An Overview of the Data Analytics for Medicine using Semi-Supervised Learning (DAMSEL) Program*

Barbara Beckerman, MBA,(PI)[1]; Robert Patton, PhD[1]; Christopher Symons, MS[1]; April McMillan, PhD[1]; Shaun Gleason, PhD[2]; Ryan Kerekes, PhD[2]; Vincent Paquit, PhD[2]; Robert Nishikawa, PhD [3]

[1]Computational Sciences and Engineering Division, ORNL; [2]Measurement Science and Systems Engineering Division, ORNL; [3] Department of Radiology, University of Chicago

ABSTRACT

Presently, knowledge discovery and cohesive decision-making capabilities for biomedical applications are hampered by significant gaps in technology for multi-modal data analytics. We are addressing this issue by developing a novel learning framework, called Data Analytics for Medicine using SEmi-supervised Learning (DAMSEL), that can intelligently combine important data-rich resources and technologies, which in turn will leapfrog current analytical capabilities in a more comprehensive, flexible, and responsive computational environment. DAMSEL is being developed using two biomedical applications: breast cancer and traumatic brain injury.

OBJECTIVE: Develop a comprehensive, flexible, and responsive computing framework that incorporates ability to analyze multi-modality data using semi-automated learning environments

Objectives/Tasks

1. Develop analytical, semi-automated learning framework and tools for processing multi-modality data
2. Address performance, portability, and scalability of framework by leveraging computing resources and restructuring computationally-intensive algorithms.
3. Validate system performance in terms of knowledge accuracy and system responsiveness using two disparate medical applications

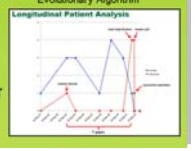
METHODS

- Text Analysis
- Image Analysis
- Semi-Supervised learning Framework
- Computational Scalability
- Integration/Analytics

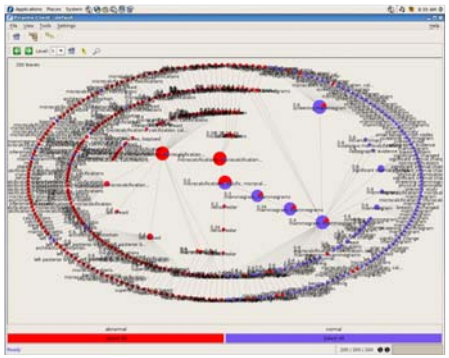
Text Analysis Approach

- Leverage *human expertise* to characterize the data
- Enhance *statistical analysis* with *evolutionary algorithms* to learn domain-specific, significant textual features

- Provide *textual feature space* in Learning Environment
- Characterize the *"language"* of the domain
- Perform *longitudinal analysis* for temporal spikes or trends



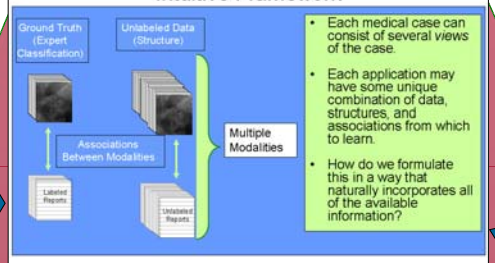
Clustering Text Data to Visualize Relationships



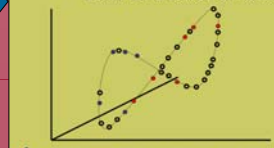
Acknowledgement:
*Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy.

MANIFOLD LEARNING

Goal: Leverage All Information in a Intuitive Framework



Data on a lower dimensional manifold

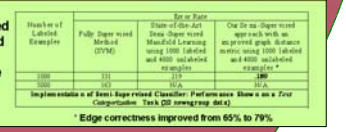


- A graph with a suitable local distance metric can be used to discover a valuable subspace for classification.
- Construction of such a graph only requires unlabeled points.

Labeled cases (where outcomes are known) are the most valuable

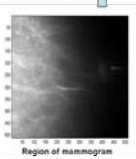
Unlabeled cases (where outcomes are unknown) can reveal structure

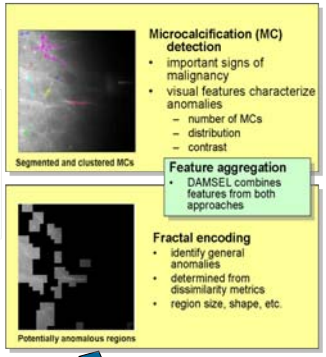
- However, the labeled points can be utilized to ensure that the graph represents the problem of interest.



* Edge correctness improved from 65% to 79%

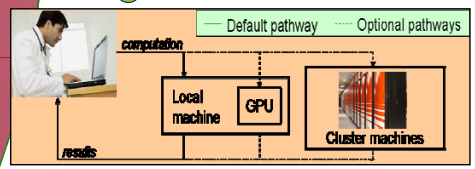
Image Analysis for Mammography Data



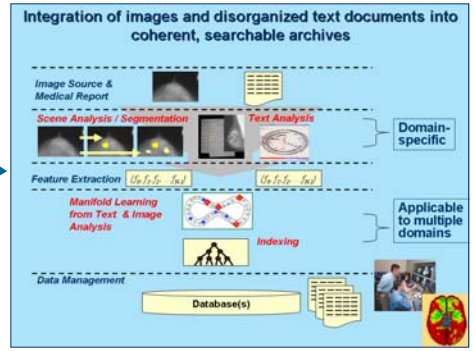


SCALABILITY

- Parallelize code
- Restructure algorithms
- Integrate Seamlessly



Anticipated Results



For additional information please contact:

Barbara Beckerman
Acting Director, Biomedical Science and Engineering Center
Computational Sciences and Engineering Division - Oak Ridge National Laboratory
Email: beckermanbg@ornl.gov; ph(865) 576-2681