High Performance Data Mining: An Essential Paradigm for Interdisciplinary Big Data Analytics

Ankit Agrawal
Research Associate Professor,
Department of Electrical Engineering and Computer Science,
Northwestern University

Collaborations:
Alok Choudhary (NU), Wei-keng Liao (NU), BP Gautham (TRDDC),
Surya Kalidindi (GaTech), Greg Olson (NU, QuesTek), NIMS, NIST,
Chris Wolverton (NU), Peter Voorhees (NU), Veera Sundararaghavan
(UMich), Marc De Graef (CMU), Jai Raman (OHSU), David Baker (JC),
Karl Bilimoria (NU), Margaret Danilovich (NU), Salman Habib (ANL)

2017 IEEE Region 4 Workshop on Big Data
October 25, 2017
Research Thrusts

Data Mining + High Performance Computing

- Materials Informatics
- Social Media Analytics
- Healthcare Informatics
• Introduction
  – Big data
  – Data mining
  – High performance computing

• High Performance Data Mining
  – Scalable clustering
  – Scalable association rule mining
  – Scalable community detection

• Interdisciplinary Applications
  – Materials informatics
  – Healthcare informatics
  – Social media analytics
  – Online Tools
• Introduction
  – Big data
  – Data mining
  – High performance computing

• High Performance Data Mining
  – Scalable clustering
  – Scalable association rule mining
  – Scalable community detection

• Interdisciplinary Applications
  – Materials informatics
  – Healthcare informatics
  – Social media analytics
  – Online Tools
Data Data Everywhere...

**Volume**: Amount of data (size)

**Velocity**: Speed with which new data is generated

**Variety**: Heterogeneity in the data

**Variability**: Inconsistency in the data

**Veracity**: How trustworthy the data is

**Value**: Knowledge hidden in big data (needle in a haystack)

**Visualization**: Ability to interpret the data and resulting insights
Paradigms of Science

1st paradigm: Empirical science

$\Delta U = Q - W$
- Change in internal energy
- Heat added to system
- Work done by system

2nd paradigm: Theoretical science

$\rho = mv$
$F = p'$
$F_{ab} = -F_{ba}$

3rd paradigm: Computational science (simulations)

4th paradigm: (Big) data-driven science

Big Data
- Variability
- Velocity
- Volume
- Variety

Intel, BMW, Systexis

1600 1950 2000
Finding Needle in a Haystack?
Data Mining

- The process of discovering information from data
  - Computationally analyzing data for extracting or “mining” knowledge
  - Goal is to discover hidden patterns and actionable insights

- Utilizes methods at the intersection of artificial intelligence, algorithms, machine learning, statistics, and database systems
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

e.g. Statistics, Query, Transformation, Visualization
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?

e.g. Predictive Modeling – NN, SVM, Decision Trees
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.

Bottom up Discovery - We don’t know the question to ask

• 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.

e.g. Relationship Mining, Clustering, Community Detection
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
Commonly Used Data Mining Techniques

- **Predictive**
  - *Classification*: Learning a model to classify new records in different categories (e.g., decision trees, NN, SVM, etc.)
  - *Regression*: Learning a real value function to model the data while minimizing the error
  - *Anomaly Detection*: Identification of outlier records that might lead to interesting discoveries

- **Descriptive**
  - *Clustering*: Discovering groups of records that have similarities
  - *Association Rule Mining*: Discovering relations between different attributes of the dataset
High Performance Computing

• Use of parallel processing to run software programs efficiently, reliably, and quickly
  – Aggregating compute power in a way that delivers much higher performance than a typical desktop computer.
  – Also known as supercomputing

• Utilizes technologies at the intersection of algorithms, computer architecture, networking, storage, and system software
Common HPC Architectures

- **Shared Memory**
  - Global address space makes data sharing across processors very easy; necessitates proper synchronization
  - Lack of scalability between memory and CPUs.
  - Programming models: Pthreads, OpenMP.

- **Distributed Memory**
  - Require a communication network to connect inter-processor memory, leading to non-uniform access.
  - Memory-CPU scalable.
  - Programming model: Message Passing Interface (MPI)

- **Hybrid Distributed-Shared Memory**
  - Commonly used in most of modern computers, e.g. multicore, CPU-GPU
  - Increased scalability as well as programming complexity
  - Programming model: MPI-OpenMP, MPI-CUDA

https://computing.llnl.gov/tutorials/parallel_comp
High Performance Data Mining

High Performance Computing

Data Mining

Big Data Analytics and Knowledge Discovery

{milk, bread} ⇒ {butter}
Outline

• Introduction
  – Big data
  – Data mining
  – High performance computing

• High Performance Data Mining
  – Scalable clustering
  – Scalable association rule mining
  – Scalable community detection

• Interdisciplinary Applications
  – Materials informatics
  – Healthcare informatics
  – Social media analytics
  – Online Tools
Density-based Data Clustering (DBSCAN)
Can identify arbitrary shaped clusters, eliminate noise

Parallelizing DBSCAN
Break its inherent sequential data access order of based on the “disjoint set data structure”

1. Domain decomposition
2. Create singleton trees for each data point in parallel
3. Merge trees based on data and parameters in parallel.
4. Resulting trees (in blue) represent clusters

Agrawal et al., 2013; Patwary et al., SC 2013; Hendrix et al., LDAV 2013; Patwary et al., SC 2012; Han et al., IPDPS ParLearning 2016; Lee et al., BigData ASH 2016; Jin et al., HPC 2015; Jin et al., BigDataService 2015; Jin et al., SC DataCloud 2013; Hendrix et al., HiPC 2012
Silverback Framework

- Probabilistic association rule mining
- Columnar probabilistic storage with minhash pruning
- Use Bloom filters for minimal database I/O
- Intuitive implementation on distributed systems
- Experiments on Facebook and Twitter data with more than 740M users, 32K items (e.g. pages/walls), and 10B user activities (e.g. likes)
- Better runtime performance with negligible accuracy loss, compared to traditional Hadoop-based approach

Xie et al., ICDE 2014; Xie et al., KAIS, 2017
Motivation and Challenges
- Community detection on large graphs is slow
- Structure of computations unknown ahead of time
- Frequent synchronization reduces parallel scope

Methodology
- A distributed memory parallel algorithm PMEP
- Data-based decomposition with duplication

Results
- Speedups of more than 200X on large graphs with millions of nodes and edges

HPDM: Scalable Community Detection
Palsetia et al., BigData 2014; Palsetia et al., ISC 2016
Outline

• Introduction
  – Big data
  – Data mining
  – High performance computing

• High Performance Data Mining
  – Scalable clustering
  – Scalable association rule mining
  – Scalable community detection

• Interdisciplinary Applications
  – Materials informatics
  – Healthcare informatics
  – Social media analytics
  – Online Tools
Materials Genome Initiative (MGI)

"To help businesses discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative.

The invention of silicon circuits and lithium ion batteries made computers and iPods and iPads possible, but it took years to get those technologies from the drawing board to the market place. We can do it faster."

-President Obama (6/11)

Goal: 2X faster and 2X cheaper

Data-driven analytics expected to play a major role

Image credit: https://www.whitehouse.gov/sites/default/files/docs/microsites/mgi/wadia_mgi_talk.pdf
Materials informatics can generate “inverse models” for optimization and design, e.g. Maximize a Property such that Structure follows some constraints.

Materials informatics can generate “forward models” for predictive analytics, e.g. Property = f(Processing, Composition, Structure).

Recent invited article: Agrawal and Choudhary, APL Materials, 4, 053208 (2016), http://dx.doi.org/10.1063/1.4946894

Interdisciplinary Applications: Materials Informatics: Steel Fatigue Strength Prediction

Significance
• Fatigue accounts for >90% of mechanical failures
• High cost and time of fatigue testing

Goal
• Data-driven forward models for fatigue strength of steels

Experimental data
• From NIMS Japan
• 371 carbon and low-alloy steels, 48 carburizing steels, and 18 spring steels

Results
• $R^2 > 0.98$ for cross-validated models
• Online tool deploying forward models

### Table: Fatigue Strength Prediction

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>% Carbon</td>
</tr>
<tr>
<td>Si</td>
<td>% Silicon</td>
</tr>
<tr>
<td>Mn</td>
<td>% Manganese</td>
</tr>
<tr>
<td>P</td>
<td>% Phosphorus</td>
</tr>
<tr>
<td>S</td>
<td>% Sulphur</td>
</tr>
<tr>
<td>Ni</td>
<td>% Nickel</td>
</tr>
<tr>
<td>Cr</td>
<td>% Chromium</td>
</tr>
<tr>
<td>Cu</td>
<td>% Copper</td>
</tr>
<tr>
<td>Mo</td>
<td>% Molybdenum</td>
</tr>
<tr>
<td>NT</td>
<td>Normalizing Temperature</td>
</tr>
<tr>
<td>THT</td>
<td>Through Hardening Temperature</td>
</tr>
<tr>
<td>THt</td>
<td>Through Hardening Time</td>
</tr>
<tr>
<td>THQCr</td>
<td>Cooling Rate for through Hardening</td>
</tr>
<tr>
<td>CT</td>
<td>Carburization Temperature</td>
</tr>
<tr>
<td>Ct</td>
<td>Carburization Time</td>
</tr>
<tr>
<td>DT</td>
<td>Diffusion Temperature</td>
</tr>
<tr>
<td>Dt</td>
<td>Diffusion time</td>
</tr>
<tr>
<td>QmT</td>
<td>Quenching Media Temperature (for Carburization)</td>
</tr>
<tr>
<td>TT</td>
<td>Tempering Temperature</td>
</tr>
<tr>
<td>Tt</td>
<td>Tempering Time</td>
</tr>
<tr>
<td>TCr</td>
<td>Cooling Rate for Tempering</td>
</tr>
<tr>
<td>RedRatio</td>
<td>Reduction Ratio (Ingot to Bar)</td>
</tr>
<tr>
<td>dA</td>
<td>Area Proportion of Inclusions Deformed by Plastic Work</td>
</tr>
<tr>
<td>dB</td>
<td>Area Proportion of Inclusions Occurring in Discontinuous Array</td>
</tr>
<tr>
<td>dC</td>
<td>Area Proportion of Isolated Inclusions</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Rotating Bending Fatigue Strength ($10^7$ Cycles)</td>
</tr>
</tbody>
</table>

Online Tool: [http://info.eecs.northwestern.edu/SteelFatigueStrengthPredictor](http://info.eecs.northwestern.edu/SteelFatigueStrengthPredictor)

Agrawal et al., IMMI 2014; Agrawal and Choudhary, CIKM 2016
Interdisciplinary Applications: Materials Informatics: Stable Compound Discovery

Density Functional Theory
- Very slow simulations
- Require crystal structure as input

Training Data
- Hundreds of thousands of DFT calculations from Open Quantum Materials Database (OQMD)

Forward models
- Can predict a compound’s formation energy with no structure information
- MAE ≈ 0.1 eV/atom, which is comparable to DFT’s own accuracy

Inverse models
- Scan combinatorial set of compositions
- Discovered 4,500 new stable compounds

Meredig and Agrawal et al., PRB 2014; Ward et al., npj Computational Materials 2016; Ward et al., PRB 2017; Liu et al., DL-KDD 2016

Online Tool: http://info.eecs.northwestern.edu/FEpredictor
Interdisciplinary Applications: Materials Informatics: Galfenol Microstructure Optimization

Galfenol
- A magnetoelastic Fe-Ga alloy

Problem
- Discover microstructure with enhanced optimal strength and magnetostriction

Forward models known
- Theoretical models well-established (homogenization)

Inverse models unknown
- Optimization problem
- Challenging due to high dimensionality of microstructure space
- Simple realization of inverse models is prohibitively expensive!
- Non-uniqueness of solutions

Data-driven optimization
- 80% faster, 20% better than traditional methods
- Multiple solutions discovered for the first time

Liu et al., Scientific Reports 2015; Liu et al., IC3 2015
Interdisciplinary Applications: Healthcare Informatics: Patient-Centered Analytics

Lung Cancer Outcome Calculator

Raw SEER data

Preprocessed cancer data

Various data mining optimizations and validations

Online Tool: [http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator](http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator)

Interdisciplinary Applications: Healthcare Informatics: Cost Analysis

Agrawal and Choudhary, KDD-DMH 2013; http://users.eecs.northwestern.edu/~ankitag/hospitalbilling
Interdisciplinary Applications: Social Media Analytics

Cough (orange) and Fever (blue) are two of the most common flu symptoms. Tweet volume mentioning cough and fever reach their peak around Jan 9-15.

Sore throat (green) peak 3-4 days before other flu symptoms like cough and fever, indicating that it is an early sign of flu.

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Count</th>
<th>Word Pair</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>#job</td>
<td>85,493</td>
<td>sick, take</td>
<td>10,504</td>
</tr>
<tr>
<td>#caregiv</td>
<td>63,833</td>
<td>mom, sick</td>
<td>9,842</td>
</tr>
<tr>
<td>#hire</td>
<td>22,247</td>
<td>burden, care</td>
<td>7,322</td>
</tr>
<tr>
<td>#carer</td>
<td>10,086</td>
<td>i’m, sick</td>
<td>4,919</td>
</tr>
<tr>
<td>#dementia</td>
<td>9,808</td>
<td>hard, take</td>
<td>2,097</td>
</tr>
<tr>
<td>#healthcar</td>
<td>8,726</td>
<td>hard, work</td>
<td>1,790</td>
</tr>
<tr>
<td>#alzheim</td>
<td>7,811</td>
<td>hard, mom</td>
<td>1,591</td>
</tr>
<tr>
<td>#nurs</td>
<td>4,655</td>
<td>burden, take</td>
<td>1,376</td>
</tr>
<tr>
<td>#endalz</td>
<td>4,399</td>
<td>sick, want</td>
<td>1,304</td>
</tr>
<tr>
<td>#health</td>
<td>3,761</td>
<td>hard, it’</td>
<td>1,210</td>
</tr>
</tbody>
</table>

Lee et al., KDD 2013; Palsetia et al., SNAM, 2014; Cheng et al., KDD 2013; Cheng et al., ICDM 2014; Lee et al., ASONAM 2015; Al-Bahrani et al., ICDM SENTIRE 2017; Lee et al., ICHI 2017
Interdisciplinary Applications: Online Tools

ThermoEL toolkit

Welcome to Seebeck coefficient predicts component of our toolkit

The calculator predicts Seebeck coefficient of a given compound at 500K, 600K, 700K, and 1000K.

Example: Calculate the Seebeck coefficient of a compound at 500K.

The calculator returns the Seebeck coefficient of the compound.

Thermoelectric Peltier effect (Peltier effect)

Welcome to Weibull modulus of a component of our toolkit

The Weibull modulus calculator predicts the Weibull modulus of a given compound at 500K, 600K, 700K, and 1000K.

Example: Calculate the Weibull modulus of a compound at 500K.

The calculator returns the Weibull modulus of the compound.

Lung Cancer Outcome Calculator

Welcome to Lung cancer mortality rates of patients in our toolkit

The lung cancer mortality rates calculator predicts the mortality rates of patients with lung cancer.

Example: Calculate the lung cancer mortality rates of patients with lung cancer.

The calculator returns the mortality rates of patients with lung cancer.

Steel Fatigue Strength Predictor

Welcome to Fatigue strength of a given steel in our toolkit

The steel fatigue strength predictor predicts the fatigue strength of a given steel.

Example: Calculate the fatigue strength of a given steel.

The calculator returns the fatigue strength of the given steel.

Get Steel Fatigue Strength Predictions

http://info.eecs.northwestern.edu
Outline

• Introduction
  – Big data
  – Data mining
  – High performance computing

• High Performance Data Mining
  – Scalable clustering
  – Scalable association rule mining
  – Scalable community detection

• Interdisciplinary Applications
  – Materials informatics
  – Healthcare informatics
  – Social media analytics
  – Online Tools
Some people think big data analytics is like drinking from a fire hose to quench the thirst.

https://tenor.com/view/familyguy-peter-simpsons-firehose-gif-7915786 Used with permission
Thank you!

Data Mining + High Performance Computing

Materials Informatics

Social Media Analytics

Healthcare Informatics

Ankit Agrawal
Research Associate Professor
Dept. of Electrical Engineering and Computer Science
Northwestern University
ankanag@eecs.northwestern.edu
www.eecs.northwestern.edu/~ankanag/