Tools, Techniques and Methods for Integrative Data Analytics

Joel Saltz MD, PhD

Director Center for Comprehensive Informatics

EMORY UNIVERSITY

Contributions

- Computer Science: Methods and middleware for analysis, classification of very large datasets from low dimensional spatio-temporal sensors; methods to carry out comparisons and change detection between sensor datasets
- Biomedical: Mine whole slide image datasets to better predict outcome and response to treatments, generate basic insights into pathophysiology and identify new treatment targets
- CFD: Quantitative characterization of spatio-temporal features generated by large scale simulations, comparisons with experimental results, uncertainty quantification

Extreme Spatio-Temporal Data Analytics

- 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

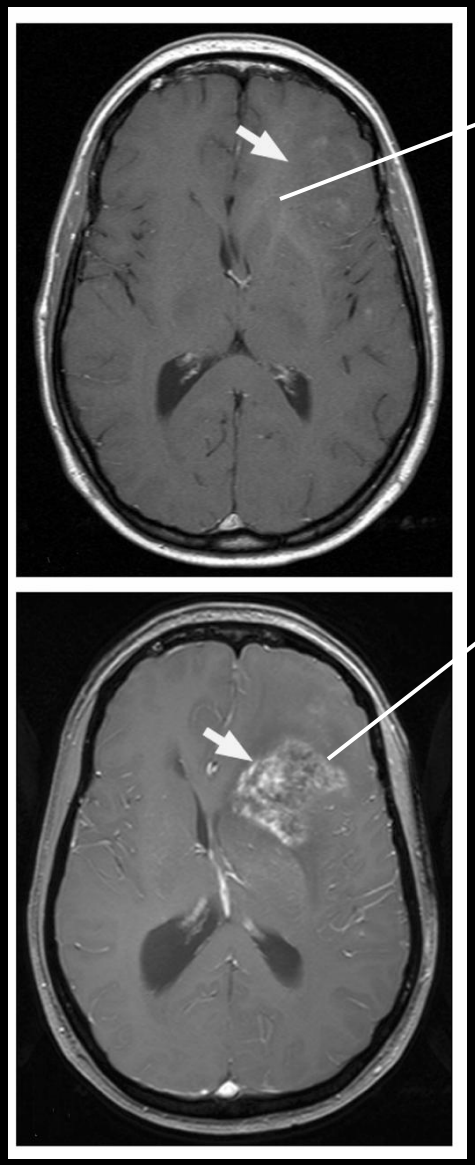
Application Targets

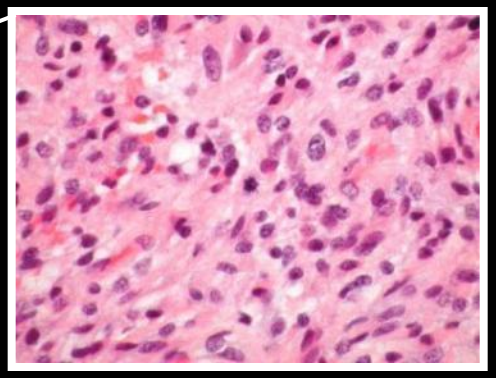
- Multi-dimensional spatial-temporal datasets
  - Microscopy image analyses
  - Biomass monitoring using satellite imagery
  - Weather prediction using satellite and ground sensor data
  - Large scale simulations
- Can we analyze 100,000+ microscopy images per hour?
- Correlative and cooperative analysis of data from multiple sensor modalities and sources
- What-if scenarios and multiple design choices or initial conditions

Core Transformations

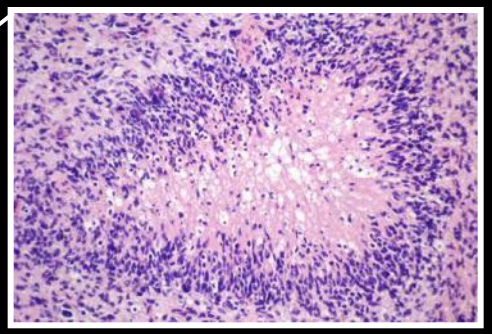
- Data Cleaning and Low Level Transformations
- Data Subsetting, Filtering, Subsampling
- Spatio-temporal Mapping and Registration
- Object Segmentation
- Feature Extraction, Object Classification
- Spatio-temporal Aggregation
- Change Detection, Comparison, and Quantification

Digital Pathology Analytics





Anaplastic Astrocytoma
(WHO grade III)

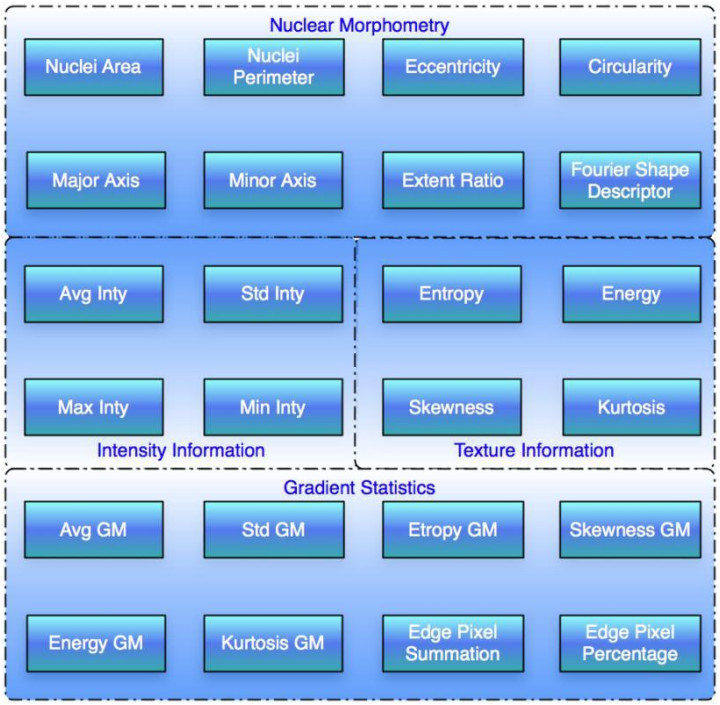


Glioblastoma
(WHO grade IV)

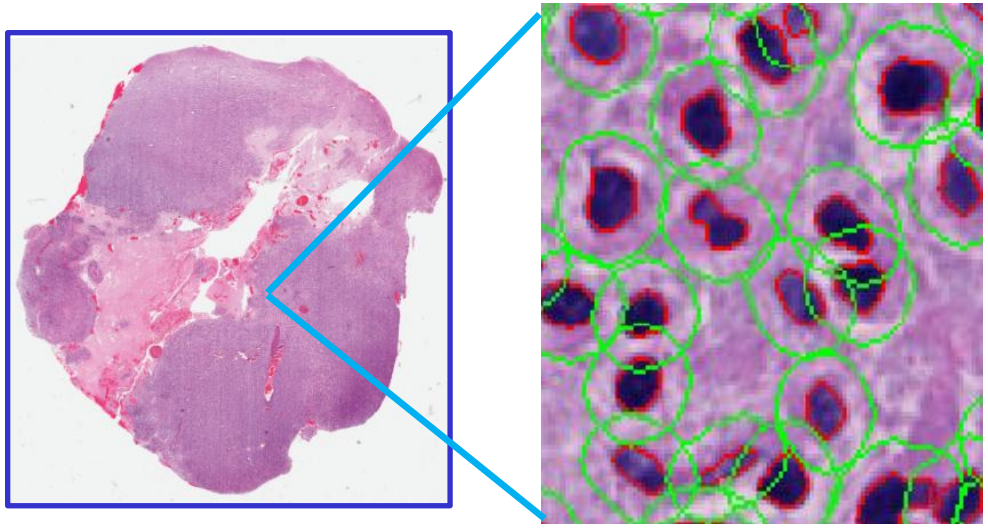
Morphological Tissue Classification

Whole Slide Imaging

Cellular Features



Nuclei Segmentation



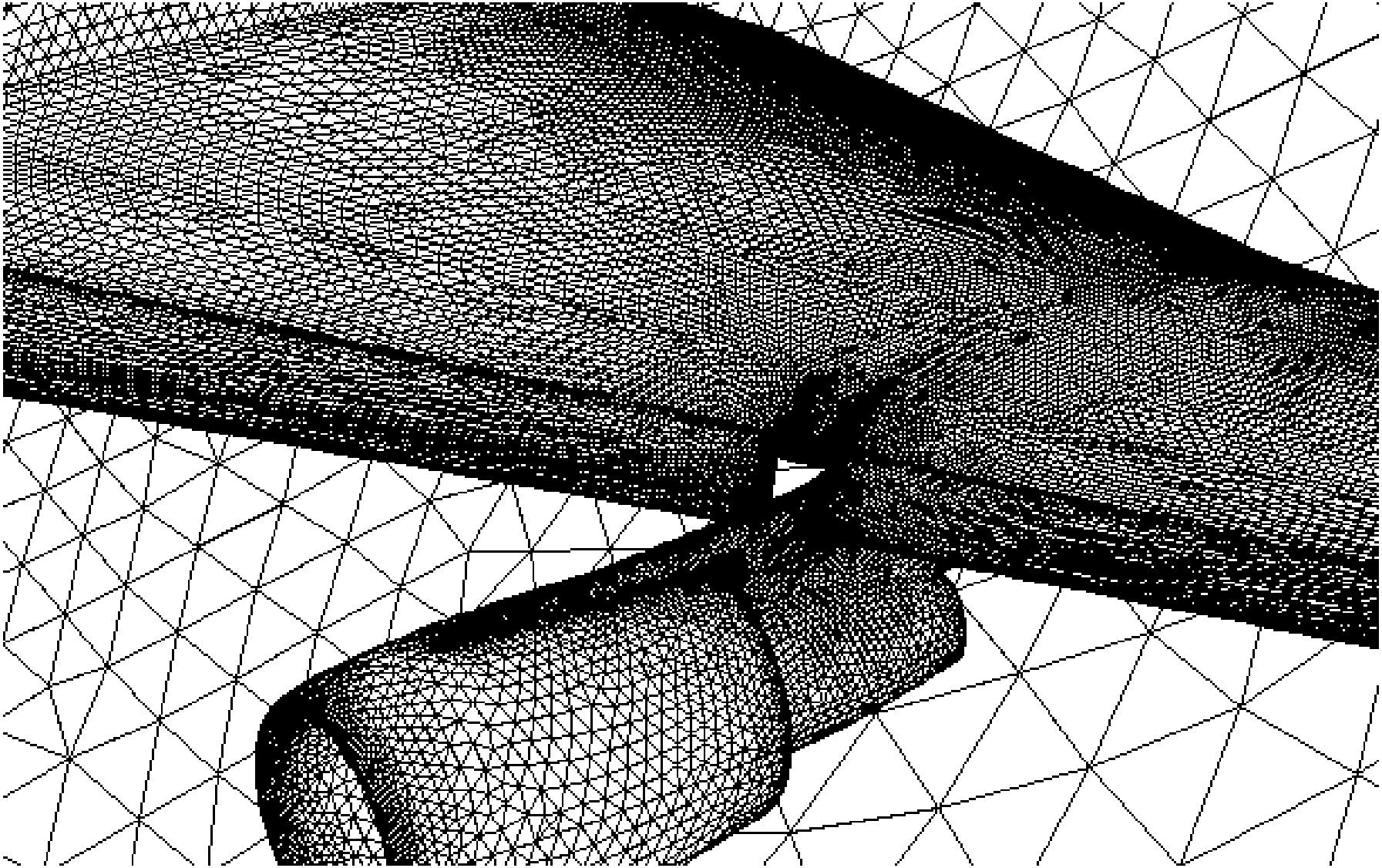
Lee Cooper,
Jun Kong

Whole Slide Imaging: Scale

	8 hrs per day*	16 hrs per day*
Average Pathology Practice 80,000 slides/yr / 250 days/yr = 320 slides/day	1.5 min per slide	3 min per slide
Large Pathology Practice 320,000 slides/yr / 250 days/yr = 1380 slides/day	21 s per slide	42 s per slide

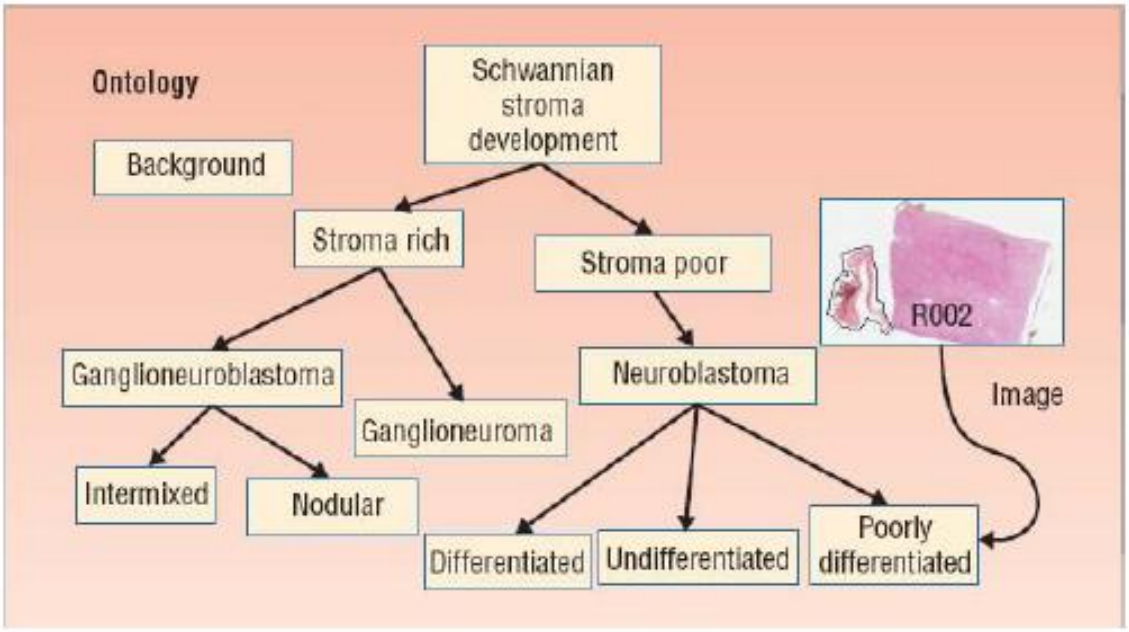
Data per slide: 500MB to 100GB
Roughly 250-500M Slides/Year in USA
Total: 0.1-10 Exabytes/year

Analysis of Computational Data; Uncertainty Quantification, Comparisons with Experimental Results



Pathology Computer Assisted Diagnosis

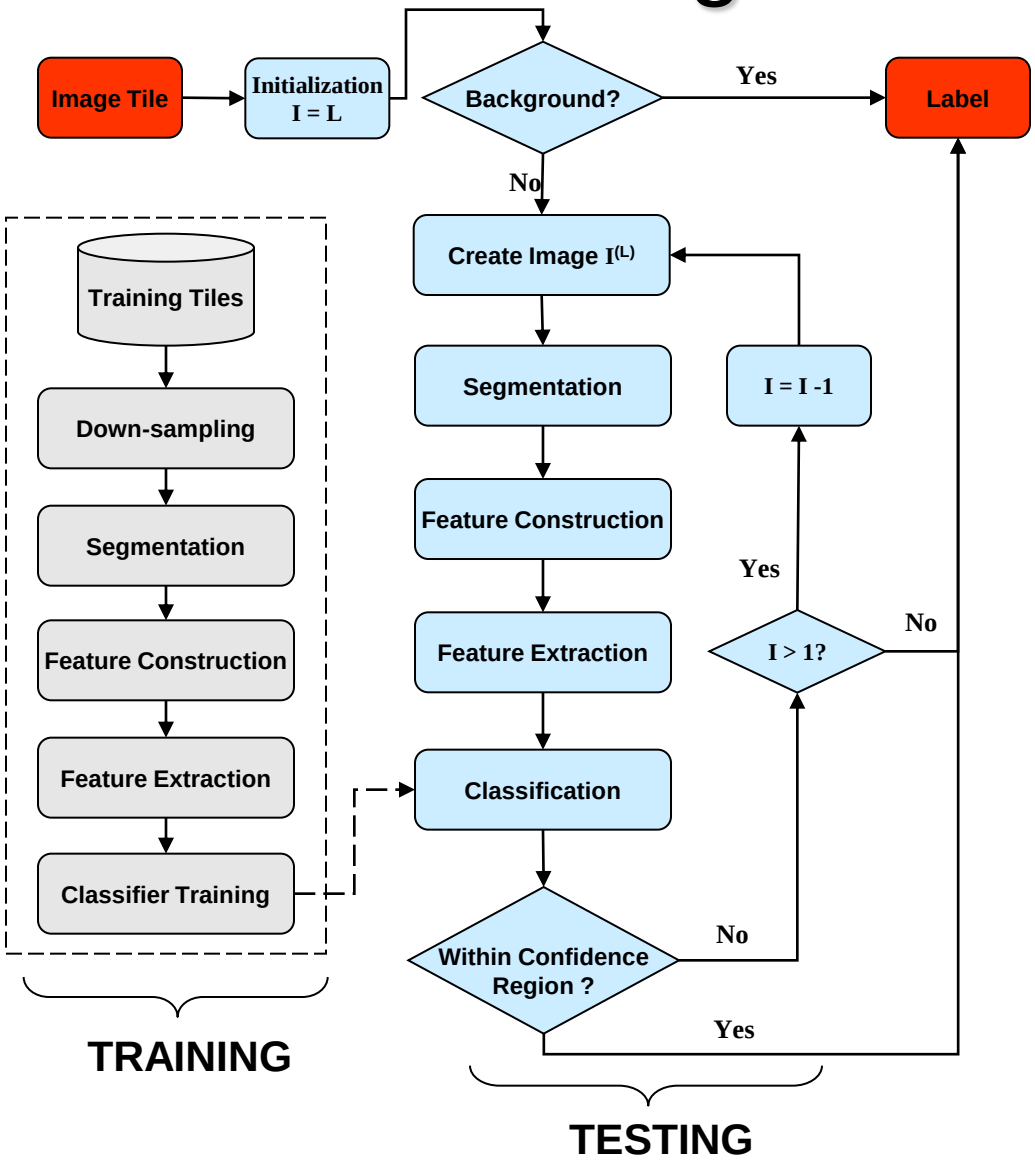
- Analyze images by computer
- Analyze the whole tissue, several slides
- Provide quantitative information to the pathologist
- Reduce inter- and intra-reader variability



Morphological characterization of tissue used for prognosis

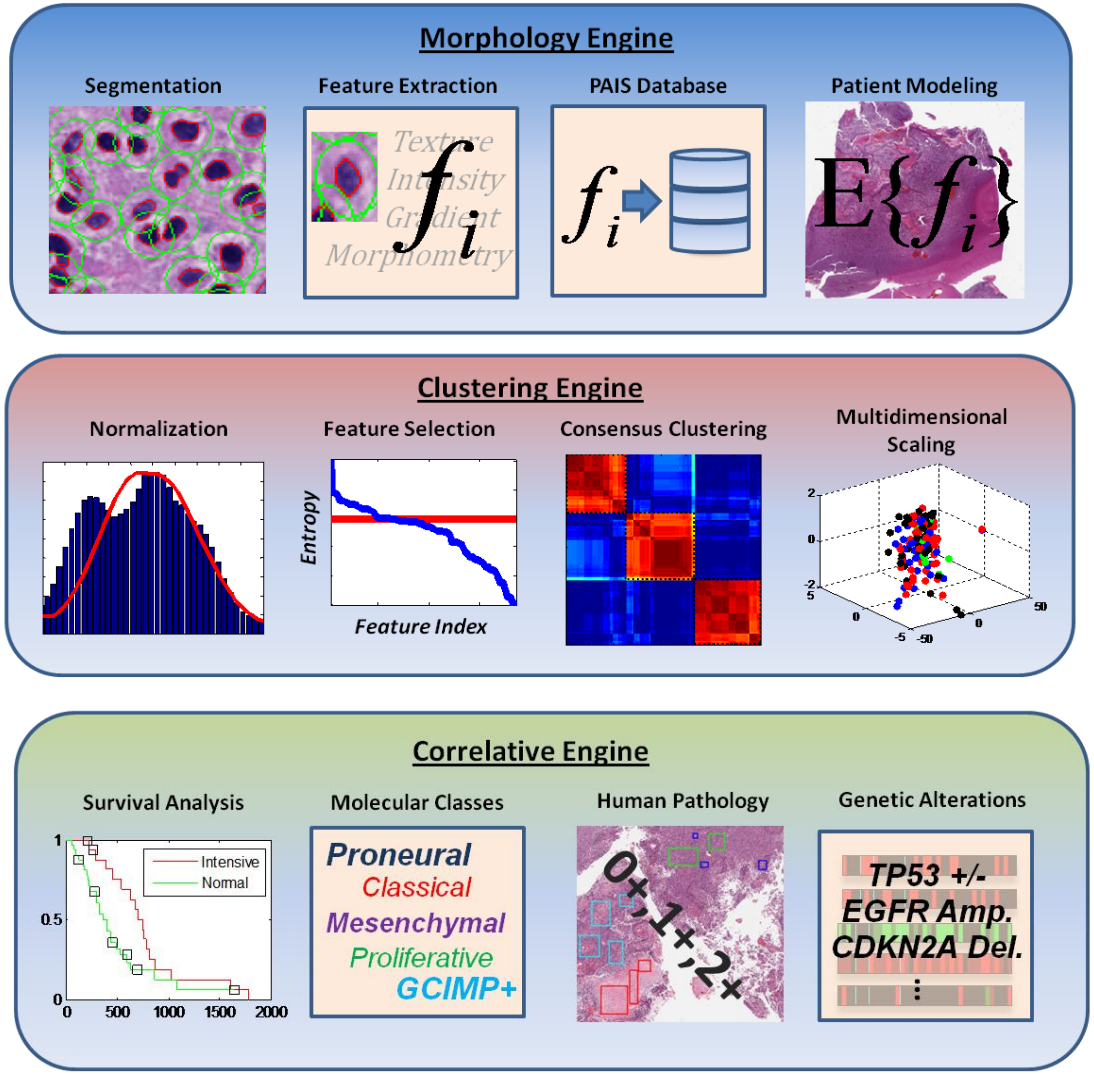
Shimada, Gurcan, Kong, Saltz

Computerized Classification System for Grading Neuroblastoma

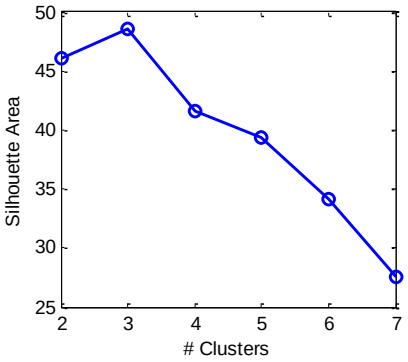


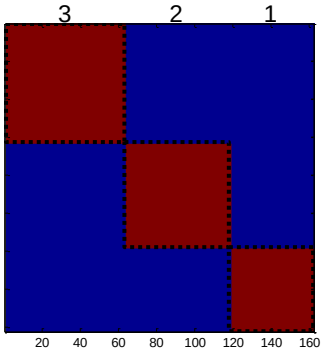
- Background Identification
- Image Decomposition (Multi-resolution levels)
- Image Segmentation (EMLDA)
- Feature Construction (2nd order statistics, Tonal Features)
- Feature Extraction (LDA) + Classification (Bayesian)
- Multi-resolution Layer Controller (Confidence Region)

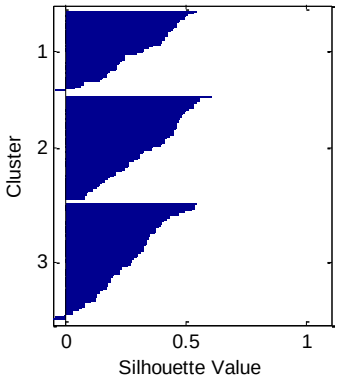
Direct Study of Relationship Between **Image Features** vs **Clinical Outcome, Response to Treatment, Molecular Information**



Nuclear Features Used to Classify GBMs







Consensus clustering of morphological signatures

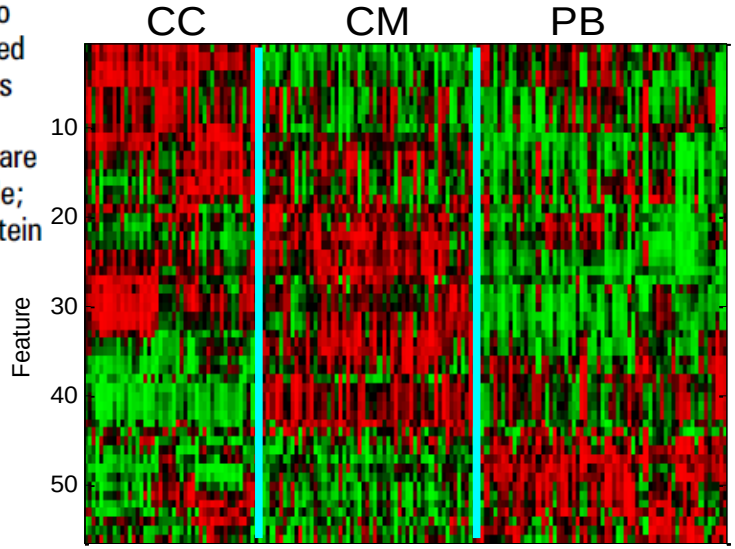
Study includes 200 million nuclei taken from 480 slides corresponding to 167 distinct patients

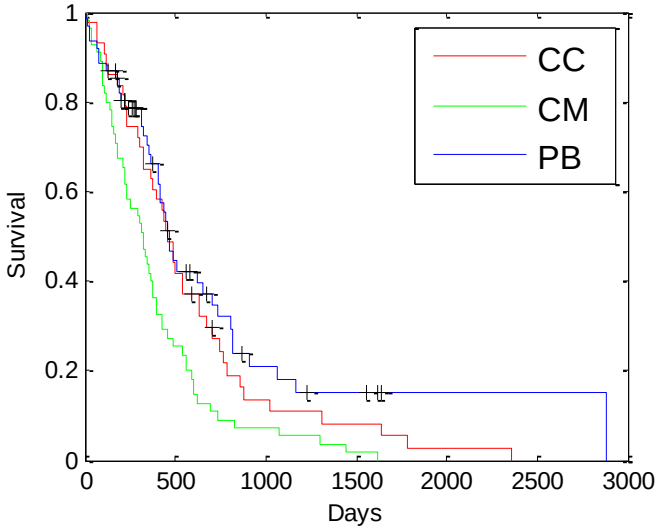
Each possibility evaluated using 2000 iterations of K-means to quantify co-clustering

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=4.5e-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.



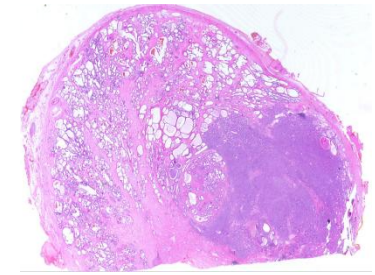


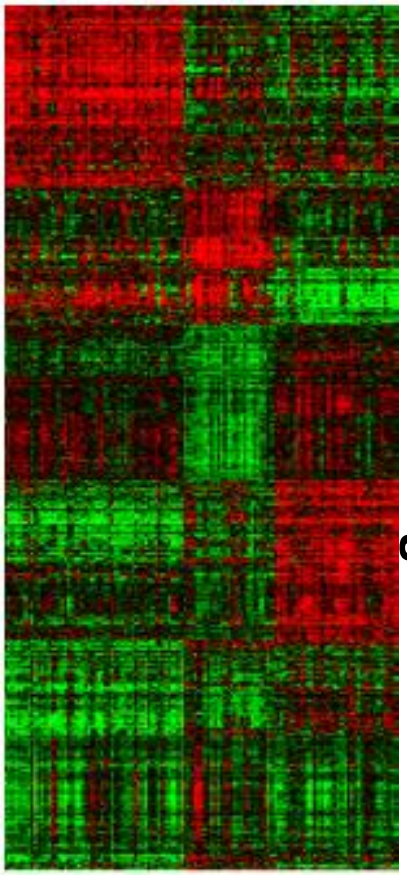
Novel Pathology Modalities

Genomics

Excellent Molecular Resolution

Limited Spatial Resolution

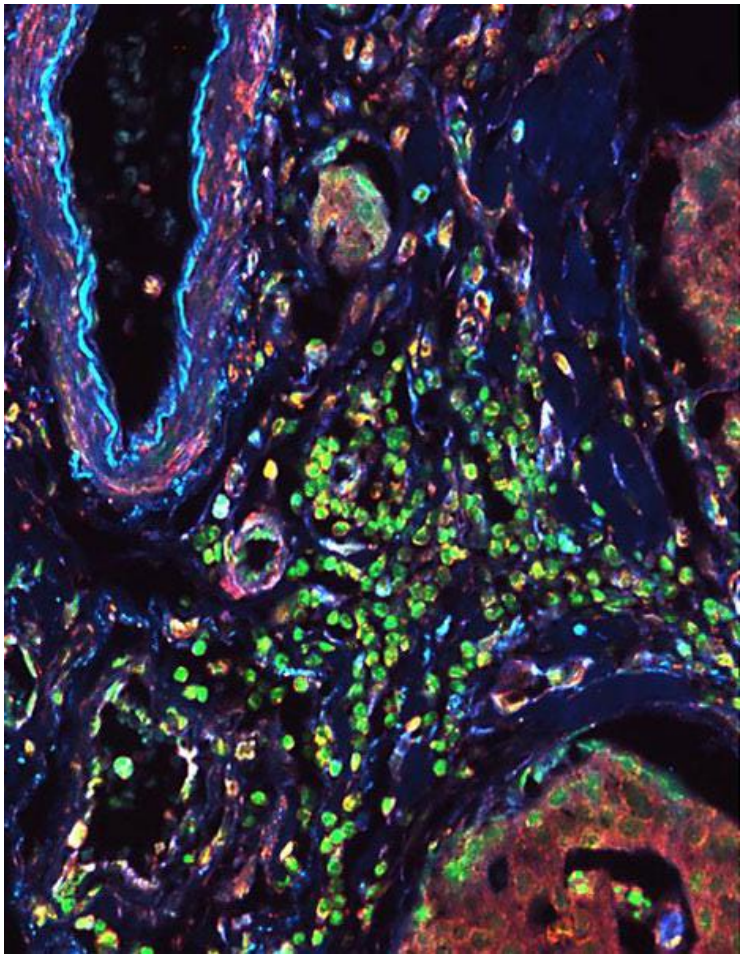




Imaging

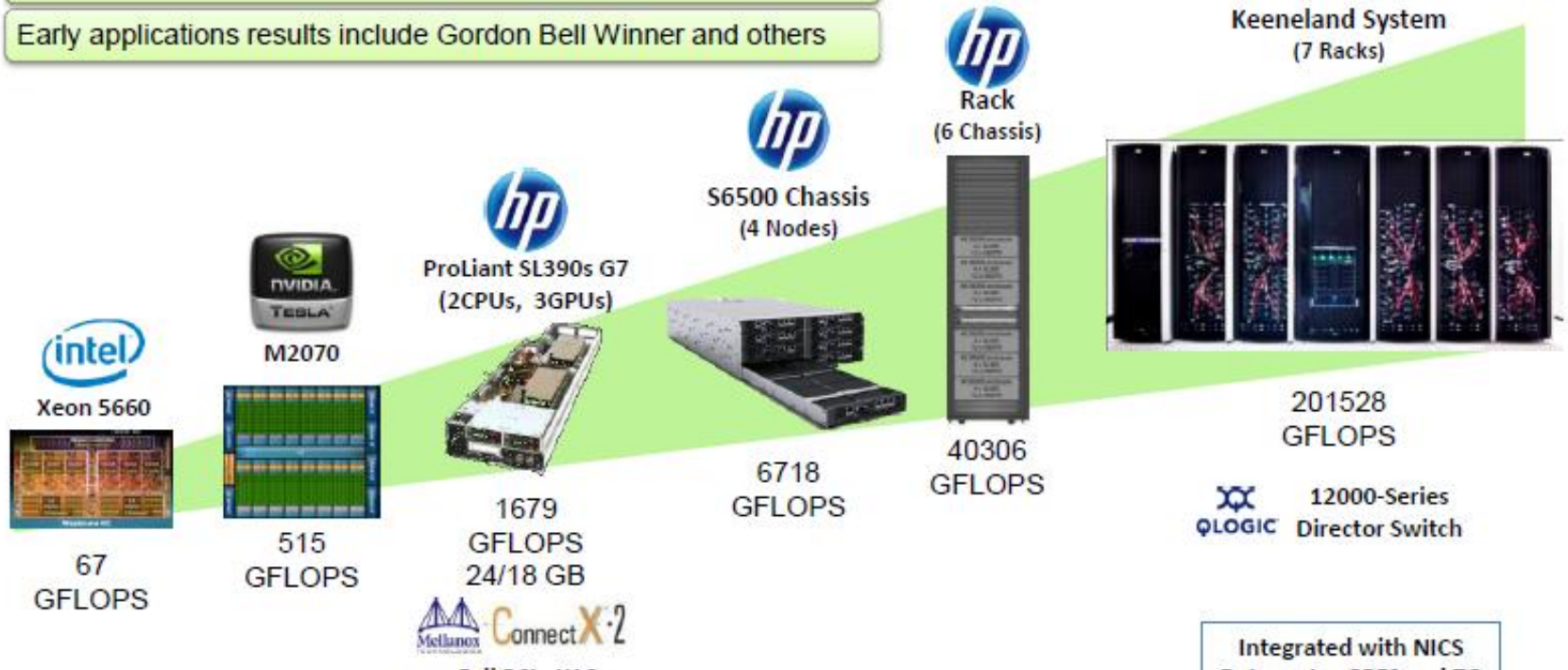
Excellent Spatial Resolution

Limited Molecular Resolution



Keeneland – Initial Delivery System

- **Initial Delivery** system installed in Oct/Nov 2010
- 201 TFLOPS in 7 racks (90 sq ft incl service area)
- 677 MFLOPS per watt on HPL (#9 on Green500, Nov 2010)
- Early applications results include Gordon Bell Winner and others



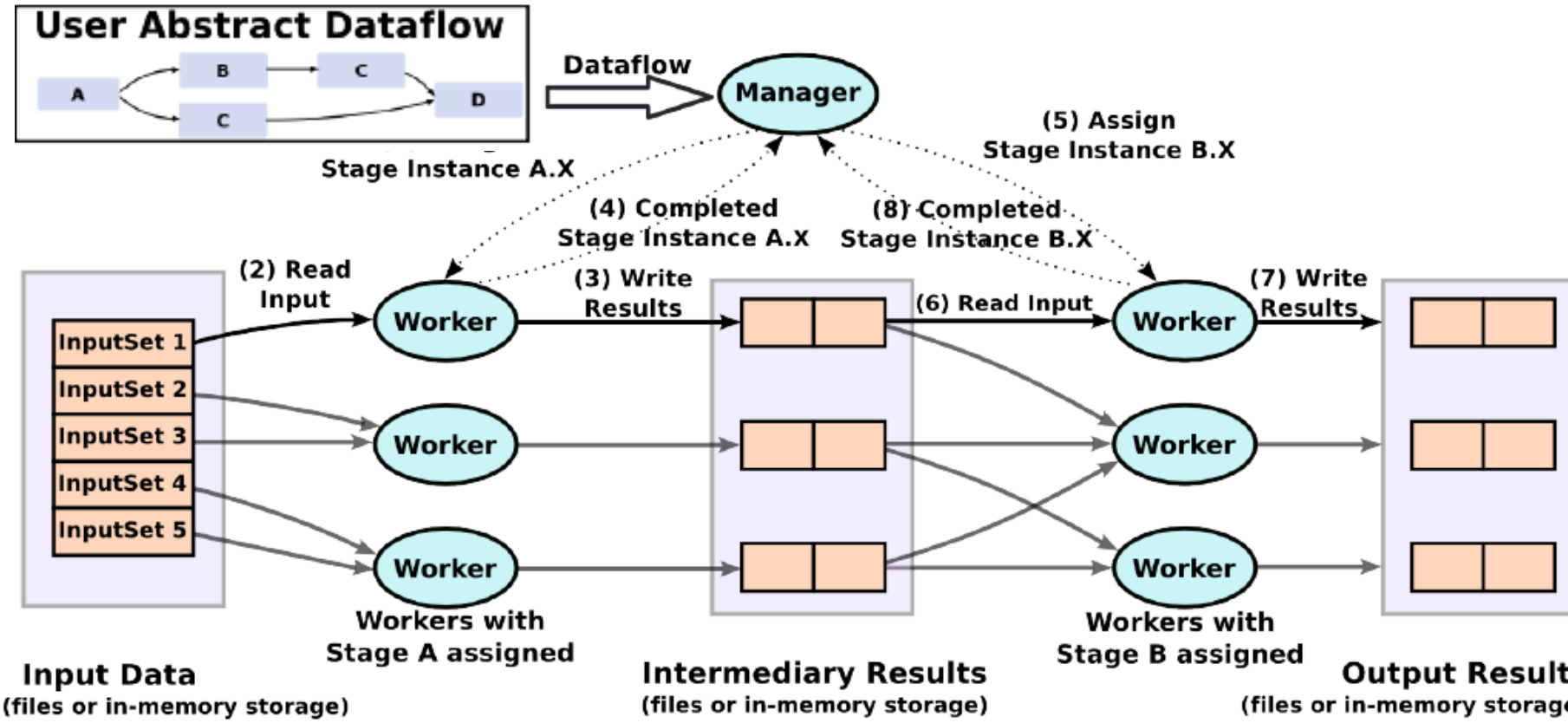
Georgia Tech NICS The University of Tennessee Oak Ridge National Laboratory NVIDIA

Extreme DataCutter Prototype

- DataCutter
  - Pipeline of filters connected though logical streams
  - In transit processing
  - Flow control between filters and streams
  - Developed 1990s-2000s; led to IBM System S
- Extreme DataCutter
  - Two level hierarchical pipeline framework
  - In transit processing
  - Coarse grained components coordinated by Manager that coordinates work on pipeline stages between nodes
  - Fine grained pipeline operations managed at the node level
  - Both levels employ filter/stream paradigm
  - Bottom line – everything ends up as DAGS

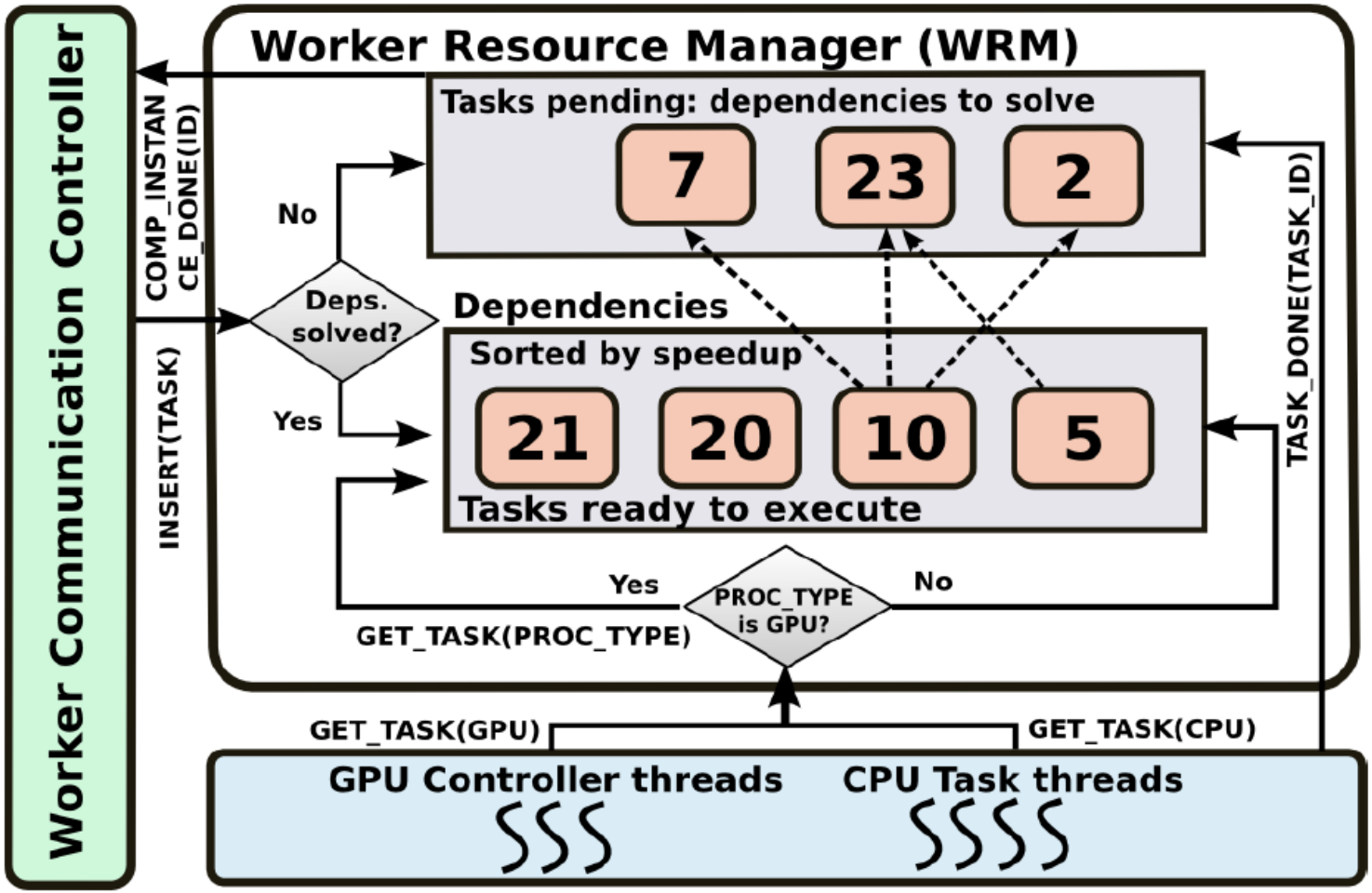
Extreme DataCutter – Two Level Model

Coarse Grained Level

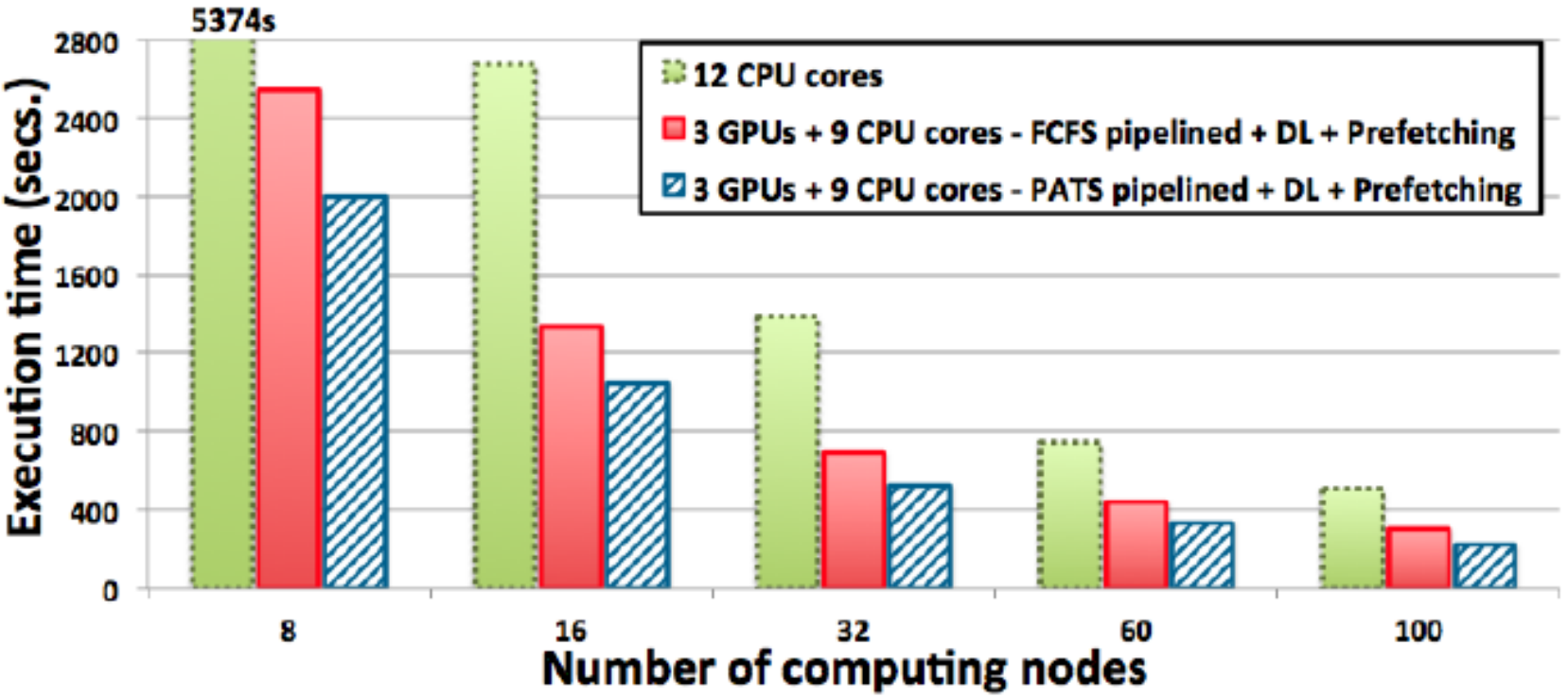


Node Level Work Scheduling

Fine Grained Level



Brain Tumor Pipeline Scaling on Keeneland (100 Nodes)



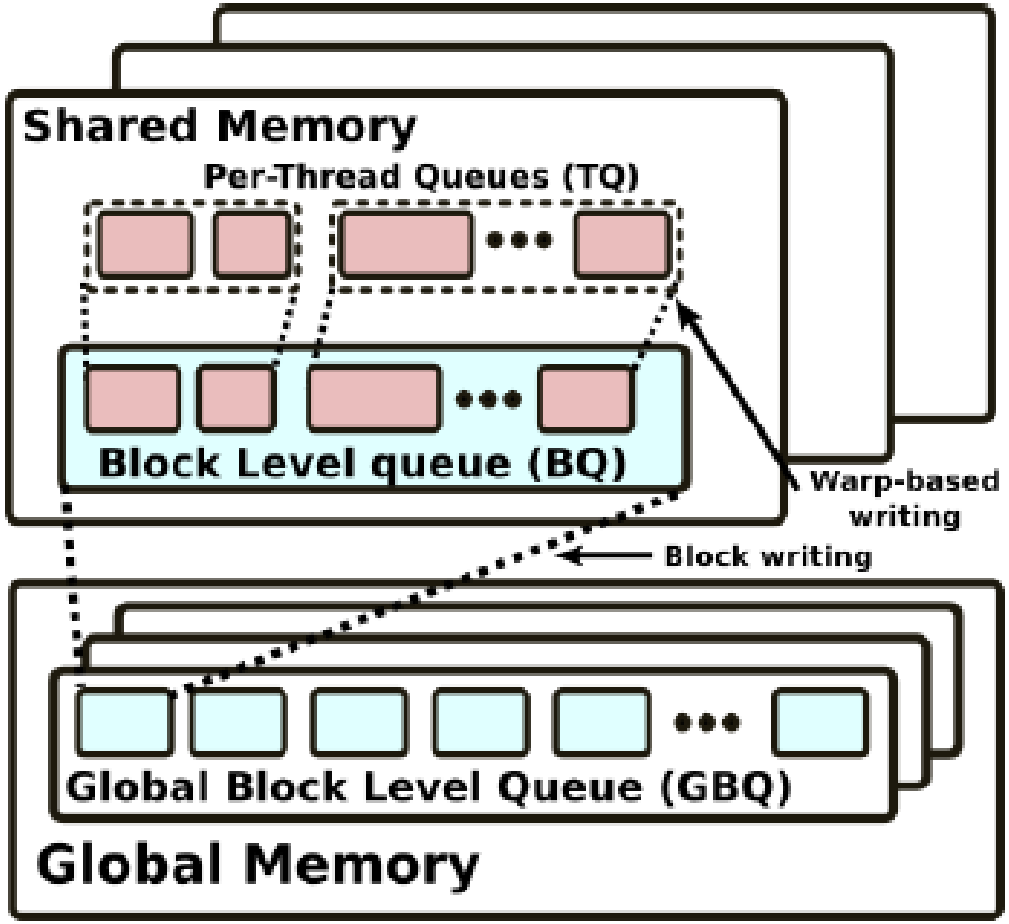
Structured/Unstructured Grid Calculations with ***Unpredictable*** Runtime Dependencies

Algorithm 1 Irregular Wavefront Propagation Pattern (IWPP)

```
1:  D ← data elements in a multi-dimensional space
2:  {Initialization Phase}
3:  S ← subset active elements from D
4:  {Wavefront Propagation Phase}
5:  while S ≠ ∅ do
6:     Extract e_i from S
7:     Q ← N_G(e_i)
8:     while Q ≠ ∅ do
9:        Extract e_j from Q
10:       if PropagationCondition(D(e_i),D(e_j)) = true then
11:          D(e_j) ← Update(D(e_i))
12:          Insert e_j into S
```

Key Kernel in Distance Transform, Morphological Reconstruction, Delaney Triagulation

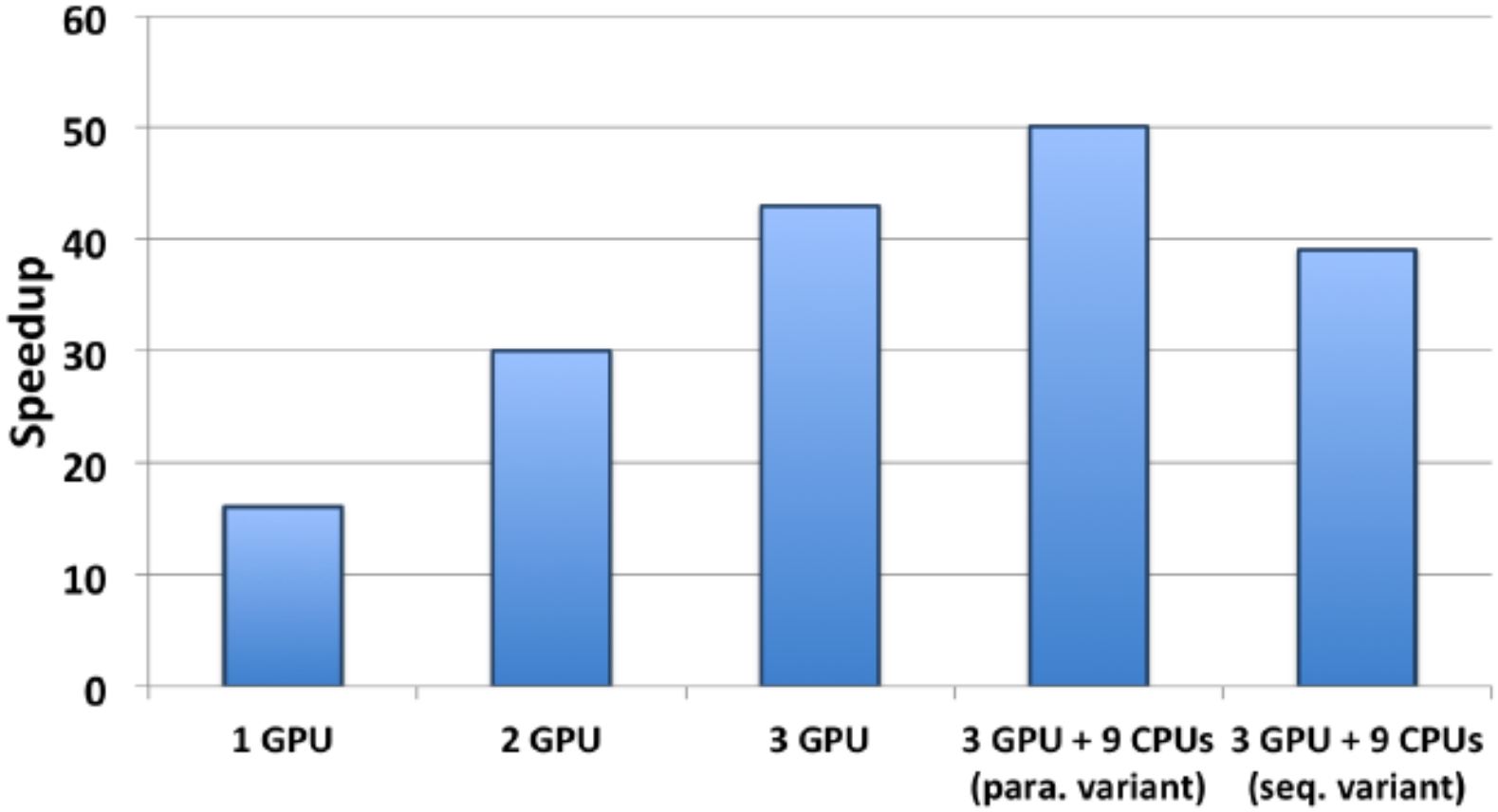
Control Structures for Handling Fine Grained/Runtime Dependent Parallelism in GPUs



Morphological Reconstruction:

8-15 Fold speedup vis one CPU core (Intel i7 2.66 GHz) on NVIDIA C2070 and GTX580 GPUs

"Speedup" relative to single CPU core



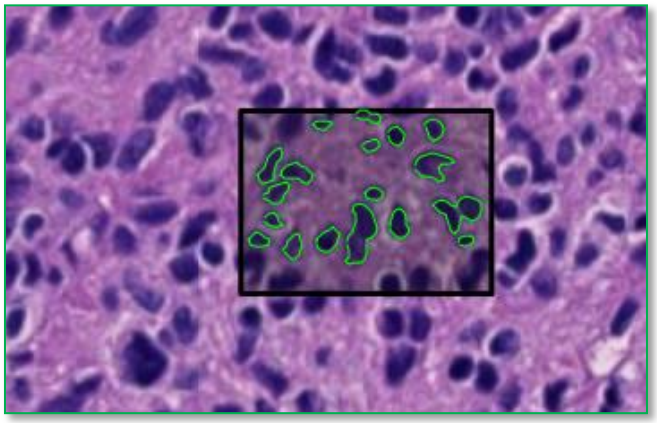
Large Scale Data Management

- Represented by a complex data model capturing multi-faceted information including markups, annotations, algorithm provenance, specimen, etc.
- Support for complex relationships and spatial query: multi-level granularities, relationships between markups and annotations, spatial and nested relationships
- Highly optimized spatial query and analyses
- Implemented in a variety of ways including optimized CPU/GPU, Hadoop/HDFS and IBM DB2

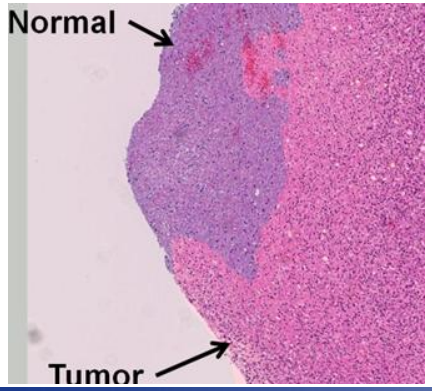
Spatial Centric – Pathology Imaging "GIS"

Point query: human marked point inside a nucleus

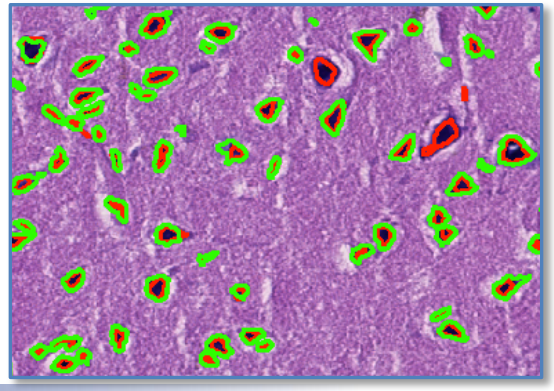
Window query: return markups contained in a rectangle

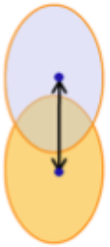


Containment query: nuclear feature aggregation in tumor regions



Spatial join query: algorithm validation/comparison

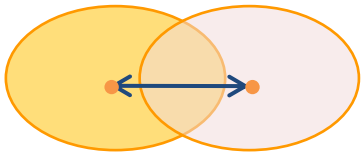


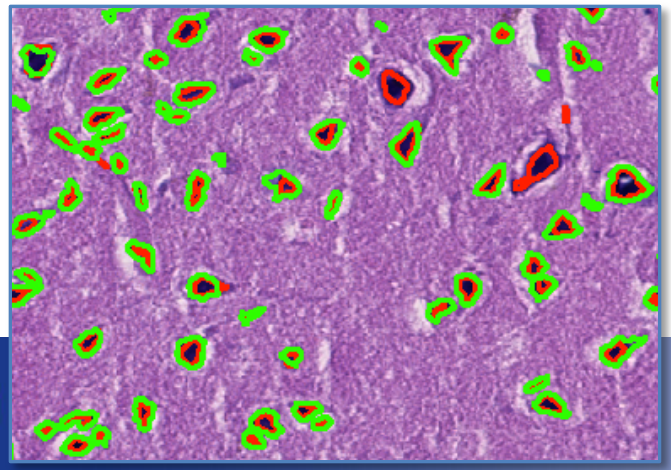


Algorithm Validation: Intersection between Two Result Sets (Spatial Join)

PAIS: Example Queries

```
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_Intersects(A.polygon, B.polygon) = 1;
```





PAIS_UID	TILE	MKPID	RATIO	DISTANCE
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,002	0.8750	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,003	0.8000	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,004	0.8064	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,005	0.8571	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,006	0.9479	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,007	0.8958	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,008	0.7903	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,009	0.8450	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,010	0.7000	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,011	0.9067	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,012	0.8953	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,013	0.9175	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,014	0.8717	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,015	0.8311	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,016	0.8623	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,017	0.8680	1.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,017	0.0000	24.52
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,018	0.8815	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,019	0.8978	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,020	0.8515	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,021	0.8255	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,022	0.8481	0.00
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,023	0.8053	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,024	0.7941	0.70
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,025	0.7721	0.50
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,026	0.2637	9.21
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,066	0.5151	2.54
astroll.1_20x_20x_NS-MORPH_1	astroll.1.ndpi-0000004096-0000004096	10,422,160,945,100,085	0.6818	0.70

VLDB 2012

Change Detection, Comparison, and Quantification

Accelerating Pathology Image Data Cross-Comparison on CPU-GPU Hybrid Systems

Kaibo Wang[1] Yin Huai[1] Rubao Lee[1] Fusheng Wang[2,3] Xiaodong Zhang[1] Joel H. Saltz[2,3]

[1]Department of Computer Science and Engineering, The Ohio State University
[2]Center for Comprehensive Informatics, Emory University
[3]Department of Biomedical Informatics, Emory University

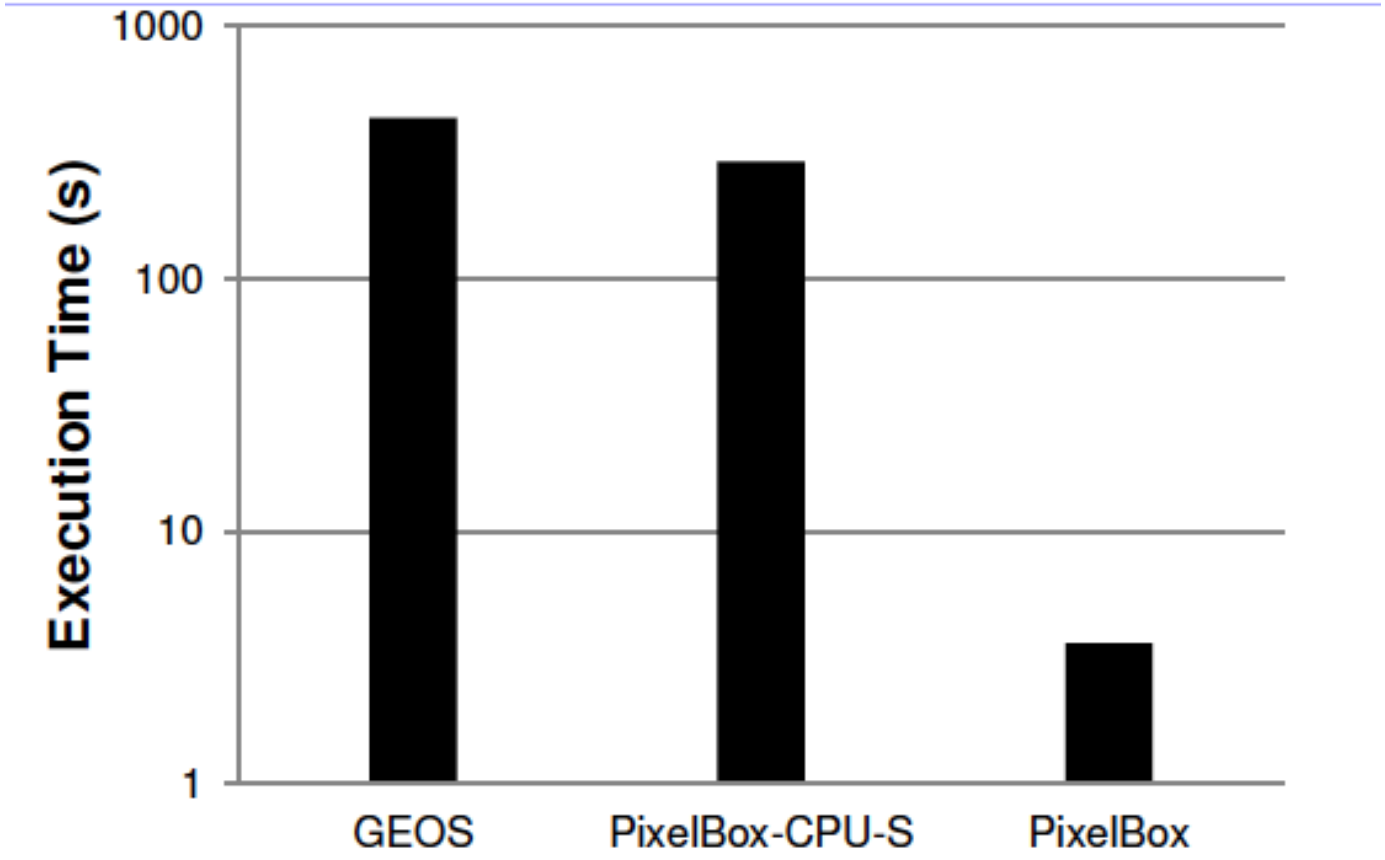
CPU/GPU Methods for Comparing Many Polygons

- Cross-compare two sets of polygons, segmented by different algorithms or the same algorithm with different parameters
- Jaccard similarity of P and Q -- two sets of polygons representing the spatial boundaries of objects generated by two methods from the same image.

$$(P \cap Q)/(P \cup Q)$$

- PixelBox accepts an array of polygon pairs as input and computes their areas of intersection and union.

Performance Improvement from PixelBox (VLDB 2012)



Summary and Perspective

- Extreme Spatio temporal data analytics
- Quantitative characterization of spatio-temporal features generated by large scale simulations, comparisons with experimental results
- Methods and tools for extreme scale data analysis pipelines
- Uncertainty quantification, comparison with experimental results

Thanks to:

- **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)**
- **NIH/in silico TCGA Imaging Group: Scott Hwang, Bob Clifford, Erich Huang, Dima Hammoud, Manal Jilwan, Prashant Raghavan, Max Wintermark, David Gutman, Carlos Moreno, Lee Cooper, John Freymann, Justin Kirby, Arun Krishnan, Seena Dehkharghani, Carl Jaffe**
- **ACTSI Biomedical Informatics Program: Marc Overcash, Tim Morris, Tahsin Kurc, Alexander Quarshie, Circe Tsui, Adam Davis, Sharon Mason, Andrew Post, Alfredo Tirado-Ramos**
- **NSF Scientific Workflow Collaboration: Vijay Kumar, Yolanda Gil, Mary Hall, Ewa Deelman, Tahsin Kurc, P. Sadayappan, Gaurang Mehta, Karan Vahi**

Thanks!