Challenges Of Big Data
In Scientific Discovery

Benjamin W. Wah
Provost, The Chinese University of Hong Kong
Emeritus, University of Illinois, Urbana-Champaign

May 2014

Outline

• Introduction
  – What is Big Data
  – Growth of Big Data and its Applications
  – National Initiatives

• Big Data Challenges
  1. Diversified Nature of Big Data
  2. Diversified Application Domains
  3. Infrastructure Supports

• Future Outlook
INTRODUCTION

What is Big Data?

Big Data and Its 5V Properties

Big Data (Wikipedia, 2012)
• Within tolerable elapsed time period
• Using existing hardware/software infrastructure
• Difficulty to capture, manage, and process data

Volume
• Large Volume
  - 40 ZB by 2020, 5.2 TB per person

Velocity
• High Velocity
  - 2.5B items/day, Over 500 TB/day

Variety
• Many Varieties

Veracity
• Difficult to Verify
  - News and rumors from twitters, SMS, and blogs

Value
• High Value
  - Great value to countries and industries
Categories of Big Data

• Data from the physical world
  – Obtained thru sensors, scientific experiments & observations
  – Biological, neural, astronomical, remote sensing data, etc.

• Data from human activities
  – Obtained through social networks, Internet, health, finance, economics, transportation, etc.

Big Data in the Spotlight

• *Nature*, Special Issue on “Big Data” (2008)
  – Challenges that come with advances on Internet, super-computing, environmental science, biological science

• *Science*, Special Issue on “Dealing with Data” (2011)
  – Opportunities in scientific and societal advances

• *Nature Physics*, Editorial on “Complexity” (2012)
  – Big Data brings opportunities to complex scientific research

All rights reserved, May 2014
Some Recent Big Data Conferences

• Big Data Conference: Washington DC, 2012 (http://www.bigdataconference.net/)

• ACM SIGMOD Conference, 2013 (http://www.sigmod.org/2013/ctcbd.shtml)

• Hadoop Summit, 2013 (http://hadoopsummit.org/san-jose/)

• IEEE 2\textsuperscript{nd} Int’l Congress on Big Data, 2013 (http://www.ieeebigdata.org/2013/)

• 2013 IEEE Int’l Conf. on Big Data (http://cci.drexel.edu/bigdata/bigdata2013/)

All rights reserved, May 2014

Big Value for Big Data

• McKinsey Report on Big Data (June 2011)

Big data can generate significant financial value across sectors

**US health care**
- $300 billion value per year
- ~0.7 percent annual productivity growth

**Europe public sector administration**
- €250 billion value per year
- ~0.5 percent annual productivity growth

**Global personal location data**
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

**US retail**
- 60+\% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

**Manufacturing**
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis
Big data is at the trough of disillusionment

- IBM
- Accel Partners
- Sumo Logic
- Trifacta
- RelateIQ
- Cloudera
- Hadoop (10 times by 2016)

INTRODUCTION

Growth of Big Data and its Applications
Jim Gray’s Fourth Paradigm

- Thousand years ago
  - Experimental Science—description of natural phenomena

- Last few hundred years
  - Theoretical Science—Newton’s Laws, Maxwell’s Equations…

- Last few decades
  - Computational Science—simulation of complex phenomena

- Today
  - Data Intensive Science—from hypothesis-driven to data-driven

Evolution of Big Data in the Last 60 Years

50’s  60-70’s  80’s  90’s  00’s  10’s  2020

- Databases
- Data Engineering
- Data Mining
- Information Science
- Information Engineering
- Knowledge Engineering
- Knowledge Discovery

Big Data?

All rights reserved, May 2014
Some Big Data Applications

- Data existing in nature (collected by modern technologies)
  - Meteorology
  - Genomics
  - Biological and environmental research (remote sensing)
  - Astronomy (200 GB/night in Sloan Digital Sky Survey)
  - Connectomics (mapping all synaptic connections in brain)

- Data existing in products of engineering work
  - Complex physical experiments & simulations
  - Large Hadron Collider (13 petabytes/year)
  - Internet search engines (Google: 20 PB data/day, > 400 PB/month; Baidu: > 100 PB)
  - Social networks (Facebook: 850 M reg. users, 1 B photos/month, > 300 TB/day)
  - Sensor networks (RFIDs, cameras, microphones, mobile sensors)
  - Electronic commerce (Taobao: 370 M users, 880 M products, >20 TB/day)
  - Software logs
  - Finance (business news, financial data, high frequency transactions)
  - Business informatics (Wal-Mart with $10^8$ transactions/hour or 2.5 PB/hour)
  - Cellular phones (~5B mobile subscribers / ~7B people)
Scientific Discoveries with Big Data

• Availability of
  – Low-cost sensors operated by many over long durations
  – Commodity computing
  – Internet connectivity, enabling sharing across disciplines

• Leading to
  – *Data archival* over time
  – *Semantic Webs*: unifying protocols to address non-uniform structure, sampling rates, and standards
  – *Data assimilation*: Integration with model assessment and forecasts
  – *Data discovery*: Derivation of scientific variables from remote sensing data
  – *Collaborative tools* in the cloud

Understanding Oceans

Understanding the complexity of oceans

• Requires documenting and quantifying a myriad collection of processes over time
• Constantly changing and interacting with each other

Source: Center for Environmental Visualization, Neptune Program, & Fourth Paradigm: Data-Intensive Scientific Discovery, Ed. Hey, Tansley, and Tolle
Ocean Observatories Initiative

Using electro-optically cabled observing systems to measure ocean activities in the northeast Pacific Ocean.

Source: Center for Environmental Visualization, Neptune Program, & Fourth Paradigm: Data-Intensive Scientific Discovery, Ed. Hey, Tansley, and Tolle

Pan-STARRS

- Panoramic Survey Telescope and Rapid Response System (Pan-STARRS)
- 2.5 PB of data each year
Environmental Science

Using densely deployed sensor networks to detect possibilities of avalanches

Source: Fourth Paradigm: Data-Intensive Scientific Discovery, Ed. Hey, Tansley, and Tolle

Health & Medicine

• Shortening the time to bring research results to medical practice—convergence in 2025
  – Massive patient information database
  – Potential to explore outcomes of patients on a new treatment across the whole population

• Bioinformatics
  – Over 5000 genome projects in 2010 (~EB size)
  – Need for storage, analysis and visualization
  – Translation into applied science

• Biological research
  – Understanding spatiotemporal data in humans
INTRODUCTION

National Initiatives

US Big Data Initiatives

• US $200m initiative announced in 2012
  – Transform ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security
  – Prepare the next generation of data scientists and engineers
  – Seeking a 100-fold increase in the ability of analysts to extract information from texts in any language

• 6 Federal departments and agencies
  – NSF, HHS/NIH, DOD, DOE, DARPA, USGS
US Big-Data Application Focuses

• Health and well-being
• Environment and sustainability
• Emergency response and disaster resiliency
• Manufacturing, robotics and smart systems
• Secure cyberspace
• Transportation and energy
• Education and workforce development

Some European Efforts

• The European Commission
  – 2-year-long Big Data Public Private Forum through their Seventh Framework Program to engage companies, academics and other stakeholders in discussing Big Data issues.
  – Define a research and innovation strategy to guide a successful implementation of Big Data economy.
  – Outcomes to be used as input for Horizon 2020, their next framework program
Big Data Applications in China

- Internet (media, service, electronic businesses)
- Telecommunications (service providers, basic infrastructure, equipment manufacturers)
- Cyberspace security
- Smart city (administration, logistics)
- Finance
- Health and medicine
- Materials and manufacturing
- Bioinformatics and pharmaceutical research

BIG DATA CHALLENGES
7D on Big Data Research

- Diversity on data properties
- Diversity on representations
- Diversity on applications
- Diversity on goals / objectives
- Diversity on algorithms
- Diversity on theoretical foundation
- Diversity on infrastructures

Big Data is a Phenomenon

Challenges

1. Diversified nature of big data
   - Multi-disciplinary, unstructured, noisy, with possibly missing components
   - Sensing, collection, storage, access, visualization

2. Diversified application domains
   - Diversity on goals, representations, & algorithms
   - Lack of general theoretical foundation

3. Infrastructure supports
CHALLENGE 1

Diversified Nature of Big Data:
• Unstructured, Noisy, and Incomplete
• Sensing, Collection, Storage, Access, Visualization

Hierarchy of Scientific Data
• Raw data (mostly unstructured/semi-structured)
  – Used by overlapping disciplines
• Derived and refined data (structured)
  – Unified and accessed through the Internet
  – Ontologies and meta-data
  – Automated tools for understanding and learning
  – Visualization tools
• Scientific literature
  – Easy access
  – Overlapping disciplines
Networks of Big Data

The Firecracker Galaxy

WWW in 1999
Nature Physics, 2012

All rights reserved, May 2014

Examples of Big Data Networks

Social Network Analysis

Connectors?
Leaders?
Experts?
Bridges?
Isolates?
Clustering?
Understand the diffusion of innovations and news

Financial Network Analysis

Business ties
Money flows
Security ownerships
Analyze systemic risk to avoid financial crash

Telecom Network Analysis

Voice calls
Video calls
SMS messages
Design better service plans
Provide more effective services (e.g., spam filtering)

Detect money laundering

All rights reserved, May 2014

Courtesy of James Cheng
Sensing, Assimilation & Visualization

- Big data networks
  - Distributed sensing
  - Minimum number of sensors
  - Maximum cover
- Abstraction and assimilation
  - Entity and causal relations
  - Complexity reduction
- Visualization and understanding
  - Trend predictions
  - Anomaly detections
  - Value judgments

Abstracting Big Data Networks

Engineering

Global Abstraction
- Global-level properties
- Global-level techniques (knowledge discovery)

Coarse Abstraction
- Cluster-level properties
- High-level techniques (machine learning, statistical analysis, data mining)

Abstracted Networks
- Macro-level properties (k-core, k-truss, clusters, etc.)
- Macro-level techniques (algorithm design, parallel computing)

Network of Big Data
- Micro-level properties (cliques, triangles, etc.)
- Micro-level algorithm design

Science
Example 1: Social Network Big Data

- Integration of physical world (cloud) and cyberspace (social media)
- User-generated Web media
  - Blogs, Twitters, Facebook, etc.
- Journal data
  - Search engine data, financial transactions, electronic commerce
- Rich media
  - Sound, video, interactive media, etc.

Example 1 (cont’d)

Properties
- Many sources
- Interactive
- Real time
- Spontaneous
- Social behavior related
- Highly noisy

Issues
- Sensing at source
- Structure identification
- Understanding & predictions
Example 2: Financial Predictions

• Meltdown modeling
  – Agent-based analysis

• Lenddo and LendUp
  – Using social media activities (Facebook, Twitter) to securitize loan applications

• Market predictions
  – Financial reports, market data, social media data, news reports

Example 3: 2012 US Election

• Data analytics (Nate Silver)
  – Use of many sources
  – Use of past to guide future
  – Extracting information
  – Understanding correlations
  – Statistical models
  – Monte Carol simulations
  – Understanding polls
  – Focus on probabilities

All rights reserved, May 2014
Example 4: PRISM

How PRISM May Work

Scenarios

Hypothetical Working Model (v 1.0) for PRISM: architecture based on currently available public information and statements. To a first approximation, PRISM is a system that allows NSA analysts to request information from named companies within the guidelines of pre-negotiated data sharing arrangements. We give two speculative scenarios A and B for how these queries may be processed.

Example 5: MUSCULAR

• Project MUSCULAR (NSA)
  – In conjunction with UK Gov’t Communications Headquarter
  – NSA collected more than 181 million records from Yahoo and Google networks in 30 days
  – Text, audio, video and metadata indicating who sent or received emails
  – Intercepting the flow of data in the fiber-optic cables linking data centers around the world
Challenge 1: Summary

• Data collection is a major issue in Big Data research
  – Impossible to do it by individuals
  – Need supports from institutions

CHALLENGE 2

Diversified Application Domains:
• Science or Engineering?
• Lack of Theoretical Foundation
**Big Data: Science or Engineering?**

- **Application dependent**
- **Science** — To discover new knowledge
  - Fundamental network properties, such as complexity
  - Partitioning and scalability, based on hierarchical networks
  - Learning and generalization
- **Engineering** — To apply knowledge to new things
  - Design and innovation
- **Challenges**
  - Unclear boundary between science and engineering
  - Difficult to generalize across applications

All rights reserved, May 2014
Theoretical Foundations

- Depends on
  - Data size
  - Data aggregation
  - Data relations
  - Interactivity
  - Intermittency

- Areas of study
  - Data complexity
  - Computational complexity
  - System complexity

Problem 1: Data Complexity

Search approach: Computing each path
(Tradeoff: time $\rightarrow$ space)

Space: $N=10^9$  Time: $2^N$
Accurate computation: impossible
Approximations: efficiency and accuracy cannot be guaranteed

Decision approach: Recording all paths
(Tradeoff: space $\rightarrow$ time)

Space: $2^{10^9}$  Time: $N$
Accurate computation: impossible
Approximations: efficiency and accuracy cannot be guaranteed

Recognition based on structure regularity:
Explore new metrics in data space for consistent reduction in time and space complexity

Data (Data Nature)

Explore pattern and structural regularity
Big Data Complexity Analysis and Models

**Existing work:**
- Observed statistical patterns
- Incomplete measurement of structural regularity, e.g., aggregation/intermittency
- Big data complexity analysis and approximations difficult without structural regularity

**Problem 2: Computational Complexity**

**Background**

Limitations of existing work:
- Limited ability in describing fuzzy and hidden features of heterogeneous data
- Unclear mapping between structural representations and computations
- Difficulty in handling complete data, and poor performance for approximations
Limitations of existing system architectures for big data:

- Focus on computations and storage, while ignoring data life cycle
- Improved concurrency through weak consistency constraints, while ignoring poor reference localities due to heterogeneity of big data
CHALLENGE 3

Infrastructure Supports

Architecture of Big Data System

- Computational Structure
- Novel Computational Model
- Sensing and Measurement
- Preprocessing, Analysis and Mining
- Storage and Management
- Security and Privacy
- Standardization
Infrastructure Supports

Network generated and managed

Centrally managed

Distributed online processing
• Sensing
• Online processing
• Distributed processing
• Abstraction of results

HW / SW architecture support
• Cloud storage
• Central processing

Cloud Storage for Big Data

Integration of Servers with Different Capability

Manage and Transmit Heterogeneous Data in Various Scales

Online/Offline Deep Analysis

System requirements
• ZB: Awareness
• EB: Efficient Storage
• PB: Real-time Computation

Data Center requirements
• Online and real-time
• Data storage
• Data transmissions
Challenge 3: Summary

• Diverse approaches
  – Network generated and managed
  – Centrally managed

• Institutional supports important

FUTURE OUTLOOK
The Future

• Big-data research alone is incomplete
  – Inter-disciplinary

• Driven by application requirements
  – All encompassing
  – Domain-specific knowledge
  – Drawing on broad data sources, including biology, chemistry, clinical medicine, computer science, and mathematical modeling
  – Including structured database records, published articles, semi-structured data, images, raw numeric data, etc.
  – Systematic workflow to manage and access data
  – Generalization and visualization

• Non-unique solutions