Big Data's Biggest Needs- Deep Analytics for Actionable Insights

Alok Choudhary  
John G. Searle Professor  
Dept. of Electrical Engineering and Computer Science  
and Professor, Kellogg School of Management  
Northwestern University  
choudhar@eecs.northwestern.edu
BIG DATA?
“Data intensive” vs “Data Driven”

Data Intensive (DI)

- Depends on the perspective
  - Processor, memory, application, storage?
- An application can be data intensive without (necessarily) being I/O intensive

Data Driven (DD)

- Operations are driven and defined by data
  - BIG analytics
    - Top-down query (well-defined operations)
    - Bottom up discovery (unpredictable time-to-result)
  - BIG data processing
  - Predictive modeling
- Usage model further differentiates these
  - Single App, users
  - Large number, sharing, historical/temporal

Very few large-scale applications of practical importance are NOT Data Intensive

In Extreme Scale Science domain, we typically focus on “Transactional” thinking
Data Mining, Analytics and Actionable Insights?

Time to Compute ➔ Time to Insights
A Poem

The Unknown

As we know,
There are known knowns.
There are things we know we know.

Conventional Wisdom

• High Humidity results in outbreak of Meningitis
• Customers switch carriers when contract is over

Validate Hypothesis

• Nuclear Reaction happens under these conditions
• Did combustion occur at the expected parameter values
• I think this location contains a black hole

© Alok Choudhary
Northwestern University
The Unknown
As we know,
There are known knowns. There are things we know we know.
We also know
There are known unknowns. That is to say
We know there are some things We do not know.

Top-Down Discovery - We know the question to ask

- Will this hurricane strike the Atlantic coast?
- What is the likelihood of this patient to develop cancer
- Will this customer buy a new smart phone?
The Unknown
As we know, There are known knowns. There are things we know we know. We also know There are known unknowns. That is to say We know there are some things We do not know. 

But there are also unknown unknowns, The ones we don't know We don't know.

• Wow! I found a new galaxy?
• Switch C fails when switch A fails followed by switch B failing
• On Thursday people buy beer and diaper together.
• The ratio $K/P > X$ is an indicator of onset of diabetes.
Who Knew?

The Unknown
As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don't know
We don't know.

—Feb. 12, 2002, Department of Defense news briefing by Donald Rumsfeld
Knowledge Discovery Life-Cycle: Transactional to Relationships – Current to Historical
From multi-dimensional data analytics to relationship mining

Anomaly time series at each node

Climate Network

Edge weights: significant correlations
Nodes in the graph: grid points on the globe

Multivariate Networks

CMIP3 $\rightarrow$ CMIP5 $\Rightarrow$ Climate BIG DATA: 10s of TBs to 10s of PBs

Climate Data
A different way of thinking: Extreme Computing + Big data analytics => Accelerating Discovery

MATERIAL SCIENCE: A “DATA DRIVEN DISCOVERY” WORTH A THOUSAND SIMULATIONS?
Discovery of stable compounds

Calculating many, known materials

Solving unknown materials structures

Dataset of materials properties

Big Data mining

Materials discovery!
Ranking – Approximation is good enough for ranking 😊 (closing the loop)

† indicates a model prediction associated with a known stable ternary compound that had was absent from DFT thermodynamic database; the prediction is thus confirmed, but no crystal structure search was necessary.
Structure-Property Optimization – Try optimization for $10^3$ dimensions

- **Microstructure Representation**
  Features that mathematically or statistically describe microstructures

- **Database Construction**
  Randomly generated microstructure-property pairs with most desired and most undesired objectives

- **Feature Selection**
  Select a small set of “critical” microstructure features

- **Global Optimization**
  Find the value of microstructure that leads to the extremal properties

**Traditional Method**

**Data Mining Method**
Accelerating Time to Insights

Experiment Result: Solution found / Performance vs. Number of Variables

- - - Time consumed
- - Optimum found
Extreme Computing + Big data: Not a single dimensional challenge

Big Data: Challenges

- Velocity
- Variety
- Volume
- Analytics Algorithms
- Visualization
- Scalability and Performance
- Storage and I/O
- Power and Energy Efficiency
- Data Management
- Software
An instrument and a discovery engine

Millions of cores
Each core is like a sensor
Each core generates data based on a model

...Millions of cores
Each core can be a data processor/analyst
Extreme scale system can be a discovery engine

NO other type of sensor can claim this capability!
BDEC: Can we do this type of analytics in-situ?

- Climate, Astronomy, Biology, Earth science
- Advanced data structure to break the inherent sequential data access order of DBSCAN
- Scalable DBSCAN identifies the clusters without sacrificing the quality of the solution
- Strong scaling on astrophysics datasets

Identifying arbitrary shaped structures using astrophysics data (http://arxiv.org/abs/1203.3695)
Right Computing infrastructure? What characteristics do typical analytics functions have?

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Benchmark of Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPECINT</td>
</tr>
<tr>
<td>Data References</td>
<td>0.81</td>
</tr>
<tr>
<td>Bus Accesses</td>
<td>0.030</td>
</tr>
<tr>
<td>Instruction Decodes</td>
<td>1.17</td>
</tr>
<tr>
<td>Resource Related Stalls</td>
<td>0.06</td>
</tr>
<tr>
<td>CPI</td>
<td>1.43</td>
</tr>
<tr>
<td>ALU Instructions</td>
<td>0.25</td>
</tr>
<tr>
<td>L1 Misses</td>
<td>0.023</td>
</tr>
<tr>
<td>L2 Misses</td>
<td>0.003</td>
</tr>
<tr>
<td>Branches</td>
<td>0.13</td>
</tr>
<tr>
<td>Branch Mispredictions</td>
<td>0.009</td>
</tr>
</tbody>
</table>

† The numbers shown here for the parameters are values per instruction.
Data Analytics/Mining applications: Do they have different characteristics?

Clear Implications on architecture, modes, memory hierarchy and other components
Identify similarities and design for co-existence
Develop scalable versions – Pay attention to I/O:
Particularly reads

Parallel hierarchical clustering
- Speedup of 18,000 on 16k processors
- I/O significant at large scale
Good News: Approximation is a TOP Option in analytics => Power aware data analytics

### Power-aware analytics
- **Reduced bit fixed-point representations**
- **Pearson correlation**
  - 2.5-3.5 times faster
  - 50-70% less energy
- **K-means**
  - ~44% less energy with an error of only 0.03% using 12-bit representation
Extreme Computing + Big Data Analytics = BDEC Knowledge Discovery Engine
Thank You!

Alok Choudhary
John G. Searle Professor
Dept. of Electrical Engineering and Computer Science
and Professor, Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu