Exascale Challenges: Space, Time, Experimental Science and Self Driving Cars

Joel Saltz MD, PhD
Emory University
February 2013
Integrate Information from Sensors, Images, Cameras

- Multi-dimensional spatial-temporal datasets
  - Radiology and Microscopy Image Analyses
  - Oil Reservoir Simulation/Carbon Sequestration/Groundwater Pollution Remediation
  - Biomass monitoring and disaster surveillance using multiple types of satellite imagery
  - Weather prediction using satellite and ground sensor data
  - Analysis of Results from Large Scale Simulations
  - Square Kilometer Array
  - Google Self Driving Car

- Correlative and cooperative analysis of data from multiple sensor modalities and sources

- Equivalent from standpoint of data access patterns – we propose a integrative sensor data mini-App
Deep Learning from Medical Imaging

<table>
<thead>
<tr>
<th></th>
<th>Average Pathology Practice</th>
<th>Large Pathology Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 hrs per day*</td>
<td>1.5 min per slide</td>
<td>21 s per slide</td>
</tr>
<tr>
<td>16 hrs per day*</td>
<td>3 min per slide</td>
<td>42 s per slide</td>
</tr>
</tbody>
</table>

- **Data per slide:** 500MB to 100GB
- **Roughly 250-500M Slides/Year in USA**
- **Total:** 0.1-10 Exabytes/year
Emory In Silico Center for Brain Tumor Research  (PI = Dan Brat, PD= Joel Saltz)
Integrative Cancer Research with Digital Pathology

High-resolution whole-slide microscopy

histology  Multiplex IHC  neuroimaging

molecular  

Integrated Analysis
Morphological Tissue Classification

Whole Slide Imaging

Nuclei Segmentation

Cellular Features

<table>
<thead>
<tr>
<th>Nuclear Morphometry</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclei Area</td>
<td>Nuclei Perimeter</td>
<td>Eccentricity</td>
<td>Circularity</td>
</tr>
<tr>
<td>Major Axis</td>
<td>Minor Axis</td>
<td>Extent Ratio</td>
<td>Fourier Shape Descriptor</td>
</tr>
<tr>
<td>Avg Inty</td>
<td>Std Inty</td>
<td>Entropy</td>
<td>Energy</td>
</tr>
<tr>
<td>Max Inty</td>
<td>Min Inty</td>
<td>Skewness</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>Intensity Information</td>
<td>Texture Information</td>
<td>Gradient Statistics</td>
<td></td>
</tr>
<tr>
<td>Avg GM</td>
<td>Std GM</td>
<td>Entropy GM</td>
<td>Skewness GM</td>
</tr>
<tr>
<td>Energy GM</td>
<td>Kurtosis GM</td>
<td>Edge Pixel Summation</td>
<td>Edge Pixel Percentage</td>
</tr>
</tbody>
</table>

Lee Cooper, Jun Kong
Direct Study of Relationship Between **Image Features vs Clinical Outcome, Response to Treatment, Molecular Information**

**Morphology Engine**
- Segmentation
- Feature Extraction
- PAIS Database
- Patient Modeling

**Clustering Engine**
- Normalization
- Feature Selection
- Consensus Clustering
- Multidimensional Scaling

**Correlative Engine**
- Survival Analysis
- Molecular Classes
- Human Pathology
- Genetic Alterations

*Lee Cooper, Carlos Moreno*
DNA sequencing costs have gone down

Moore's Law: A long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years.

Source: National Institutes of Health

Reuters graphic/Van Tsui 05/01/12
Laser Captured Microdissection -- Spatio-temporal "omic" studies

NIH researchers use three different types of laser-capture microdissection (LCM) technologies to dissect tissue such as the human prostate gland shown here. From top, the traditional LCM and the more automated systems—spatially invariant vector quantization (SIVQ)—LCM and expression microdissection (xMD).
Associate genotype with phenotype

Big science experiments on cancer, heart disease, pathogen host response

Tissue specimen -- 1 cm³
0.3 µm resolution – roughly \(10^{13}\) 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: \(10^{18}\) bytes – exabyte per big science experiment
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
Complex Structure/Function Interactions – Very, Very Active Materials
Core Transformations

- 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
- Change Detection, Comparison, and Quantification
A Data Intense Challenge: The Instrumented Oil Field of the Future
ITR Proposal

A Data Intense Challenge:
The Instrumented Oilfield of the Future

Participants:

i. University of Texas at Austin
   • CSM: Wheeler, Dawson, Peszynska
   • IG: Sen, Stoffa
   • PGE: Torres-Verdin

ii. University of Chicago—CS: Stevens, Papka

iii. University of Maryland—CS: Sussman

iv. Ohio State—CS: Saltz, Kurc

v. Rutgers—ECE: Parashar

vi. MIT—Engineering: Haines
The Tyranny of Scale

(Tinsley Oden - U Texas)

field scale

process scale

simulation scale

km

cm

µm

pore scale
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:
    - $10^6 \times 10^6 \times 10^4$ mesh points, $10^2$ bytes per mesh point, $10^6$ timesteps --- $10^{24}$ bytes (Yottabyte) of data!!!
Detect and track changes in data during production

Invert data for reservoir properties

Detect and track reservoir changes

Assimilate data & reservoir properties into the evolving reservoir model

Use simulation and optimization to guide future production
Coupled Ground Water and Surface Water Simulations

Multiple codes -- e.g. fluid code, contaminant transport code
Different space and time scales
Data from a given fluid code run is used in different contaminant transport code scenarios
Abiotic reactions compete with microbes, reduce extent of biodegradation.

- Microbe colonies (magenta)
- Dissolved NAPL (blue)
- Mineral oxidation products (green)
Core Transformations (Again)

- 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
- Change Detection, Comparison, and Quantification
Sensor Integration and Analysis -- “Mini-app” and Co-Design Vehicle

- Tremendous commonality between applications that compose and analyze information from multiple sensors/cameras/scientific simulations

- **Recommendation** – define and publicize “abstract application class” that captures essential aspects

- Spatio-temporal data is often accompanied by chemical species, ‘omics’ or appropriate domain specific chemical information

- Computationally similar applications are currently independently developed by many research communities

- **Biology/Medicine is not a special case!!!**

- Mini-app is quite doable and would be tremendously useful if accompanied by domain area buy-in
Runtime Support

- Hierarchical dataflow/task management
- Programming model and runtime support need to work together: specify/extract tasks, dependencies
- Concurrent Collections (CnC), ParalleX Execution Model, Region Templates
Runtime Support Objectives

- *Coordinated* mapping of data and computation to complex memory hierarchies
- *Hierarchical* work assignment with *flexibility* capable of dealing with data dependent computational patterns, fluctuations in computational speed associated with power management, faults
- Linked to *comprehensible* programming model – model *targeted at abstract application class but not to application domain* (In the sensor, image, camera case -- Region Templates)
- Software stack including *coordinated compiler/runtime support/autotuning* frameworks
Coarse Grain Workflows: Interoperability

- Pipelines are complex and written in multiple languages and designed to run on multiple environments
- Key components needed to tackle problems may be purpose built to run on particular environments, may be proprietary
- Patient privacy and HIPPA issues can constrain portions of computations to institutional or highly secure, HIPPA certified environments
- Last mile bandwidth issues, performance/storage availability where data needs to be staged. Large number of tactical pitfalls which erode researcher productivity
- Data generation is cheap and often local, still expensive to move multiple TB/PB data to supercomputer centers
Exascale Hardware/Software Architecture

- Need to stage very large datasets for relatively short periods of time -- large aggregate bandwidth to non volatile scratch storage -- distributed flash and disk
- Globally addressed/indexed persistent data collections -- e.g. DataSpaces, Region Templates (GIS analogy), persistent PGAS
- Intelligent I/O with in-transit processing, data reduction (e.g. ADIOS)
- Visualizations need to be carried out interactively and in situ as data is produced and as computations proceed – efficient streaming data
Data Representation and Query

- Complex data models capturing multi-faceted information including markups, annotations, algorithm provenance etc.
- Much data modeling can be re-used in sensor image camera application class
- Efficient implementations of data models
- ADIOS, in transit processing – data, format transformations, reductions, summarizations close to data
- Key-value stores capable of supporting efficient application data access in deep, complex storage hierarchies