Squeezing Information from Data at Exascale

Joel Saltz
Emory University
Georgia Tech
Squeezing Information from Temporal Spatial Datasets

- Leverage exascale data and computer resources to squeeze the most out of image, sensor or simulation data
- Run lots of **different** algorithms to derive **same features**
- Run lots of algorithms to derive **complementary features**
- Data models and data management infrastructure to manage data products, feature sets and results from classification and machine learning algorithms
- Much can be done at “data staging time”
Overview

• Integrative biomedical informatics analysis – feature sets obtained from Pathology and Radiology studies

• This is the same CS problem as what we have seen in Oil Reservoir/Seismic analyses, astrophysics and in Computational Fluid Dynamics

• Techniques, tools and methodologies for derivation, management and analysis of feature sets

• Ideas for how to move to exascale
## Examples

<table>
<thead>
<tr>
<th><strong>Astrophysics</strong></th>
<th><em>Which portions of a star’s core are susceptible to implosion over time period</em> $[t_1, t_2]$?</th>
<th>Compute streamlines on vector field $v$ within grid points $[(x_1, y_1) - (x_2, y_2)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Material Science</strong></td>
<td><em>Is crystalline growth likely to occur within range</em> $[p_1, p_2]$ of pressure conditions?</td>
<td>Compute likelihood of local cyclic relationships among nanoparticles within a frame</td>
</tr>
<tr>
<td><strong>Cancer studies</strong></td>
<td><em>Which regions of the tumor are undergoing active angiogenesis in response to hypoxia?</em></td>
<td>Determine image regions where (blood vessel density &gt; 20) and (nuclei and necrotic region are within 50 microns of each other)</td>
</tr>
</tbody>
</table>
Typical data analysis scenario

Transformation of raw image data

- Normalization: illumination.
- Spatial Alignment: displacements
- Stitching: seamless image mosaic
- Warping: standard template / canonical atlas
- ...

Analysis

- Pixel-based computing
- Color decomposition
- Correcting for non uniform staining

- Segmentation
- Feature extraction, classification

- Annotation of data
- Semantic querying
- Image mining

Data volume decreases; Data complexity & domain specificity increase
INTEGRATIVE BIOMEDICAL INFORMATICS ANALYSIS

Reproducible anatomic/functional characterization at gross level (Radiology) and fine level (Pathology)
Integration of anatomic/functional characterization with multiple types of “omic” information
Create categories of jointly classified data to describe pathophysiology, predict prognosis, response to treatment

*In Silico Center – Application Driven Computer Science (with National Cancer Institute flavor)*
Specific Aims:

1. Influence of necrosis/hypoxia on gene expression and genetic classification.

2. Molecular correlates of high resolution nuclear morphometry.

3. Gene expression profiles that predict glioma progression.

Integration of heterogeneous multiscale information

- Coordinated initiatives Pathology, Radiology, “omics”
- Exploit synergies between all initiatives to improve ability to forecast survival & response.
Oligodendroglioma

Astrocytoma

Nuclear Qualities
Vessel Characterization

- Bifurcation detection
Progression to GBM

Anaplastic Astrocytoma (WHO grade III)

Glioblastoma (WHO grade IV)
Astrocytoma vs Oligodendroglioma
Overlap in genetics, gene expression, histology

- Assess nuclear size (area and perimeter), shape (eccentricity, circularity major axis, minor axis, Fourier shape descriptor and extent ratio), intensity (average, maximum, minimum, standard error) and texture (entropy, energy, skewness and kurtosis).
Whole slide scans from 14 TCGA GBMS (69 slides)
7 purely astrocytic in morphology; 7 with 2+ oligo component
399,233 nuclei analyzed for astro/oligo features
Cases were categorized based on ratio of oligo/astro cells

TCGA Gene Expression Query: c-Met overexpression
# Classification Performance

**SFFS + 10% Filtering + 100 runs**

<table>
<thead>
<tr>
<th>Neoplastic Astrocyte</th>
<th>Neoplastic Oligodendrocyte</th>
<th>Reactive Endothelial</th>
<th>Reactive Astrocyte</th>
<th>Junk</th>
</tr>
</thead>
<tbody>
<tr>
<td>91.89%</td>
<td>1.82%</td>
<td>2.88%</td>
<td>2.25%</td>
<td>1.16%</td>
</tr>
<tr>
<td>1.53%</td>
<td>95.60%</td>
<td>1.10%</td>
<td>0.14%</td>
<td>1.62%</td>
</tr>
<tr>
<td>4.87%</td>
<td>0.53%</td>
<td>88.96%</td>
<td>2.18%</td>
<td>3.47%</td>
</tr>
<tr>
<td>5.37%</td>
<td>1.54%</td>
<td>6.21%</td>
<td>85.62%</td>
<td>1.27%</td>
</tr>
<tr>
<td>2.86%</td>
<td>1.34%</td>
<td>5.24%</td>
<td>0.64%</td>
<td>89.93%</td>
</tr>
</tbody>
</table>
Which features carry most prognostic significance?
Which features correlate with genetic alterations?
Pipeline for Whole Slide Feature Characterization

- $10^{10}$ pixels for each whole slide image
- 10 whole slide images per patient
- $10^{8}$ image features per whole slide image
- 10,000 brain tumor patients
- $10^{15}$ pixels
- $10^{13}$ features
- Hundreds of algorithms
- Annotations and markups from dozens of humans
Feature Management and Query Framework
Data Models to Represent Feature Sets and Experimental Metadata

PAIS |pās|: Pathology Analytical Imaging Standards

• Provide semantically enabled data model to support pathology analytical imaging
• Data objects, comprehensive data types, and flexible relationships
• Reuse existing standards
• Data models (in general) likely route to integrating staging, immediate on line analyses and full scale analyses
• Semantic models/annotations
• Semantic directed runtime compilation that embedded various partitioners (work with Kennedy, Fox)
Compute Intersection Ratio and Distance Between Markups from Two Segmentation Algorithms

```sql
INSERT INTO PAIS.VALIDATION_PRECOMPUTE(pais_uid, tilename, markup_id, AREA_OVERLAP_RATIO, centroid_distance)
SELECT A.pais_uid, A.tilename, A.markup_id,
    CAST(db2gse.ST_Area(db2gse.ST_Intersection(a.polygon, b.polygon))/db2gse.ST_Area(db2gse.ST_Union(a.polygon, b.polygon)) AS DECIMAL(4,2)) AS area_ratio,
    CAST(db2gse.ST_Distance(db2gse.ST_Centroid(b.polygon), db2gse.ST_Centroid(a.polygon)) AS DECIMAL(5,2)) AS centroid_distance
FROM pais.markup_polygon A, pais.markup_polygon B
WHERE A.pais_uid = 'oligoIII.2_20x_20x_NS-MORPH_1' AND
    A.tilename = 'oligoIII.2.ndpi-0000090112-0000024576' AND
    B.pais_uid = 'oligoIII.2_20x_20x_NS-MORPH_2' AND
    B.tilename = 'oligoIII.2.ndpi-0000090112-0000024576' AND
    db2gse.ST_Overlaps(A.polygon, B.polygon) = 1;
```
Example TCGA Query: Mean Feature Vector and Feature Covariance

- Mean feature vector for each slide and tumor subtype

```
SELECT AVG(area), AVG(sum_canny_pixel), AVG(mean_canny_pixel)
FROM pais.ccalculation_flat c, tctga.patient_characteristic pc, pais.patient p
WHERE p.patientid = pc.patient_id AND p.pais_uid = c.pais_uid
GROUP BY c.pais_uid, pc.subtype;
```

- Covariance between features

```
SELECT COVARIANCE(PERIMETER, AREA) AS COV_PERIMETER_AREA,
     COVARIANCE(PERIMETER, ECCENTRICITY) AS COV_PERIMETER_ECCENTRICITY
FROM pais.ccalculation_flat
WHERE PAIS_UID = 'TCGA-06-0152-01Z-00-DX7_20x_20x_NS-MORPH_1';
```
Analysis framework architecture

**Workflow design**

- Application workflow
- Metadata

**Description module**
- Ontology representations of datasets
- Application structure
- Application behavior
- System components

**Execution module**
- Runtime support for multidimensional data
  - Data management, I/O abstraction
- Workflow engines, filter streaming middleware, batch schedulers

**Trade-off module**
- Time constraints, accuracy requirements (application-level QoS)
- Map high-level queries to low-level execution plans
Execution Module: Runtime support for multidimensional data

- Customize for specific domains
  - Out-of-core Virtual Microscope

- Out-of-core data?
  - Data stored as a collection of chunks
  - Chunk: unit of data management (disk I/O, indexing and compression)

- Data model
  - Data **spatially partitioned** into chunks
  - Chunks distributed across nodes in a shared-nothing environment

- Semi-streaming programming model
  - Leverages lightweight filter-streaming, buffer management by streaming middleware (e.g., DataCutter, IBM System $S$)
Mediators: I/O abstraction layer

- **Compute Nodes**
  - Cluster 1 (m nodes)
  - M
  - A: Application filters
  - M: Mediator filters

- **Active Storage Nodes**
  - M
  - M

- **Archival Nodes**
  - Cluster 2 (p nodes)
  - M
  - Cluster 3 (n nodes)

September 8, 2010 Oak Ridge National Laboratory
In Transit Processing using DataCutter Spatial Crossmatch

- Mapping to atlas and 3-D reconstruction frequently rely on spatial crossmatch
- We have studied spatial crossmatch with LLNL initially in an astronomy context
- Large Synoptic Survey Telescope (LSST) -- 3.2 Gigapixel camera that captures field of view every 15 seconds
- Catalog roughly 50 billion objects in 10 years
- Netezza (active disk) implementation vs two DataCutter based distributed mySQL implementations
- Benchmarked on Netezza and small (16 node) cluster
Semantic Workflows (Wings)
Collaborative Work with Yolanda Gil, Mary Hall

- A systematic strategy for composing application components into workflows
- Search for the most appropriate implementation of both components and workflows
- Component optimization
  - Select among implementation variants of the same computation
  - Derive integer values of optimization parameters
  - Only search promising code variants and a restricted parameter space
- Workflow optimization
  - Knowledge-rich representation of workflow properties
Adaptivity
Heuristics determine more favorable chunks at an earlier point of time

- Tune ‘order of execution’ of chunks and ‘data resolution’ parameter per chunk

Query: “Maximize average classification confidence within time $t$”
Multiple Granularity Workflows
Map Images into Atlas, Measure Gene Expression

Fuse components into metacomponents
Tasks associated with metacomponent managed by execution module Pegasus, DataCutter, Condor used to support multiple grained workflow
Performance Impact of Combined Coarse and Fine Grained Workflows
Data Science Research Challenges Driven by In Silico Discovery Research

- Data integration that targets multiple data sources with conflicting metadata and conflicting data
- Efficient methods for semantic query that targets questions involving complex multi-scale features associated with petascale and exascale ensembles of highly annotated images
- Computer assisted annotation and markup for very large datasets
- Systems to support combinations of structured and irregular accesses to exascale datasets
Data Science Research Challenges

• Structural and semantic metadata management: how to manage tradeoff between flexibility and curation
• Data and semantic modeling infrastructures and policies able to scale to handle distributed systems with an aggregate of 10^9 or more data models/concepts
• Three dimensional (time dependent) reconstruction, feature detection and annotation of 3-D microscopy imagery
• Workflow infrastructure for large scale data intensive computations
Final Data Science Challenge: Large Dataset Size

- Basic small mouse is 10 cm³
- 1 µ resolution – very roughly $10^{13}$ bytes/mouse
- Molecular data (spatial location) multiply by $10^2$
- Vary genetic composition, environmental manipulation, systematic mechanisms for varying genetic expression; multiply by $10^3$

Total: $10^{18}$ bytes per big science animal experiment
Thanks to:

- Tahsin Kurc, Vijay Kumar
- In silico center team: Dan Brat (Science PI), Tahsin Kurc, Ashish Sharma, Tony Pan, David Gutman, Jun Kong, Sharath Cholleti, Carlos Moreno, Chad Holder, Erwin Van Meir, Daniel Rubin, Tom Mikkelsen, Adam Flanders, Joel Saltz (Director)
- caGrid Knowledge Center: Joel Saltz, Mike Caliguiri, Steve Langella co-Directors; Tahsin Kurc, Himanshu Rathod Emory leads
- caBIG In vivo imaging team: Eliot Siegel, Paul Mulhern, Adam Flanders, David Channon, Daniel Rubin, Fred Prior, Larry Tarbox and many others
- In vivo imaging Emory team: Tony Pan, Ashish Sharma, Joel Saltz
- Emory ATC Supplement team: Tim Fox, Ashish Sharma, Tony Pan, Edi Schreibmann, Paul Pantalone
- Digital Pathology R01: Foran and Saltz; Jun Kong, Sharath Cholleti, Fusheng Wang, Tony Pan, Tahsin Kurc, Ashish Sharma, David Gutman (Emory), Wenjin Chen, Vicky Chu, Jun Hu, Lin Yang, David J. Foran (Rutgers)
Thanks!