Spatio-temporal Sensor Integration, Analysis, Classification or

Can Exascale Cure Cancer?
EXASCALE CHALLENGES IN INTEGRATIVE MULTI-SCALE SPATIO-TEMPORAL ANALYSIS
“Domain”: Spatio-temporal Sensor Integration, Analysis, Classification
Big Data Extreme Computing 2014

• Multi-scale material/tissue structural, molecular, functional characterization. Design of materials with specific structural, energy storage properties, brain, regenerative medicine, cancer
• Integrative multi-scale analyses of the earth, oceans, atmosphere, cities, vegetation etc – cameras and sensors on satellites, aircraft, drones, land vehicles, stationary cameras
• Digital astronomy
• Hydrocarbon exploration, exploitation, pollution remediation
• Aerospace – wind tunnels, acquisition of data during flight
• Solid printing integrative data analyses
• Autonomous vehicles, e.g. self driving cars
• Data generated by numerical simulation codes – PDEs, particle methods
Typical Computational/Analysis Tasks
Spatio-temporal Sensor Integration, Analysis, Classification

- Data Cleaning and Low Level Transformations
- Data Subsetting, Filtering, Subsampling
- Spatio-temporal Mapping and Registration
- Object Segmentation
- Feature Extraction
- Object/Region/Feature Classification
- Spatio-temporal Aggregation
- Diffeomorphism type mapping methods (e.g. optimal mass transport)
- Particle filtering/prediction
- Change Detection, Comparison, and Quantification
Data-Centric View

Extreme Computing

Instruments
Sensors

Science Community

Analysis
Integrative Analysis: OSU BISTI NBIB Center
Big Data (2005)

Associate genotype with phenotype
Big science experiments on cancer, heart disease, pathogen host response
  - Tissue specimen -- 1 cm\(^3\)
  - 0.1 μ resolution – roughly $10^{15}$ bytes
Molecular data (spatial location) can add additional significant factor; e.g. $10^2$
  - Multispectral imaging, laser captured microdissection, Imaging Mass Spec, Multiplex QD
Multiple tissue specimens; another factor of $10^3$

Total: $10^{20}$ bytes -- 100 exabytes per big science experiment
The Tyranny of Scale
(Oil Reservoir Management
Tinsley Oden - U Texas)
Why Applications Get Big

• Physical world or simulation results
• Detailed description of two, three (or more) dimensional space
• High resolution in each dimension, lots of timesteps
  • e.g. oil reservoir code -- simulate 100 km by 100 km region to 1 km depth at resolution of 100 cm:
    – 10^6*10^6*10^4 mesh points, 10^2 bytes per mesh point, 10^6 timesteps --- 10^24 bytes (Yottabyte) of data!!!
Center for Multi Scale Cancer Informatics (Sept 2014)

• Stony Brook
• Oak Ridge National Labs
• Emory
• Yale

• Cancer Research meets HPC, Material Science, "omics"
• Vector Valued "omics"
Reconstruction of Cellular Biological Structures from Optical Microscopy Data

Kishore Mosaliganti, Student Member, IEEE, Lee Cooper, Richard Sharp, Member, IEEE, Raghu Machiraju, Member, IEEE, Gustavo Leone, Kun Huang, Member, IEEE, and Joel Saltz, Senior Member, IEEE
Integrative Cancer Research with Digital Pathology

High-resolution whole-slide microscopy

Integrated Analysis

histology  Multiplex IHC  neuroimaging

molecular  clinical\pathology

Integrated Analysis
Direct Study of Relationship Between Image Features vs Clinical Outcome, Response to Treatment, Molecular Information

Lee Cooper, Carlos Moreno
Clustering identifies three morphological groups

- Analyzed 200 million nuclei from 162 TCGA GBMs (462 slides)
- Named for functions of associated genes: Cell Cycle (CC), Chromatin Modification (CM), Protein Biosynthesis (PB)
- Prognostically-significant (logrank $p=4.5\times10^{-4}$)

Figure 2  Glioblastoma (GBM) clusters, survival, and relationship to molecular subtypes. (A) Means-based analysis of GBM morphology reveals three patient clusters. (B) Survival differences between these clusters are statistically significant. CC, cell cycle; CM, chromatin modification; PB, protein biosynthesis.
MR Imaging Predictors of Molecular Profile and Survival: Multi-institutional Study of the TCGA Glioblastoma Data Set

David A. Gutman, MD, PhD, Lee A. D. Cooper, PhD, Scott N. Hwang, MD, PhD, Chad A. Holder, MD, JingJing Gao, PhD, Tarun D. Aurora, BS, William D. Dunn, Jr, BS, Lisa Scarpace, MS, Tom Mikkelsen, MD, Rajan Jain, MD, Max Wintemark, MD, MAS, Manal Jilwan, MD, Prashant Raghavan, MD, Erich Huang, PhD, Robert J. Clifford, PhD, Pattanasak Mongkolwat, PhD, Vladimir Kleper, BS, John Freymann, BA, Justin Kirby, BS, Pascal O. Zinn, MD, Carlos Moreno, PhD, Carl Jaffe, MD, Rivka Colen, MD, Daniel L. Rubin, MD, MS, Joel Saltz, MD, PhD, Adam Flanders, MD and Daniel J. Brat, MD, PhD
Complex image analysis, feature extraction, machine learning pipelines
Spatio-temporal Sensor Integration, Analysis, Classification
Programming Tools

- Multi level computational pipeline management
- Region Templates – abstraction for multi-scale spatio-temporal computations
- “Domain” specific language
Database Tools (Fusheng Wang)
Spatial Queries and Analytics

• Feature based descriptive queries
  – Feature based filtering or feature aggregation
  – Spatial relationship based queries
  – Spatial join (two- or multi-), window, point-in-polygon
  – Polygon overlay or spatial cross-matching
  – Distance based queries
  – Nearest neighbors

• Spatial analytics
  – Density based spatial patterns: find clusters, hotspots, and anomalies
  – Spatial relationship modeling, e.g., geographically weighted regression model (GWR)
Vector Valued “omics”
Data Scale
Whole Slide Imaging: Scale

Data per slide: 500MB to 100GB
Roughly 250-500M Slides/Year in USA
Total: 0.1-10 Exabytes/year
Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore’s law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.

Cost of genome sequencing.

Moore’s law for computing costs.

Next generation sequencers enter the market.

The price of sequencing a whole human genome hovers around $5,000 and is expected to drop even lower.

Nature News and Comment -
$1000 Genome
Genomics: The single life

Sequencing DNA from individual cells is changing the way that researchers think of humans as a whole.

ONE GENOME FROM MANY
Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

- **Standard genome sequencing**
  - A sample containing thousands to millions of cells is isolated.
  - DNA is extracted from all the nuclei.
  - DNA is broken into fragments and then sequenced.
  - The sequences are assembled to give a common, ‘consensus’ sequence.

- **Single-cell sequencing**
  - A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.
  - The DNA is extracted and amplified, during which errors can creep in.
  - Amplified DNA is sequenced.
  - Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

Brian Owens
Imprinted genes are under greater selective pressure than normal genes. This is because only one copy is active at a time. Any variations in that copy will be expressed. There is no "back-up copy" to mask its effects. As a result, imprinted genes evolve more rapidly than other genes. And imprinting patterns -- which genes are silenced in the eggs and sperm -- also evolve quickly. They can be quite different in closely related species.

Lions and tigers don't normally meet in nature. But they can get along very well in captivity, where they sometimes produce hybrid offspring. The offspring look different, depending on who the mother is. A male lion and a female tiger produce a liger - the biggest of the big cats. A male tiger and a female lion produce a tigon, a cat that is about the same size as its parents.

The difference in size and appearance between ligers and tigons is due in part to the parents' differently imprinted genes. Other animals can also hybridize, with similar results. For example, a horse and a donkey can produce a mule or a hinny.
acetyl groups
OPEN!
for gene expression & synaptogenesis
methyl groups
CLOSED!

University of Utah Epigenetics Training Site
The epigenetic product portfolio represents a defined and well positioned series of drug candidates and discovery opportunities. Given the interest in the epigenetics space, our package should attract a number of spin out options.

Progen Pharmaceuticals Ltd
Human Phenotyping and Population Health

- Patient History
- Physical Exams
- Chemistries
- Hematology
- CT Scans
- MRI Scans
- PET Scans
- Path specimens
- Genomics
- Proteomics
- Metabolomics
- Lipidomics

Integrative Predictive Analytics
Clinical Phenotype Characterization and the Emory Analytic Information Warehouse

• Example Project: Find hot spots in readmissions within 30 days
  – What fraction of patients with a given principal diagnosis will be readmitted within 30 days?
  – What fraction of patients with a given set of diseases will be readmitted within 30 days?
  – How does severity and time course of co-morbidities affect readmissions?
  – Geographic analyses

• Compare and contrast with UHC Clinical Data Base
  – Repeat analyses across all UHC hospitals
  – Are we performing the same?
  – How are UHC-curated groupings of patients (e.g., product lines) useful?

Andrew Post, Sharath Cholleti, Doris Gao, Michel Monsour, Himanshu Rathod
# 30-Day Readmission Rates for Derived Variables

**Emory Health Care**

<table>
<thead>
<tr>
<th>Patient Population</th>
<th>Number of Encounters</th>
<th>Number of Readmissions</th>
<th>Readmission Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Emory</td>
<td>202181</td>
<td>36734</td>
<td>15%</td>
</tr>
<tr>
<td>Multiple MI</td>
<td>4414</td>
<td>1506</td>
<td>36% (Single MI 15%)</td>
</tr>
<tr>
<td>ESRD</td>
<td>18445</td>
<td>5036</td>
<td>27% (CKD 23%)</td>
</tr>
<tr>
<td>&gt;=4 readmissions</td>
<td>19510</td>
<td>10707</td>
<td>55%</td>
</tr>
<tr>
<td>Multiple MI and &gt;=4 readmissions</td>
<td>997</td>
<td>520</td>
<td>52%</td>
</tr>
<tr>
<td>CKD and &gt;=4 readmissions</td>
<td>7865</td>
<td>4110</td>
<td>52%</td>
</tr>
<tr>
<td>Uncontrolled diabetes</td>
<td>12219</td>
<td>2573</td>
<td>21% (Diabetes 19%)</td>
</tr>
<tr>
<td>Uncontrolled diabetes &amp; pressure ulcer</td>
<td>648</td>
<td>201</td>
<td>31%</td>
</tr>
<tr>
<td>Uncontrolled diabetes &amp; ESRD</td>
<td>1645</td>
<td>531</td>
<td>32%</td>
</tr>
<tr>
<td>Sickle cell crisis</td>
<td>1809</td>
<td>663</td>
<td>37% (Sickle cell anemia 34%)</td>
</tr>
<tr>
<td>MRSA</td>
<td>1565</td>
<td>410</td>
<td>26%</td>
</tr>
<tr>
<td>Stroke and MRSA</td>
<td>42</td>
<td>16</td>
<td>38% (Stroke 24%)</td>
</tr>
<tr>
<td>MI and MRSA</td>
<td>140</td>
<td>43</td>
<td>31% (MI 15%)</td>
</tr>
</tbody>
</table>
Geographic Analyses
UHC Medicine General Product Line (#15)

\[
\text{Readmission } O: E = \frac{\# \text{ of 30-day readmits in the census tract}}{\# \text{ of 30-day readmits overall}} \times \frac{\# \text{ of encounters in the census tract}}{\# \text{ of encounters overall}}
\]

Income Levels:
- $< 25000$
- $25000 - 50000$
- $50000 - 75000$
- $75000 - 100000$
- $> 100000$
Predictive Modeling for Readmission

- Random forests (ensemble of decision trees)
  - Create a decision tree using a random subset of the variables in the dataset
  - Generate a large number of such trees
  - All trees vote to classify each test example in a training dataset
  - Generate a patient-specific readmission risk for each encounter

- Rank the encounters by risk for a subsequent 30-day readmission

Sharath Cholleti
Predictive Modeling for **180 UHC Hospitals, 35 Million Patients**

*Identify High Risk Patients!*

*Readmission fraction of top 10% high risk patients*
DSRIP

Delivery System Reform Incentive Payment (DSRIP) Program
What is DSRIP?

• **8 billion dollar grant from CMS to NY State**
  
  – 25% reduction over five years in avoidable hospitalizations and ER visits in the Medicaid and uninsured population
  
  – Collaborative effort to implement innovative projects focused on
    
    • System transformation
    
    • Clinical improvement
    
    • Population health improvement
5 YEAR GOALS

• Create integrated care delivery system anchored by safety net providers

• Engage partners across the care delivery spectrum to create a county wide network of care

• After five years transition this network to an ACO which will contract with insurance providers on an at risk basis
The projects

- The chosen projects must address the most significant healthcare issues in the Suffolk County Medicaid and uninsured population and address healthcare disparities—some examples
  - Behavioral health: BH and primary care integration
  - Adults: COPD, diabetes, HTN, renal failure
  - Children: Asthma
  - Hi risk OB/neonates—Esp. Hispanic and African American communities
DATA ANALYTICS

• County wide healthcare data will be collected
• Near real-time data analytics will be used to drive healthcare improvements
  – Analyzing success and failures to create fast turn-around improvement opportunities
  – Analyze trends in disease and wellness population wide
    – Continuous analysis of outcomes
• Testbed for Machine Learning
Conclusions

• Major application areas
  – Exascale++
  – Impact – “cure cancer”

• “Domains”
  – Spatio-temporal Sensor Integration, Analysis, Classification
  – Integrative Predictive Analytics

• Agile extreme scale computing