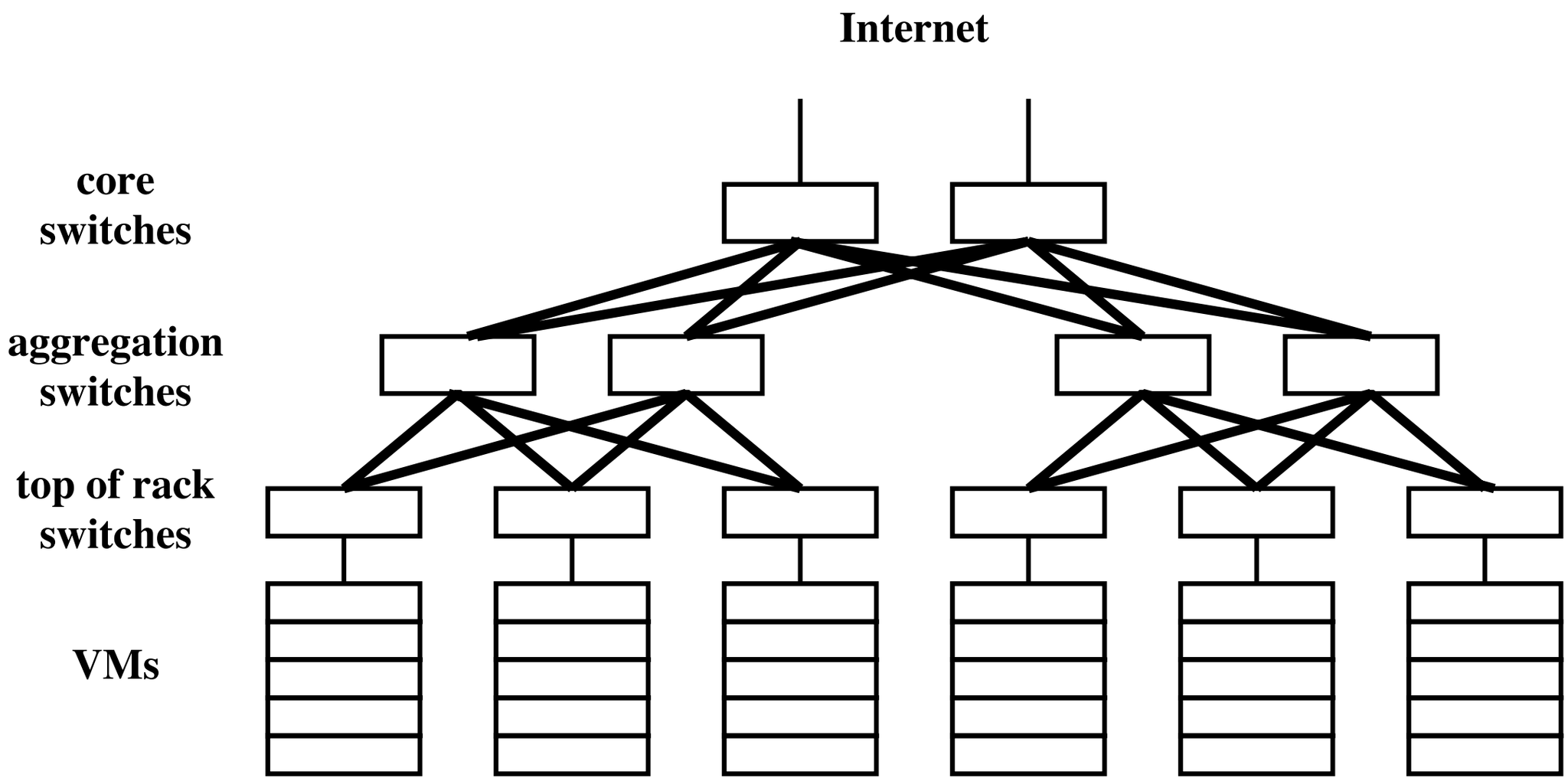
- big data processing used to be limited to big organizations that could collect, manage, and analyze data in-house
- now, anyone can easily use big data with cloud services
- package tools are available for collecting and analyzing online customer behaviors
- customer information can be easily used for marketing with minimal initial investment

types of cloud services

- public/private/hybrid
- SaaS/PaaS/IaaS



typical cloud network topology



age of data

- technological innovations known as the data revolution are occurring in every field
- previously difficult applications become possible
  - access to huge amounts 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: paradigm shift?
  - theory
  - experiment
  - simulation (enabled by computer)
  - data-driven discovery (enabled by big data)
- SC community starts to realize that they should invest for data-sharing rather than computation-sharing

big data technologies

- data collection
  - increasing data sources (e.g., sensors, social media)
- data storage
  - distributed storage, NoSQL database
- data processing
  - distributed/cloud computing (e.g., MapReduce)
- data understanding
  - data mining, machine learning, statistical analysis

big data computing

- use of cloud services
  - utility computing: pay-per-use model
- example: Amazon EC2
  - part of Amazon Web Seervices (AWS)
  - VM (linux or windows, from $0.02/hour)
  - data transfer (from $0.12/GB for download)

computation models

- MapReduce: parallel batch processing
  - e.g., Hadoop, Facebook Puma/Ptail
- Bulk Synchronous Parallel (BSP)
  - e.g., Google Pregel, Apache Giraffe/ Hama
- Complex Event Processing (CEP): realtime parallel processing for stream data
  - e.g., Twitter Storm, Yahoo! S4

storage/database systems

- advances in disk technologies
  - capacity
  - access time
  - use of nonvolatile solid-state memory (SSD, etc)
- new perspectives (for measurement)
  - most measurement data is write-once
    - consistency/locking can be relaxed
    - large block size
  - most data types are simple: e.g., key-value

file systems

- distributed file systems
  - fault-tolerance
  - scalability
  - huge file support
  - e.g., GFS, HFS, GlusterFS, MogileFS, NFSv4.1

NoSQL database

- key-value store: (simple key-value type)
  - e.g., Dynamo, Redis, Voldemorte, Membase
- column-oriented DB: (optimized for column access)
  - e.g., BigTable, Hbase, Cassandra
- document oriented DB: (schemaless)
  - e.g., MongoDB, CouchDB, SimpleDB
- graph-oriented DB: (index-free adjacency)
  - e.g., InfoGrid, Neo4j

UNIX shell and pipe

- old technology, but still most useful
  - becomes much more efficient on multi-core machines
  - most of data analysis can be done on a single machine
    - with appropriate pre-processing

- my favorite and most often used method!

data analysis is merely a tool

- recent big data trends focus too much on tools and methods but data analysis is merely a tool
- data analysis is an iterative process
  - forming a hypothesis, verifying it with data
  - if the results are unexpected, you find new questions
  - repeating the process will uncover interesting facts
- analysis without purpose ends up with useless numbers
- If you identify what to get from data, you will see a path forward

privacy issues

- increasing risk of privacy breaches by big data and data mining
  - anonymization and de-identification of personal info
    - personal info: health, location, electricity use, online activity
- there are technical, legal, moral aspects
- hard to assess risks in the future

issues in the data age

- shortage of data specialists
  - needs of data scientists with field-specific knowledge and data analysis skills
  - challenge conventional thinking and interpretations, establish issues clearly, use statistics and data as tools to resolve problems
- data as assets
  - if based on the same data, success depends on analytical ability
  - but if data quality varies, big advantage with better quality data

issues in the data age (cont'd)

- data sharing
  - social benefits of data sharing and third-party verification
  - a balance of social benefit and risk of privacy breaches
- social consensus making
  - how far private/public organizations can track individuals
  - how information (e.g., personal medical records) will be shared for social benefit

recipient literacy

- info recipients also need to understand and question data
  - as there are more and more questionable data and dubious theories based on data
- we tend to want to see things as either black or white
  - but most things are grey; determining black/white is to draw a line just for convenience
  - seeking a black or white answer is to avoid own judgment and responsibility
- we should accept grey as grey, and make our own judgement

fundamental change to creative thinking process?

- data-driven decision making has been always important
- but, ICT pushes it to a completely different level (in quality, quantity, expressions)
- now, we can literally interact with data (data-human interaction)

summary

- big data is a hype, still some part is useful for Internet measurement
  - to attract people and money
  - useful tools
- increasing importance of data
  - it will change how we think
  - our experiences from Internet measurement would be useful for others