CISC 520: Data Engineering and Mining

Daqing Yun
Harrisburg University

https://goo.gl/Awz54i

The 4th Paradigm
Extreme Data-Intensive Scientific Computing

- Introduction to Big Data

Introduction to Big Data

Some of the these slides have been provided through the courtesy of Prof. Chase Q. Wu at New Jersey Institute of Technology

Motivation

Oak Ridge National Laboratory

solving the BIG PROBLEMS

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

Motivation

BIG DATA is GOING TO IMPACT US SOON!
CALL BRUCE WILLIS!!
DATA
www.nuxeo.com
Soulcié

What Is Big Data?

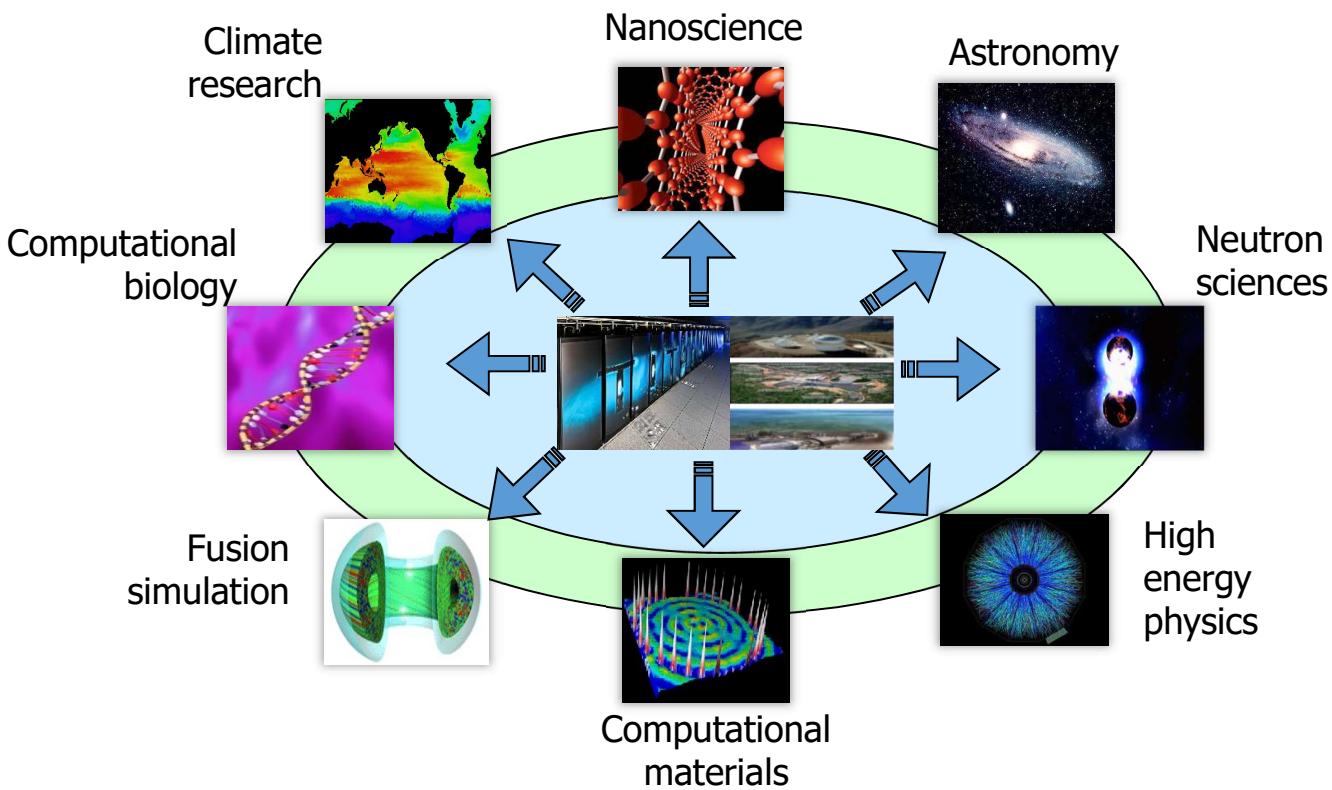
Big data is similar to *small* data, but only bigger?

Having data bigger consequently requires different approaches

- Techniques, tools, and architectures
- With an aim to solve new problems
- And old problems in a better way

Where Are The Data From?

Large-Scale Scientific Applications

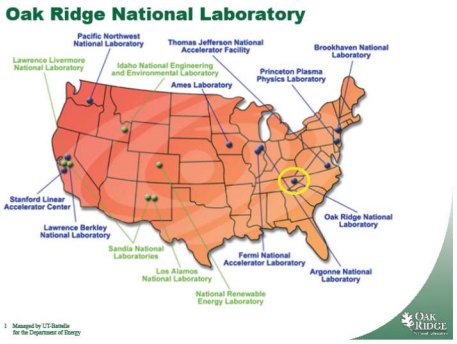


Where Are The Data From?

Large-Scale Scientific Applications

- *Simulation*: Astrophysics, climate modeling, combustion research, etc.
- *Experimental*: Spallation Neutron Source, Large Hadron Collider, microarray, genome sequencing, protein folding, etc.
- *Observational*: Large-scale sensor networks, astronomical imaging/radio devices (Dark Energy Camera, Five-hundred-meter Aperture Spherical Telescope - FAST), etc.

Where Are The Data From?



Science Is Changing…

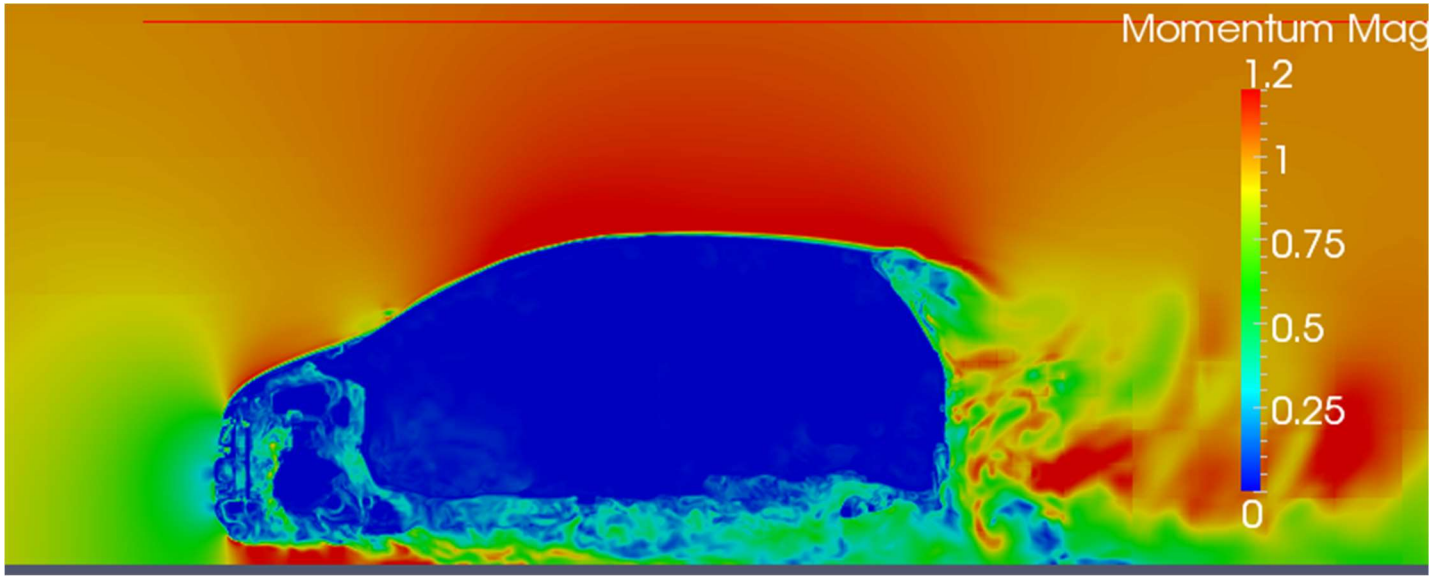
Thousand years ago, science was empirically describing natural phenomena…

Science Is Changing…

Last few hundred years, theoretical branch using models, generalizations…

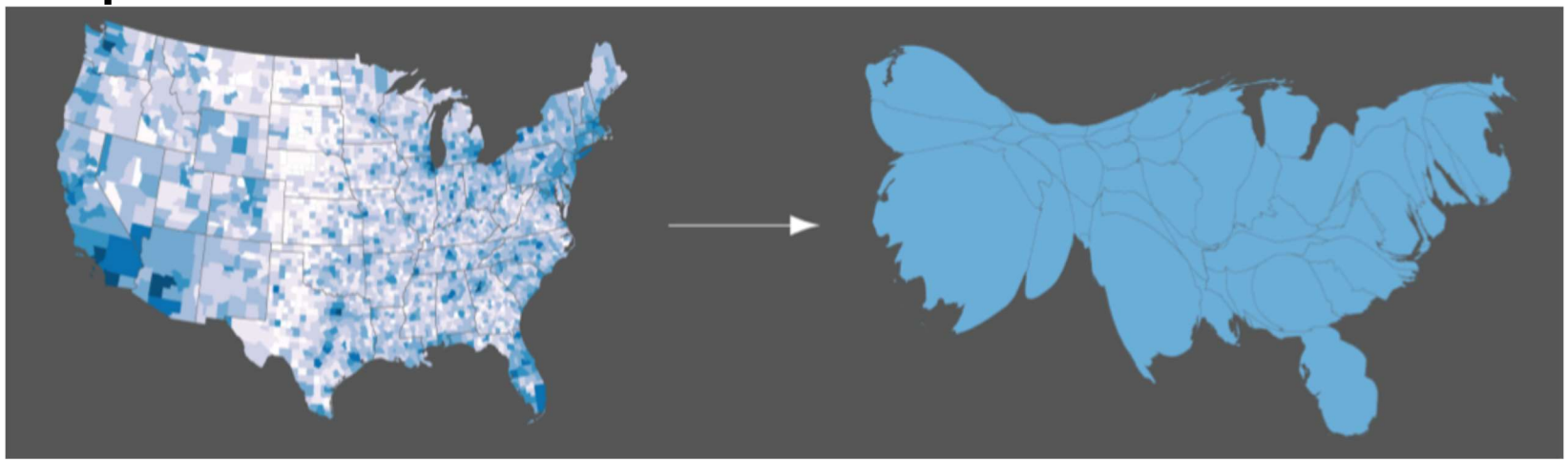
Science Is Changing…

Last few decades, a computational branch simulating complex phenomena…



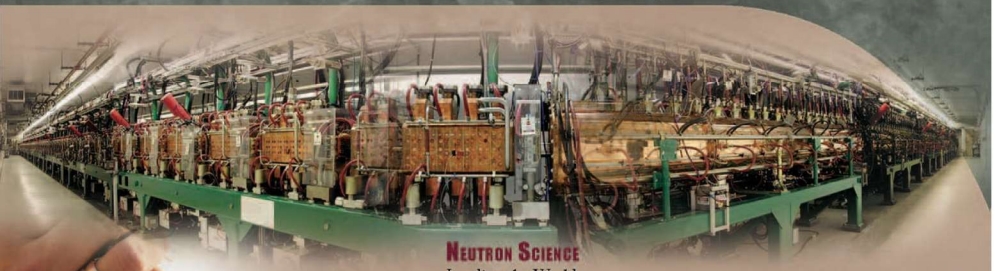
Science Is Changing…

Today, data-intensive science, synthesizing theory, experiment and computation with statistics, i.e., new way of thinking required…



Six Scientific Themes

Born of necessity. Inspired by our quest to know. We have always been called upon to address America's greatest scientific challenges.



Biological Systems
Developing New Options

Whether converting biomass to fuel or understanding the impacts of climate change, biological research at ORNL is helping develop new options for energy, environmental protection, and human health. **mannre@ornl.gov**

Neutron Science
Leading the World

The Spallation Neutron Source and the High Flux Isotope Reactor together make Oak Ridge the world's foremost center for neutron science. **andersonian@ornl.gov**

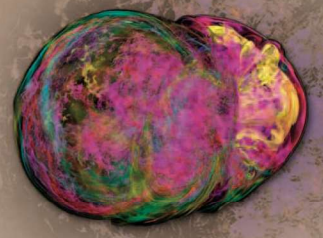
Advanced Materials
Strengthening American Industry

With DOE's first Nanoscience Center, the world's most powerful electron microscope, and the High Temperature Materials Laboratory, Oak Ridge plays a critical role in American industrial competitiveness. **buchananrw@ornl.gov**

"Men love to wonder, and that is the seed of science."
. . . Ralph Waldo Emerson

National Security
Guarding the Gates

From biochemical sensors to stopping the proliferation of nuclear weapons, technologies that make America safer are among the laboratory's top research priorities.
akersfhjr@ornl.gov



High Performance Computing
Tackling the Big Problems

With unmatched computational capacity for open scientific research, Oak Ridge is on a path by 2009 to reach a petaflop, or 1 quadrillion mathematical calculations per second, making it possible to model the most complex scientific problems.
zachariat@ornl.gov

Energy
Providing Energy Alternatives

Increased production, improved transmission, reduced consumption: Oak Ridge is addressing our energy challenges on all fronts, from safer nuclear power to more energy-efficient cars and homes.
christensend@ornl.gov

Examples

All science is becoming data-driven
Terascale Supernova Initiative (TSI)

Oak Ridge National Laboratory

2 Managed by UT-Battelle
for the Department of Energy

Examples

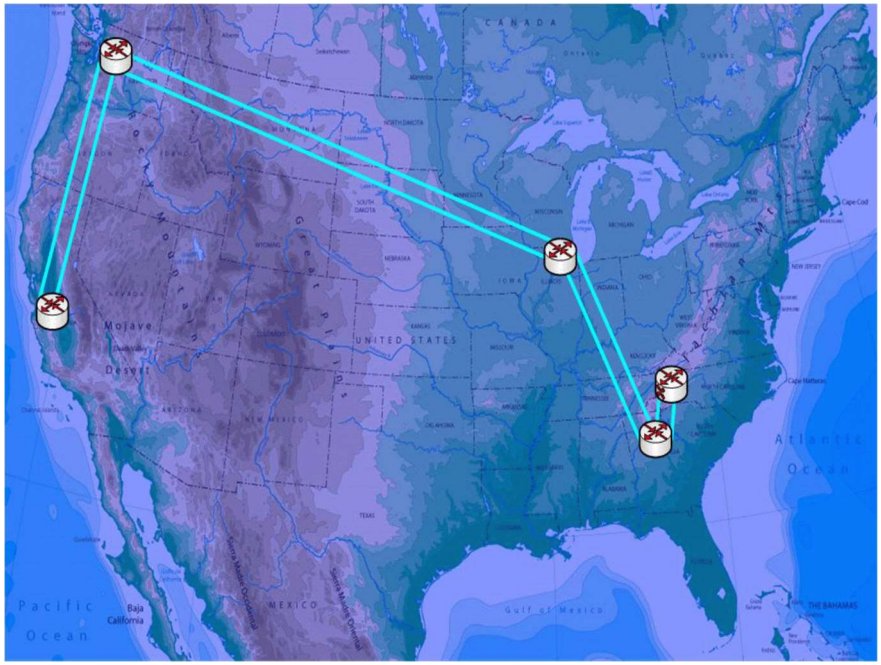
All science is becoming data-driven
Terascale Supernova Initiative (TSI)

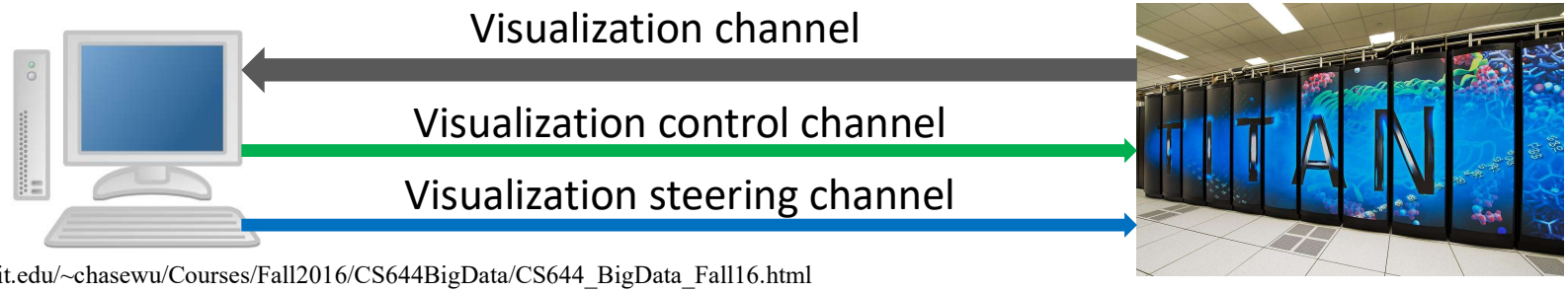
Collaborative project, supernova explosion

- TSI simulation, 1 terabyte per day with a small portion of parameter
- From TSI to PSI

Transfer to remote sites

- Interactive distributed visualization
- Collaborative data analysis
- Computation monitoring
- Computation steering



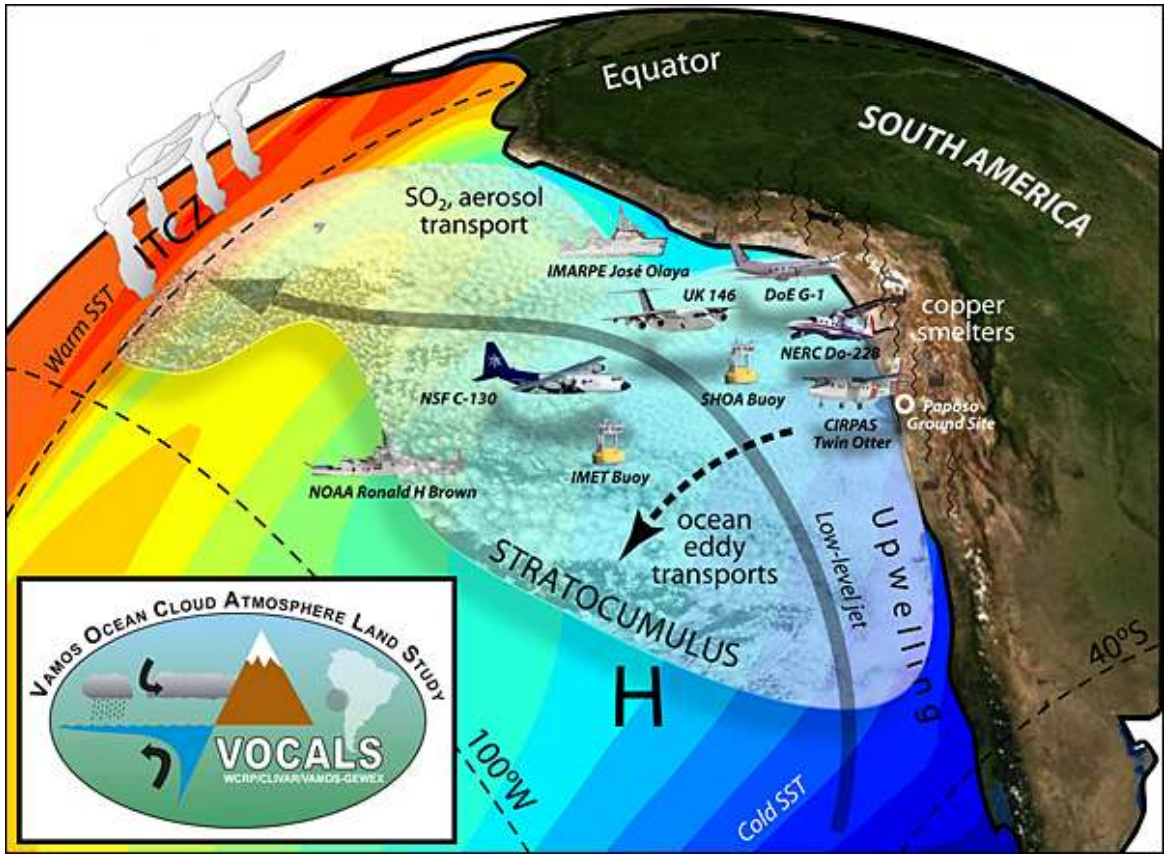


https://web.njit.edu/~chasewu/Courses/Fall2016/CS644BigData/CS644_BigData_Fall16.html

Examples

Convergence of physical and life sciences through Big Data (statistics and computing) and a scientific revolution in how discovery takes place, which is a rare and unique opportunity

Brookhaven National Laboratory (BNL) Climate Science Research



https://www.bnl.gov/newsroom/news.php?a=184

DATA NEVER SLEEPS

How Much Data Is Generated Every Minute?

Big data is not just some abstract concept used to inspire and mystify the IT crowd; it is the result of an avalanche of digital activity pulsating through cables and airwaves across the world. This data is being created every minute of the day through the most innocuous of online activity that many of us barely even notice. But with every website browsed, status shared, or photo uploaded, we leave digital trails that continually grow the hulking mass of big data. Below, we explore how much data is generated in one minute on the Internet.

EVERY MINUTE of the DAY

THE MOBILE WEB RECEIVES **217** NEW USERS.

YOUTUBE USERS UPLOAD **48** HOURS OF NEW VIDEO.

EMAIL USERS SEND **204,166,667** MESSAGES.

GOOGLE RECEIVES OVER **2,000,000** SEARCH QUERIES.

FACEBOOK USERS SHARE **684,478** PIECES OF CONTENT.

WORDPRESS USERS PUBLISH **347** NEW BLOG POSTS.

CONSUMERS SPEND **$272,070** ON WEB SHOPPING.

571 NEW WEBSITES ARE CREATED.

TWITTER USERS SEND OVER **100,000** TWEETS.

FOURSQUARE USERS PERFORM **2,083** CHECK-INS.

APPLE RECEIVES ABOUT **47,000** APP DOWNLOADS.

FLICKR USERS ADD **3,125** NEW PHOTOS.

INSTAGRAM USERS SHARE **3,600** NEW PHOTOS.

TUMBLR BLOG OWNERS PUBLISH **27,778** NEW POSTS.

BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE **34,722** "LIKES."

WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING

These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

The global Internet population grew 6.59 percent from 2010 to 2011 and now represents **2.1 BILLION PEOPLE.**

These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.

SOURCES: HTTP://NEWS.INVESTORS.COM/, ROYAL.PINGDOM.COM, BLOG.GROVO.COM, BLOG.HUBSPOT.COM, SIMPLYZESTY.COM, PCWORLD.COM, BIZTECHMAGAZINE.COM, DIGBY.COM

Where Are The Data From?

Business

- Financial transactions: Wal-Mart Headquarter, stock traders

Where Are The Data From?

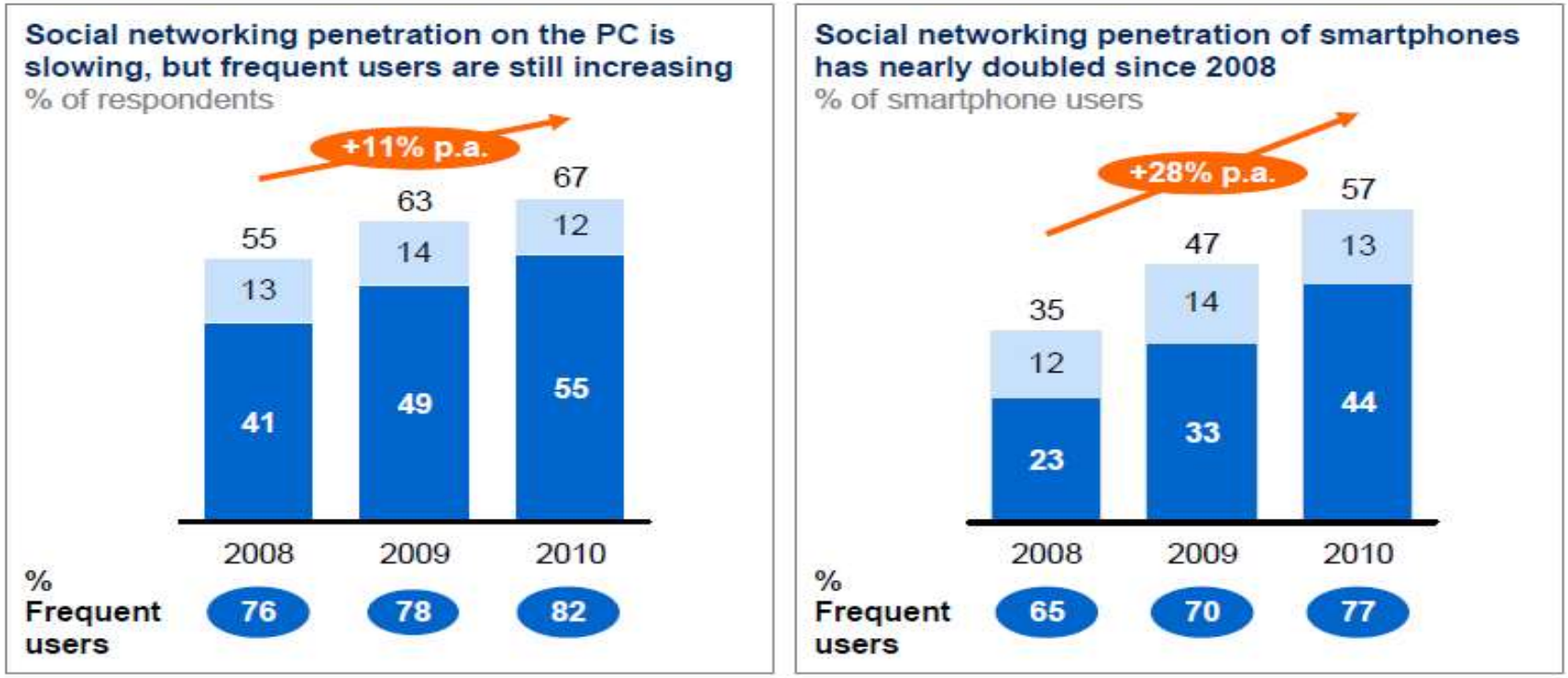
Social media

- Weblogs, twitter feeds, click streams, WeChat/Blog, etc.

Social Networks and Mobile Devices

The penetration of social networks is increasing online and on smartphones; frequent users are increasing as a share of total users[1]

Frequent user[2]



1 Based on penetration of users who browse social network sites. For consistency, we exclude Twitter-specific questions (added to survey in 2009) and location-based mobile social networks (e.g., Foursquare, added to survey in 2010).
2 Frequent users defined as those that use social networking at least once a week.

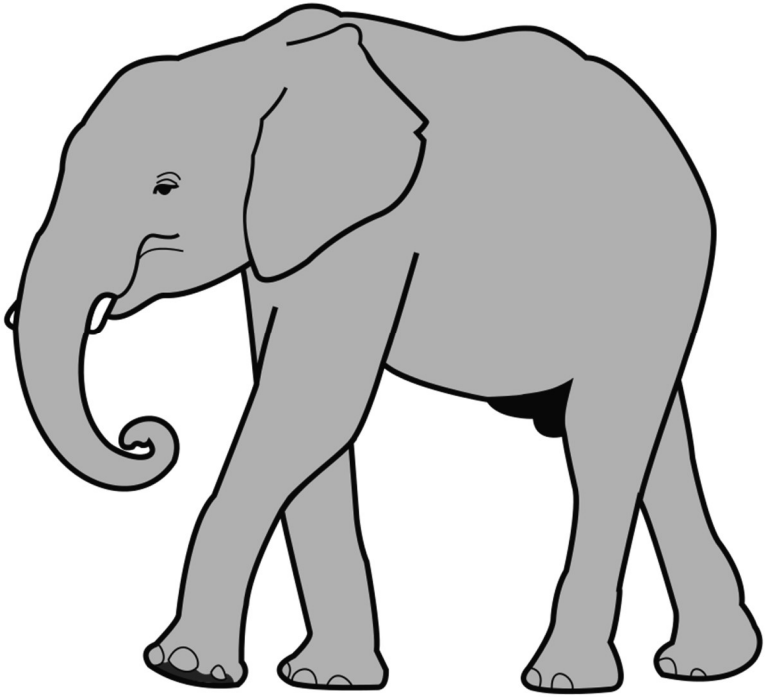
SOURCE: McKinsey iConsumer Survey

The Problem

No matter which type of data is considered, we need a high-performance end-to-end computing solution to support data *generation*, *storage*, *transfer*, *processing*, and *analysis*

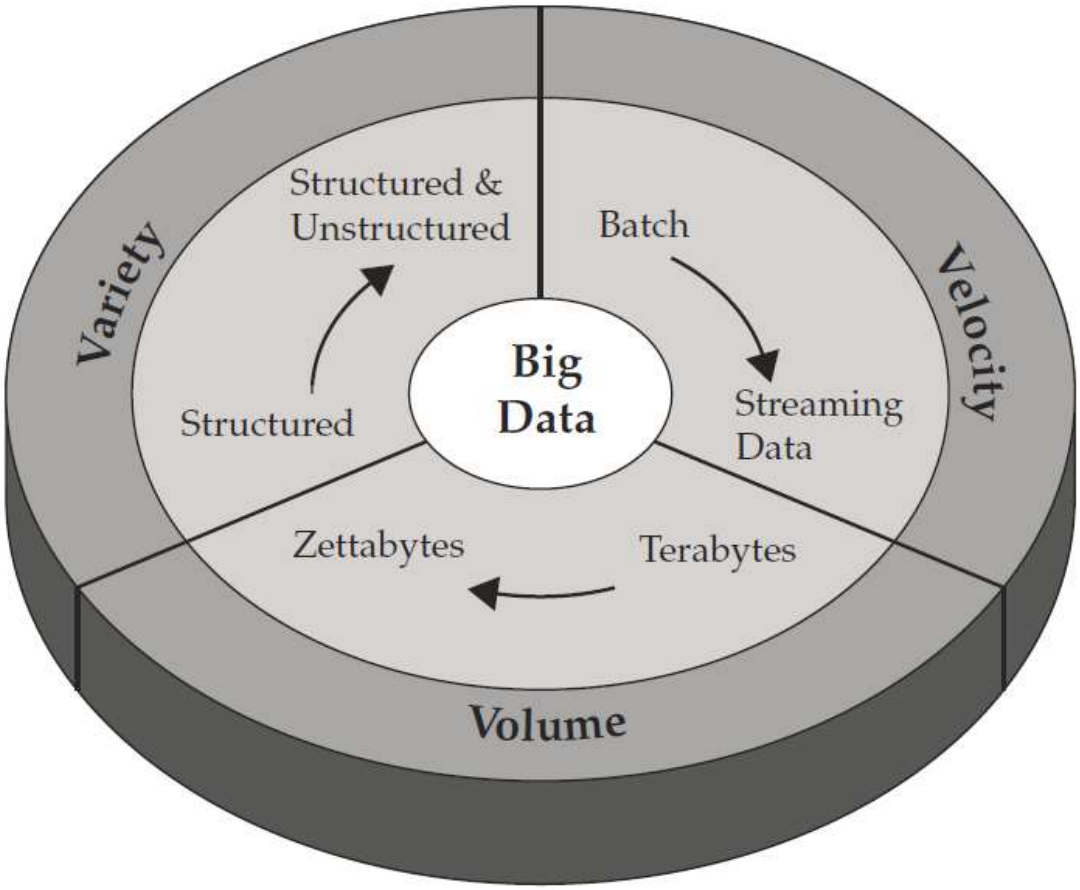
From *T*(erabyte) to *P*(etabyte), to *E*(xabyte), to *Z*(ettabyte), to *Y*(ottabyte), and beyond…

v.s.



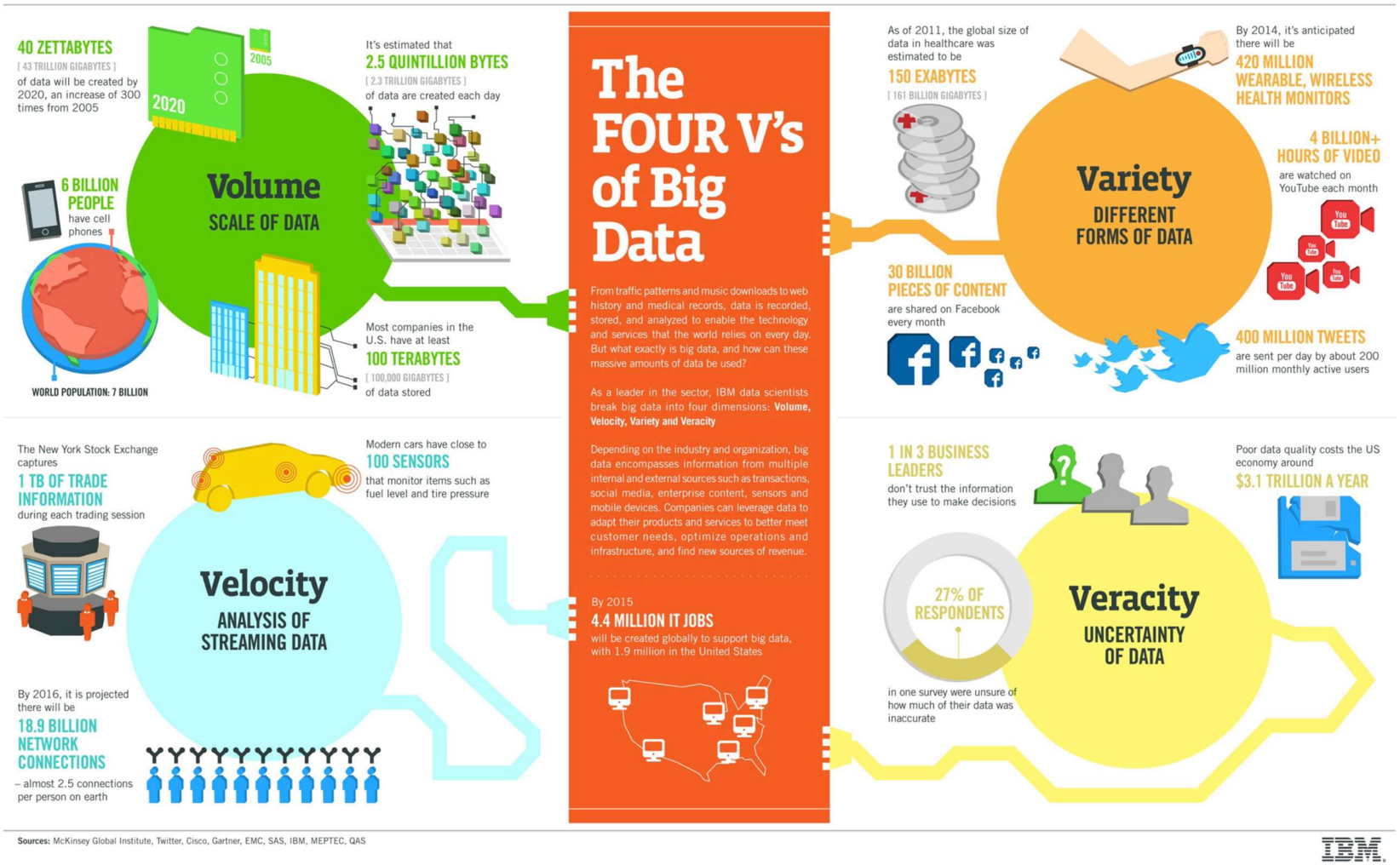
Characterization of Big Data

Types, Quality, Relationships, and Processing of Data



Characterization of Big Data

Four V's of Big Data



Big Data Workflows

Require massively distributed resources

- Hardware: computing facilities, storage systems, special rendering engines, display devices (tiled display, powerwall, etc.), network infrastructures, etc.
- Software: domain-specific data analytics/processing tools, programs, etc.
- Data: real-time, archival

Feature different complexities

- Simple case: linear pipeline (a special case)
- Complex case: Directed Acyclic Graph (DAG)-structured graph

Different application types have different performance requirements

- Interactive: minimize total end-to-end delay for fast response
- Streaming: maximize frame rate to achieve smooth data flow

Why Big Data?

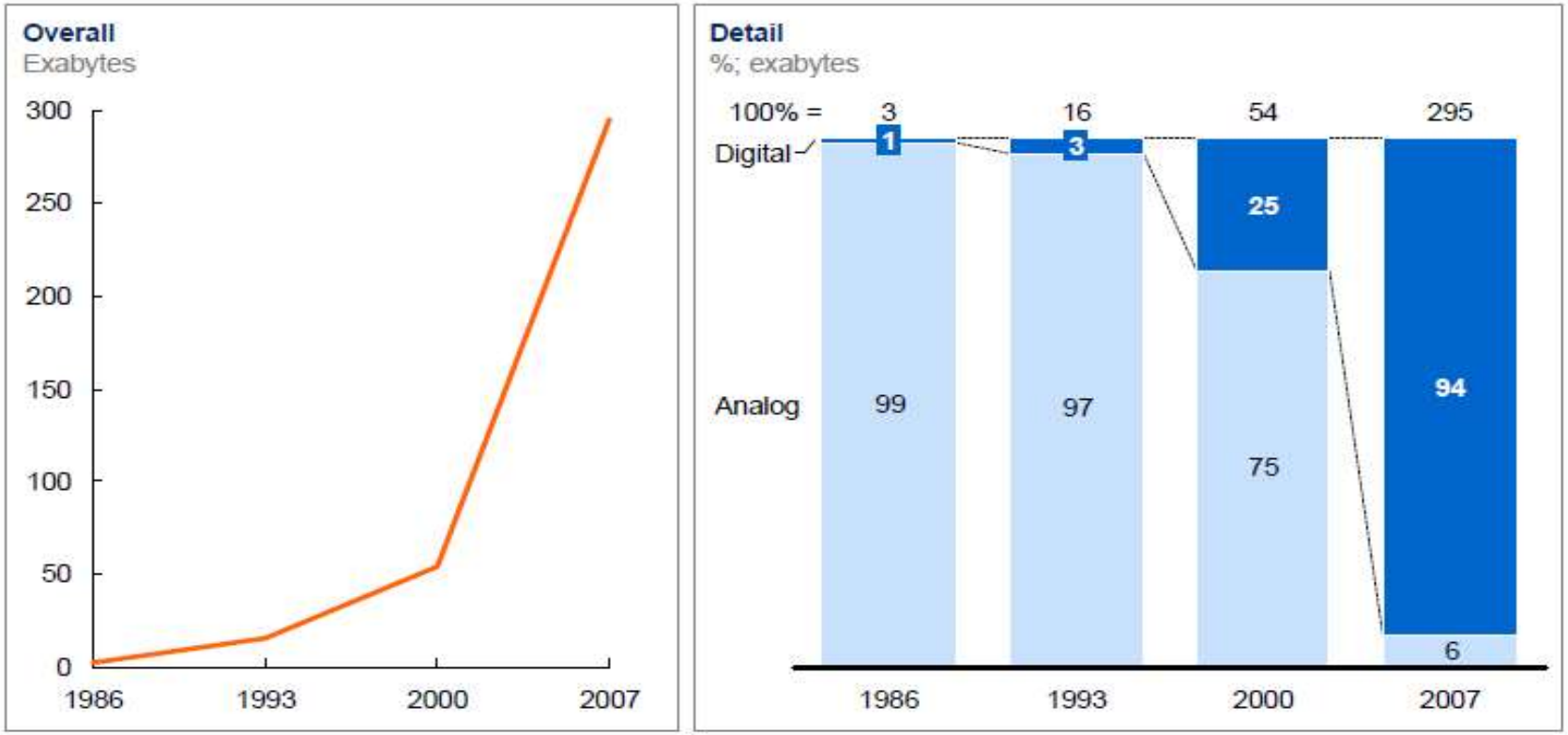
Key enablers for the growth of data

- Increase of storage capacities
- Increase of computing power
- Availability of data

Increase of Storage Capacities

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Increase of Storage Capacities (cont'd)

161 EB captured and replicated in 2006

281 EB captured and replicated in 2007

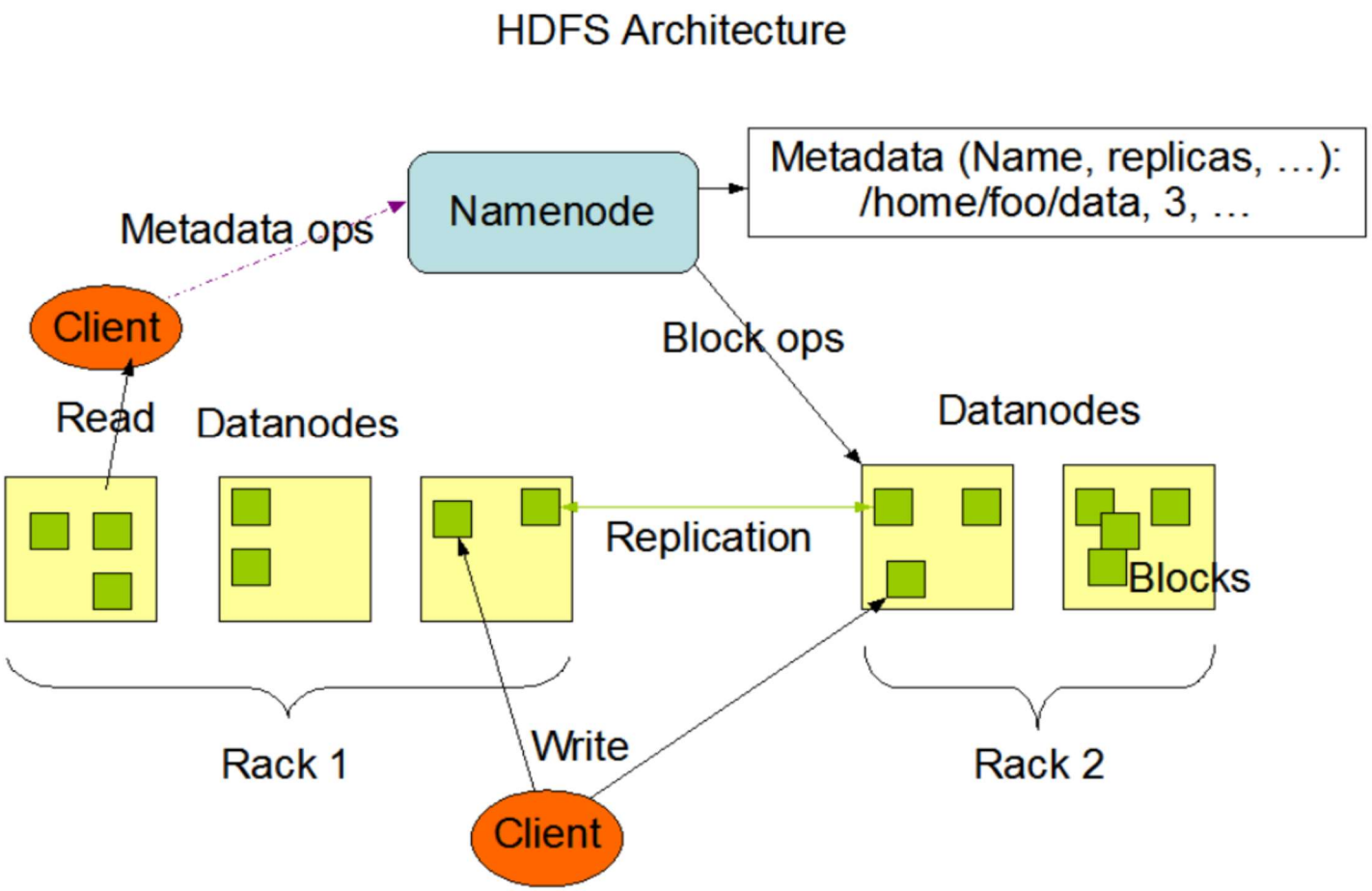
264 EB of available storage in 2007

1800 EB captured and replicated in 2011

IDC
Analyse the Future

Increase of Storage Capacities (cont'd)

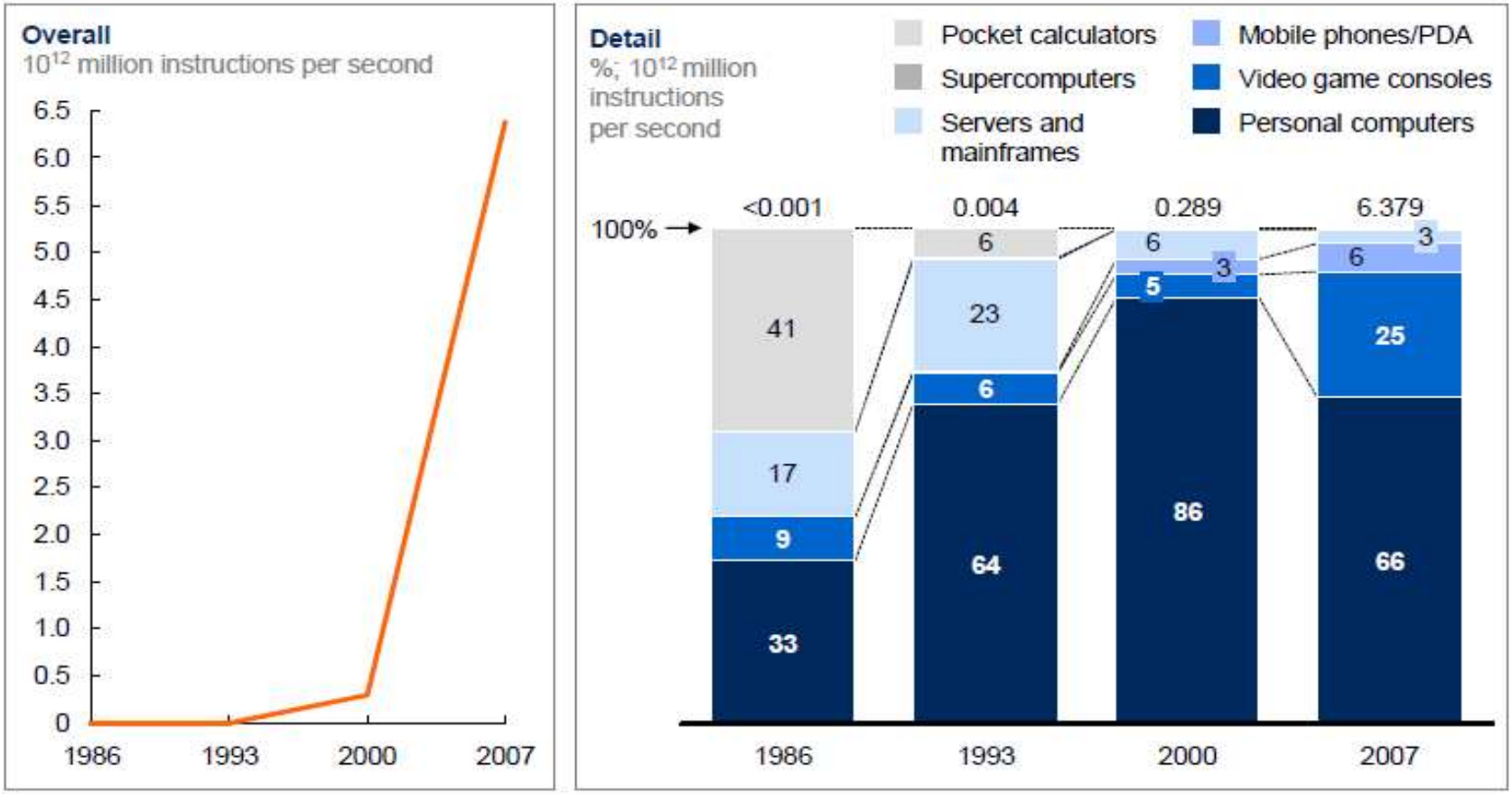
Hadoop Distributed File System (HDFS)



Computing Power

Computation capacity has also risen sharply

Global installed computation to handle information



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

World Class Facilities

World Class Facilities

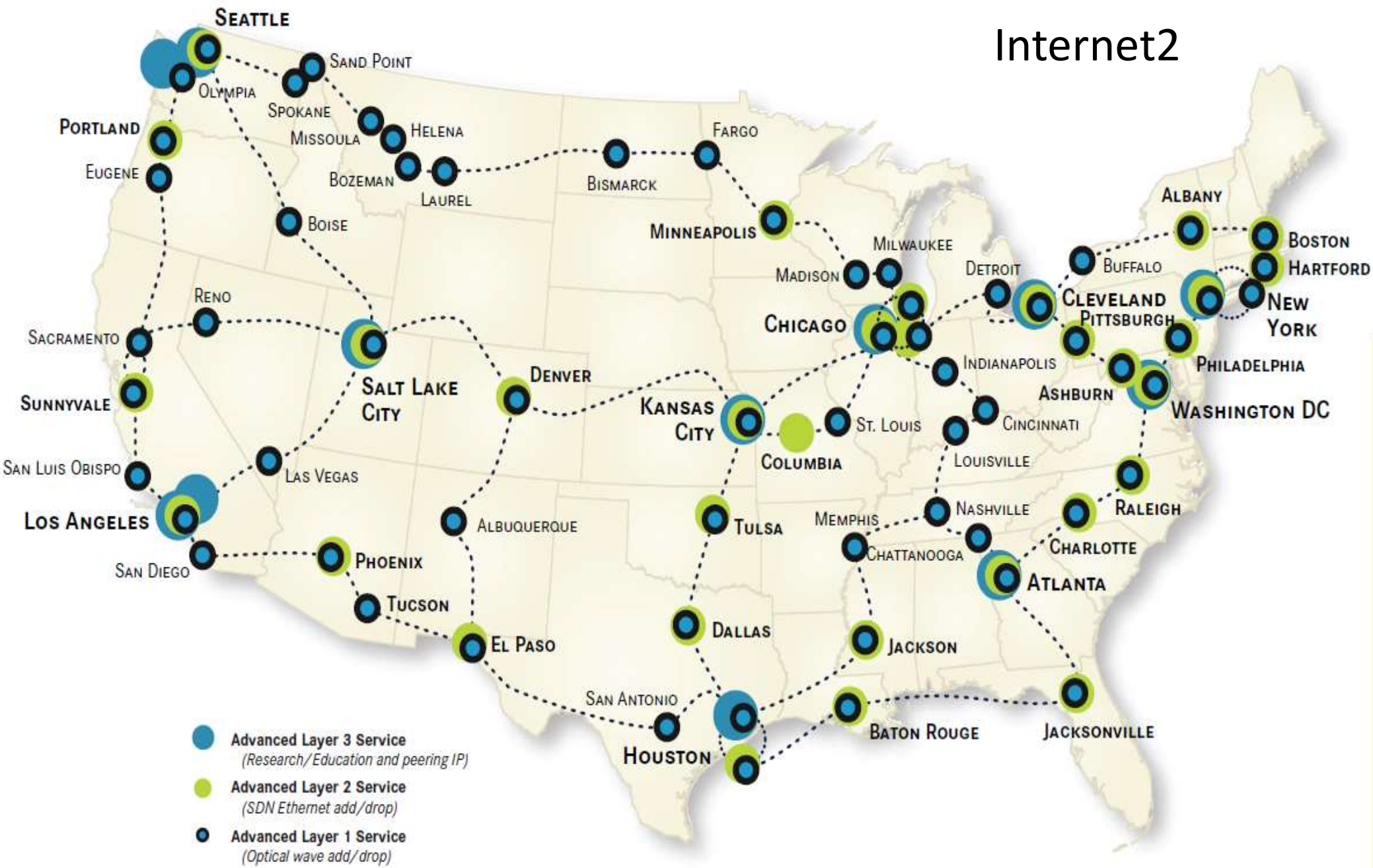
World Class Facilities

Sunway, Taihu Light

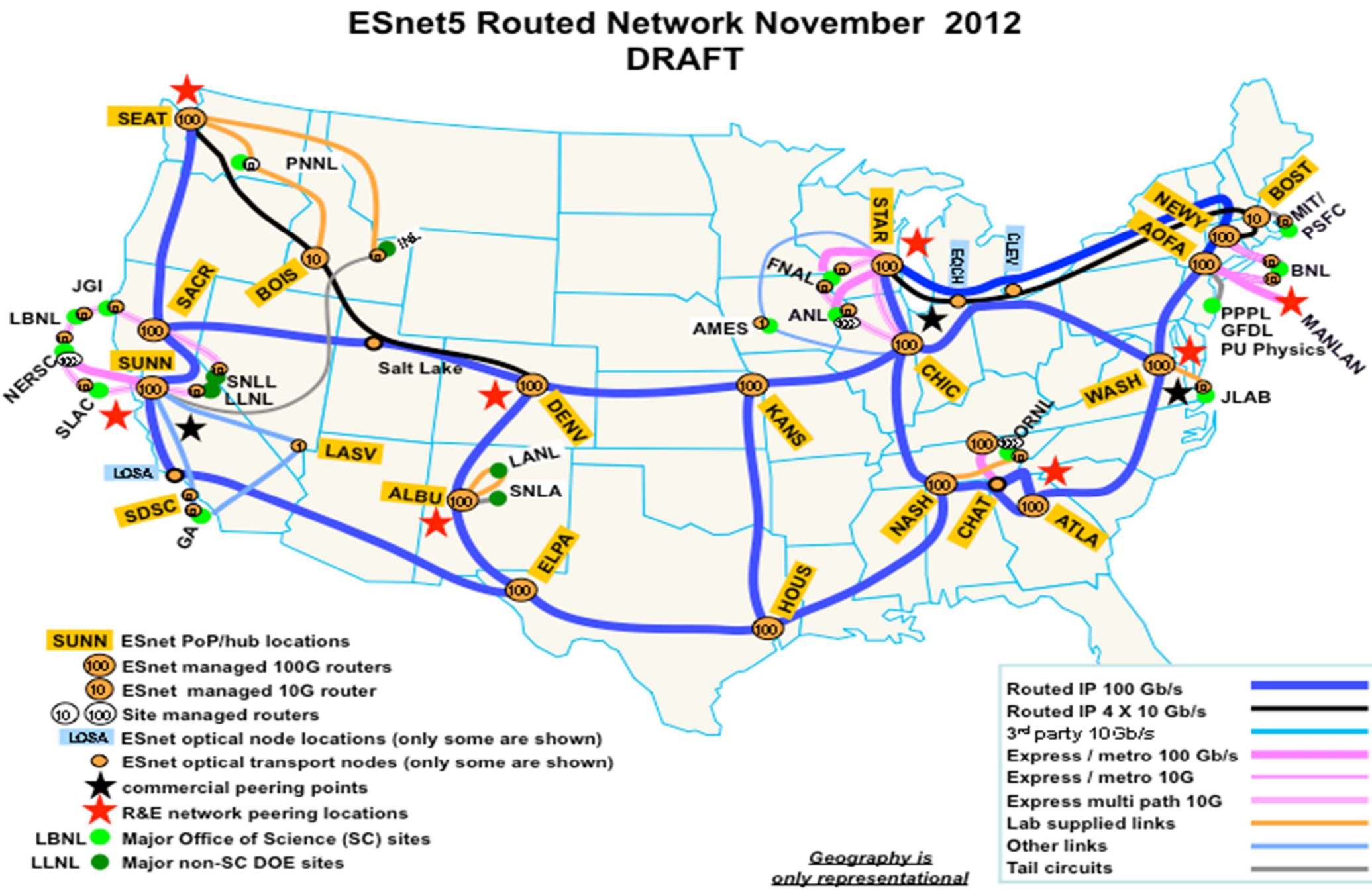
PC Cluster

http://aggregate.org/press001122.html

High-performance Networks



High-performance Networks



Computing Power (cont'd)

Why do we care about computing power?

Computer Security: Exhaustive Key Search

- Two types of security
  - Computational security
  - Unconditional security
- Attack on computational security
  - Always possible to simply try every day
  - Most basic attack, proportional to key size
  - Assume either know / recognize plaintext

Computing Power (cont'd)

Why do we care about computing power?

Computer Security: Exhaustive Key Search

- Two types of security
  - Computational security
  - Unconditional security
- Attack on computational security
  - Always possible to simply try every day
  - Most basic attack, proportional to key size
  - Assume either know / recognize plaintext
- Question?
  - How long does a brute force approach take on average to crack a key of 56 bits using a PC that is capable of performing 1 decryption per microsecond?

Computing Power (cont'd)

Why do we care about computing power?

Computer Security: Exhaustive Key Search

- Two types of security
  - Computational security
  - Unconditional security
- Attack on computational security
  - Always possible to simply try every day
  - Most basic attack, proportional to key size
  - Assume either know / recognize plaintext
- Question?
  - How long does a brute force approach take on average to crack a key of 56 bits using a PC that is capable of performing 1 decryption per microsecond? (*1142 years*)

Computing Power (cont'd)

Why do we care about computing power?

Computer Security: Exhaustive Key Search

- Two types of security
  - Computational security
  - Unconditional security
- Attack on computational security
  - Always possible to simply try every day
  - Most basic attack, proportional to key size
  - Assume either know / recognize plaintext
- Question?
  - How long does a brute force approach take on average to crack a key of 56 bits using a PC that is capable of performing 1 decryption per microsecond? (*1142 years*)
  - How about using a supercomputer that is capable of performing 10^6 decryptions per microsecond?

Computing Power (cont'd)

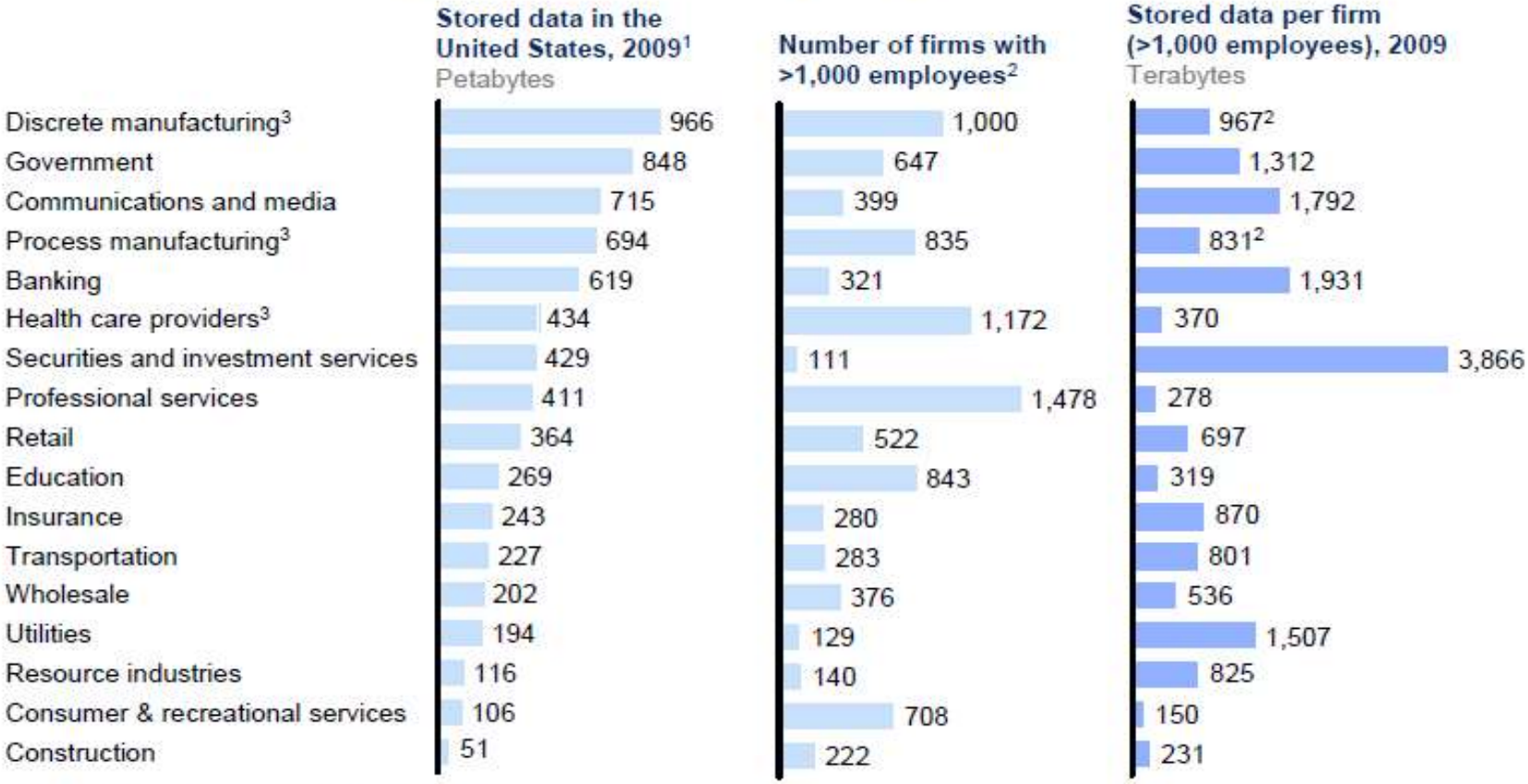
Why do we care about computing power?

Computer Security: Exhaustive Key Search

- Two types of security
  - Computational security
  - Unconditional security
- Attack on computational security
  - Always possible to simply try every day
  - Most basic attack, proportional to key size
  - Assume either know / recognize plaintext
- Question?
  - How long does a brute force approach take on average to crack a key of 56 bits using a PC that is capable of performing 1 decryption per microsecond? (*1142 years*)
  - How about using a supercomputer that is capable of performing 10^6 decryptions per microsecond? (*10.01 hours*)

Availability of Data

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

Types of Available Data

The type of data generated and stored varies by sector[1]

	Video	Image	Audio	Text/ numbers
Banking	Medium	Medium	Medium	High
Insurance	Low	Low	Low	High
Securities and investment services	Low	Low	Low	High
Discrete manufacturing	Medium	Medium	Low	High
Process manufacturing	Medium	Medium	Low	High
Retail	Medium	Low	Low	High
Wholesale	Low	Low	Low	High
Professional services	Medium	Medium	Medium	High
Consumer and recreational services	Medium	Low	Medium	Low
Health care	Low	High	Low	High
Transportation	Low	Low	Low	High
Communications and media[2]	High	Medium	High	High
Utilities	Low	Low	Low	High
Construction	Low	Medium	Low	Low
Resource industries	Medium	Medium	Low	High
Government	High	Medium	High	High
Education	High	Medium	High	Low

Penetration

- High
- Medium
- Low

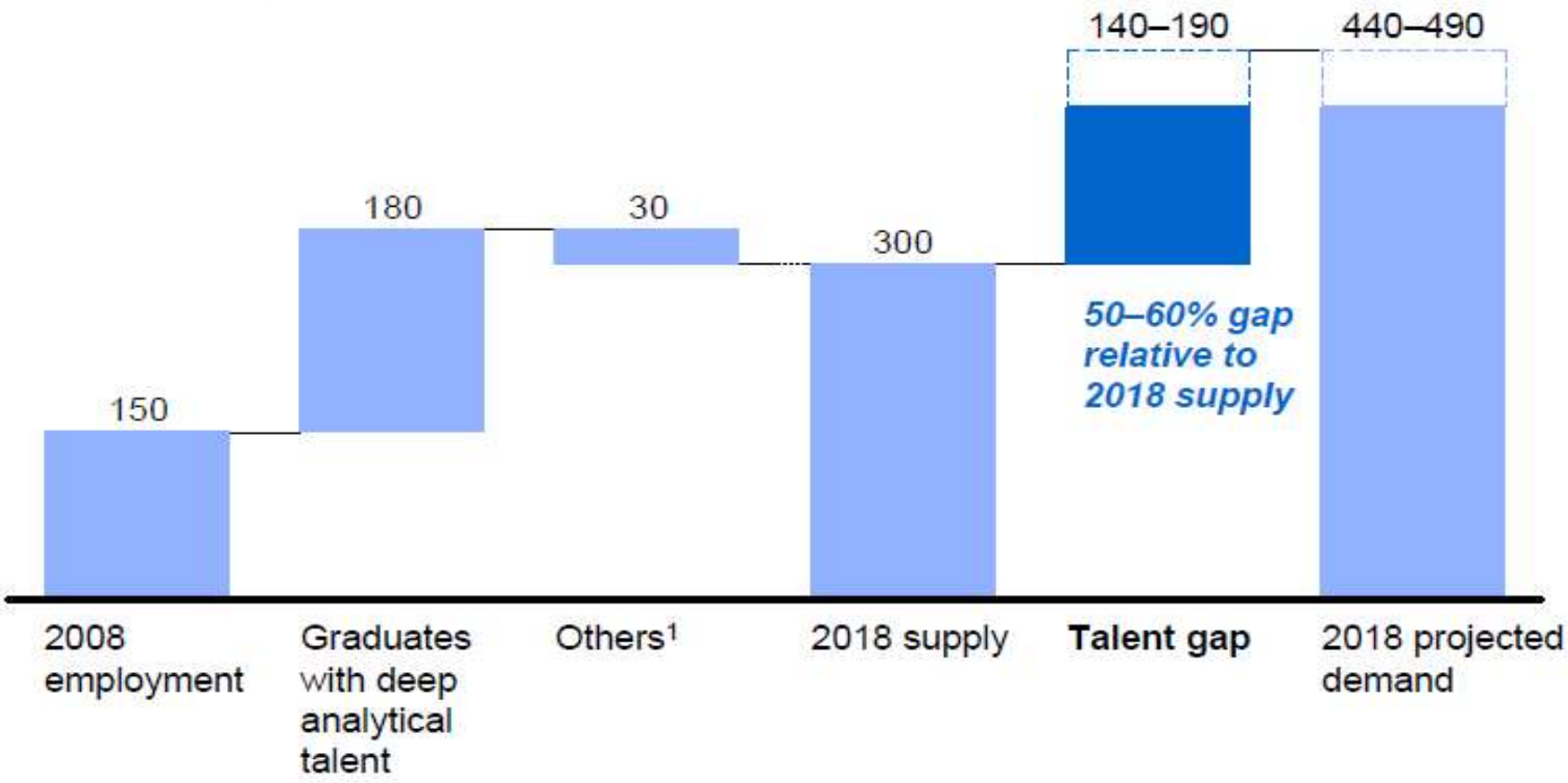
1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

Platform, Techniques, and Tools

Predicted Lack of Talent for Data Mining

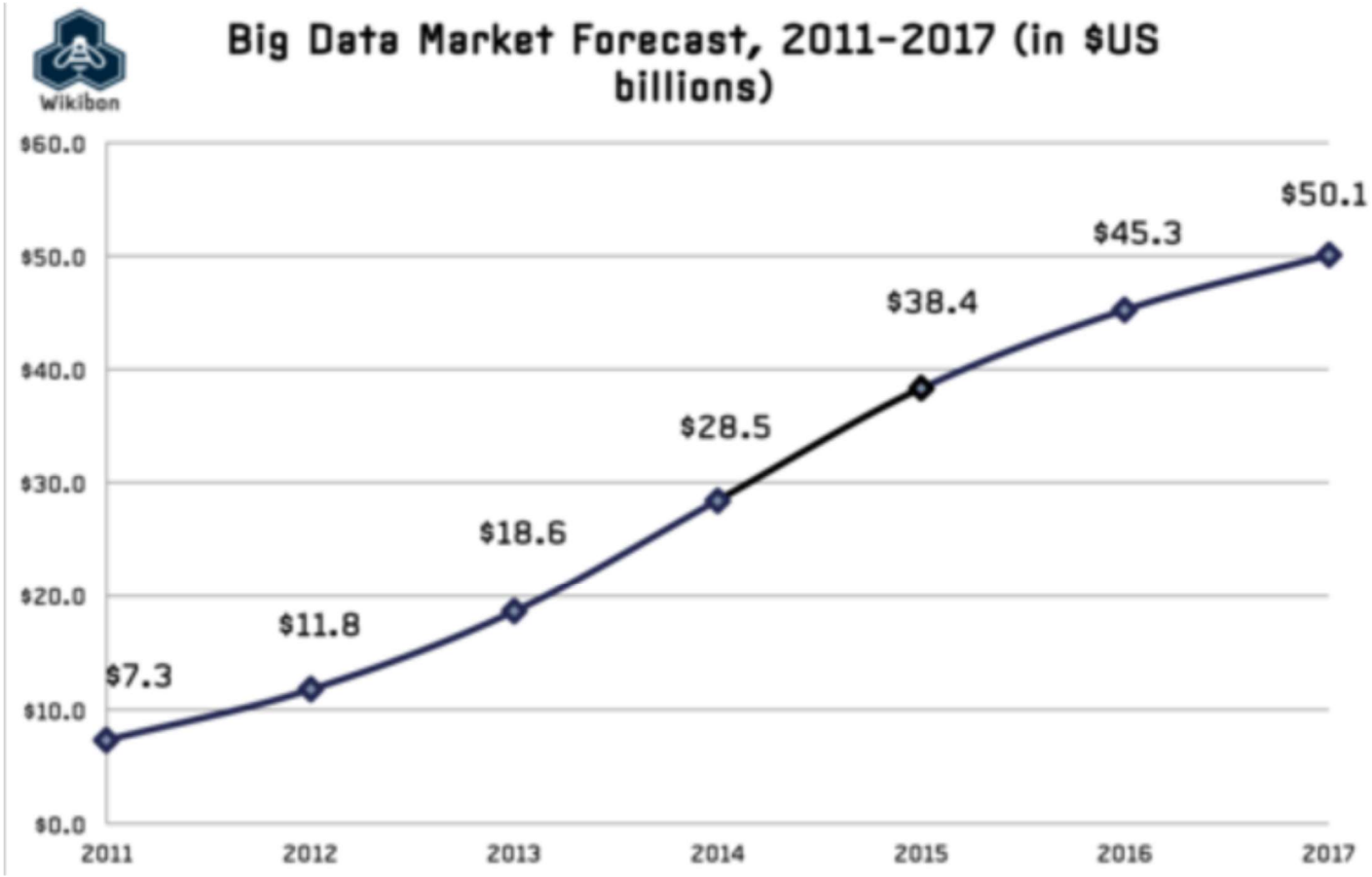
Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people

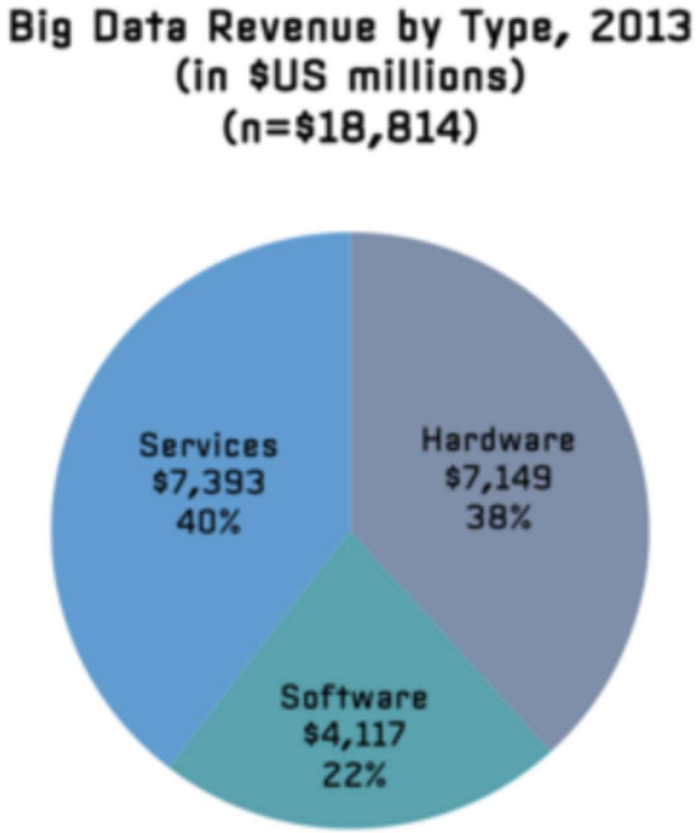


1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

The Market


Wikibon
Big Data Market Forecast, 2011-2017 (in $US billions)
$60.0
$50.0
$40.0
$30.0
$20.0
$10.0
$0.0
$7.3
$11.8
$18.6
$28.5
$38.4
$45.3
$50.1
2011
2012
2013
2014
2015
2016
2017

The Market (cont'd)



Typical Operation on Data

Smart sampling of data

- Reducing the original data while not losing the statistical properties of data

Finding similar items

- Efficient multidimensional indexing

Incremental updating of the models

- (vs. building models from scratch)
- Crucial for streaming data

Distributed linear algebra

- Dealing with large sparse matrices

Supervised learning

- classification, regression, etc.

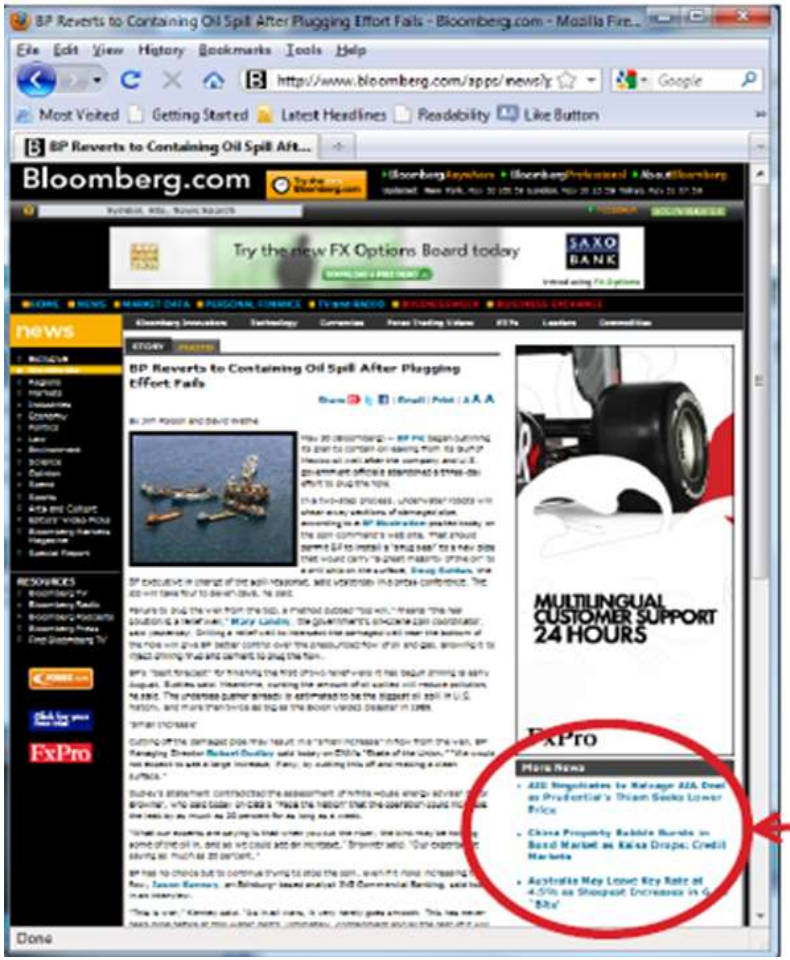
Non-supervised learning

- Clustering, different types of decompositions, etc.

…

Applications

Recommendation



- Good recommendations can make a big difference when keeping a user on a web site
  - ...the key is how rich the context model a system is using to select information for a user
  - Bad recommendations <1% users, good ones >5% users click
  - 200clicks/sec

Contextual personalized recommendations generated in ~20ms

Applications (cont'd)

The context of each click on the web site used for recommendation

- Domain
- Sub-domain
- Page URL
- URL sub-directories

- Page meta tags
- Page title
- Page content
- Named entities

- Has query
- Referrer query

- Referring domain
- Referring URL
- Outgoing URL

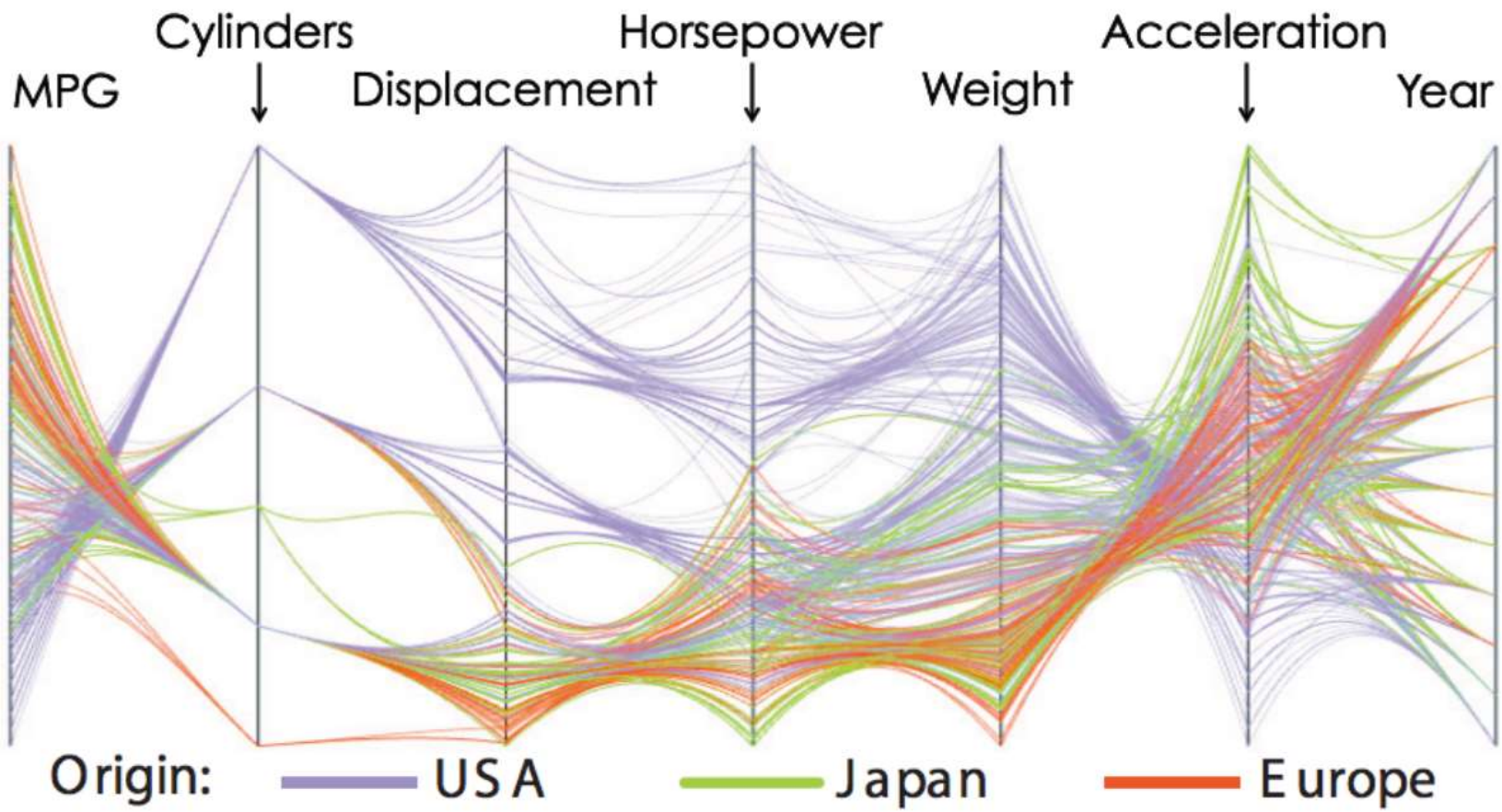
- GeoIP country
- GeoIP state
- GeoIP city

- Absolute data
- Day of the week
- Day period
- Hour of the day
- User agent

- Zip code
- State
- Income
- Age
- Gender
- Country
- Job title
- Job industry

Applications (cont'd)

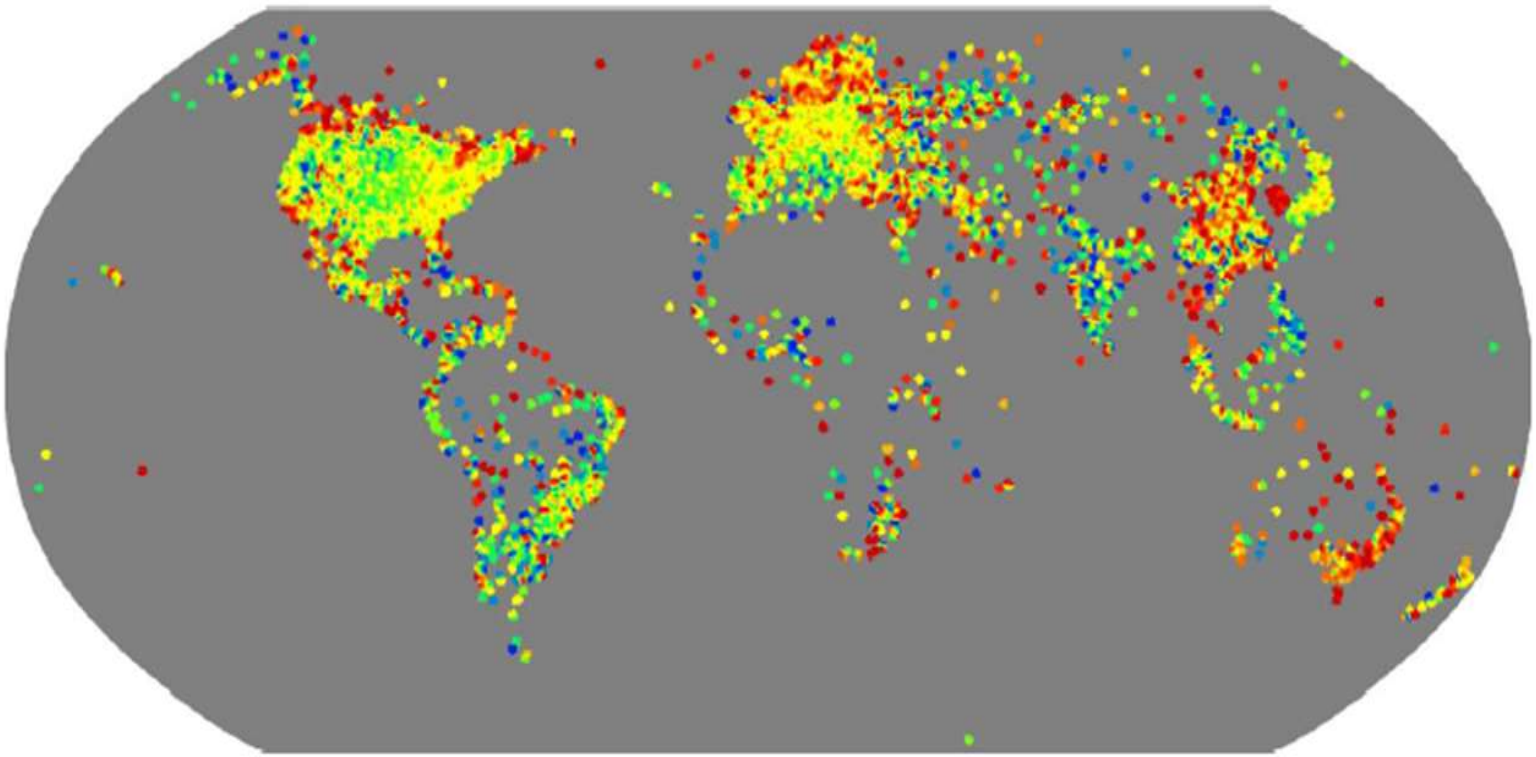
Multi-dimensional data



Applications (cont'd)

Geography and communication

- Count the number of users logging in from particular location on the earth



Applications (cont'd)

Tweets text mining

- An Introduction to Text Mining using Twitter Streaming API and Python
- http://adilmoujahid.com/posts/2014/07/twitter-analytics/

Ongoing Research Topics

Data-intensive Computing

- Scientific Workflow Optimization
  - Mapping, scheduling, and modeling
  - Various network environments, e.g., HPN, grids, clouds, PC clusters, etc.

High-performance Networking

- Bandwidth scheduling
- Transport control
- Transport profiling
- Big data movement

Review of the Calendar

For Today

Project proposal

- Introduction
  - What you plan to do in your course project
- Source datasets
  - What datasets you'd like to analyze/process
- Data analysis/processing/mining tools
  - You certainly can write your *own* code
- Timeline schedule
- Guideline:
  - https://goo.gl/0rm8uj
  - Course project road map

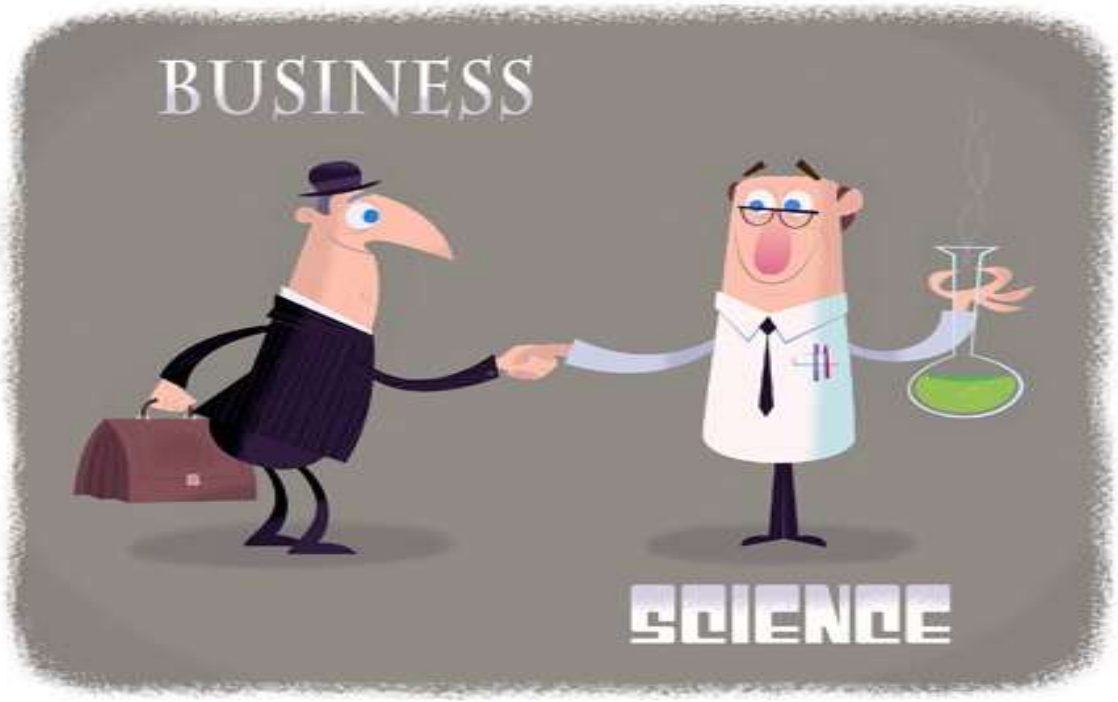
For Today

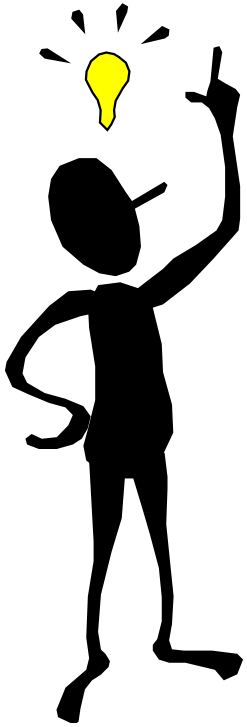
Team information

- By next meeting on Thursday
- 23:59 p.m. (EST), May 18, 2017

Conclusion

- Enjoy! ;-)





Thanks ! ☺

Questions ?