Machine Learning

Prabhat
HPC User Forum
April 12, 2016
Imagenet ILSVRC Challenge

Error rate

Deep learning techniques

human performance

Slide Courtesy of Nervana Systems
(2012) This is all great, but...

- Is Machine Learning relevant to science?

- Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?
(2012) This is all great, but…

• Is Machine Learning relevant to science?
  – Success stories are for images and audio, but how about scientific data?

• Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?
  – Our applied mathematicians are content with formulating and solving PDEs
  – The NNSA folks care about Uncertainty Quantification
  – Our data ‘analytics’ folks are happy dealing with meshes, computational geometry, topology
(2016) The writing is on the wall

- O(B) $ worth of investment by industry
- Machine Learning and Statistics are established as key disciplines for this decade
  - Deep Learning has taken off as the most promising ML technique
(2012) Revisited..

• Is Machine Learning relevant to science?

• Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?
(2010-2016): The Rise of Data-Intensive Science

- Astronomy
- Genomics
- Climate
- Physics
- Light Sources
## 4 V’s of Scientific Big Data

<table>
<thead>
<tr>
<th>Science Domain</th>
<th>Variety</th>
<th>Volume</th>
<th>Velocity</th>
<th>Veracity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy</td>
<td>Multiple Telescopes, multi-band/spectra</td>
<td>O(100) TB</td>
<td>100 GB/night – 10 TB/night</td>
<td>Noisy, acquisition artefacts</td>
</tr>
<tr>
<td>Light Sources</td>
<td>Multiple imaging modalities</td>
<td>O(100) GB</td>
<td>1 Gb/s - 1 Tb/s</td>
<td>Noisy, sample preparation/acquisition artefacts</td>
</tr>
<tr>
<td>Genomics</td>
<td>Sequencers, Mass-spec, proteomics</td>
<td>O(1-10) TB</td>
<td>TB/week</td>
<td>Missing data, errors</td>
</tr>
<tr>
<td>High Energy Physics</td>
<td>Multiple detectors</td>
<td>O(100) TB –  O(10) PB</td>
<td>1-10 PB/s reduced to GB/s</td>
<td>Noisy, artefacts, spatio-temporal</td>
</tr>
<tr>
<td>Climate</td>
<td>Simulations Multi-variate, spatio-temporal</td>
<td>O(10) TB</td>
<td>100 GB/s</td>
<td>‘Clean’, need to account for multiple sources of uncertainty</td>
</tr>
</tbody>
</table>
Does Machine Learning matter?

• **Is Machine Learning relevant to science?**
  – Yes!

• **Why should HPC facilities care about Machine Learning, Deep Learning, Statistics?**
  – Analytics is *the* key step for gaining scientific insights
  – The nature of questions in data-intensive science are inferential
  – Statistics and Machine Learning deal with inference in presence of noise and errors
Creating a catalog of all objects in the Universe
Characterizing Extreme Weather in a Changing Climate
Knowledge Extraction from Scientific Literature
Top 10 Data Analytics Problems

5 Understanding Speech Production
Top 10 Data Analytics Problems

6 Quantitative and Predictive Biology
7. Understanding the Genetic Code
Top 10 Data Analytics Problems

8

Personalized Toxicology
Top 10 Data Analytics Problems

9 Designer Materials

The New Alchemists
How supercomputers are transforming innovation in materials design
10 Fundamental Constituents of Matter
Towards Synthesis (and maybe Convergence)

• What is the landscape of Machine Learning problems in science?
  – Bewildering array of taxonomy and domain-specific terminology

• What are the key computational motifs?
  – Need to have a productive conversation with HPC software, hardware vendors
<table>
<thead>
<tr>
<th></th>
<th>Astronomy</th>
<th>Cosmology</th>
<th>Climate</th>
<th>Systems Biology</th>
<th>Neuroscience</th>
<th>EM/X-Ray Imaging</th>
<th>Mass-spect Imaging</th>
<th>Personalized Toxicology</th>
<th>Materials</th>
<th>Particle Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Dimensionality Reduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Inference</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Model Estimation</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Design of Experiments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Semantic Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Feature Learning</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Astronomy</td>
<td>Cosmology</td>
<td>Climate</td>
<td>Systems Biology</td>
<td>Neuroscience</td>
<td>EM/X-Ray Imaging</td>
<td>Mass-spec Imaging</td>
<td>Personalized Toxicology</td>
<td>Materials</td>
<td>Particle Physics</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------------</td>
<td>--------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------------</td>
<td>-----------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Classification</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimensionality</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design of Experiments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature Learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Machine Learning Research Strategy

## Science Applications
- Astronomy, Cosmology, Climate, BRAIN, BioImaging, HEP

## Scientific Analysis
- **Pattern/Anomaly Discovery**
- **Large Scale Inference**
- **Clustering, Dimensionality Reduction**
- **Data Fusion**
- **Genome Assembly**

## Scalable Algorithms
- **Deep Learning**
- **Sparse Coding**
- **Stochastic Variational Inference**
- **DBSCAN CUR/CX**
- **Distributed MCMC**
- **Direct Graph Kernel Computation**

## Big Data Motif
- **Dense/Sparse Linear Algebra**
- **MapReduce**
- **Optimization (Stochastic)**
- **Randomized Linear Algebra**
- **Graph Methods (BFS, DFS,...)**

## Optimized Libraries
- **ScaLAPACK, BLAS, PCL-DNN**
- **TECA**
- **TensorFlow, SpearMint**
- **RandLA**
- **GraphLab**

## Hardware
- Many-Core Chipset, Deep Memory Hierarchy, Reducing Data Movement, Power Efficiency
Machine Learning Research Strategy

Science Applications

Scientific Analysis

Scalable Algorithms

Big Data Motif

Optimized Libraries

Hardware

MANTISSA

Co-Design

Science Applications
- Astronomy, Cosmology, Climate, BRAIN, Bioimaging, HEP
- Pattern/Anomaly Discovery
- Large Scale Inference
- Dimensionality Reduction
- Data Fusion
- Genome Assembly

Scientific Analysis
- Deep Learning
- Sparse Coding
- Stochastic Variational Inference
- DBSCAN
- CUR/CX
- Distributed MCMC
- Direct Graph Kernel Computation

Scalable Algorithms
- Dense/Sparse Linear Algebra
- Optimization (Stochastic)
- Randomized Linear Algebra
- Graph Methods (BFS, DFS, ...)

Optimized Libraries
- ScaLAPACK, BLAS
- PCL-DNN
- TECAT
- RandLA
- GraphLab

Hardware
- Many-Core Chipset, Deep Memory Hierarchy, Reducing Data Movement, Power Efficiency

Machine Learning Research Strategy

Machine Learning Research Strategy

Machine Learning Research Strategy

Machine Learning Research Strategy

Machine Learning Research Strategy
Machine Learning: Challenges

• Cultural
  – ML doesn’t cleanly ‘fit’ within Computer Science or Applied Math
  – Statistics, CS (Machine Learning, HPC) taxonomy
  – Mindshare
    • Attracting the best academic and industry talent is hard

• Technical
  – Big Data ecosystem has evolved independently of HPC
  – Aspirations of Convergence (Software, Hardware)
    • HPC institutions need to do a better job of characterizing their Data Analytics requirements
Machine Learning: Opportunities

• **HPC community is uniquely positioned**
  – Storage and Compute Hardware
  – Meaningful scientific problems

• **Software (Research and Production) is wide open**

• **Most exciting discoveries happen at the intersection of domain sciences and methods**
  – We don’t know the limits of Deep Learning methods
Thanks!

Contact: prabhat@lbl.gov