Challenges Of Big Data In Scientific Discovery

Benjamin W. Wah
The Chinese University of Hong Kong

December 2013
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

• 973 project on Networked Big Data

December 2013
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

### 5V Properties

<table>
<thead>
<tr>
<th><strong>Volume</strong></th>
<th><strong>Velocity</strong></th>
<th><strong>Variety</strong></th>
<th><strong>Veracity</strong></th>
<th><strong>Value</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Volume</td>
<td>High Velocity</td>
<td>Many Varieties</td>
<td>Difficult to Verify</td>
<td>Great value to countries and industries</td>
</tr>
<tr>
<td>40 ZB by 2020, 5.2 TB per person</td>
<td>2.5B items/day, Over 500 TB/day</td>
<td>News and rumors from twitters, SMS, and blogs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

December 2013
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
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/)
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

Databases

Data Engineering

Data Mining

Information Science

Information Engineering

Knowledge Engineering

Knowledge Discovery

Big Data?

December 2013
• World market size of IDC in 2012 is $22.6B (21.5% growth)
• China IDC market size is RMB 17.1B (67.1% growth)

• $34B in IT spending (2013)
• 4.4M new jobs related to Big Data (2015)
• Gartner predicted that Big Data will be a traditional industry by 2020
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; Beidu: > 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 $10^6$ transactions/hour or 2.5 PB/hour)
  – Cellular phones (~5B mobile subscribers / ~7B people)

December 2013
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 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

December 2013
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

December 2013
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

Network Analysis

WWW in 1999
Nature Physics, 2012

December 2013
Examples of Big Data Networks

Social Network Analysis

- Leaders?
- Experts?
- Bridges?
- Isolates?
- Connectors?
- Clusters?

Understand the diffusion of innovations and news
Examine the spread of rumours
Improve organization performance

Financial Network Analysis

- Business ties
- Money flows
- Security ownerships

Analyze systemic risk to avoid financial crash
Detect money laundering

Telecom Network Analysis

- Voice calls
- Video calls
- SMS messages

Design better service plans
Provide more effective services (e.g., spam filtering)

Courtesy of James Cheng

December 2013
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

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
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

December 2013
Example 4: PRISM

**HOW PRISM MAY WORK**

**Scenarios**

Hypothetical / Working Model (<1.0) for PRISM’s 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.

1. **NSA Analyst** identifies target, makes request, receives data via PRISM
2. **NSA Targeting Feedback Loop**
   - METADATA
   - UPSTREAM
   - OTHER INTEL
3. **NAMED COMPANY**
   - Scenario A: EMPLOYEE(S) RECEIVES & RUNS QUERY, PUSHES DATA TO NSA
   - Scenario B: EMPLOYEE(S) FACILITATE COLLECTION or DIVERT DATA INTO ‘DROP BOX’ FOR NSA
4. **NSA Database**

*Possible padding or similar techniques (e.g., a broad query rather than a specific name) to obscure the precise target of the investigation (Speculation)*

Feedback Welcome

**V1.0 June 13 2013 @ashk4n & semipr0**

http://ashkansoltani.org/2013/06/13/prism-solving-for-x/
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 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
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)

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

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

Space: K = logN, Time: O(K^r)
Structural computation

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

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

Explore pattern and structural regularity
Existing work:

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

**Data reduction methods**
- Kleinberg discovered the small-world phenomenon in large-scale social networks
- Faloutsos et al. reported cascading behavior in large blog graphs
- Pentland et al. observed a power relationship when inferring friendship network structure by using mobile phone data
- Mahoney et al. proposed CUR decompositions for large compressed matrices
- Jordan et al. proposed distributed algorithms and selections of hyperparameters for big data

**Computational theories and algorithms**
- Cervantes et al. proposed SVM minimum enclosing ball clustering for large data sets
- Blei developed online Bayesian algorithms for large-scale topic modeling
- Jordan et al. proposed distributed algorithms and selections of hyperparameters for big data

**Data regularity analysis**
- Kleinberg discovered the small-world phenomenon in large-scale social networks
- Faloutsos et al. reported cascading behavior in large blog graphs
- Pentland et al. observed a power relationship when inferring friendship network structure by using mobile phone data
- Mahoney et al. proposed CUR decompositions for large compressed matrices
- Jordan et al. proposed distributed algorithms and selections of hyperparameters for big data

**Background and challenges**

Problems and methodology

December 2013
Big Data Complexity Analysis and Models

Explore the intrinsic property of big data, study complexity measures, structural reduction, and incremental reduction based on kernel.

Explore the patterns and intrinsic connections of data
- Distributions and emergence
- Intrinsic structure and connections
- Connections across time and space, common structure

Propose application-dependent structural reduction theories

Find complexity measures and establish unified theories
- Dimension of data
- Structural diversity
- Pattern intermittency
- Uncertainty

Propose computation models based on structures and learning
- Stochastic
- Non IID
- Probabilistic graphical models
- Partial incremental updates

December 2013
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

Background

Problem 2: Computational Complexity

Computational theories and algorithms

Data reduction methods

Data representation methods

- Motwani et al.’s first research on stream data
- Vasilescu et al. applied tensor decompositions to facial image ensembles
- Cervantes et al. proposed SVM minimum enclosing ball clustering for large data sets
- Mahoney et al. proposed CUR decompositions for large compressed matrices
- Chakrabarti et al. proposed general graph compression
- Han proposed Graph OLAP
- Blei developed online Bayesian algorithms for large-scale topic modeling
- Jordan et al. proposed distributed algorithms and selections of hyperparameters for big data
- Phan et al. proposed tensor decomposition for complex data (images, music)

December 2013
Life-Cycle Aware Computational Complexity

Simple Computations

- Discover kernel data (K) from existing data (N) in the Life Cycle of networked big data
- Implement structure expressions and simple computations based on kernel data

Incremental Computations

Based on simple computation and real-time rapidly changing data, propose incremental computation theory to support OLAP analysis for streaming data

Existing data: N

Kernel data: K

Computation

Reduction

Existential data

Offline data

Streaming data

Existing data

Sensing

Storage

OLAP analysis

Simple expressions

December 2013
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
Life-Cycle Aware System Complexity

- Traditional data processing systems focus on sensing, storage, and computations in an isolated fashion but not their interactions.
- Big data processing needs to consider the life cycle of data.

Requirements:
- Life-Cycle Aware
- Flexible
- Adaptive
- Real-Time

Design? Evaluation? Optimization?

Interactions

Sensing

Computations

Storage

Sensing

Storage

Computation
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

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

Centrally managed

- Sensing
- Communications

HW / SW architecture support
- Cloud storage
- Central processing

December 2013
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

December 2013
CHINA NATIONAL KEY RESEARCH DEVELOPMENT PLAN (973 PROJECT)

Theory and Applications of Network Big Data Computations
Mining and Utilizing Network Big Data

Sense and Mine
Learn and Predict

Life Cycle of Network Big Data
Sensing → Storage → Computation

Modeling Relations of Network Big Data in Life Cycle
Existing Data ↔ Streaming Data ↔ Offline Data
Unstructured Data ↔ Semi-structured Data ↔ Structured Data

Properties of Network Big-Data Structure
Topology: Agminated
Messages: Paroxysmal
Relation: Heterogeneous
Distance: Diversiform

Network Big Data
Web Media
Log Data
Rich Media

Various Types
Dynamic In Nature
Questionable Source

Computable? Algorithm?
Computational Complexity
Measurement? Representation?
Content Complexity

System Complexity
Design? Optimization?

Our National Key Basic Research Program (973 Program)

December 2013
Research Tasks

Problems

- Content Complexity
- Structure
- Regularity

- Computational Complexity
- Kernel

- System Complexity
- Life Cycle

Major Tasks to be Studied

**Fundamental Theory**

1) Analysis of Computational Model and Complexity

2) Life-Cycle-Aware System Architecture

**Key Technologies**

3) Sensing and Representation

4) Content Modeling and Semantics Understanding

5) Pattern Recognition and Effect Analysis

**System**

6) Demonstration and Application of Integrated Software-Hardware Engine

December 2013
Application 1: Early Forensic Warning Based on Clues from Network

Develop early warning system based on forensics, using real-time sensing and online/offline analysis of big data from logs, forums, BBS, and micro-blog.

System Capability
- **Sequential Data**
  - 10 TB / day
- **Incremental Data**
  - 30 TB / day
- **Sensing, Storage and Analyze**
  - Petabyte level

Goal: Improve Processing Capability by 2 Orders of Magnitude

Analyze Big Data, Discover Criminality

Distributed Network Big Data Platform

December 2013
Application 2: Mining and Predicting Social Development Index

Based on big data source from Xinhua News Agency and Aliyun, deploy distributed platform for mining and predicting Social Development Index, including health, education, price, and pollution index.

- **Content**
  - News
  - Public Opinion
  - Financial
  - Trading

- **Property**
  - **Worldwide** Real-time News Data (Xinhua News Agency)
  - **National** Opinion Data (Xinhua Public Opinion)
  - Nationally Largest Scale Financial Data (Xinhua 08)
  - Nationally Largest Scale Online Trading Data (Aliyun)

December 2013
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