Integrative Multi-Scale Biomedical Informatics

Joel Saltz MD, PhD

Director Center for Comprehensive Informatics

EMORY UNIVERSITY

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

Outline

- Integrative biomedical informatics analysis – feature sets obtained from Pathology, Radiology and Large Population Healthcare studies
- *Same CS problems as what we have seen in Oil Reservoir/Seismic analyses and in Computational Fluid Dynamics*
- Techniques, tools and methodologies for derivation, management and analysis of feature sets
- Ideas for how to move to exascale

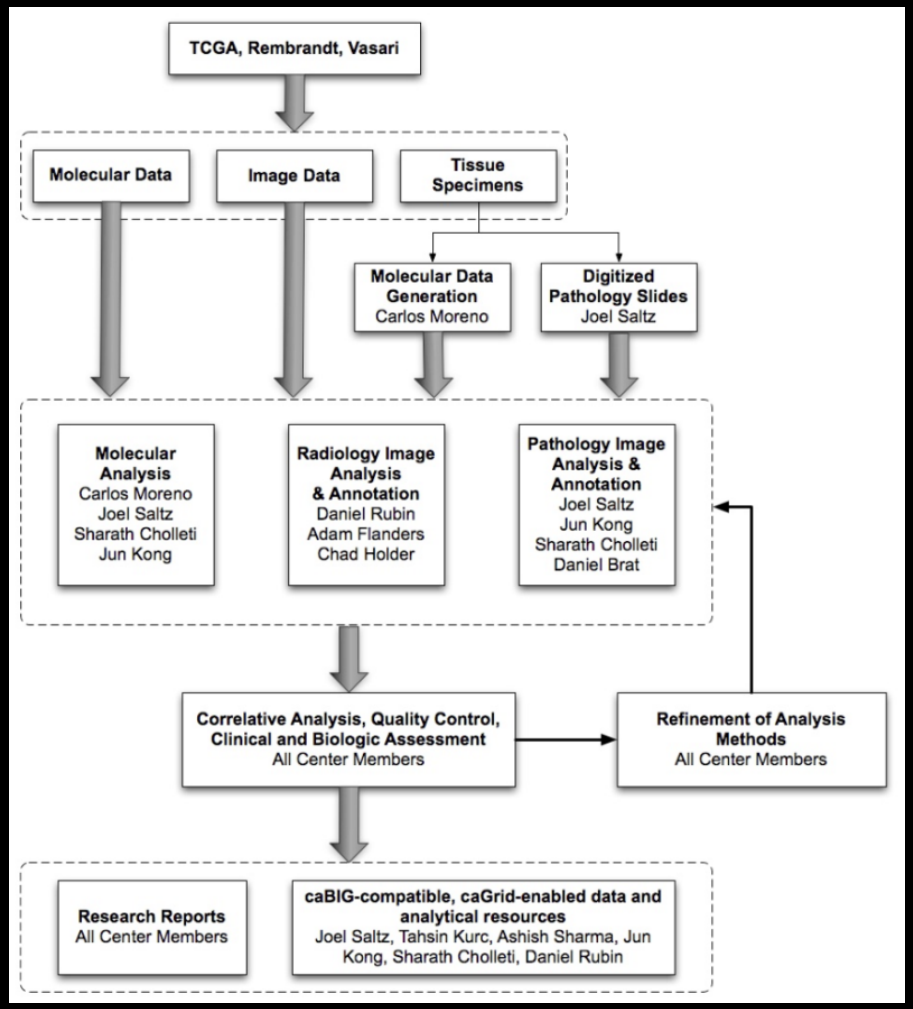
INTEGRATIVE BIOMEDICAL INFORMATICS ANALYSIS

- Reproducible anatomic/functional characterization at fine level (Pathology) and gross level (Radiology)
- 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
- Emulate classifications made by experts
- Characterize clinical phenotype

Outline of Talk

- New GBM subtypes through high throughput whole slide analysis of cellular features
- Characterization of GBM tumor microenvironment systems biology and integrated analyses of molecular, Pathology, imaging information
- Pathology data analytic tools and techniques
- High throughput characterization of clinical phenotype

In Silico Center for Brain Tumor Research



Molecular correlates of high resolution nuclear morphometry

Influence of necrosis/ hypoxia on gene expression and genetic classification

Molecular correlates of MRI enhancement patterns

In Silico Center for Brain Tumor Research

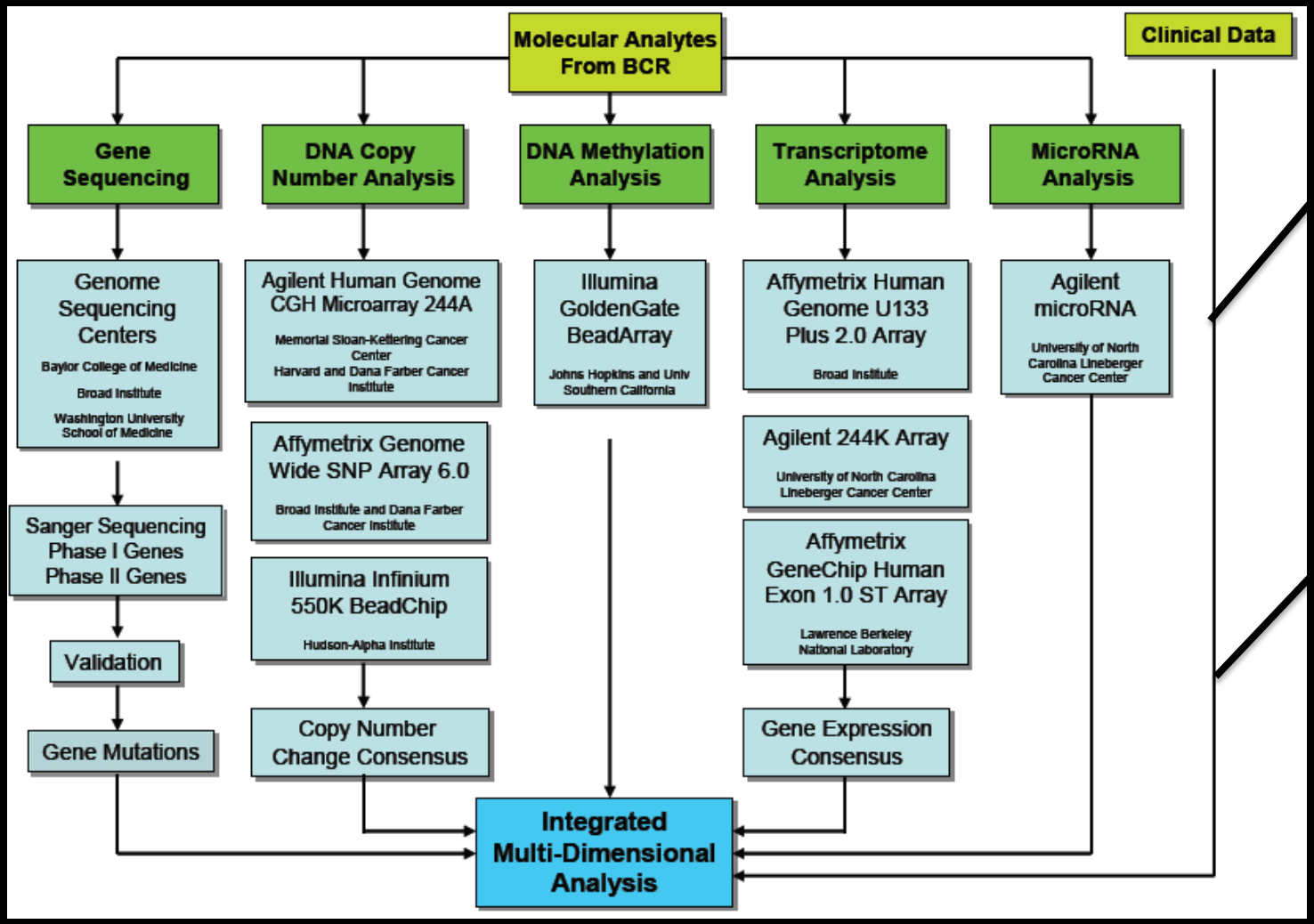
Key Data Sets

REMBRANDT: Gene expression and genomics data set of all glioma subtypes

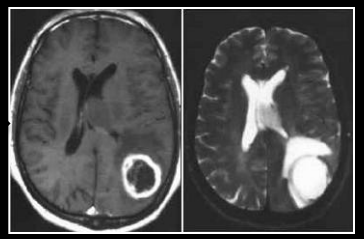
The Cancer Genome Atlas (TCGA): Rich “omics” set of GBM, digitized Pathology and Radiology

Pathology and Radiology Images from Henry Ford Hospital, Emory, Thomas Jefferson U, MD Anderson and others

TCGA Research Network



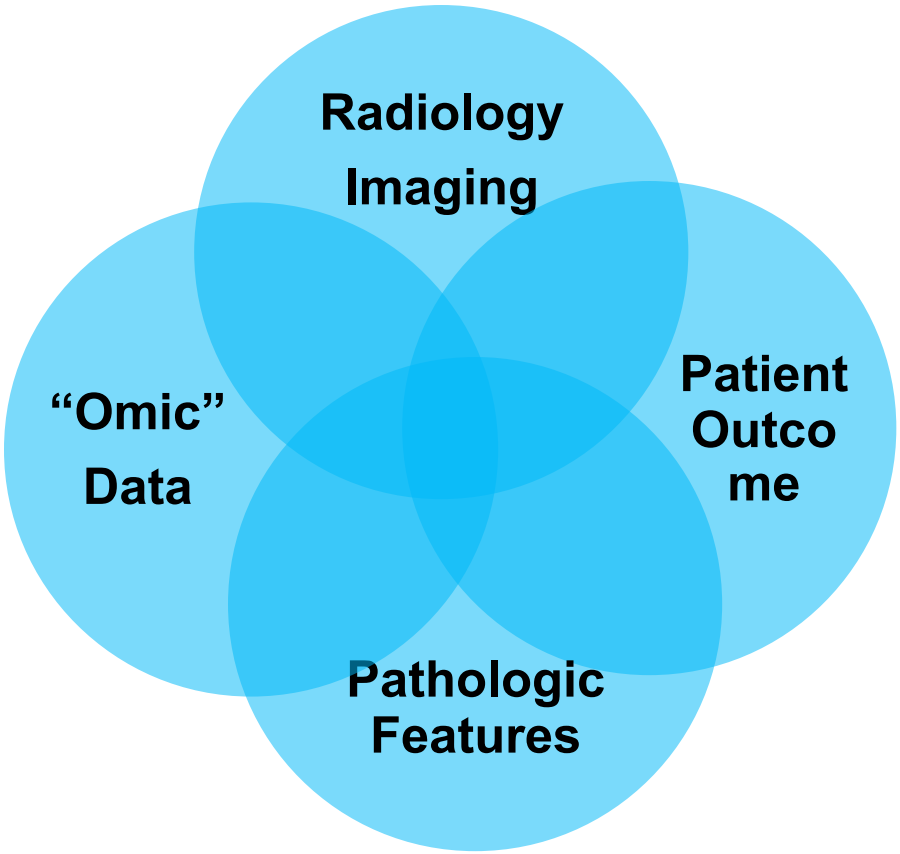
Digital Pathology



Neuroimaging

Integration of heterogeneous multiscale information

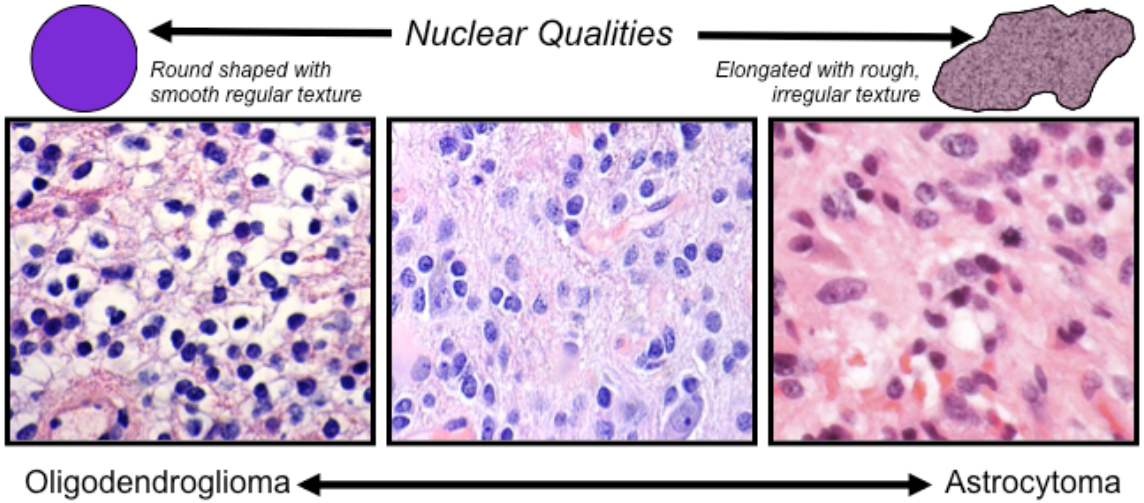
- **Coordinated initiatives Pathology, Radiology, "omics"**
- **Exploit synergies between all initiatives to improve ability to forecast survival & response.**



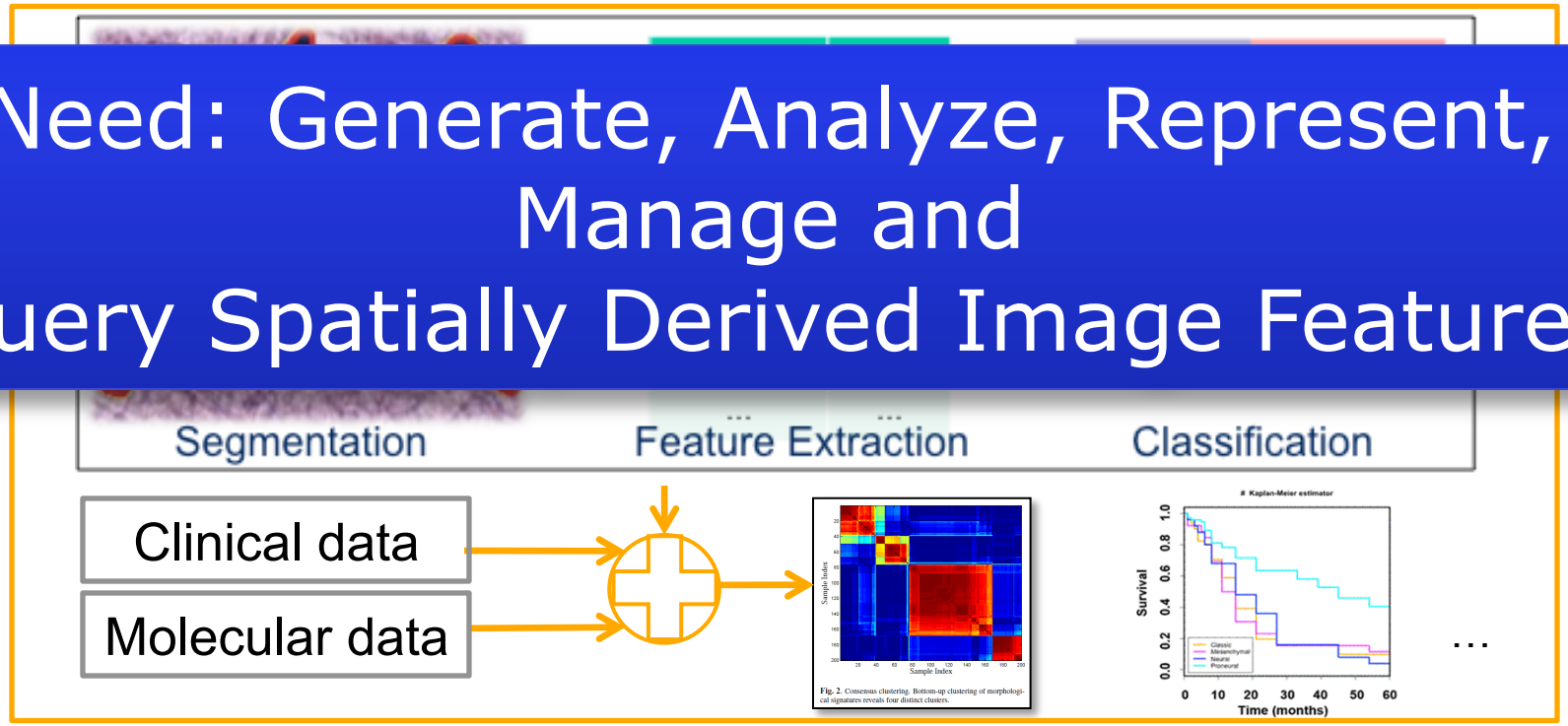
EMORY UNIVERSITY

Distinguishing Characteristics in Gliomas

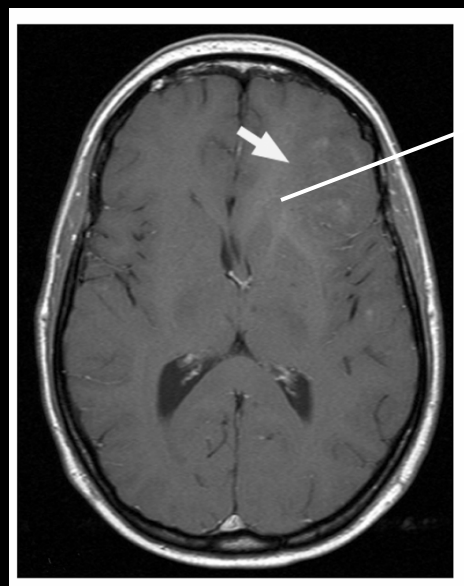
Image morphology to discover genetic alterations and forecast survival and treatment response

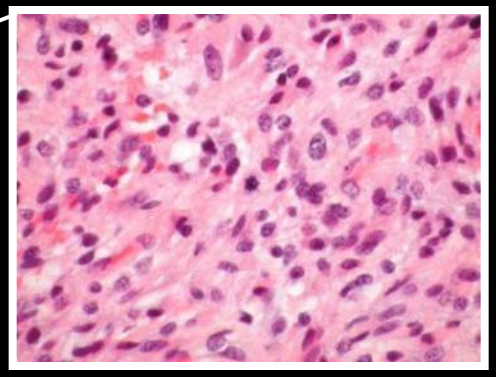


Need: Generate, Analyze, Represent, Manage and Query Spatially Derived Image Features

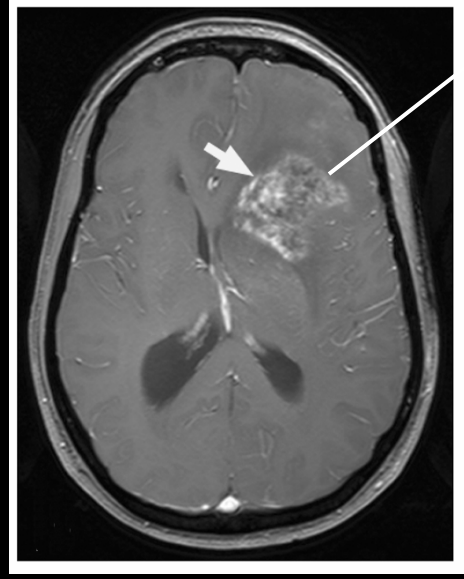


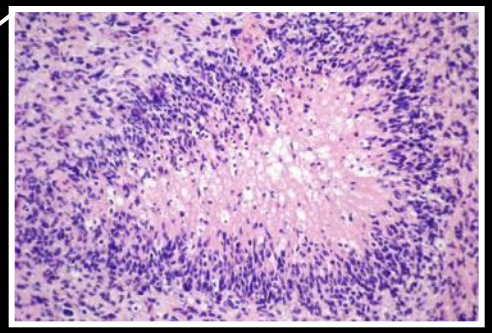
Brain Tumor Radiology and Pathology (Multiscale)





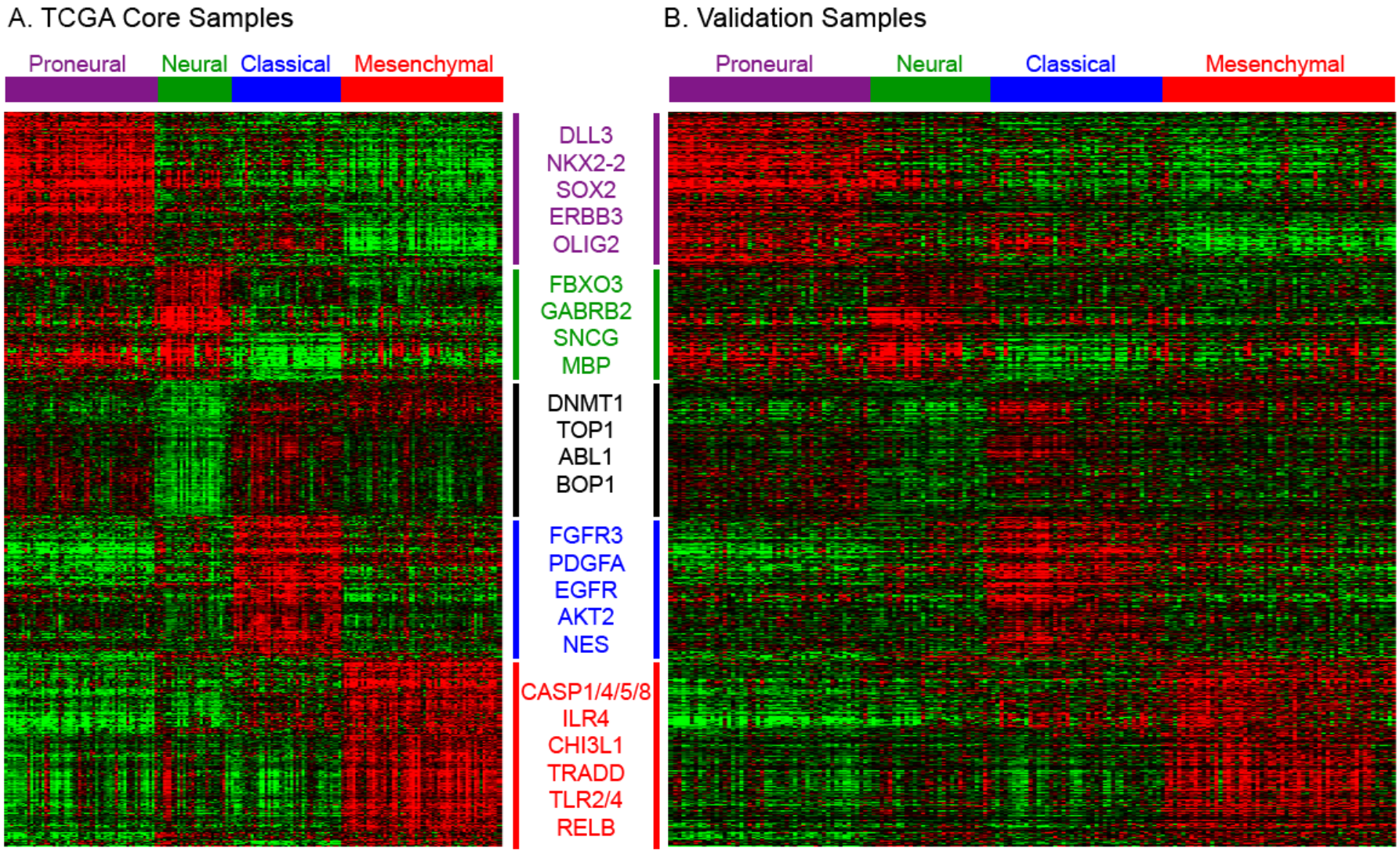
Anaplastic Astrocytoma (WHO grade III)



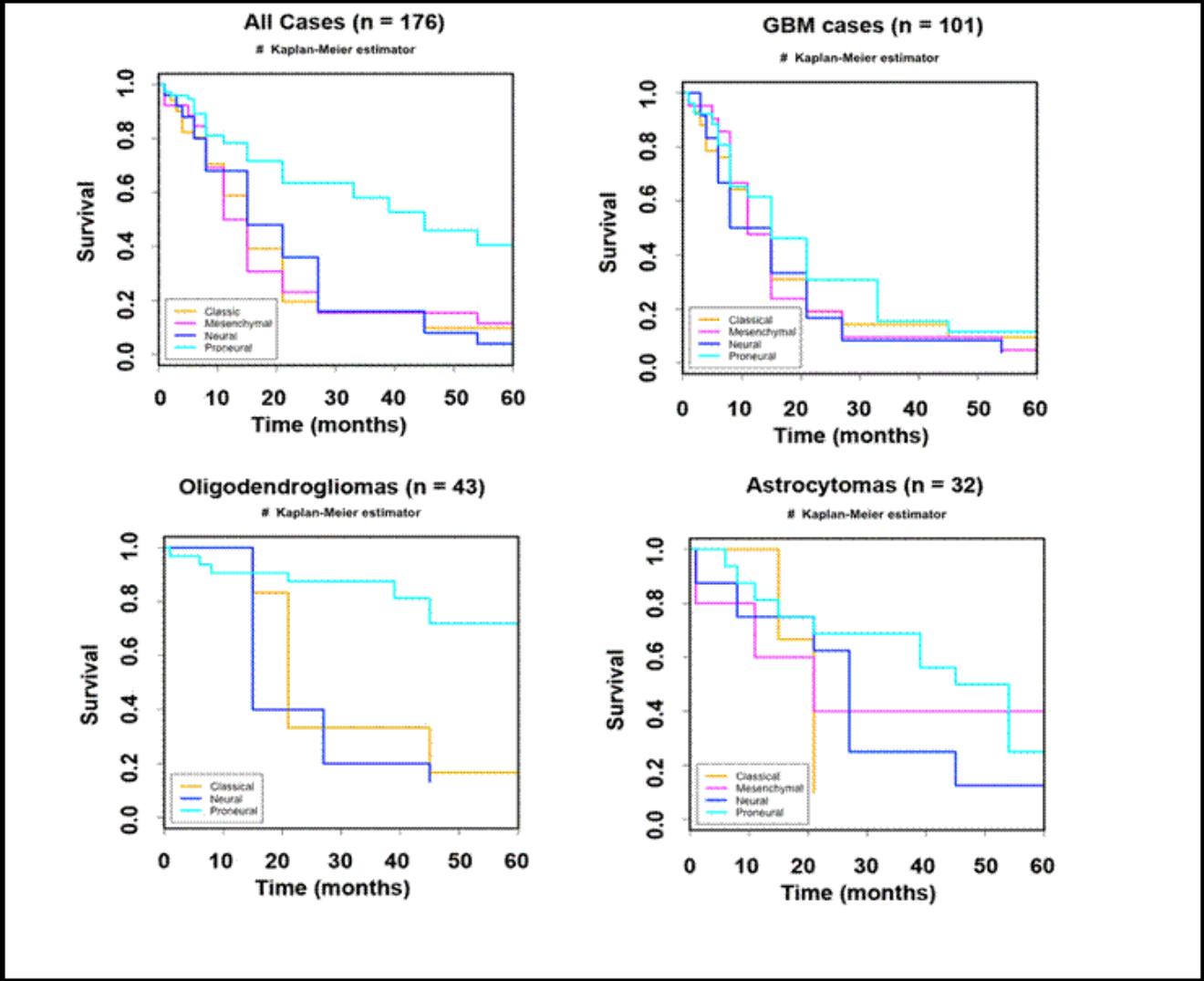


Glioblastoma (WHO grade IV)

Gene Expression Families



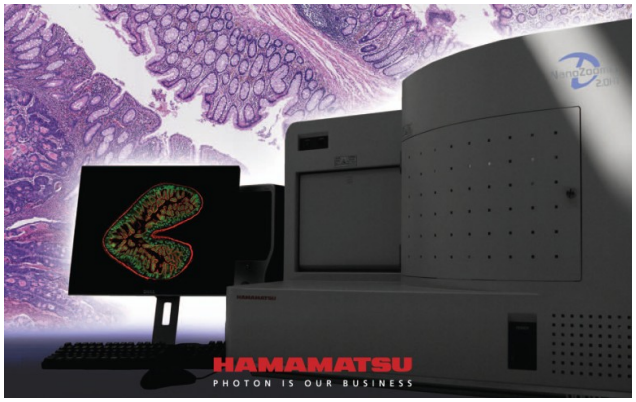
Example: Pathology and Gene Expression Joint Predictors of Recurrence/Survival



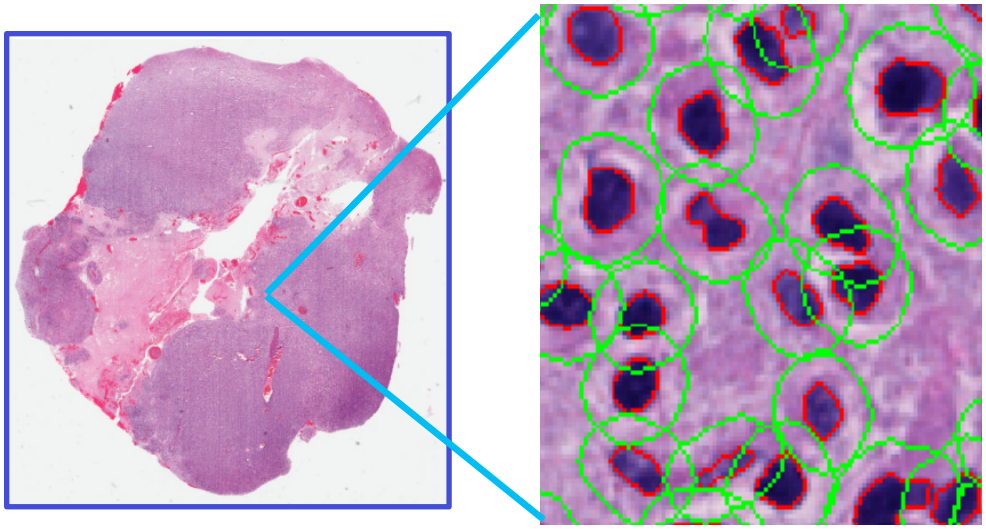
Lee Cooper Carlos Moreno

Morphological Tissue Classification

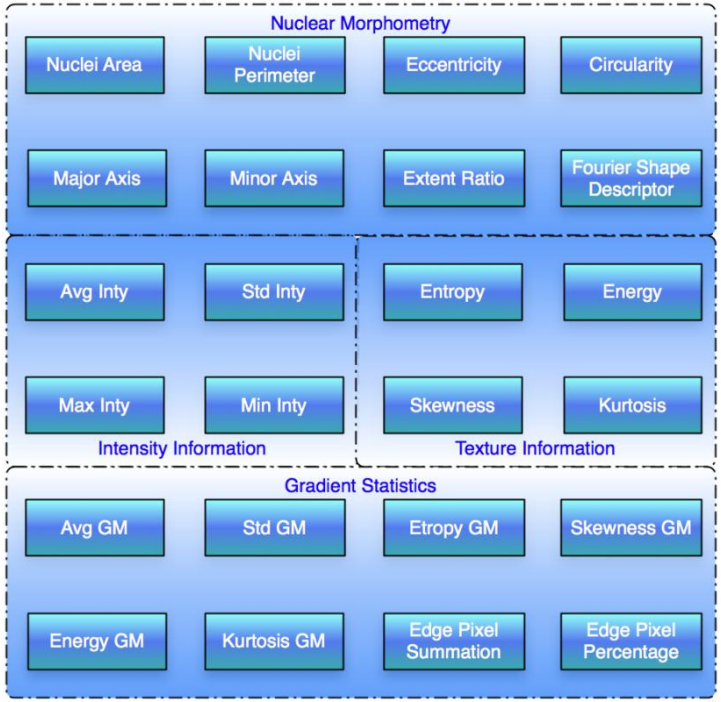
Whole Slide Imaging



Nuclei Segmentation



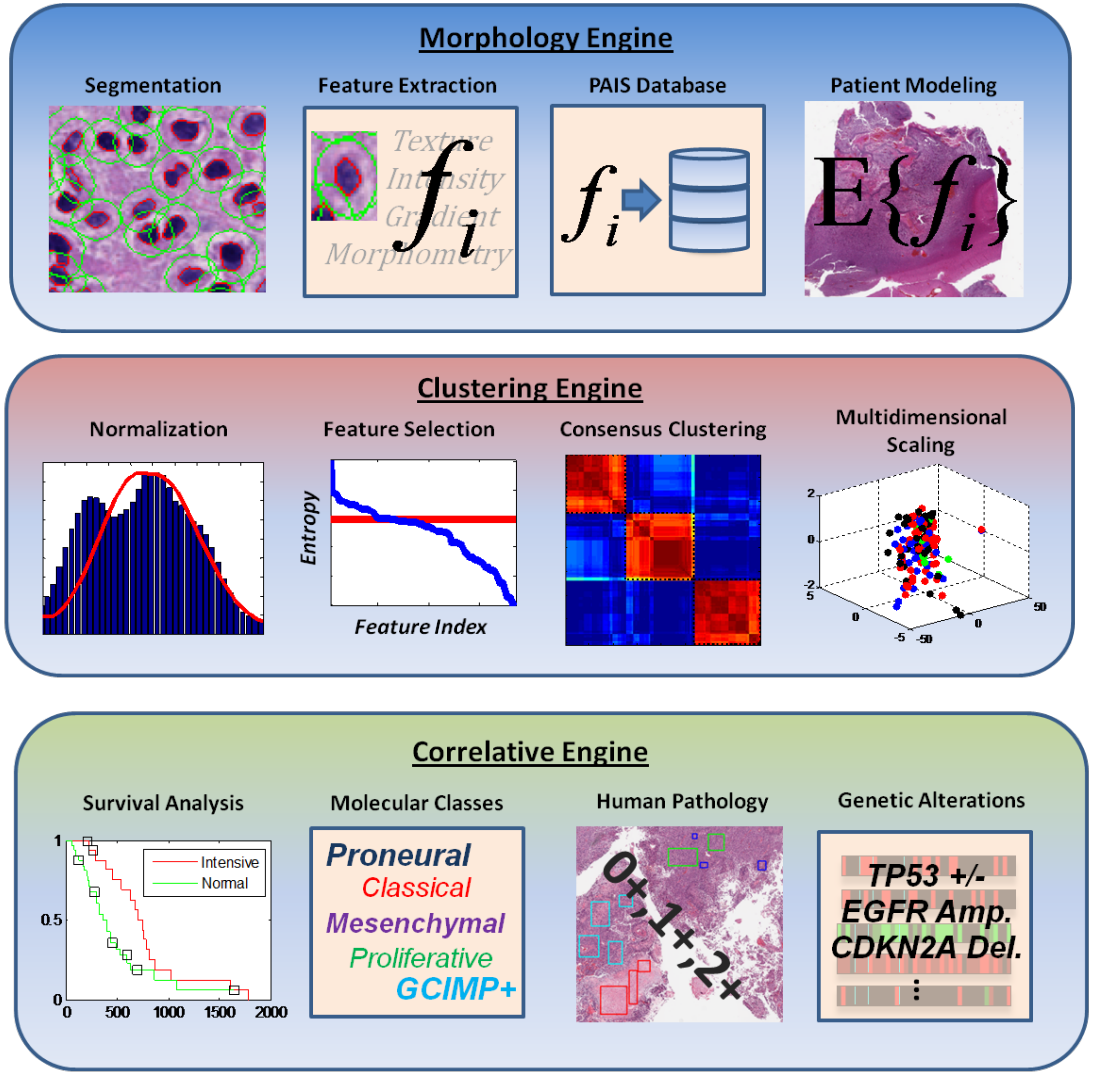
Cellular Features



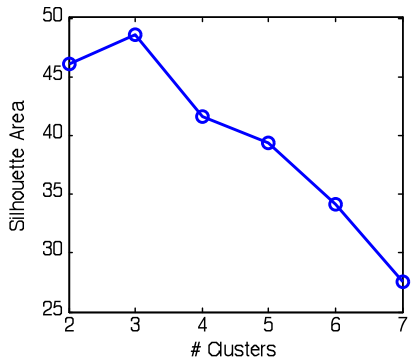
Lee Cooper, Jun Kong

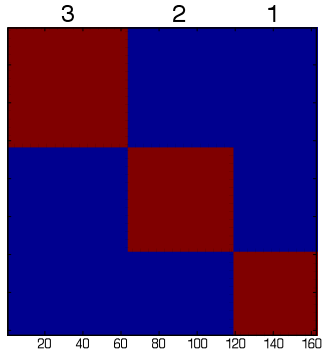
Center for Comprehensive Informatics

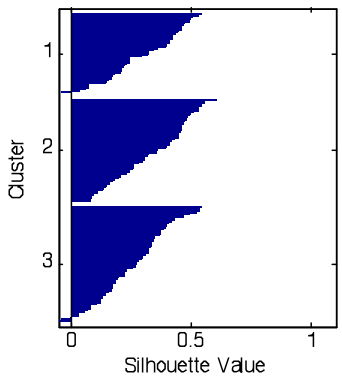
Next Study: Unsupervised Morphological Clustering – Outcome and Molecular Correlates



Nuclear Features Used to Classify GBMs







Consensus clustering of morphological signatures

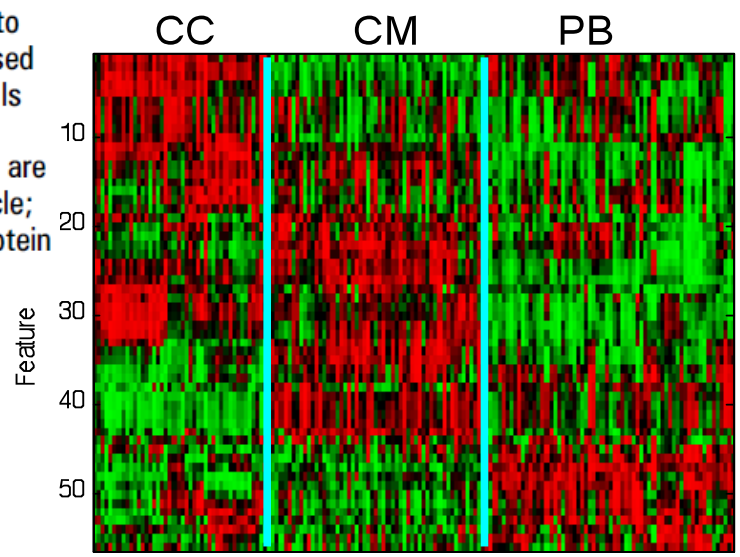
Study includes 200 million nuclei taken from 480 slides corresponding to 167 distinct patients

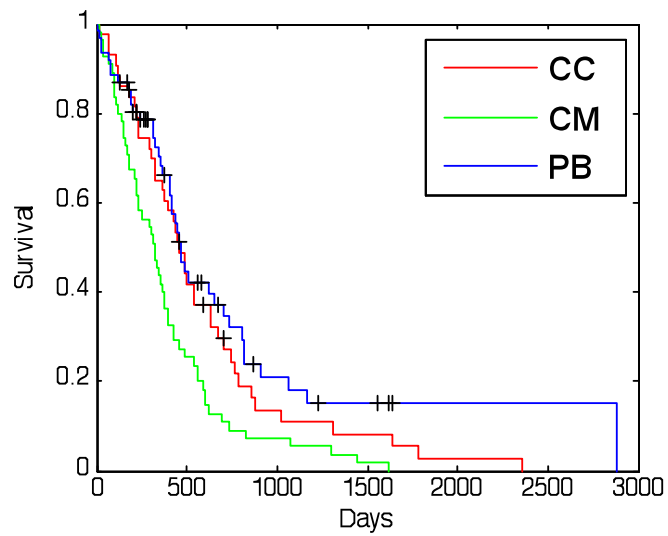
Each possibility evaluated using 2000 iterations of K-means to quantify co-clustering

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)

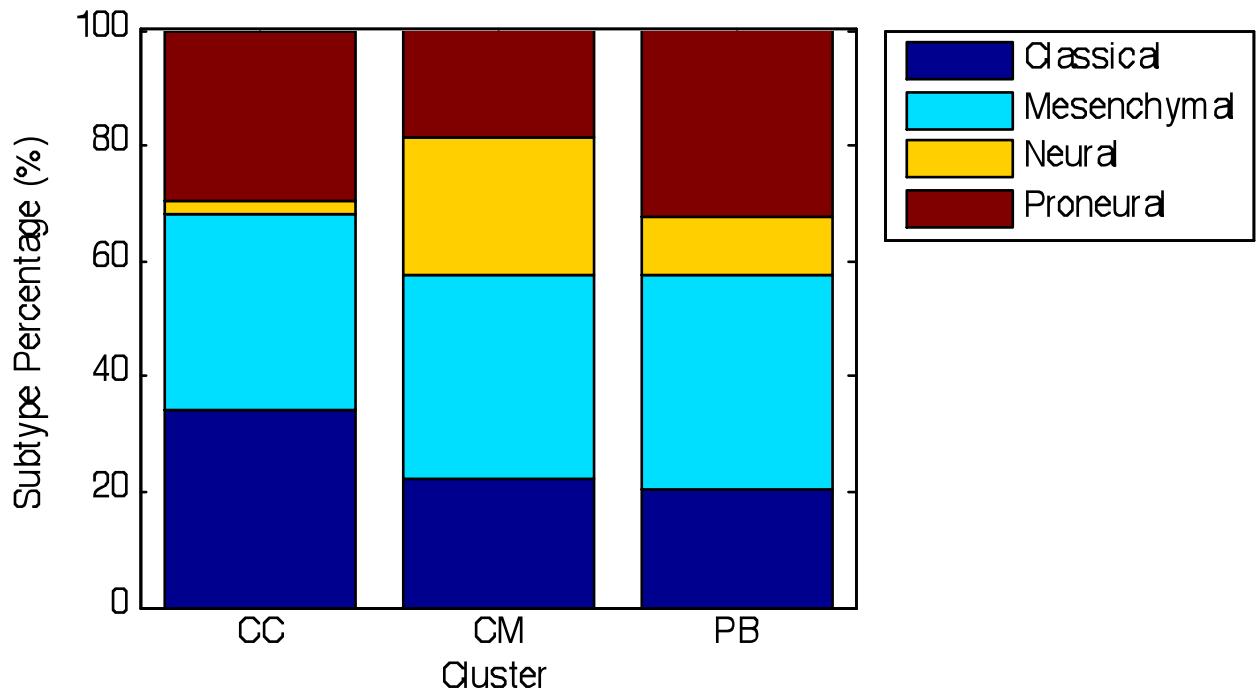
Figure 2 Glioblastoma (GBM) clusters, survival, and relationship to molecular subtypes. (A) Means-based analysis of GBM morphology reveals three patient clusters. (B) Survival differences between these clusters are statistically significant. CC, cell cycle; CM, chromatin modification; PB, protein biosynthesis.





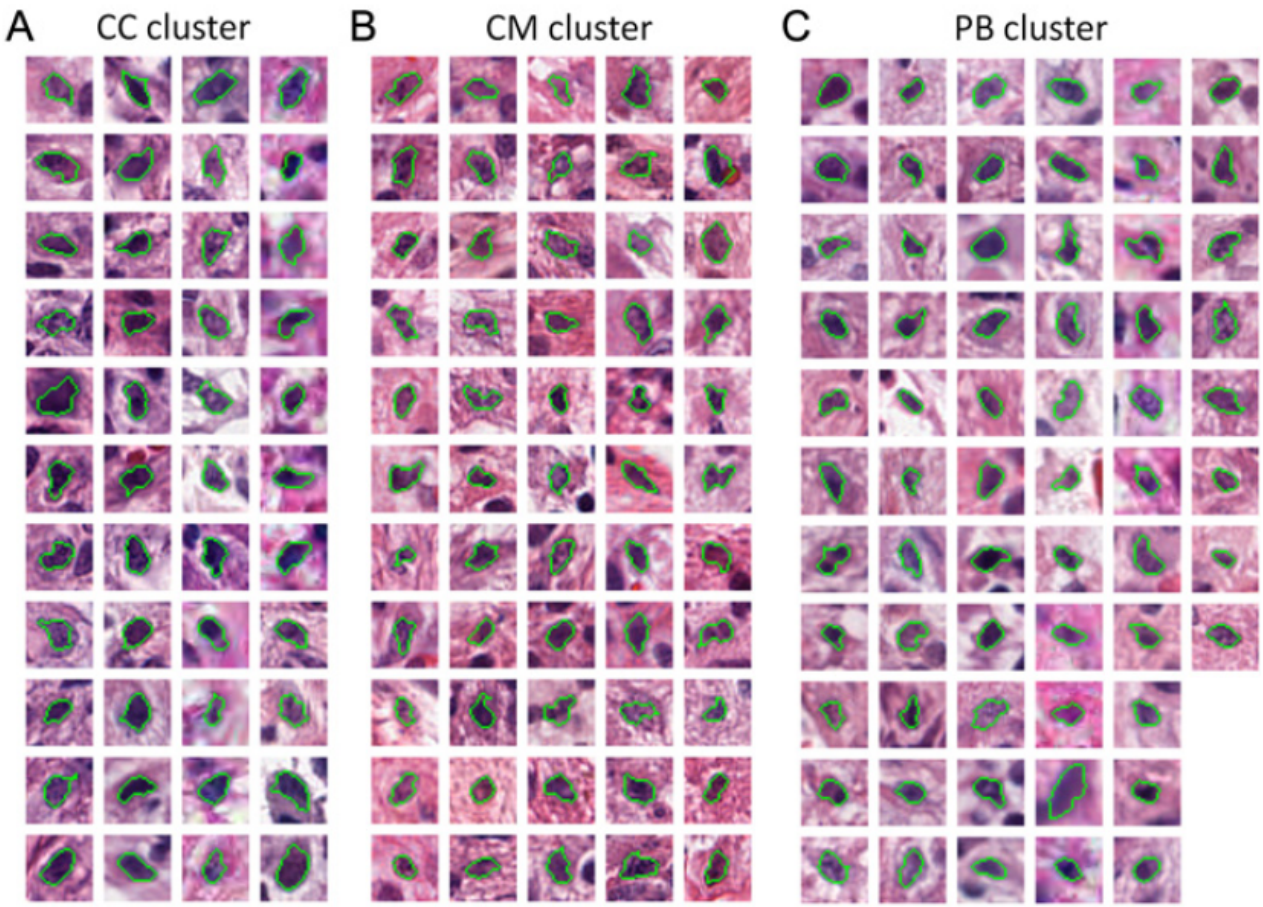
Gene Expression Class Associations

- Cox proportional hazards
  - Gene expression class not significant p=0.58
  - Morphology clustering p=5.0e-3



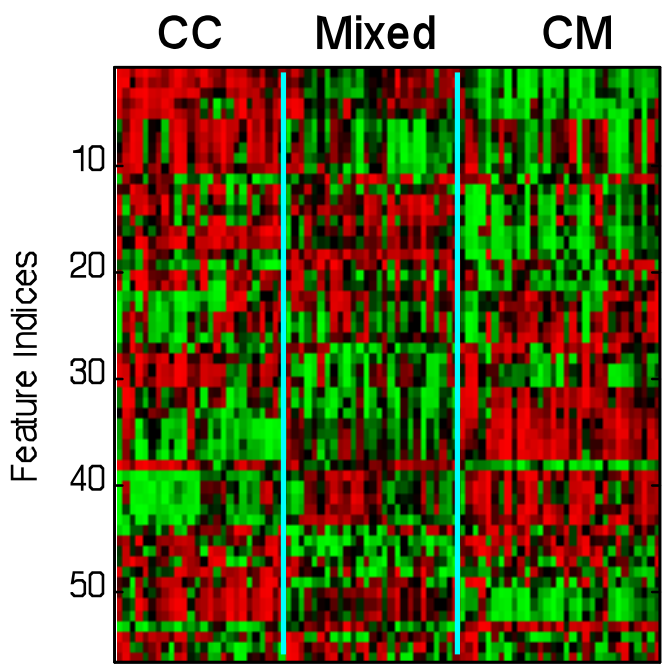
Representative nuclei

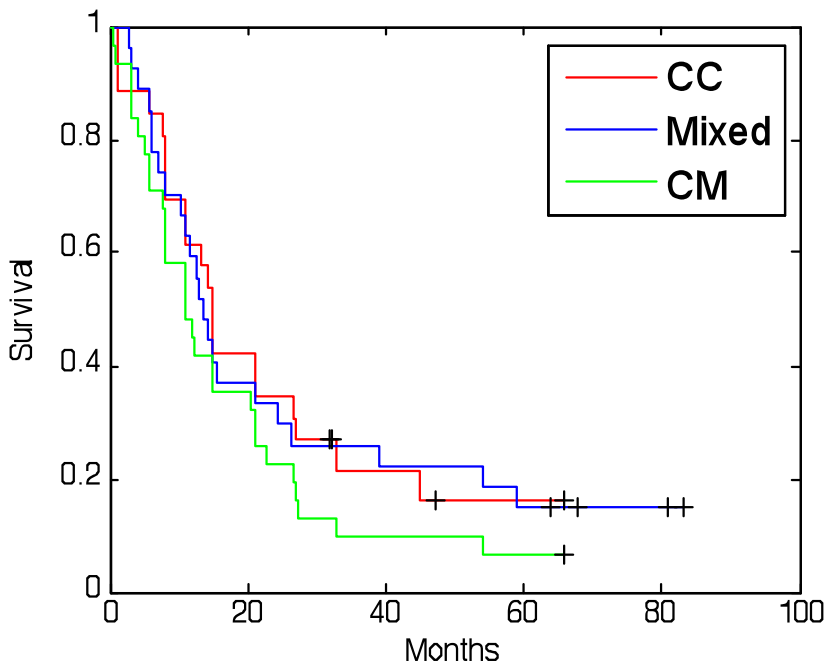
Figure 3 Signature nuclei for: (A) cell cycle (CC), (B) chromatin modification (CM), and (C) protein biosynthesis (PB) clusters. Cluster morphology is visualized by selecting the most representative nucleus from each patient. The selection is defined by the cell with the shortest distance in feature space to the patient's morphology signature.



Clustering Validation

- Separate set of 84 GBMs from Henry Ford Hospital
- ClusterRepro: CC p=7.2e-3, CM p=1.3e-2





Associations between genetics, genomics, epigenetics and imaging class

Table 1 Survival, genomic, and pathology correlates of morphology clusters

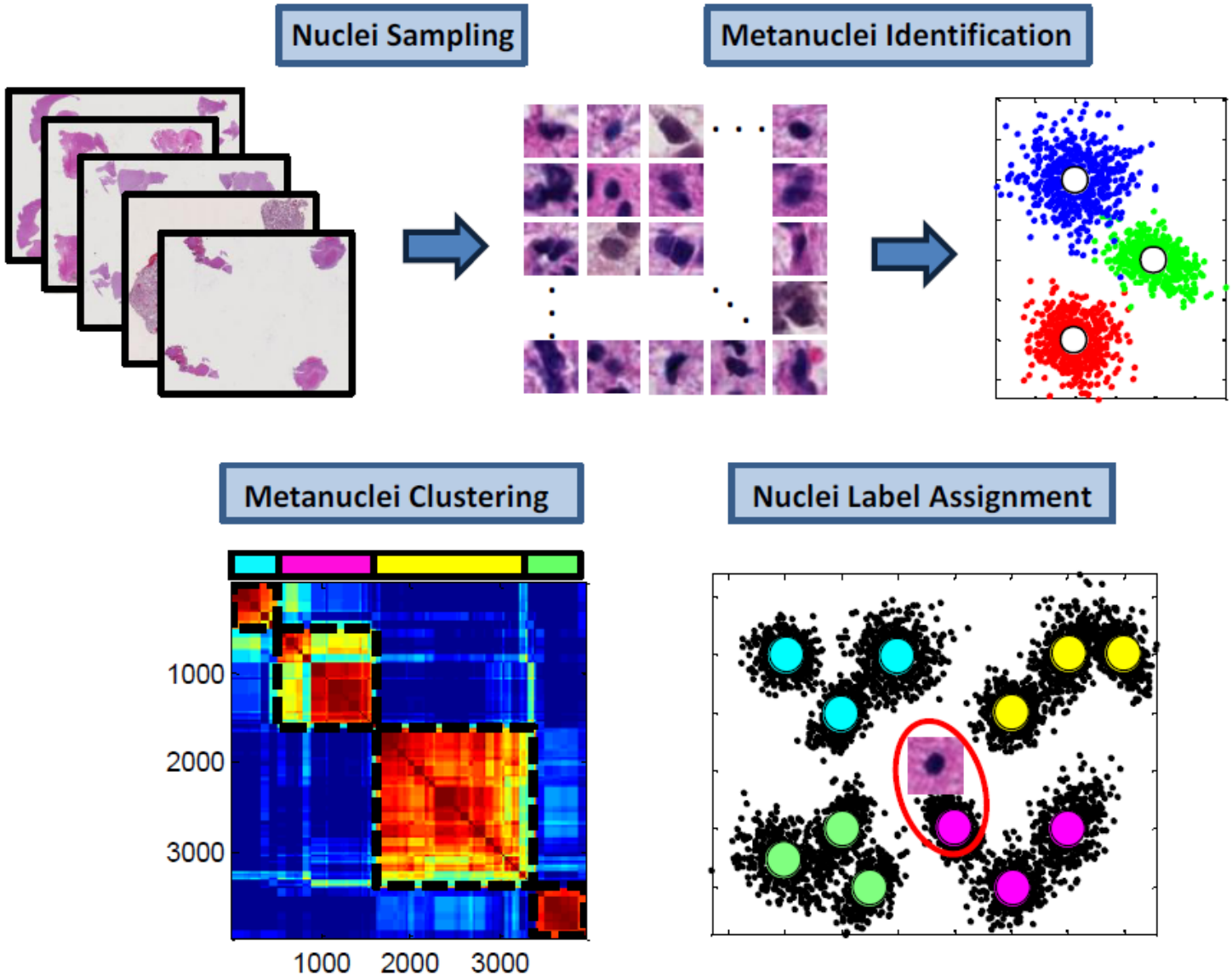
	CC cluster	CM cluster	PB cluster
Prognosis	Average	Poor	Better
Subtype associations	Neural depleted	Neural enriched Proneural depleted	GCIMP depleted
Pathology	Small cells enriched	Lymphocytes enriched	Inflammation depleted
Genetics	*NF1* mutant depleted *TP53* mutant depleted	None	None
Pathways	*TP53* signaling *Wnt* signaling *β-catenin*	*Wnt* signaling *β-catenin*	*NF-κB* signaling *ATM* checkpoint
Differential Expression	2740/663 genes up/down 97/100 miRNAs up/down	200/463 genes up/down 121/81 miRNAs up/down	0/188 genes up/down 15/5 miRNAs up/down
Differential methylation	69 Genes hypermethylated	244 Genes hypermethylated	45 Genes hypomethylated
Copy number	1068 deletions 38 amplifications	301 deletions 5 amplifications	399 deletions 7 amplifications
Expression mapping	23 mapped to methylated sites 595 mapped to CNV sites	8 mapped to methylated sites 27 mapped to CNV sites	1 mapped to methylated sites 19 mapped to CNV sites

CC, cell cycle; CM, chromatin modification; CNV, copy number variation; GCIMP, CpG island methylator phenotype; PB, protein biosynthesis.

Requirements for High Resolution 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*

High Resolution Methods

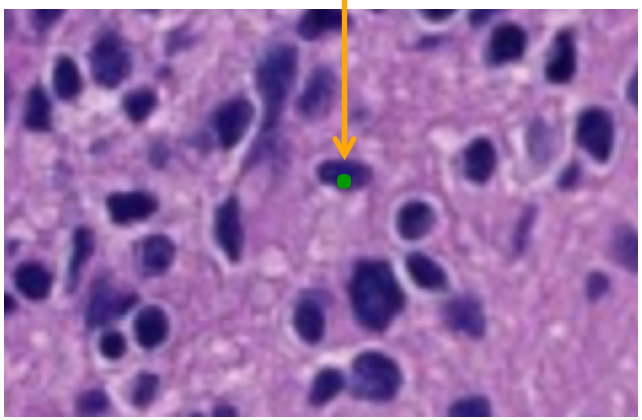


Large Scale Data Management

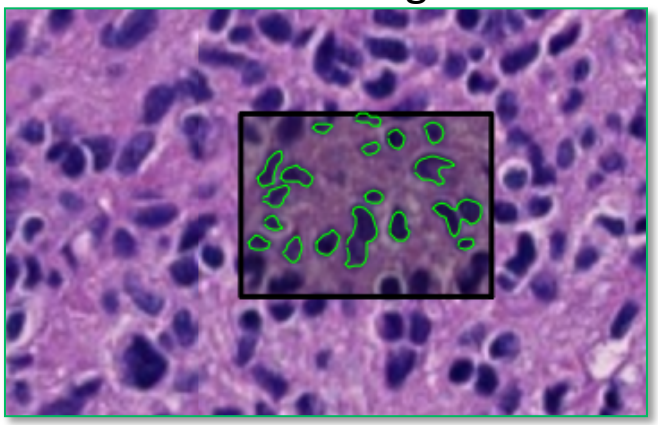
- Implemented with IBM DB2 for large scale pathology image metadata (~million markups per slide)
- Represented by a complex data model capturing multi-faceted information including markups, annotations, algorithm provenance, specimen, etc.
- Support for complex relationships and spatial query: multi-level granularities, relationships between markups and annotations, spatial and nested relationships

Spatial Centric – Pathology Imaging "GIS"

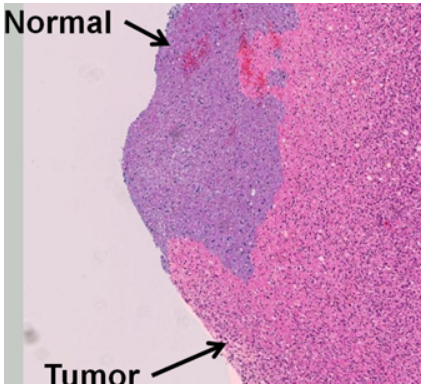
Point query: human marked point inside a nucleus



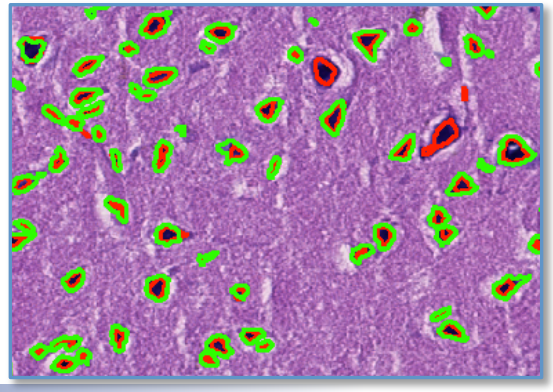
Window query: return markups contained in a rectangle

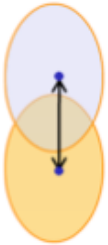


Containment query: nuclear feature aggregation in tumor regions

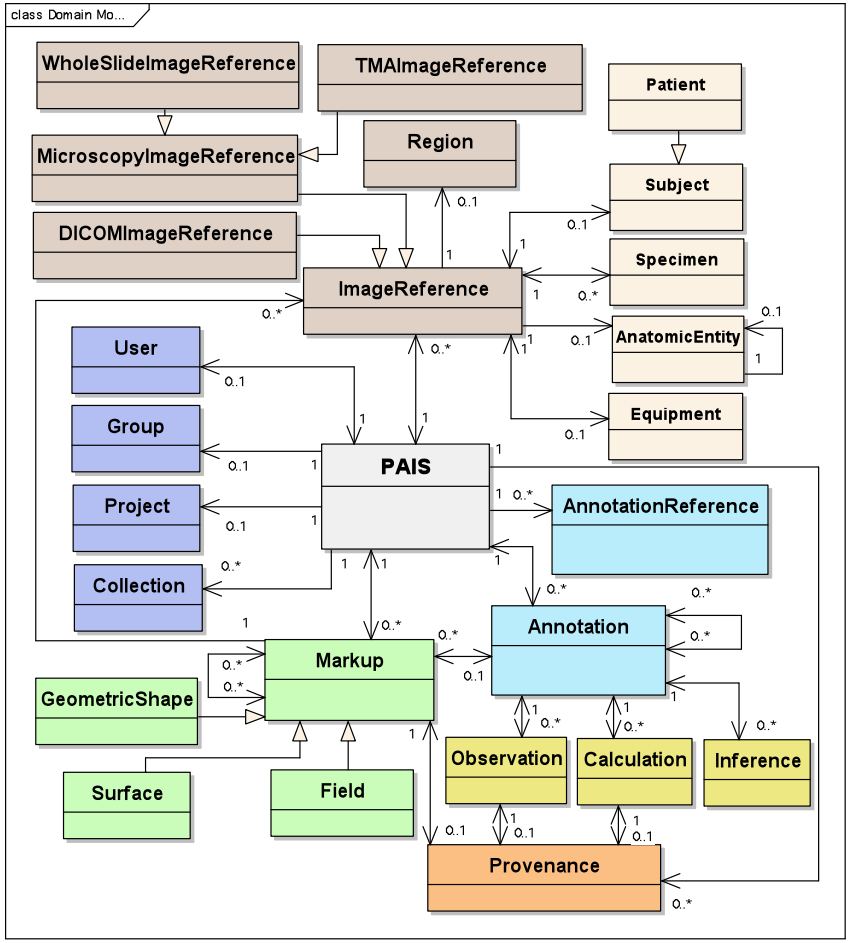


Spatial join query: algorithm validation/comparison



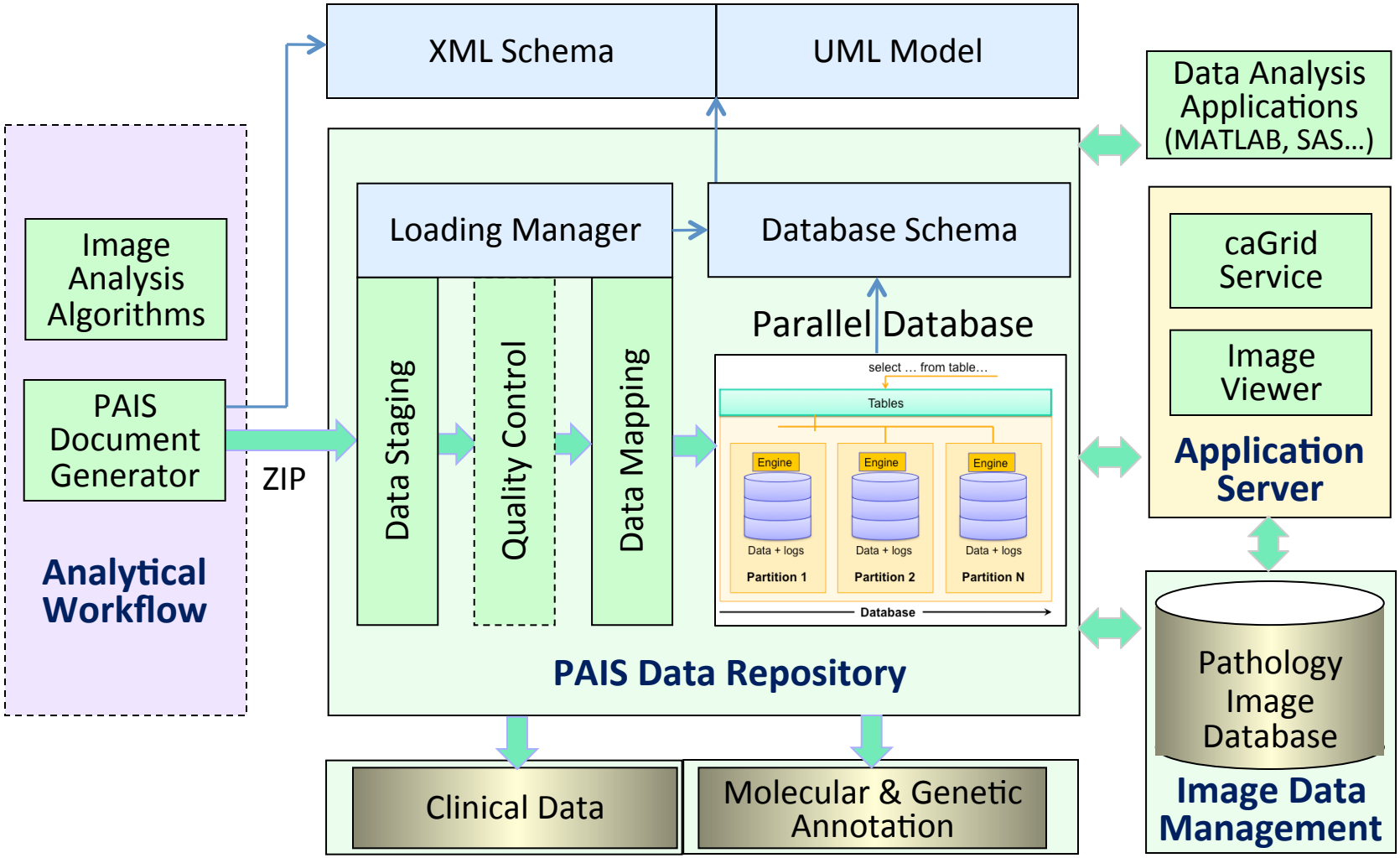


PAIS (Pathology Analytical Imaging Standards) Supported by caBIG, R01 and ACTSI



- PAIS Logical Model
  - 62 UML classes
  - markups, annotations, imageReferences, provenance
- PAIS Data Representation
  - XML (compressed) or HDF5
- PAIS Databases
  - loading, managing and querying and sharing data
  - Native XML DBMS or RDBMS + SDBMS

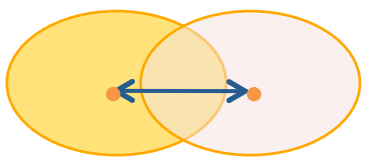
PAIS Relational Data Management Architecture

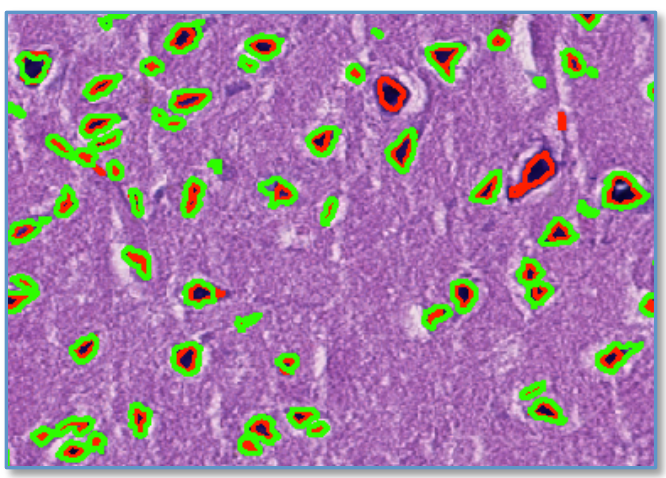


Algorithm Validation: Intersection between Two Result Sets (Spatial Join)

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;
```



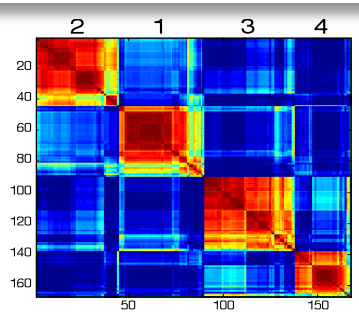


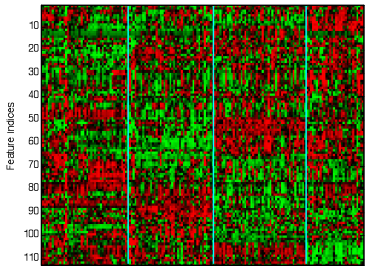
PAIS_UID	TILE	MKPID	RATIO	DISTANCE
astroII.1_20x_20x_NS-MORPH_1	astroII.1.ndpi-0000004096-0000004096	10,422,160,945,100,002	0.8750	0.50
astroII.1_20x_20x_NS-MORPH_1	astroII.1.ndpi-0000004096-0000004096	10,422,160,945,100,003	0.8000	0.50
astroII.1_20x_20x_NS-MORPH_1	astroII.1.ndpi-0000004096-0000004096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

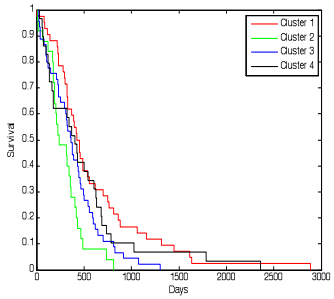
Example Query for Integrative Studies

- Find mean nuclear feature vector and covariance on tumor regions for each patient grouped by tumor subtype

```
SELECT c.pais_uid, pc.subtype,
   AVG(area), AVG(perimeter), AVG(eccentricity),
   COVARIANCE(area, perimeter), COVARIANCE(area, eccentricity)
FROM pais.calculation_flat c,TCGA.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;
```

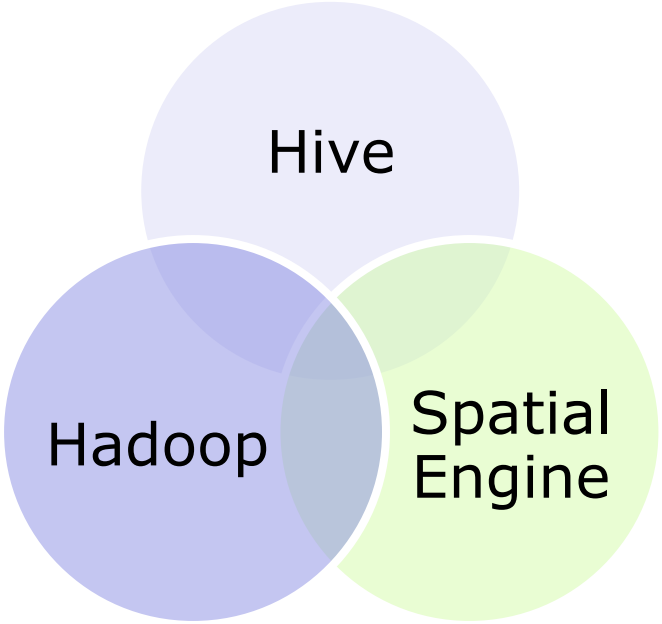




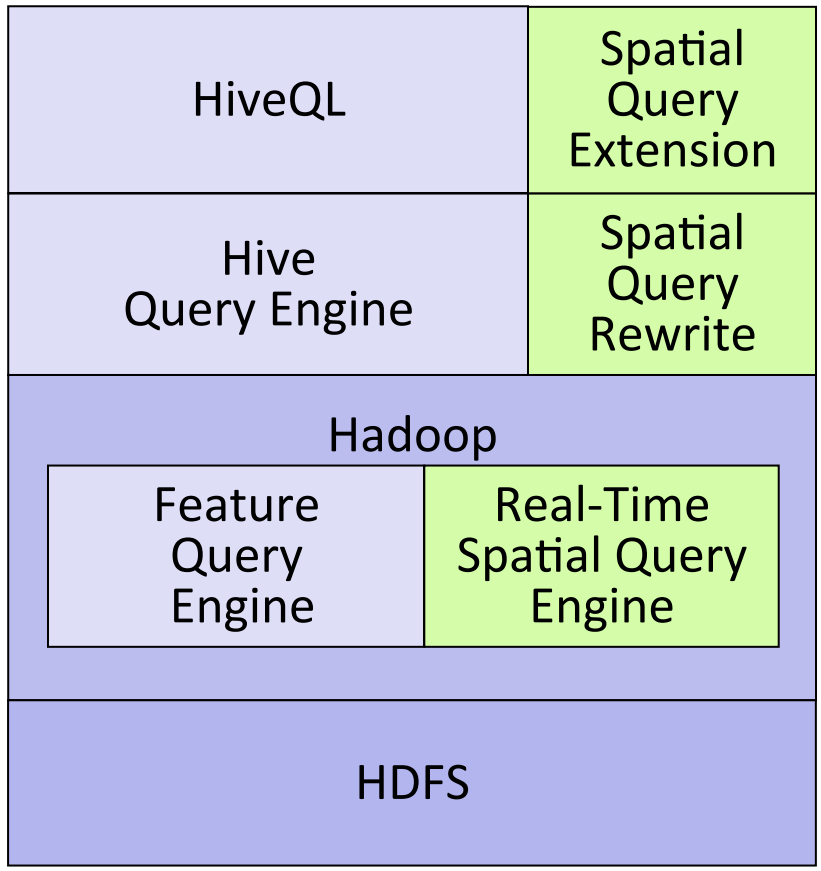


Ongoing: Hadoop-GIS

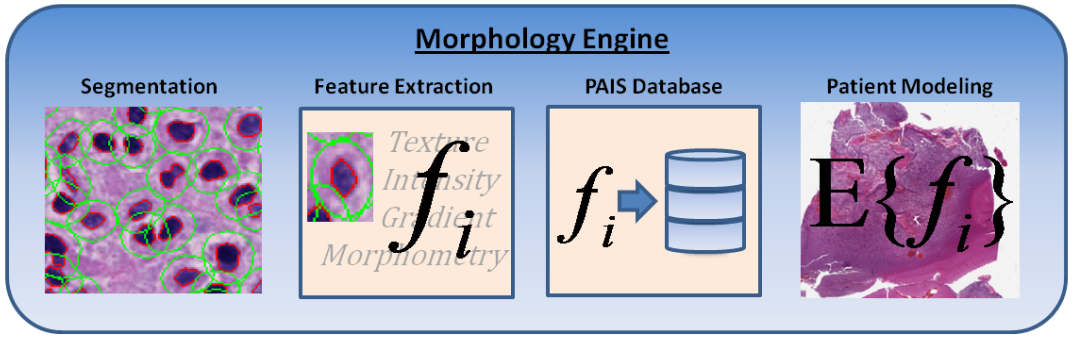
- Hadoop – scalability
- Hive – expressiveness + scalability
- Real-time spatial query engine – spatial/GIS queries

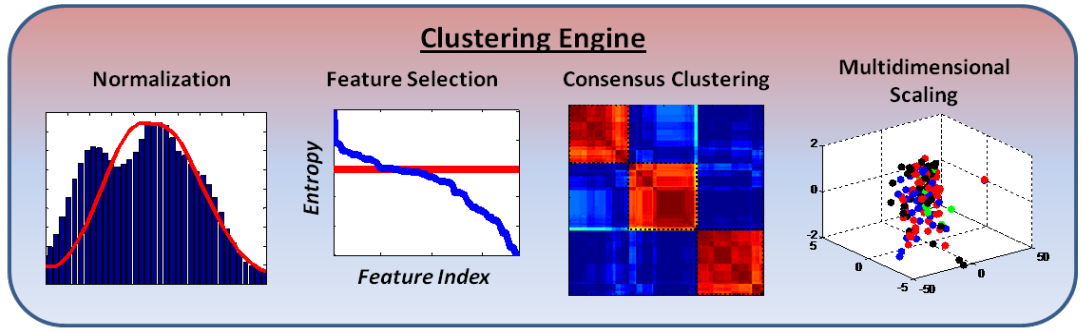


Hadoop-GIS



Multicore/GPU Implementation of Morphology and Clustering Pipelines:



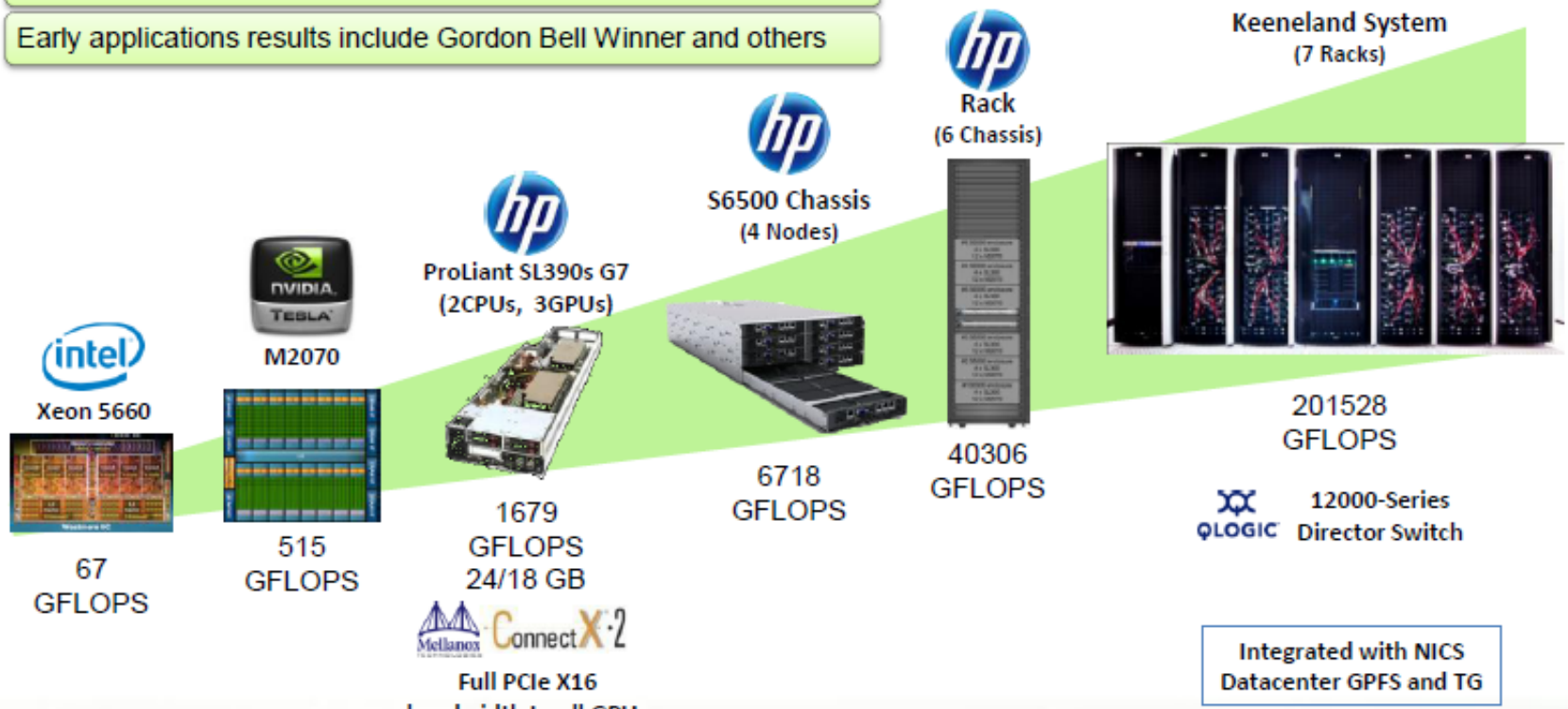


- Nuclear/cell segmentation
- Feature extraction
- Generate clusters of nuclear/cell Features
- Classify nuclei/cells by feature cluster
- Classify patients by populations of classified nuclei/cells

EMORY UNIVERSITY

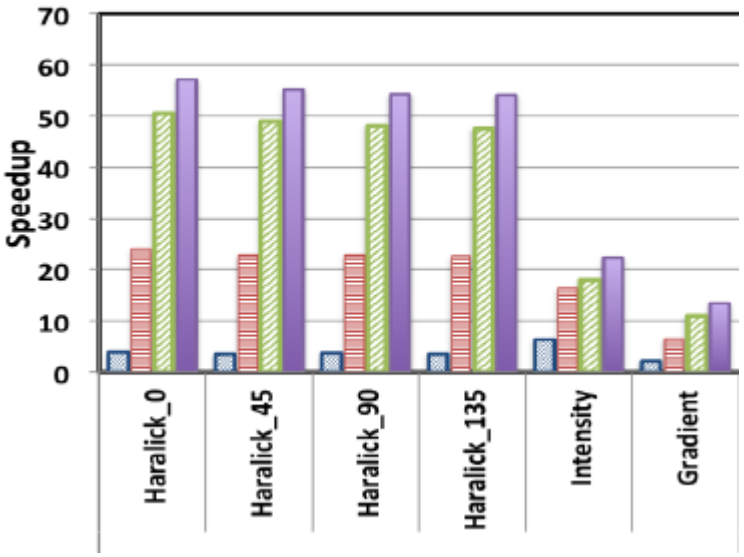
Keeneland – Initial Delivery System

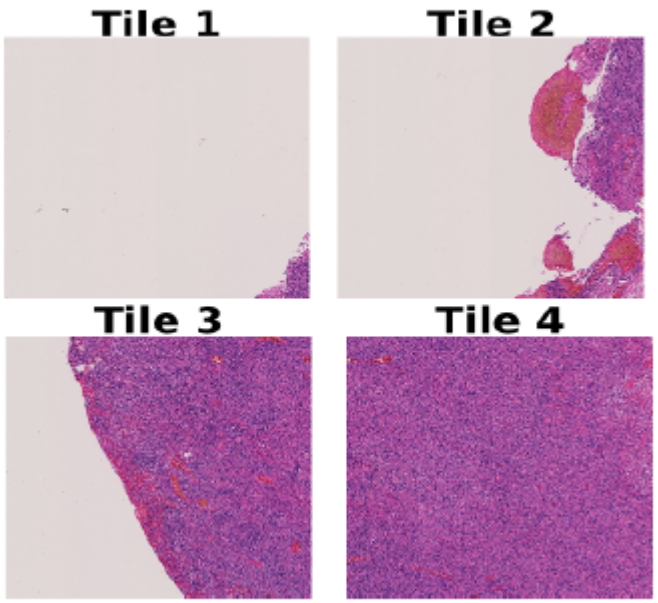
- **Initial Delivery** system installed in Oct/Nov 2010
- 201 TFLOPS in 7 racks (90 sq ft incl service area)
- 677 MFLOPS per watt on HPL (#9 on Green500, Nov 2010)
- Early applications results include Gordon Bell Winner and others



Georgia Tech NICS THE UNIVERSITY of TENNESSEE OAK RIDGE National Laboratory NVIDIA

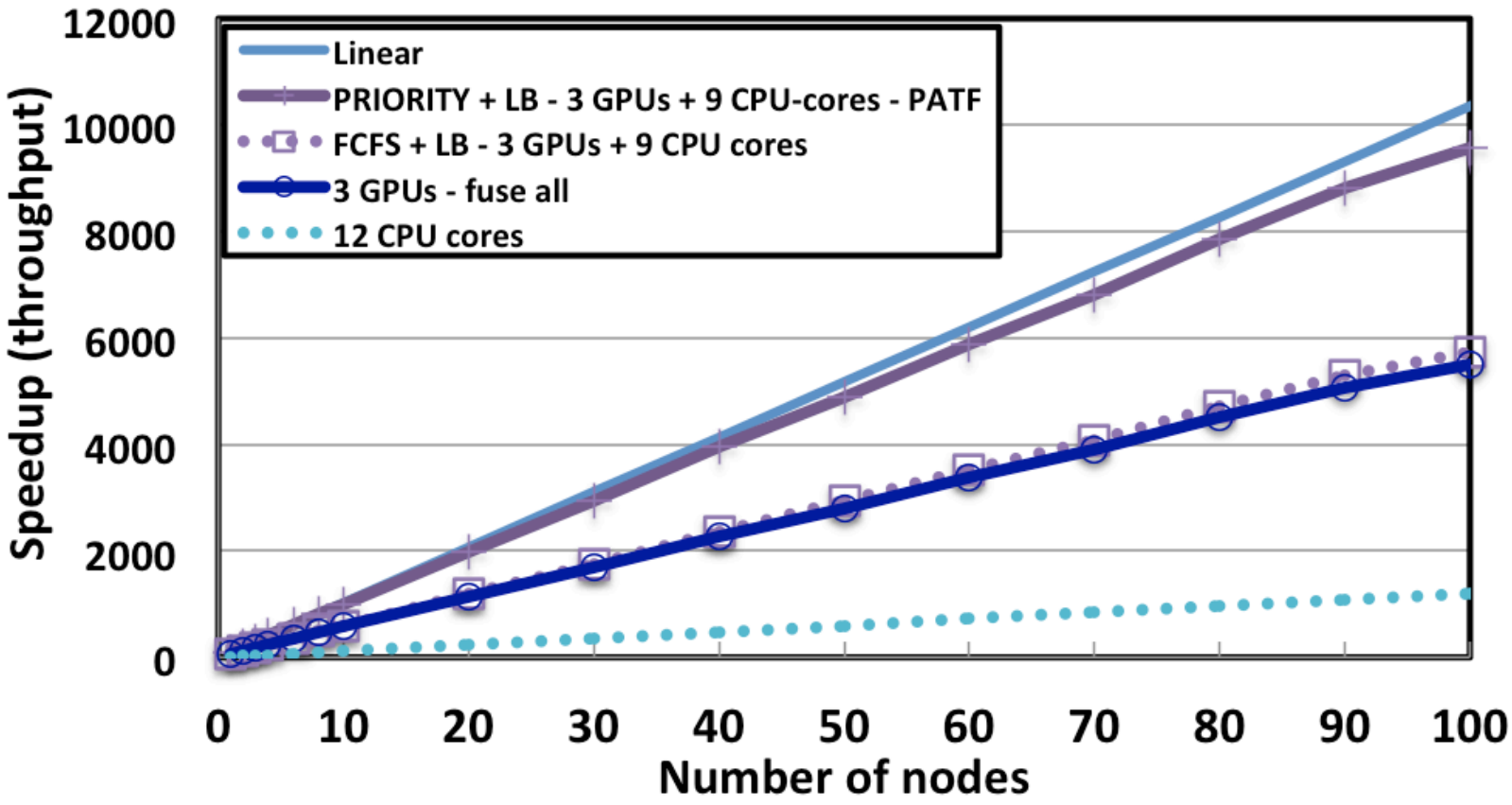
Optimizing CPU/GPU Irregular Problem Performance



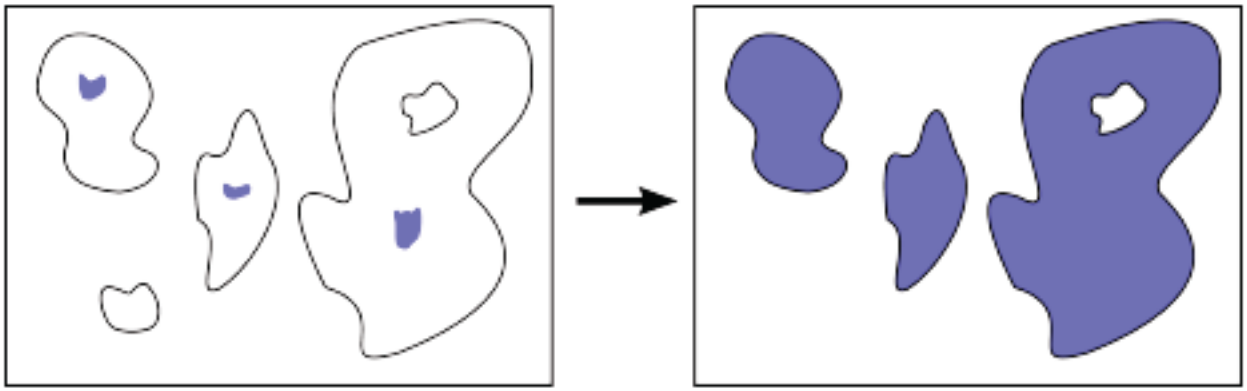


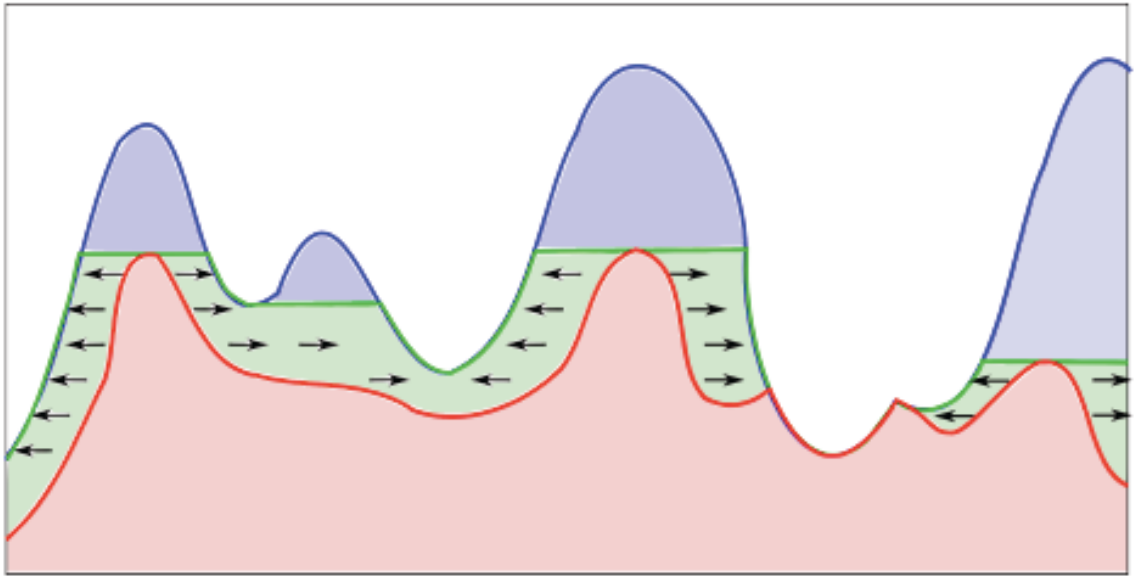
(a) Tiles with different tissue coverage.

Keeneland CPU/GPU Performance for Feature Extraction



Morphological Reconstruction

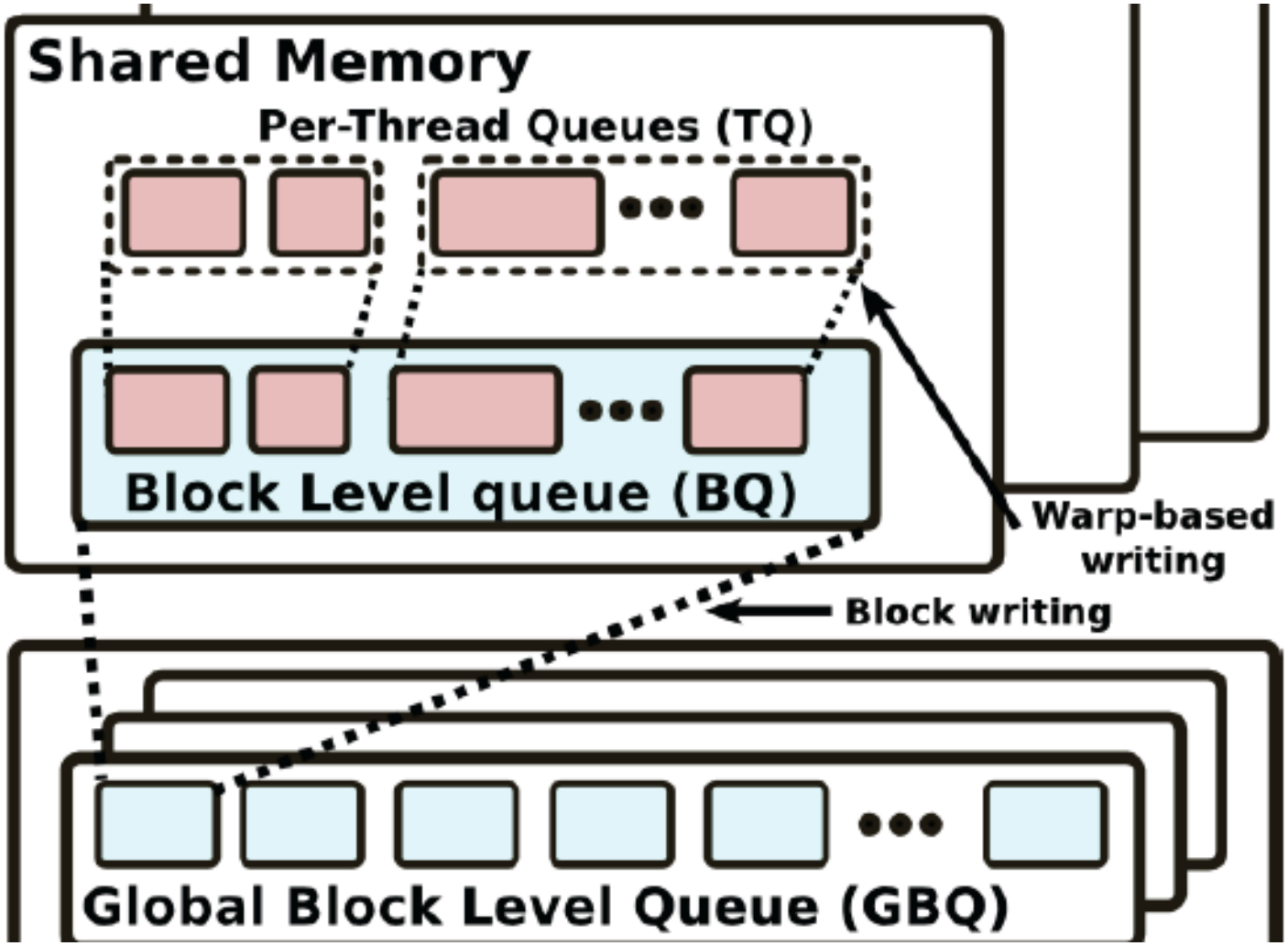


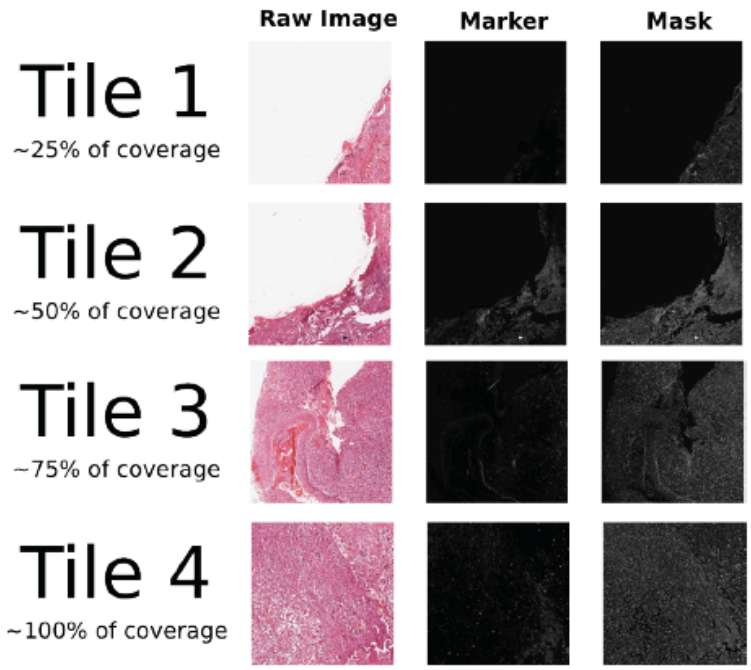


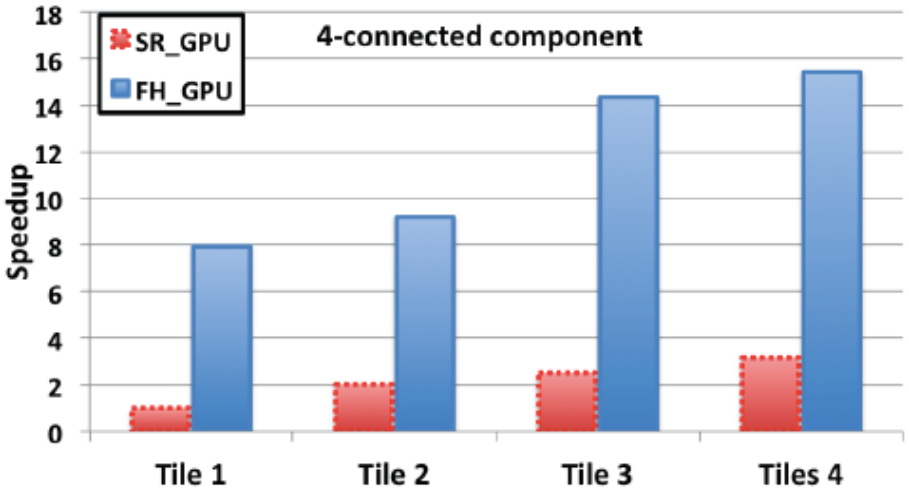
Algorithms

- **Sequential Reconstruction (SR)**: Pixel value propagation in the marker image is computed by alternating raster and anti-raster scans
- **Queue-based Reconstruction (QB)**: A first-in first-out (FIFO) queue is initialized with pixels in the regional maxima. Remove a pixel from the queue, scan the pixel's neighborhood, and queue the neighbor pixels that changed.
- **Fast Hybrid Reconstruction (FH)**:incorporates the characteristics of the SR and QB algorithms--first makes one pass using the raster and anti-raster orders, then continues the computation using a FIFO queue as in QB.

GPU Fast Hybrid Reconstruction using parallel queues: Concurrency, memory hierarchy at the cost of modest increase in operation count







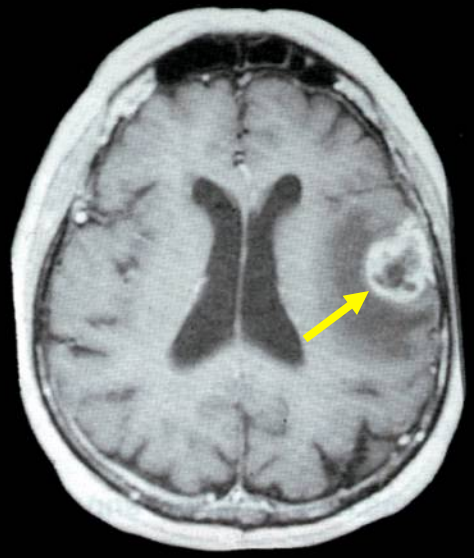
(a) 4-connected grid.

Multiscale Systems Biology

Employ multi-resolution methods to characterize necrosis, angiogenesis and correlate these with "omics"

No enhancement
Normal Vessels
Stable lesion

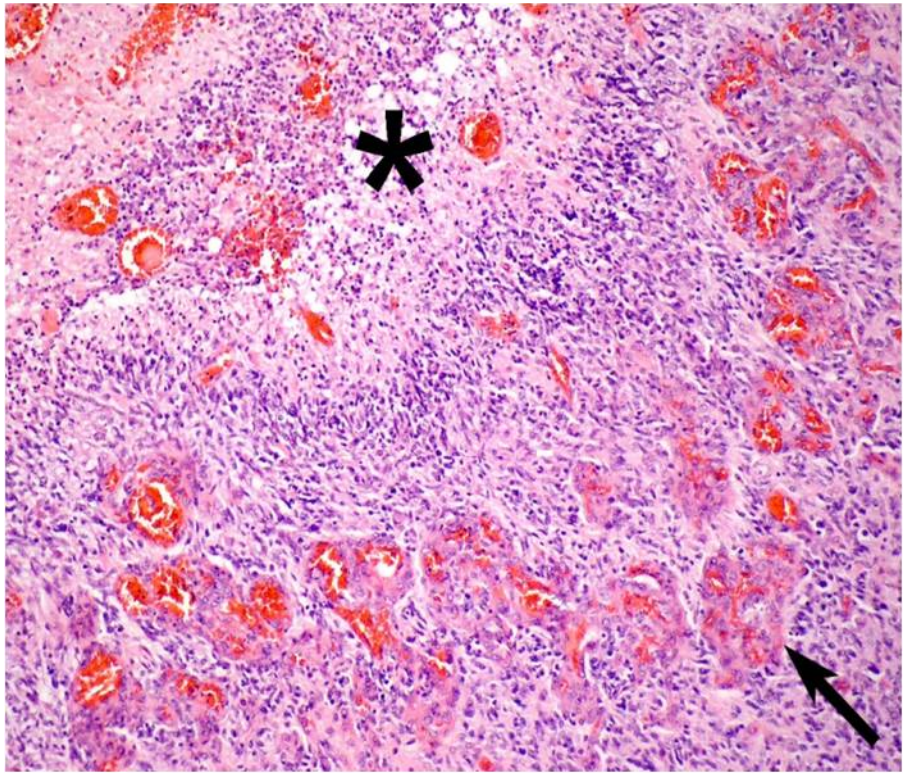
?



Rim-enhancement
Vascular Changes
Rapid progression

Correlation of Necrosis, Angiogenesis and "omics"

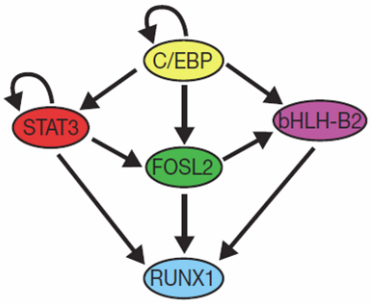
- GBMs display variable and regionally heterogeneous degrees of necrosis (asterisk) and angiogenesis
- These factors may impact gene expression profiles



Genes Correlated with Necrosis include Transcription Factors Identified as Regulators of the Mesenchymal Transition

- Frozen sections from 88 GBM samples marked to identify regions of necrosis and angiogenesis
- Extent of both necrosis and angiogenesis calculated as a percentage of total tissue area

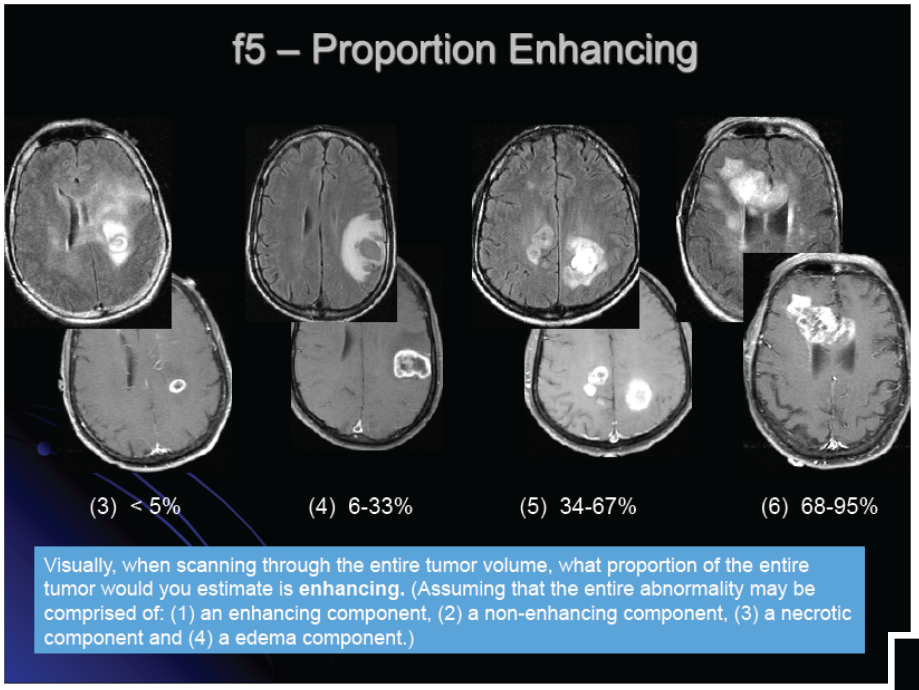
Gene Symbol	SAM q-value (Corrected p-value)
C/EBPB	< 0.000001
C/EBPD	< 0.000001
FOSL2	< 0.000001
STAT3	0.0047
RUNX1	0.0082



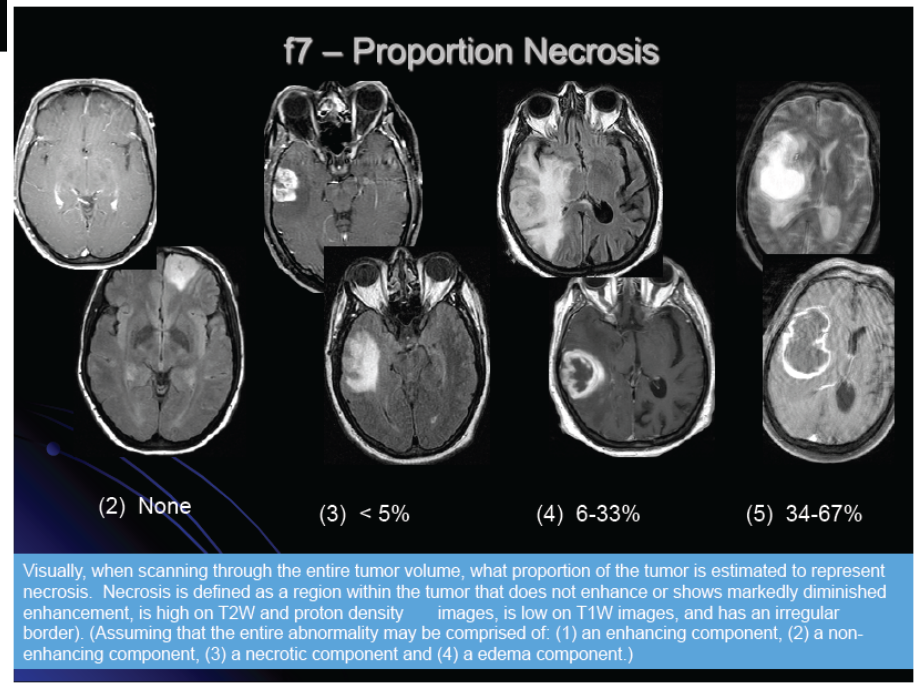
Carro MS, et al. Nature 263: 318-25, 2010

Center for Comprehensive Informatics

EMORY UNIVERSITY



VASARI Feature Set



Chad Holder
Scott Hwang
Adam Flanders
David Gutman

Defining Rich Set of Qualitative and Quantitative Image Biomarkers

- Community-driven ontology development project; collaboration with ASNR
- Imaging features (5 categories)
  - **Location** of lesion
  - **Morphology of lesion margin** (definition, thickness, enhancement, diffusion)
  - **Morphology of lesion substance** (enhancement, PS characteristics, focality/multicentricity, necrosis, cysts, midline invasion, cortical involvement, T1/FLAIR ratio)
  - **Alterations in vicinity of lesion** (edema, edema crossing midline, hemorrhage, pial invasion, ependymal invasion, satellites, deep WM invasion, calvarial remodeling)
  - **Resection features** (extent of nCE tissue, CE tissue, resected components)

Center for Comprehensive Informatics

EMORY UNIVERSITY

Imaging Predictors of survival and molecular profiles in the TCGA Glioblastoma Data set

The TCGA glioma working group

[1]Emory University Hospital, Atlanta, GA [2]National Cancer Institute, Bethesda, MD. [3]Thomas Jefferson University Hospital, Philadelphia, PA. [4]Henry Ford University Hospital, Detroit, Michigan. 5National Institute of Health, Bethesda, MD. [6]Boston University School of Medicine, Boston, MA. [7]SAIC-Frederick, Inc., Frederick, MD. [8]University of Virginia, Charlottesville, VA. [9] Northwestern University Chicago, IL

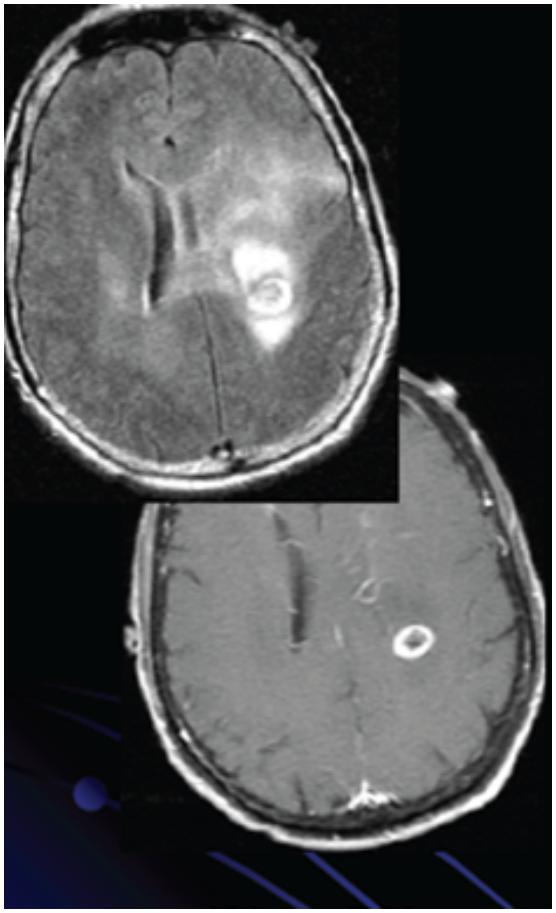
Emory	TJU/CBIT/NCI	UVA/Northwestern	Henry Ford
David A Gutman[1]	Adam Flanders[3]	Max Wintermark[8]	Lisa Scarpace[4]
Lee Cooper[1]	Eric Huang[2]	Manal Jilwan[8]	Tom Mikkelsen[4]
Scott N Hwang[1]	Robert J Clifford[2]	Prashant Raghavan[8]	
Chad A Holder[1]	Dina Hammoud[3]	Pat Mongkolwat[9]	
Doris Gao[1]	John Freymann[7]		
Carlos Moreno[1]	Justin Kirby[7]		
Arun Krishnan[1]	Carl Jaffe[6]		
Jun Kong[1]			
Seena Dehkharghani[1]			
Joel Saltz[1]			
Dan Brat[1]			

Correlative Imaging Results Link Radiology, Pathology, "omics"

Minimal enhancing tumor (≤5%) strongly associated with Proneural classification (p=0.0006).

>5% proportion of necrosis and the presence of microvascular hyperplasia in pathology slides (p=0.008).

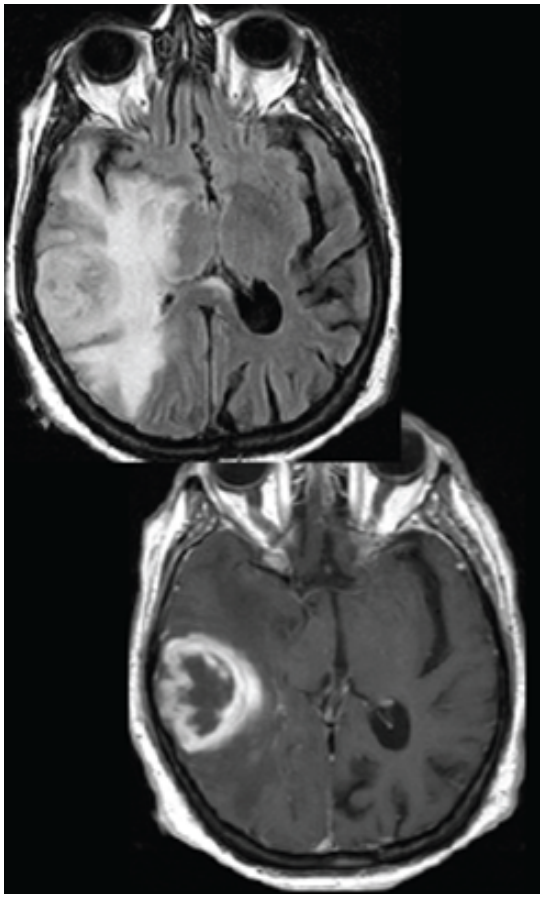
Greater maximum tumor dimension (T2 signal) associated with present/ abundant microvascular hyperplasia (p=0.001).



< 5% Enhancement

Correlative Imaging Results

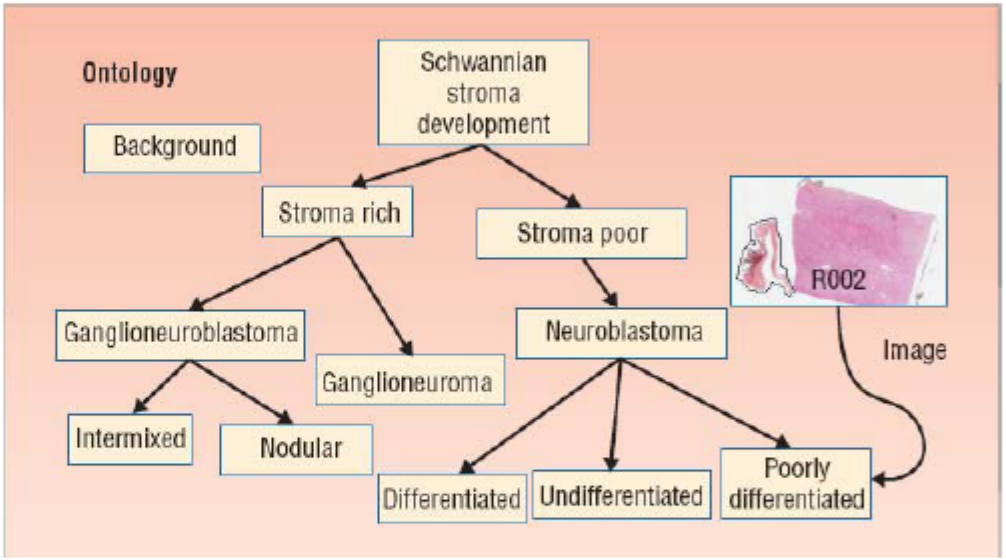
- *TP53* mutant tumors had a smaller mean tumor sizes (p=0.002) on T2-weighted or FLAIR images.
- *EGFR* mutant tumors were significantly larger than *TP53* mutant tumors (p=0.0005).
- High level *EGFR* amplification was associated with >5% enhancement and >5% proportion of necrosis (p < 0.01).



> 5% Necrosis

Same Methods Used to Directly Emulate Pathologists and Provide Decision Support

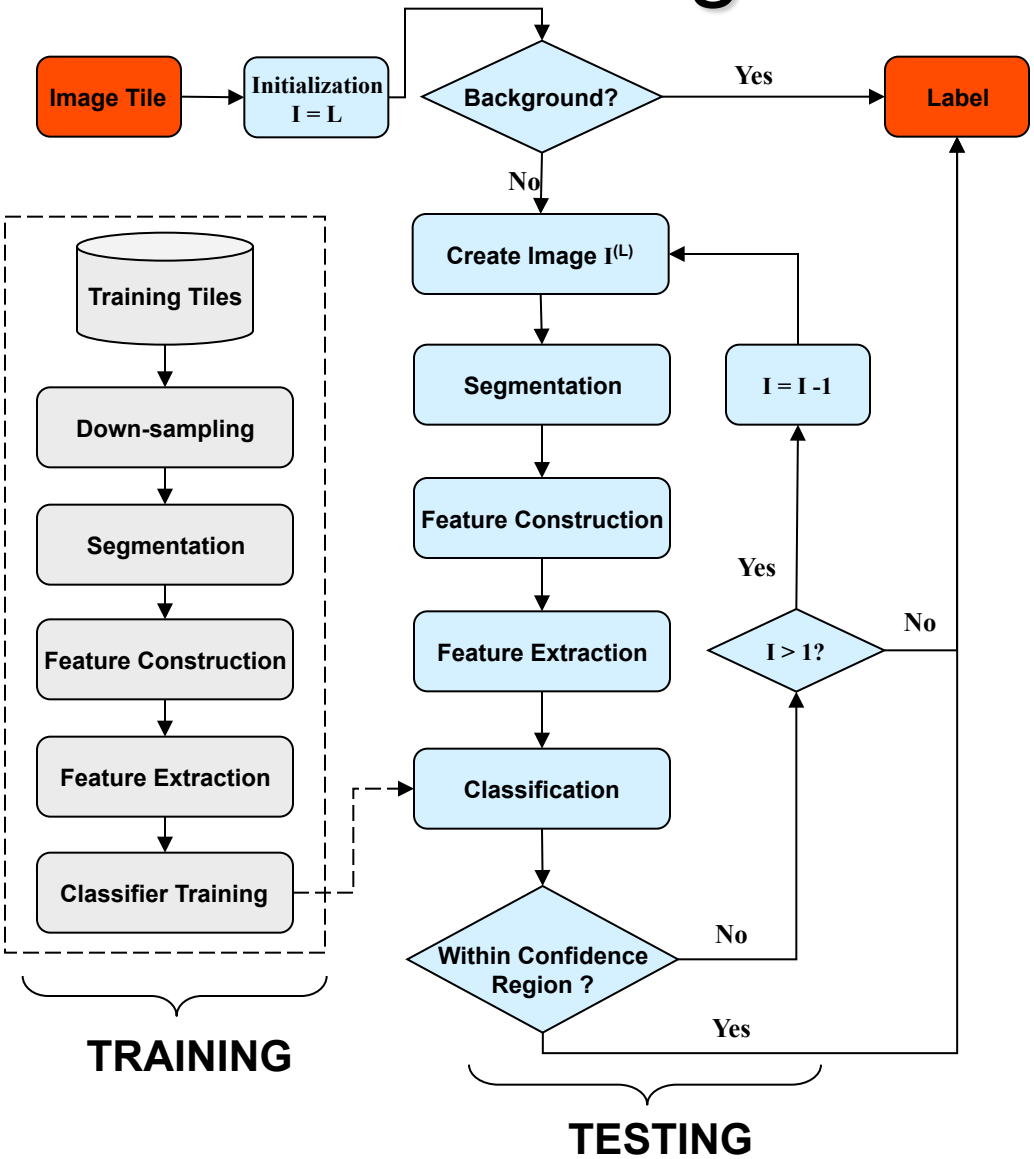
- Analyze images by computer
- Analyze the whole tissue, several slides
- Provide quantitative information to the pathologist
- Reduce inter- and intra-reader variability



Morphological characterization of tissue used for prognosis

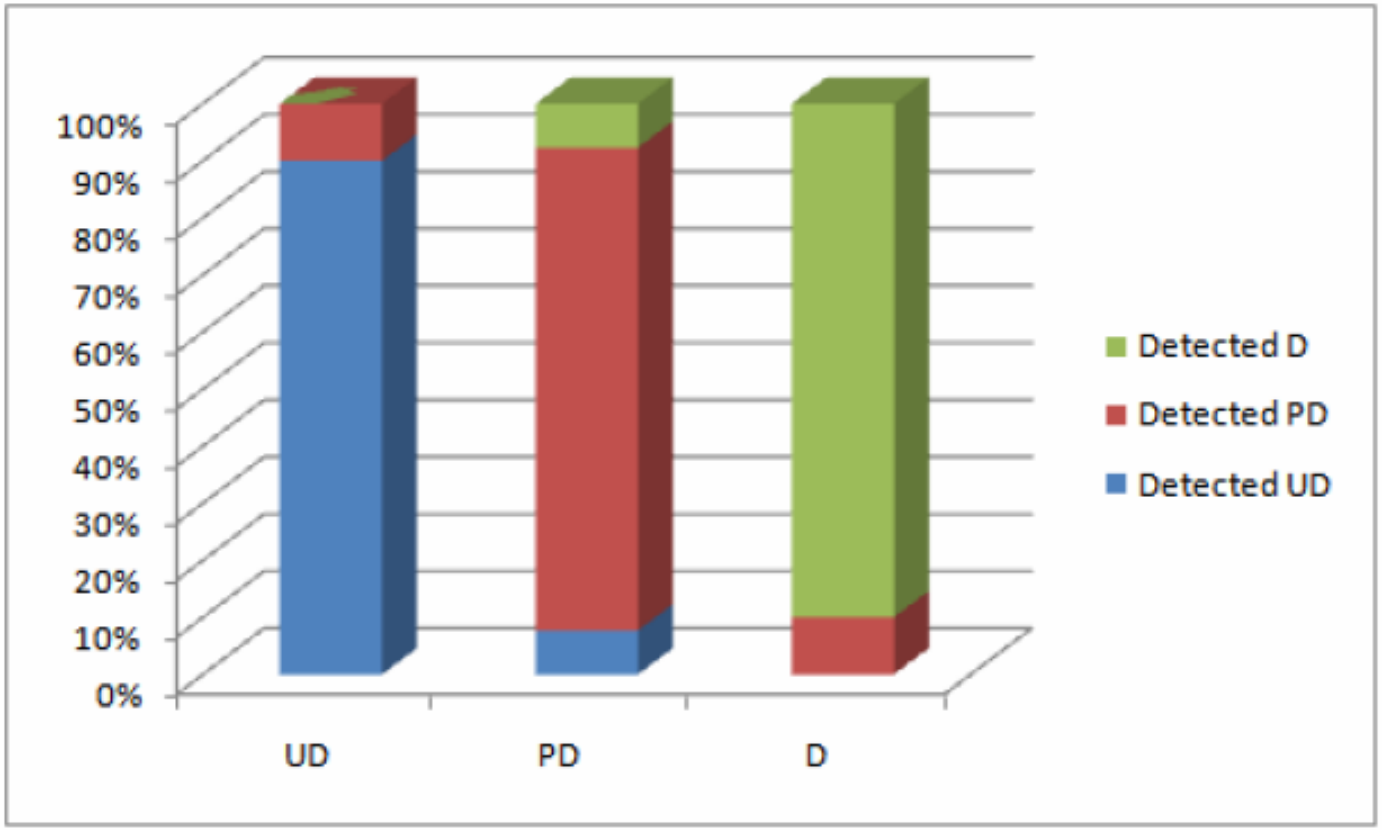
Neuroblastoma – Shimada Classification (Gurcan-OSU, Shimada – LA Children's)

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)

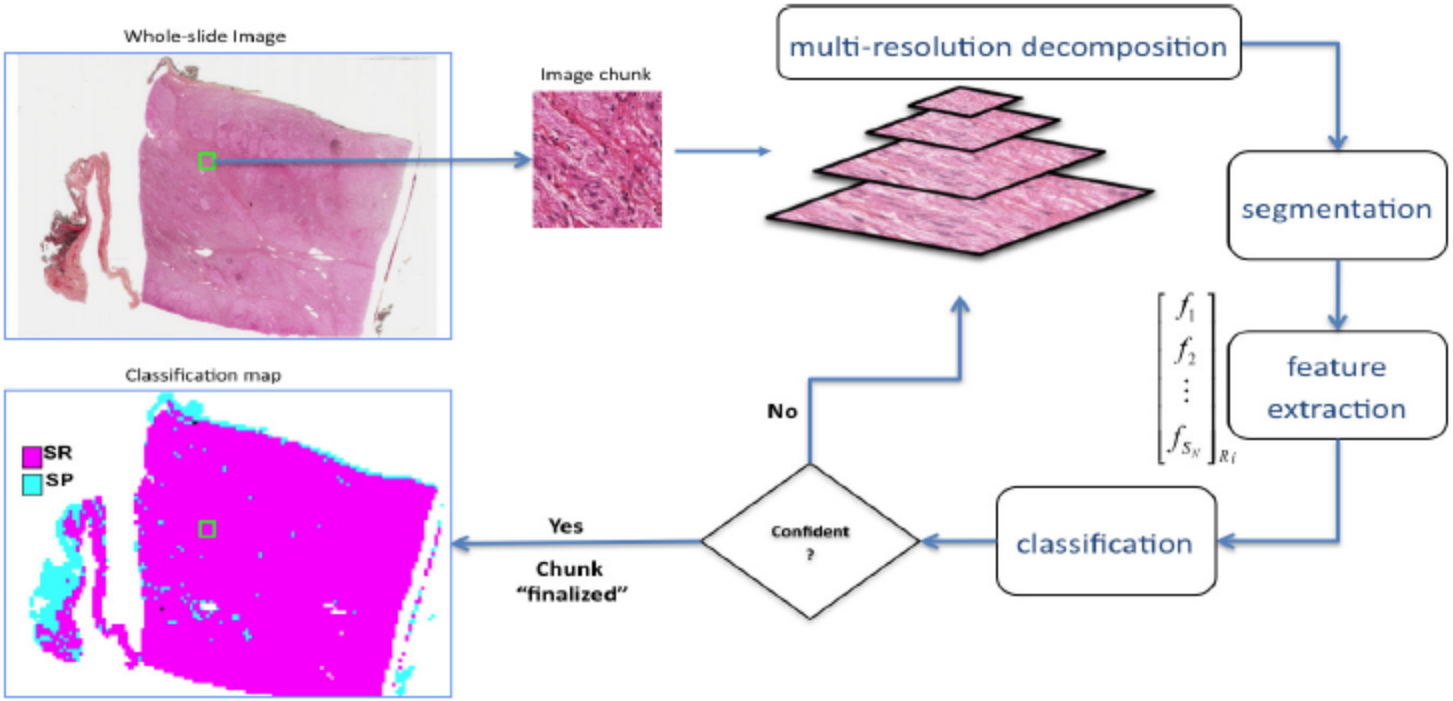
Example Algorithm Results: Neuroblastoma Grade of Differentiation



UD: Undifferentiating
PD: Partially differentiating
D: Differentiating

Provide best classification under time constraint

Leverage semantic workflow management system (WINGS)



Center for Comprehensive Informatics

Fram

An Integrated Framework for Parameter-based Optimization of Scientific Workflows*

Vijay S. Kumar, P. Sadayappan
Department of Computer Science and Engineering
Ohio State University
Columbus, Ohio 43210
{kumarvi,saday}@cse.ohio-state.edu

Gaurang Mehta, Karan Vahi, Ewa Deelman
Advanced Systems Division
Information Sciences Institute
University of Southern California
Marina del Rey, California 90292
{gmehta,vahi,deelman}@isi.edu

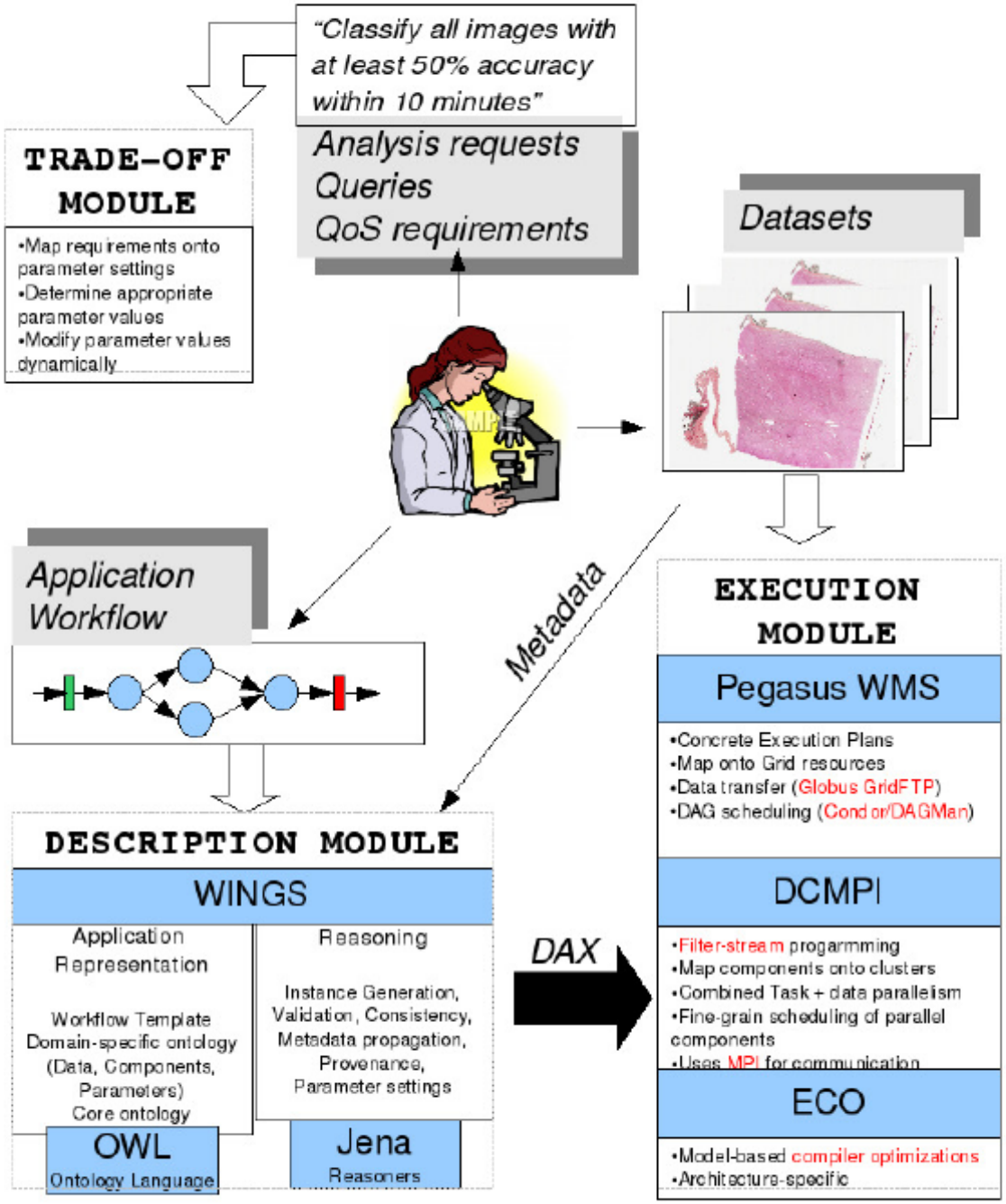
Varun Ratnakar, Jihie Kim, Yolanda Gil
Intelligent Systems Division
Information Sciences Institute
University of Southern California
Marina del Rey, California 90292
{varunr,jihie,gil}@isi.edu

Mary Hall
School of Computing
University of Utah
Salt Lake City, Utah 84112
mhall@cs.utah.edu

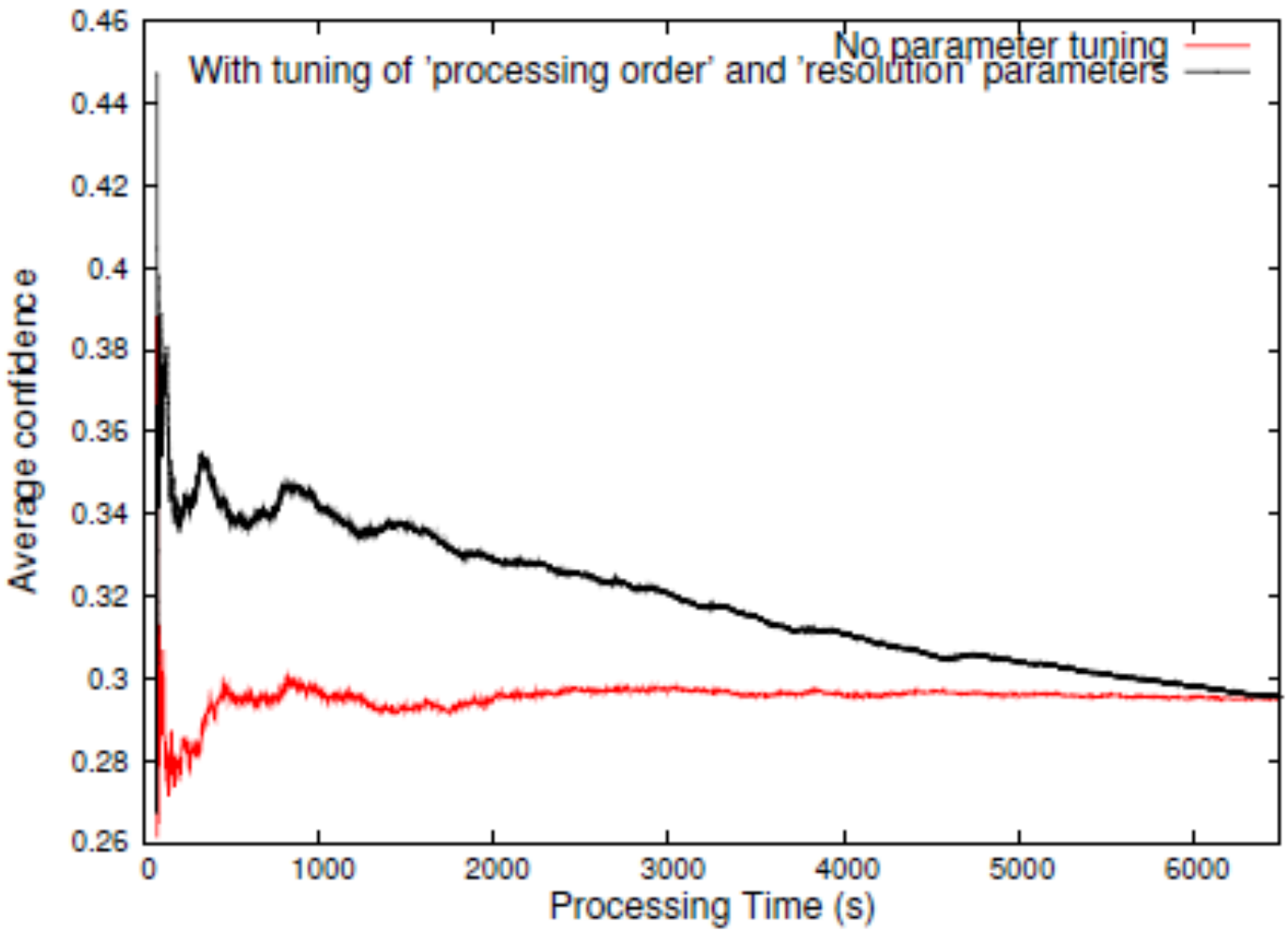
Tahsin Kurc, Joel Saltz
Center for Comprehensive Informatics
Emory University
Atlanta, Georgia 30307
{tkurc,jhsaltz}@emory.edu

- **Description Module** (Wings): Describe application workflow using semantics of workflow components
- **Execution module** (Pegasus, DataCutter, Condor): Maps to resources, generates and places fine grained filter/stream pipelines
- **Tradeoff Module**: Schedules execution based on application level QOS

Center for Comprehensive Informatics



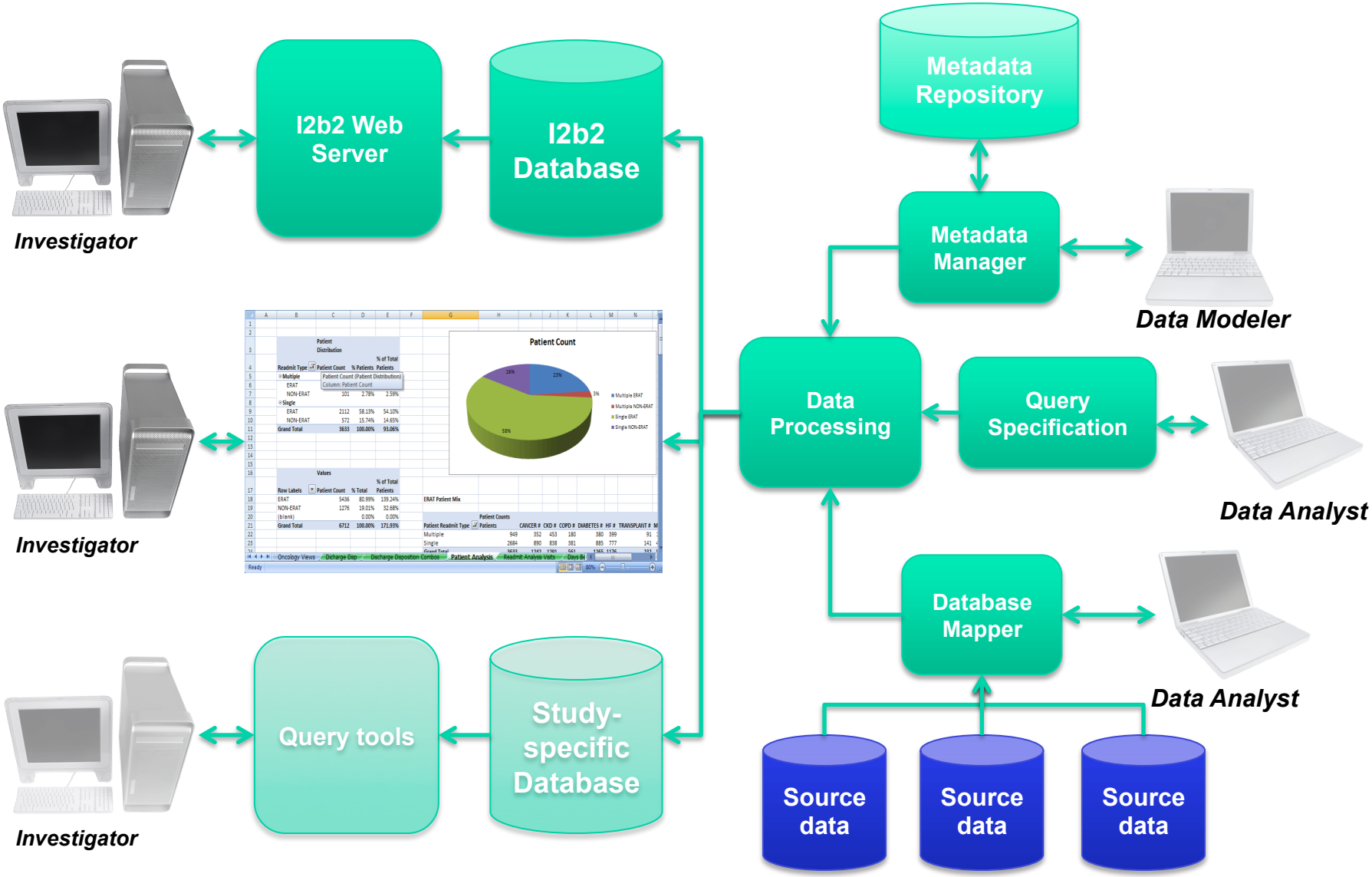
Impact


No parameter tuning
With tuning of 'processing order' and 'resolution' parameters
Average confidence
Processing Time (s)
0.46
0.44
0.42
0.4
0.38
0.36
0.34
0.32
0.3
0.28
0.26
0
1000
2000
3000
4000
5000
6000

Predictive Modeling and Clinical Phenotype Characterization

- 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?
- Need a repeatable process that we can apply identically to both local and UHC data

Overall System



5-year Datasets from Emory and University Healthcare Consortium

- EUH, EUHM and WW (inpatient encounters)
- Removed encounter pairs with chemotherapy and radiation therapy ***readmit*** encounters (CDW data)

- Encounter location (down to unit for Emory)
- Providers (Emory only)
- Discharge disposition
- Primary and secondary ICD9 codes
- Procedure codes
- DRGs
- Medication orders (Emory only)
- Labs (Emory only)
- Vitals (Emory only)
- Geographic information (CDW only + US Census and American Community Survey)

Analytic Information Warehouse

Using Emory & UHC Data to Find Associations With 30-day Readmits

- Problem: "Raw" clinical and administrative variables are difficult to use for associative data mining
  - Too many diagnosis codes, procedure codes
  - Continuous variables (e.g., labs) require interpretation
  - Temporal relationships between variables are implicit
- Solution: Transform the data into a much smaller set of variables using heuristic knowledge
  - Categorize diagnosis and procedure codes using code hierarchies
  - Classify continuous variables using standard interpretations (e.g., high, normal, low)
  - Identify temporal patterns (e.g., frequency, duration, sequence)
  - Apply standard data mining techniques

Derived Variables

- 30-day readmit
- The 9 Emory Enhanced Risk Assessment Tool diagnosis categories
- UHC product lines
- Variables derived from a combination of codes and/or laboratory test results
  - Obesity
  - Diabetes/uncontrolled diabetes
  - End-stage renal disease (ESRD)
  - Pressure ulcer
  - Sickle cell disease/sickle cell crisis
- Temporal variables derived over multiple encounters
  - Multiple MI
  - Multiple 30-day readmissions
  - Chemotherapy within 180 (or 365) days before surgery
  - Previous encounter within the last 90 (or 180) days

30-Day Readmission Rates for Derived Variables

Emory Health Care

Patient Population	Number of Encounters	Number of Readmissions	Readmission Rate
All-Emory	202181	36734	15%
Multiple MI	4414	1506	36% (Single MI 15%)
ESRD	18445	5036	27% (CKD 23%)
>=4 readmissions	19510	10707	55%
Multiple MI *and* >= 4 readmissions	997	520	52%
CKD *and* >=4 readmissions	7865	4110	52%
Uncontrolled diabetes	12219	2573	21% (Diabetes 19%)
Uncontrolled diabetes & pressure ulcer	648	201	31%
Uncontrolled diabetes & ESRD	1645	531	32%
Sickle cell crisis	1809	663	37% (Sickle cell anemia 34%)
MRSA	1565	410	26%
Stroke and MRSA	42	16	38% (Stroke 24%)
MI and MRSA	140	43	31% (MI 15%)

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

Emory Readmission Rates for High and Low Risk Groups Generated with Random Forest

ERAT Category	30-Day Readmission Rate (%)		
	Baseline	Predictive Modeling High Risk Group	Predictive Modeling Low Risk Group
Cancer	18.46	41.59	4.45
CKD	22.03	47.32	7.04
COPD	17.48	45.52	7.02
Diabetes	16.87	42.47	8.72
Transplant	22.77	41.86	11.51
MI	13.76	40.41	2.97
Pulmonary Hypertension	21.06	47.83	3.72
Sickle Cell Anemia	34.03	46.36	19.42
Stroke	12.28	48.40	0.88

Status of Clinical Phenotype Characterization

- Integrative dataset analysis can leverage patient information gathered over many encounters
- Temporal analyses can generate derived variables that appear to correlate with readmissions
- Predictive modeling has promise of providing decision support
- Ongoing analyses involve characterization of clinical phenotype in GWAS, biomarker and quality improvement efforts
- Co-lead (with Bill Hersh) of CTSA CER Informatics taskforce dedicated to this issue

Summary and Perspective

- Large scale integrative data analytic methods and tools to integrate clinical, molecular, Pathology, Radiology data
- Characterize new cancer subtypes and biomarkers, predict outcome, treatment response
- Generate and manage nuanced temporal summary of patients health status, co-morbidities, treatment, treatment response
- Importance: generate basic insights into pathophysiology, clues to new treatments, better ways of evaluating existing treatments and core infrastructure needed for comparative effectiveness research studies

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**

The AIW Team

- Stakeholders
  - Joel Saltz, MD, PhD – CCI Director and ACTSI BIP Director
  - William Bornstein, MD, PhD – Emory Healthcare Chief Quality Officer
  - Dee Cantrell, RN – Emory Healthcare CIO
  - Marc Overcash – Emory Deputy CIO of Research and Health Sciences IT
- Project Team
  - Andrew Post, MD, PhD – AIW Project Lead & CCI Clinical Informatics Architect
  - Terry Willey, RN – IS Director of Business Strategy/Planning
  - Richie Willard – Project Manager
  - Tahsin Kurc, PhD – CCI Chief Software Architect
  - Sharath Cholleti, PhD – Research Scientist
  - Jingjing Gao, PhD – Biostatistician
  - Michel Mansour – Software Engineer
  - Himanshu Rathod – Software Engineer
  - Mike Torian – Data Warehouse Engineer
  - Michael Brown – Software Engineer
  - Geoff Milton – Software Engineer
  - Akshatha Kalsanka Pai – Software Engineer

Thanks!