Stony Brook University

SUNY

Machine Learning and Deep Contemplation of Data

Joel Saltz

Department of Biomedical Informatics

Stony Brook University

CCDSC

October 5, 2016

Stony Brook University

SUNY

From BDEC: "Domain": Spatio-temporal Sensor Integration, Analysis, Classification

- 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
- Solid printing integrative data analyses
- Data generated by numerical simulation codes – PDEs, particle methods

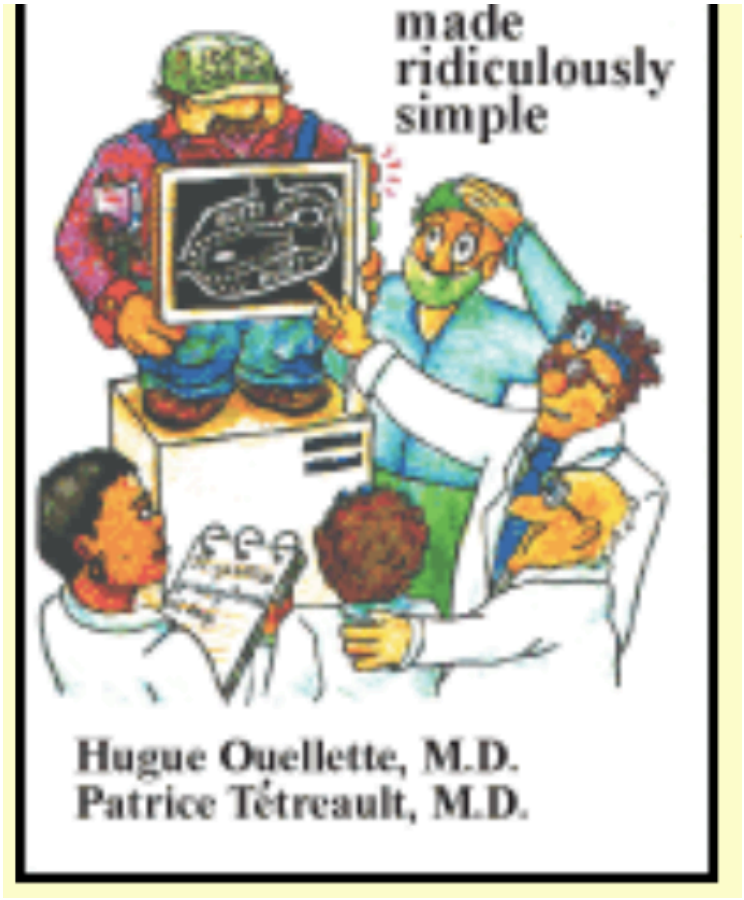
Things that Need to be Done with Spatio Temporal Data

- Generation of Features
- Sanity Checking and Data Cleaning
- Qualitative Exploration
- Descriptive Statistics
- Classification
- Identification of Interesting Phenomena
- Prediction
- Control
- Save Data for Later (Compression)

Precision Medicine Meta Application

- Predict treatment outcome, select, monitor treatments
- Reduce inter-observer variability in diagnosis
- Computer assisted exploration of new classification schemes
- Multi-scale cancer simulations

Multi-Scale Precision Medicine



Imaging and Precision Medicine - Pathomics, Radiomics

Identify and segment trillions of objects – nuclei, glands, ducts, nodules, tumor niches … from Pathology, Radiology imaging datasets

Extract features from objects and spatio-temporal regions

Support queries against ensembles of features extracted from multiple datasets

Statistical analyses and machine learning to link Radiology/Pathology features to "omics" and outcome biological phenomena

Principle based analyses to bridge spatio-temporal scales – linked Pathology, Radiology studies

Things that Need to be Done with Spatio Temporal Data

- Generation of Features
- Sanity Checking and Data Cleaning
- Qualitative Exploration
- Descriptive Statistics
- Classification
- Identification of Interesting Phenomena
- **Prediction**
- **Control**
- Save Data for Later (Compression)

Current Driving Applications

- Checkpoint Inhibitors – when to use, when to stop
- Pathology, Imaging data obtained prior to and during treatment
- Integration of “omics”, tissue and imaging to manage treatment
- Non Small Cell Lung Cancer, Melanoma, Brain
- Virtual Tissue Respository
- SEER Cancer Epidemiology
- 500K Cancer Patients per year
- DOE/NCI pilot involving text
- Our co-located companion Virtual Tissue Repository pilot targets SEER images

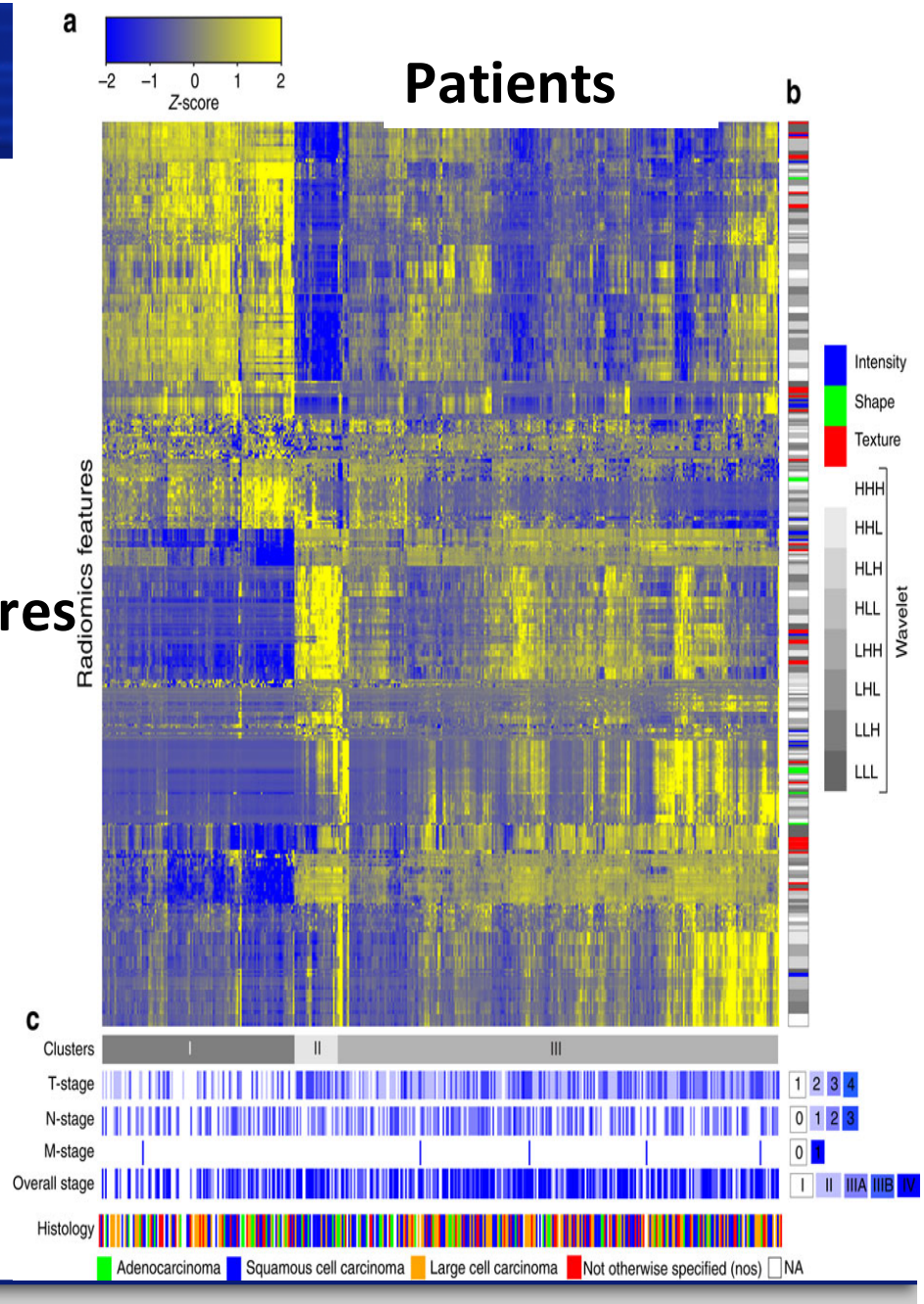
Radiomics

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

Hugo J. W. L. Aerts et. Al.

Nature Communications **5**, Article number: 4006
doi:10.1038/ncomms5006

Features



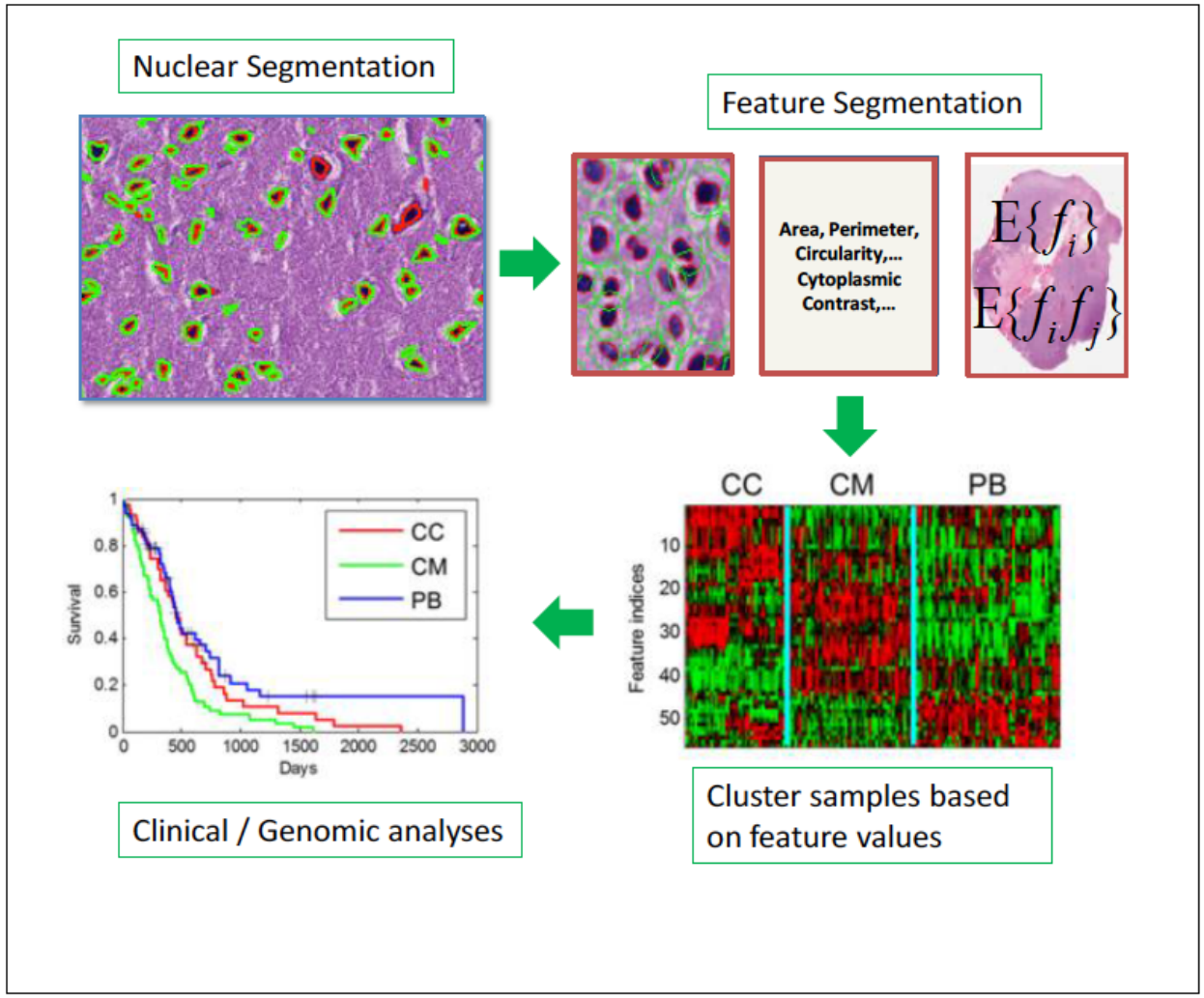
Pathomics

Integrative Morphology/"omics"

Quantitative Feature Analysis in Pathology: Emory In Silico Center for Brain Tumor Research (PI = Dan Brat, PD= Joel Saltz)

NLM/NCI: Integrative Analysis/Digital Pathology R01LM011119, R01LM009239 (Dual PIs Joel Saltz, David Foran)

J Am Med Inform Assoc. 2012 **Integrated morphologic analysis for the identification and characterization of disease subtypes**.

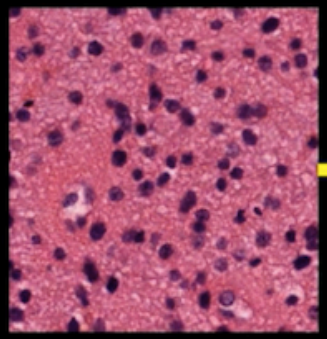


Lee Cooper, Jun Kong

Things that Need to be Done with Spatio Temporal Data

- **Generation of Features**
- Sanity Checking and Data Cleaning
- Qualitative Exploration
- Descriptive Statistics
- Classification
- Identification of Interesting Phenomena
- Prediction
- Control
- Save Data for Later (Compression)

Whole Slide Images (WSI)



Segmentation Parameters

Compute Cluster

Process the images for subjects selected

Compute object-level (nucleus-level) image features

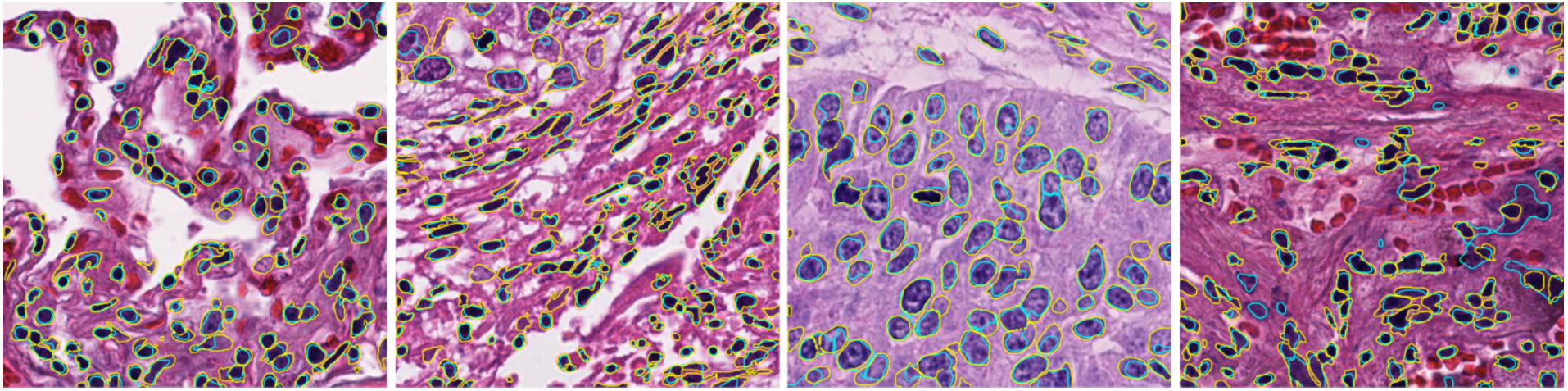
Compute aggregated patient-level image features from object-level features

FeatureDB

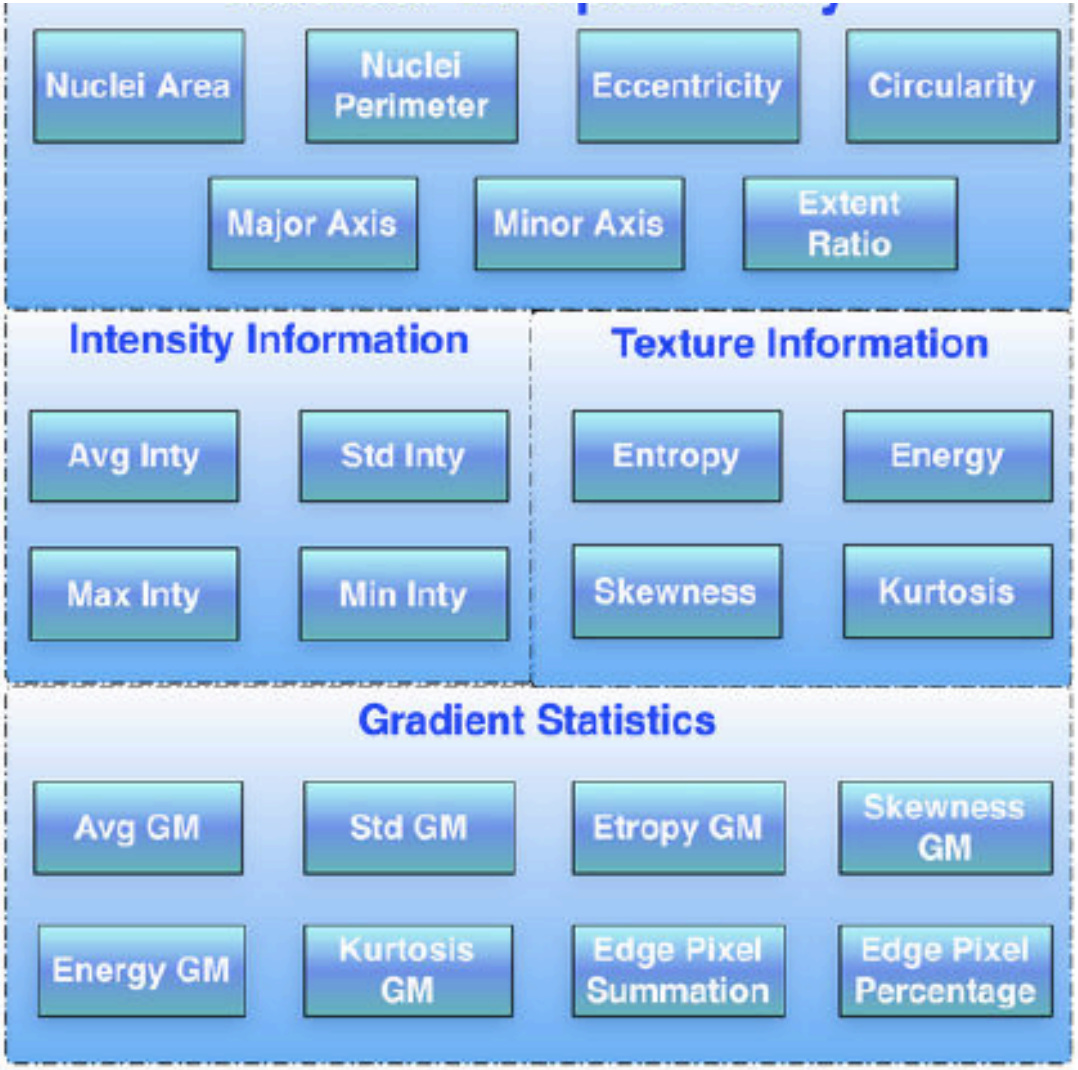
- Load object-level imaging features and segmentation results
- Load patient-level imaging features along with a selected subset of clinical and genomic data (e.g. gene mutations, days to death, vital status)

Robust Nuclear Segmentation

- Robust ensemble algorithm to segment nuclei across tissue types
- Optimized algorithm tuning methods
- Parameter exploration to optimize quality
- Systematic Quality Control pipeline encompassing tissue image quality, human generated ground truth, convolutional neural network critique
- Yi Gao, Allen Tannenbaum, Dimitris Samaras, Le Hou, Tahsin Kurc



Cell Morphometry Features

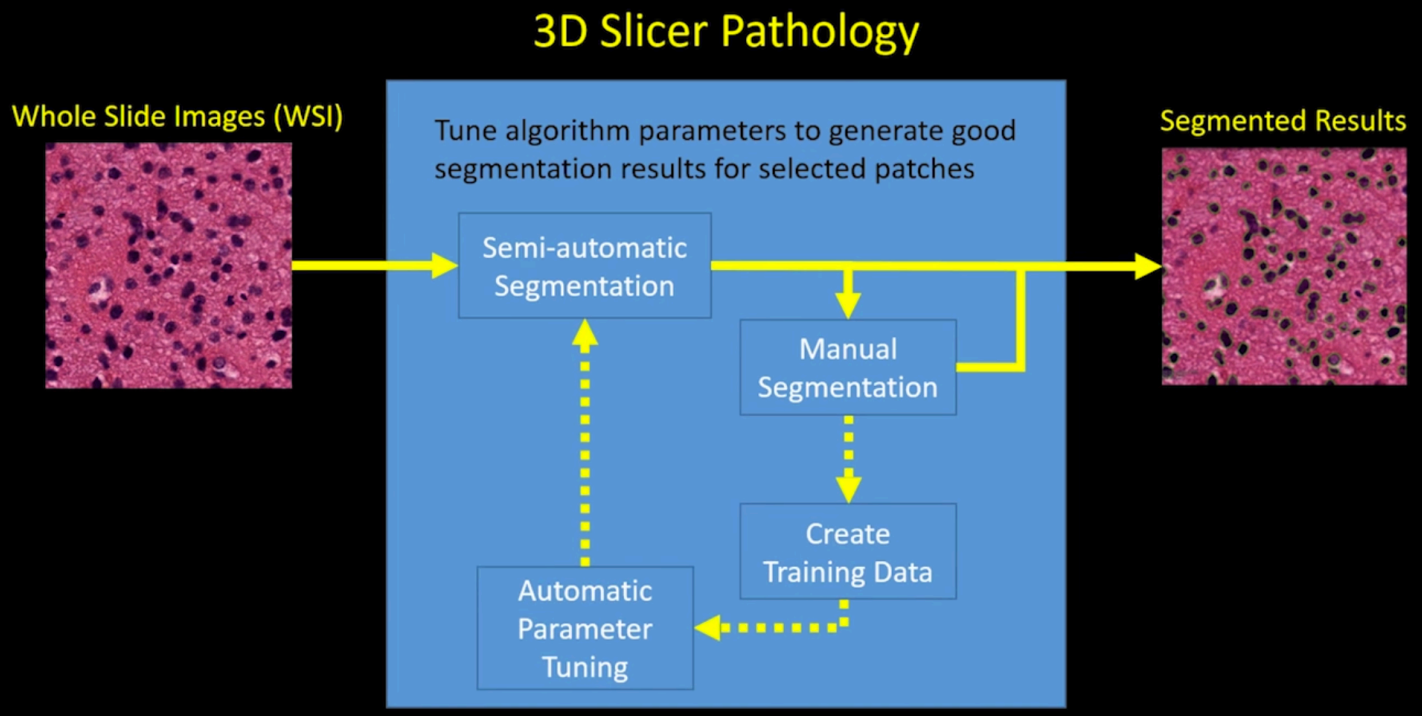


Things that Need to be Done with Spatio Temporal Data

- Generation of Features
- **Sanity Checking and Data Cleaning**
- Qualitative Exploration
- Descriptive Statistics
- Classification
- Identification of Interesting Phenomena
- Prediction
- Control
- Save Data for Later (Compression)

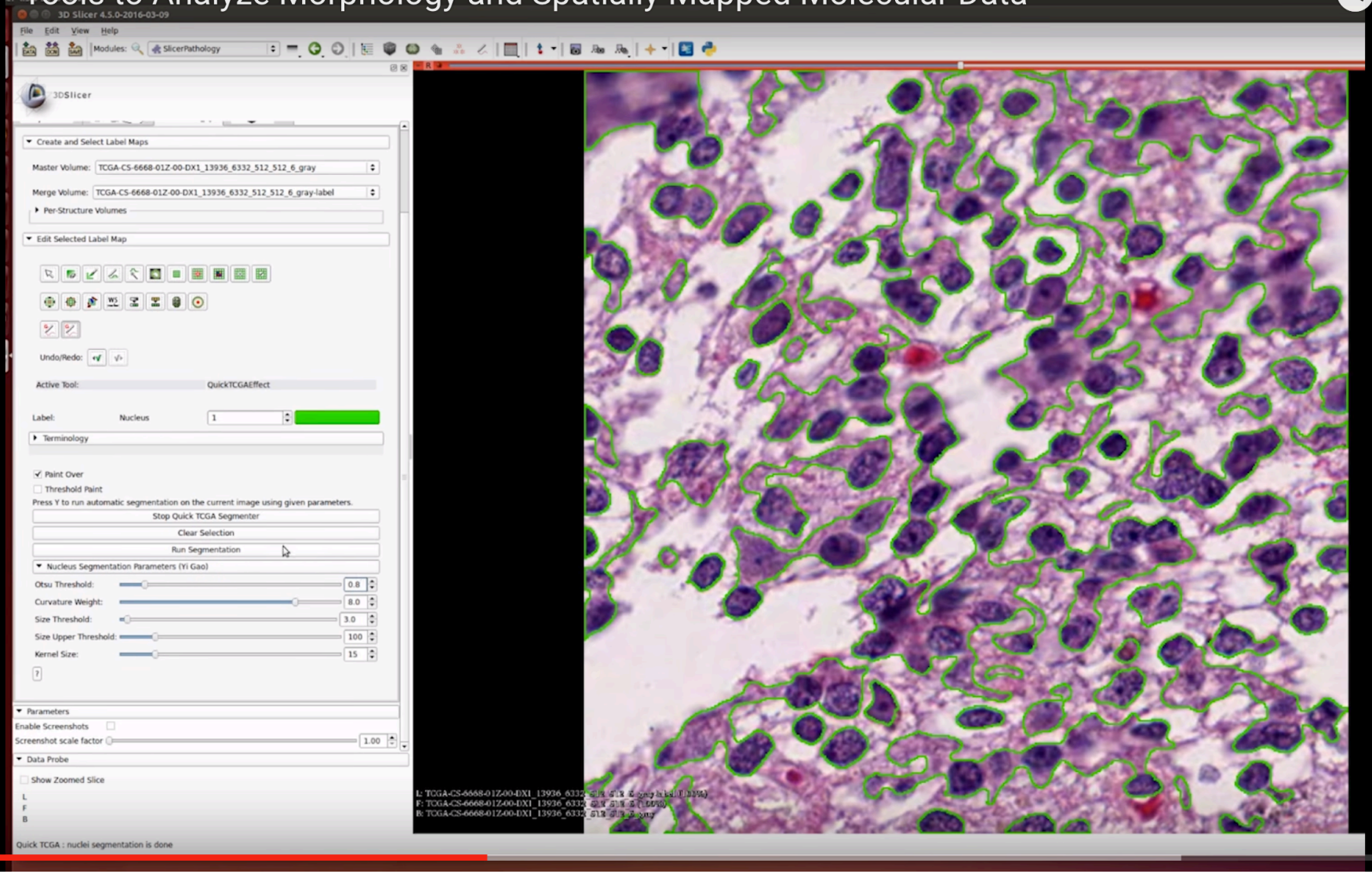
3D Slicer Pathology– Generate High Quality Ground Truth

ITCR - Tools to Analyze Morphology and Spatially Mapped Molecular Data

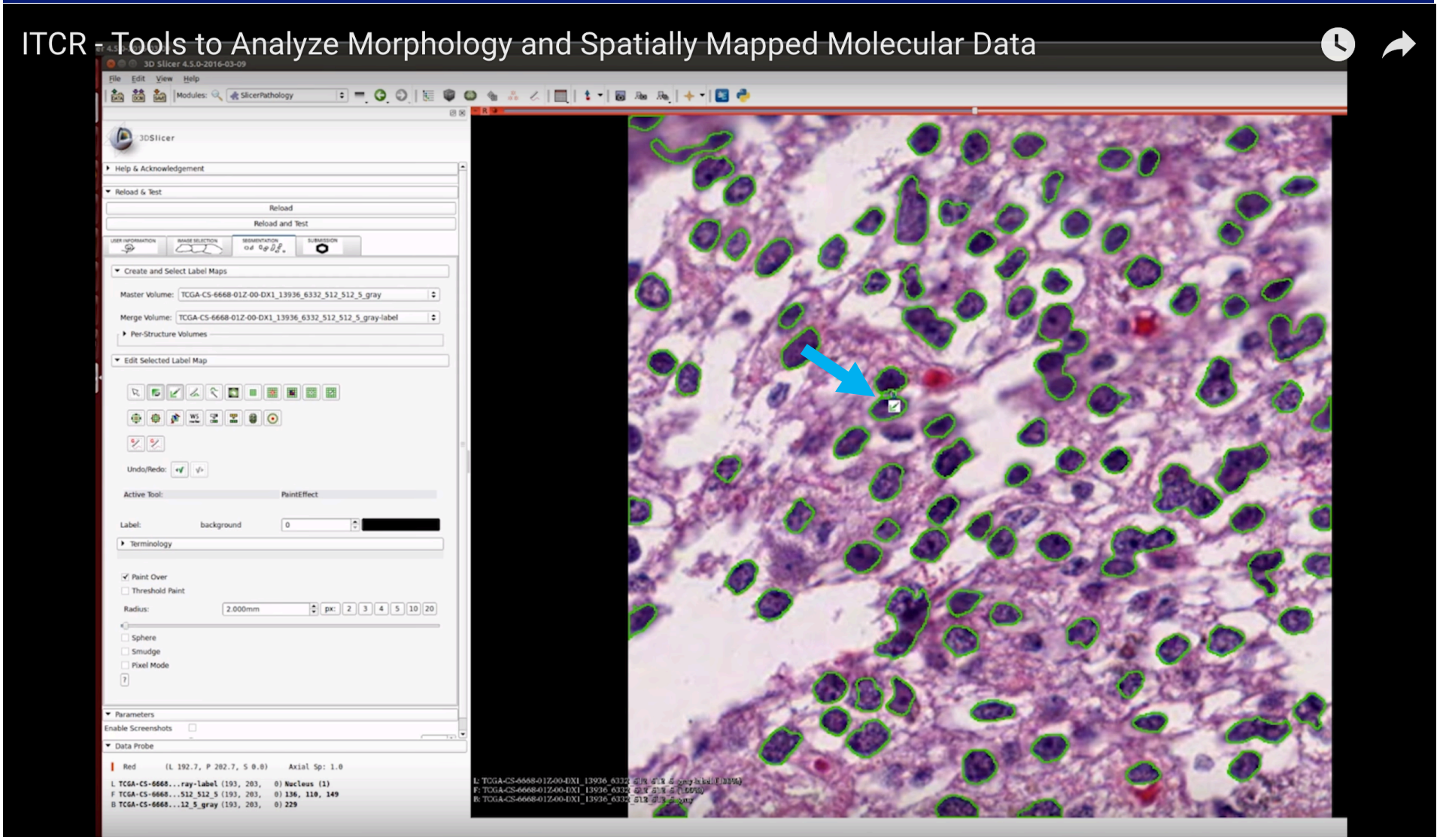


Apply Segmentation Algorithm

ITCR - Tools to Analyze Morphology and Spatially Mapped Molecular Data



Adjust algorithm parameters, manual fine tuning

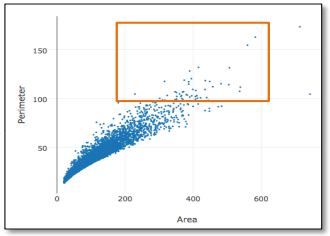


Sanity Check Features

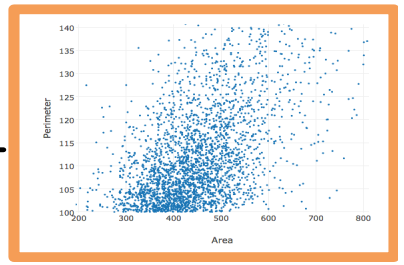
Relationship Between Image and Features

Step 1: Choose a case from the TCGA atlas (case #20)

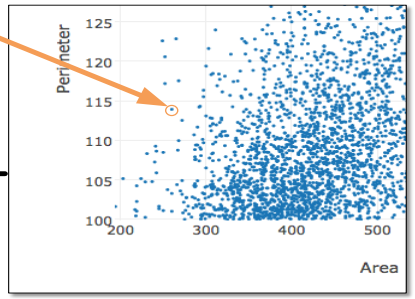
Step 2: Select two features of interest; X axis (*area*), Y axis (*perimeter*)



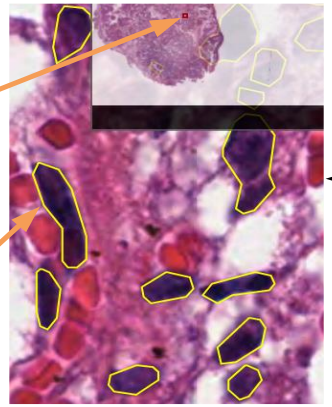
Step 3: Zoom in on region of interest



Step 4: Pick a specific nucleus of interest. Each dot represents a single nucleus



Step 5: Evaluate the features selected in the context of the specific nucleus and where this nucleus is located within the whole slide image



The tool provides visual context for feature evaluation. This technique maps both intuitive features (i.e. size, shape, color) and non-intuitive features (i.e. wavelets, texture) to the ground truth of source images through an interactive web-based user interface.

FeatureScape: u24 case preview prototype

Preview of using featurescape to explore different patient case ids.
For an interactive visualization of Pathology results (including links to the cBio reccord) see TCGAscope.

preview Case IDs

1. TCGA-34-2605-01Z-00-DX1, (cbio) random seed:**0.559**, feature sample size: 1000 featurescape of sampled features
2. TCGA-38-4625-01Z-00-DX1, (cbio) random seed:**0.628**, feature sample size: 1000 featurescape of sampled features
3. TCGA-38-4626-01Z-00-DX1, (cbio) random seed:**0.700**, feature sample size: 1000 featurescape of sampled features
4. TCGA-38-4628-01Z-00-DX1, (cbio) random seed:**0.016**, feature sample size: 1000 featurescape of sampled features
5. TCGA-38-4629-01Z-00-DX1, (cbio) random seed:**0.185**, feature sample size: 1000 featurescape of sampled features
6. TCGA-38-6178-01Z-00-DX1, (cbio) random seed:**0.317**, feature sample size: 1000 featurescape of sampled features
7. TCGA-38-A44F-01Z-00-DX1, (cbio) random seed:**0.906**, feature sample size: 1000 featurescape of sampled features
8. TCGA-50-5044-01Z-00-DX1, (cbio) random seed:**0.055**, feature sample size: 1000 featurescape of sampled features
9. TCGA-50-5045-01Z-00-DX1, (cbio) random seed:**0.946**, feature sample size: 1000 featurescape of sampled features
10. TCGA-50-5045-01Z-00-DX2, (cbio) random seed:**0.551**, feature sample size: 1000 featurescape of sampled features
11. TCGA-50-5055-01Z-00-DX1, (cbio) random seed:**0.127**, feature sample size: 1000 featurescape of sampled features
12. TCGA-34-5232-01Z-00-DX1, (cbio) random seed:**0.208**, feature sample size: 1000 featurescape of sampled features
13. TCGA-50-5055-01Z-00-DX2, (cbio) random seed:**0.321**, feature sample size: 1000 featurescape of sampled features
14. TCGA-50-5066-01Z-00-DX1, (cbio) random seed:**0.711**, feature sample size: 1000 featurescape of sampled features
15. TCGA-50-5066-02Z-00-DX1, (cbio) random seed:**0.008**, feature sample size: 1000 featurescape of sampled features
16. TCGA-50-5942-01Z-00-DX1, (cbio) random seed:**0.031**, feature sample size: 1000 featurescape of sampled features
17. TCGA-50-5946-01Z-00-DX1, (cbio) random seed:**0.768**, feature sample size: 1000 featurescape of sampled features
18. TCGA-50-6590-01Z-00-DX1, (cbio) random seed:**0.668**, feature sample size: 1000 featurescape of sampled features
19. TCGA-50-6591-01Z-00-DX1, (cbio) random seed:**0.498**, feature sample size: 1000 featurescape of sampled features

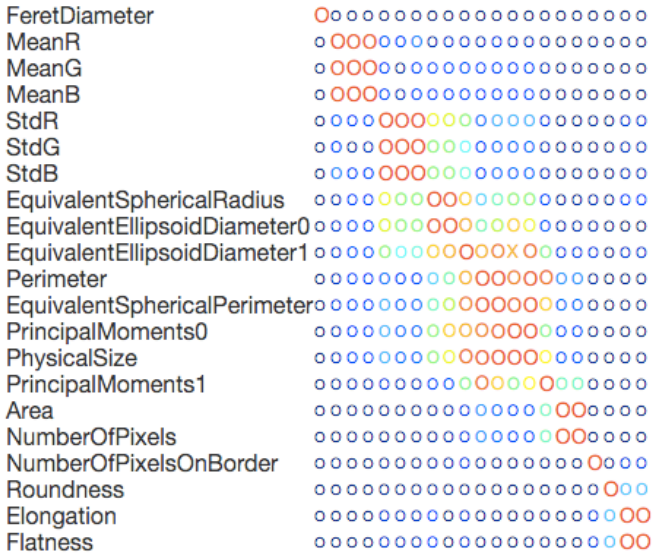
Select Feature Pair – dots correspond to nuclei

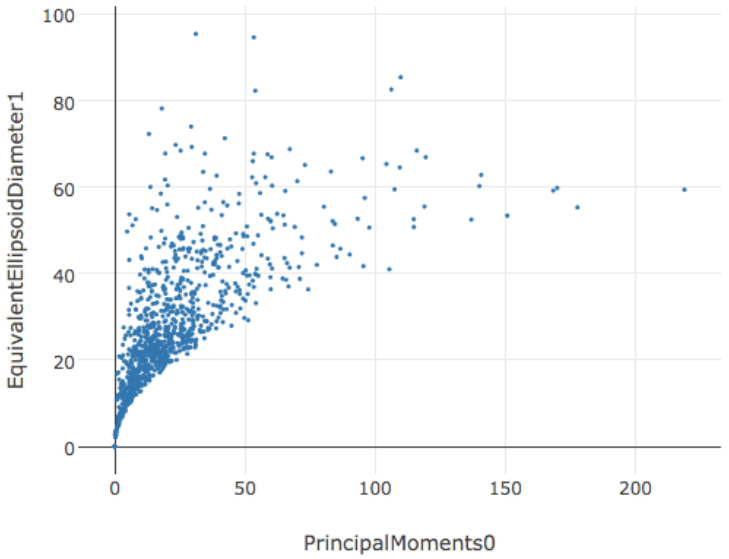
FeatureScape ← Preliminary demo of integrative use of multiple FeatureScape tools

1000 entries sampled from https://tahsin175.informatics.stonybrook.edu:4500/?limit=1000&find={%22randval%22:{%22$gte%22:0.149},%22provenance.analysis_execution_id%22:%22lung-features-v4%22,%22image.caseid%22:%22TCGA-38-4628-01Z-00-DX1%22}

+ Load Data

Cross-tabulated feature correlations





Pearson correlation between

- PrincipalMoments0
- EquivalentEllipsoidDiameter1

Resample from selected region (under development)

Subregion selected – form of gating analogous to flow cytometry

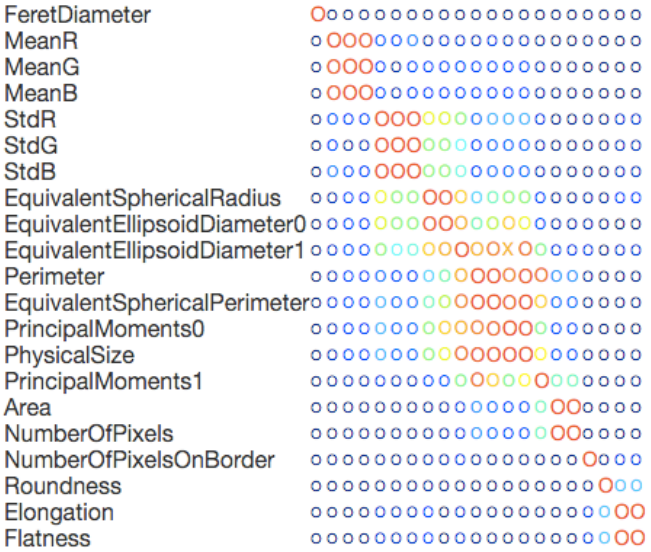
FeatureScape

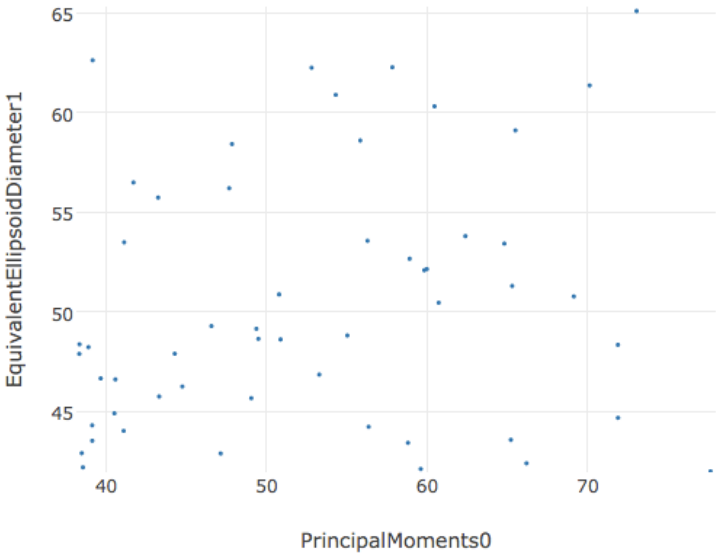
Preliminary demo of integrative use of multiple FeatureScape tools

1000 entries sampled from https://tahsin175.informatics.stonybrook.edu:4500/?limit=1000&find={%22randval%22:{%22$gte%22:0.149},%22provenance.analysis_execution_id%22:%22lung-features-v4%22,%22image.caseid%22:%22TCGA-38-4628-01Z-00-DX1%22}

+ Load Data

Cross-tabulated feature correlations





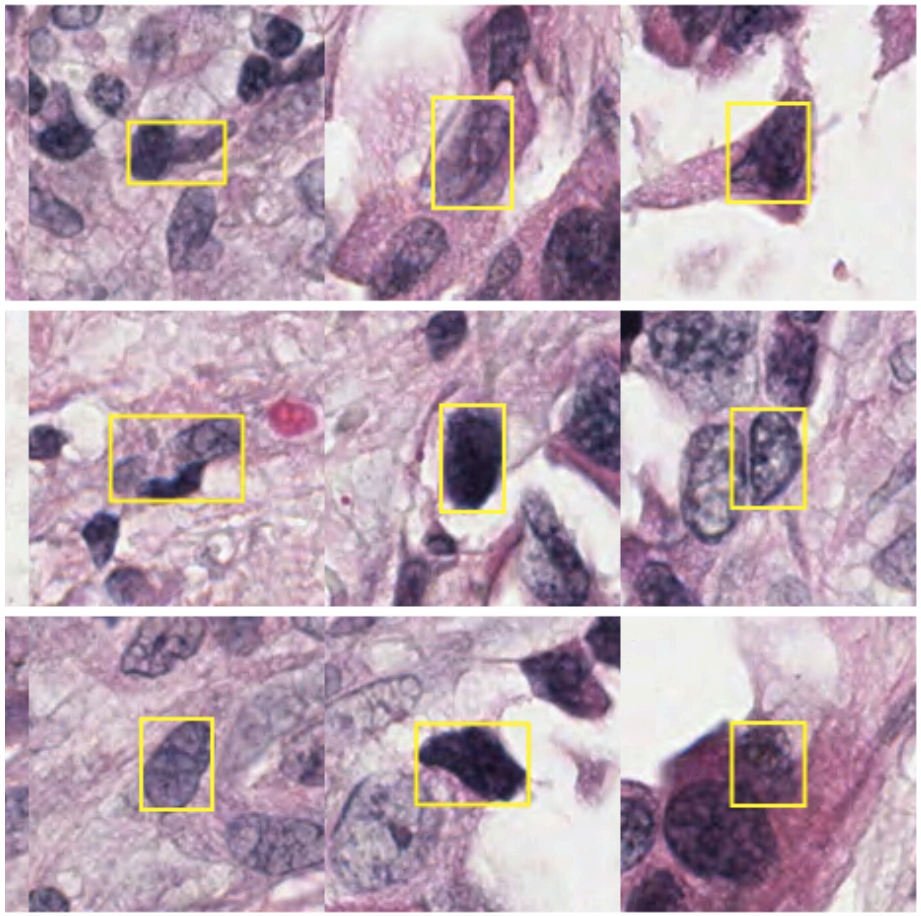
Pearson correlation between

- PrincipalMoments0
- EquivalentEllipsoidDiameter1

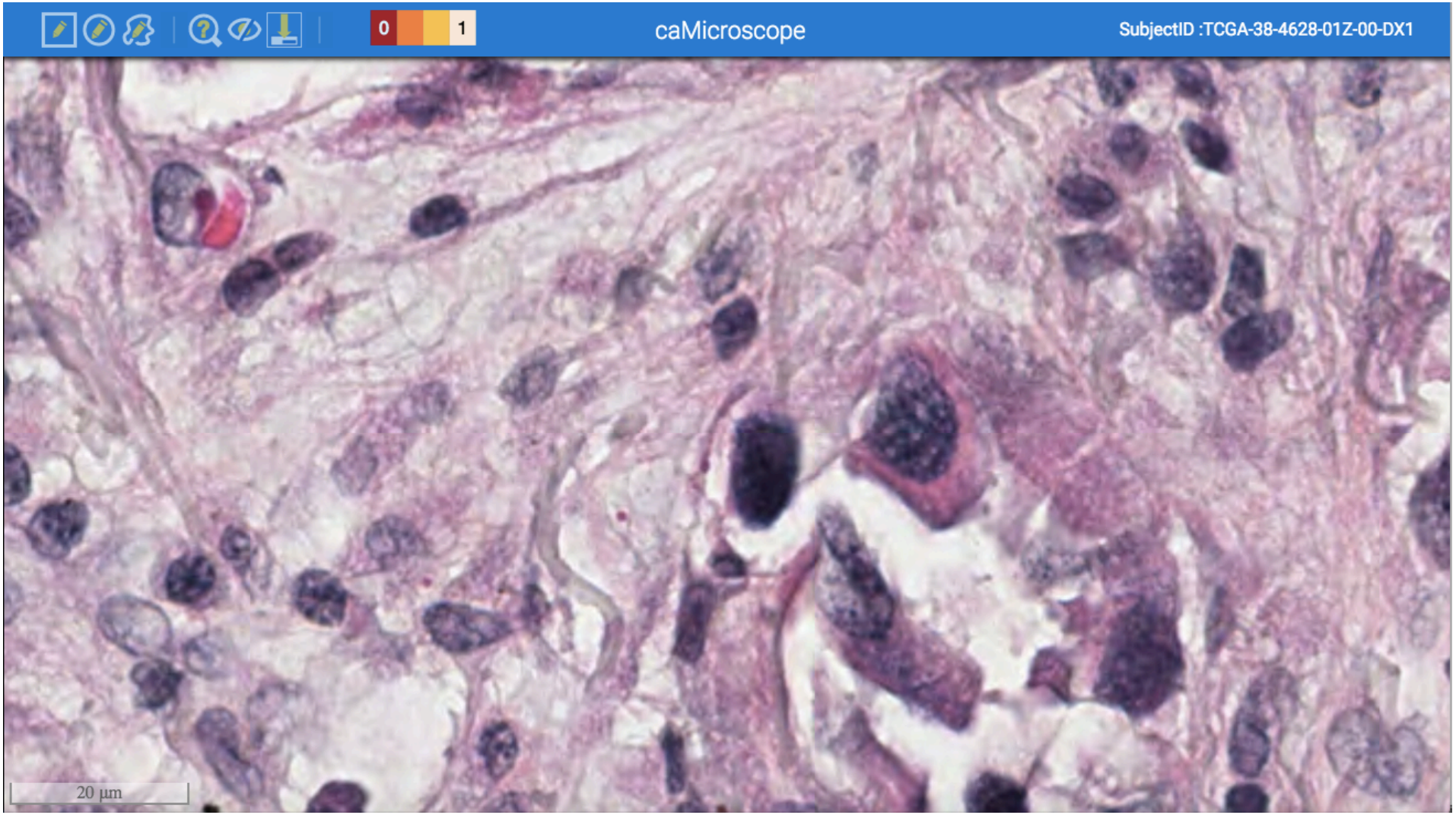
Resample from selected region (under development)

Stony Brook Medicine

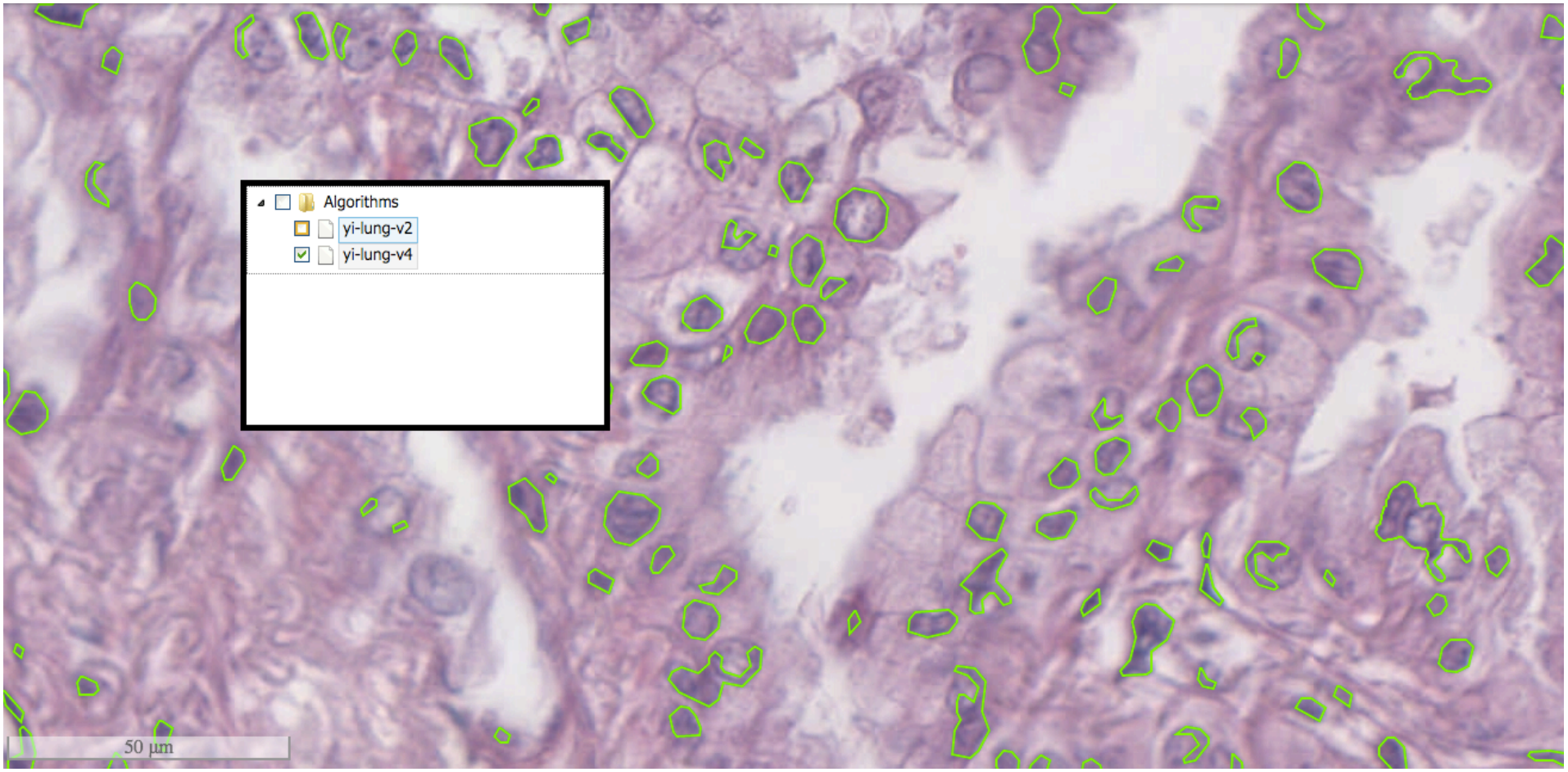
Sample Nuclei from Gated Region



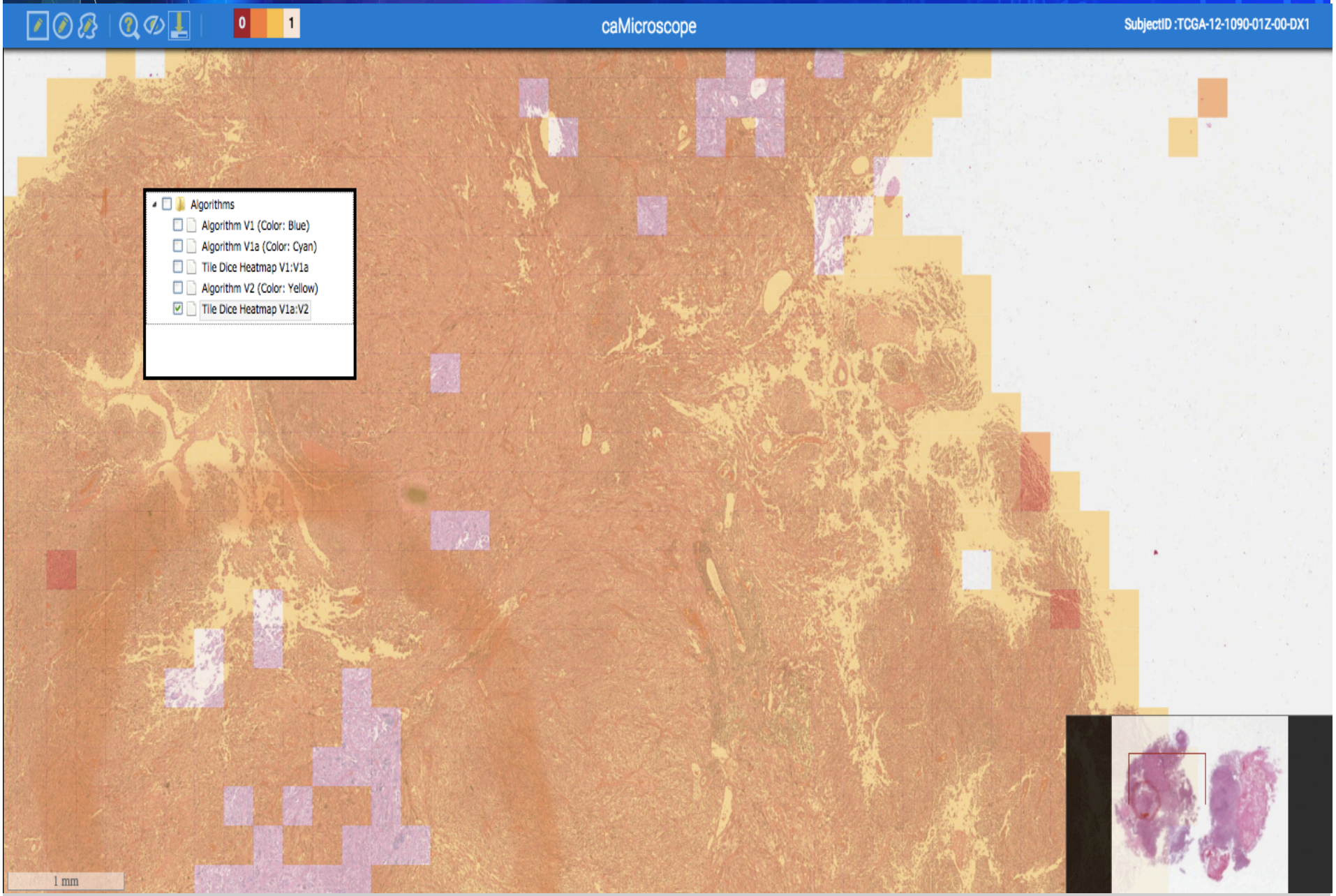
Gated Nuclei in Context



Compare Algorithm Results



Heatmap – Depicts Agreement Between Algorithms

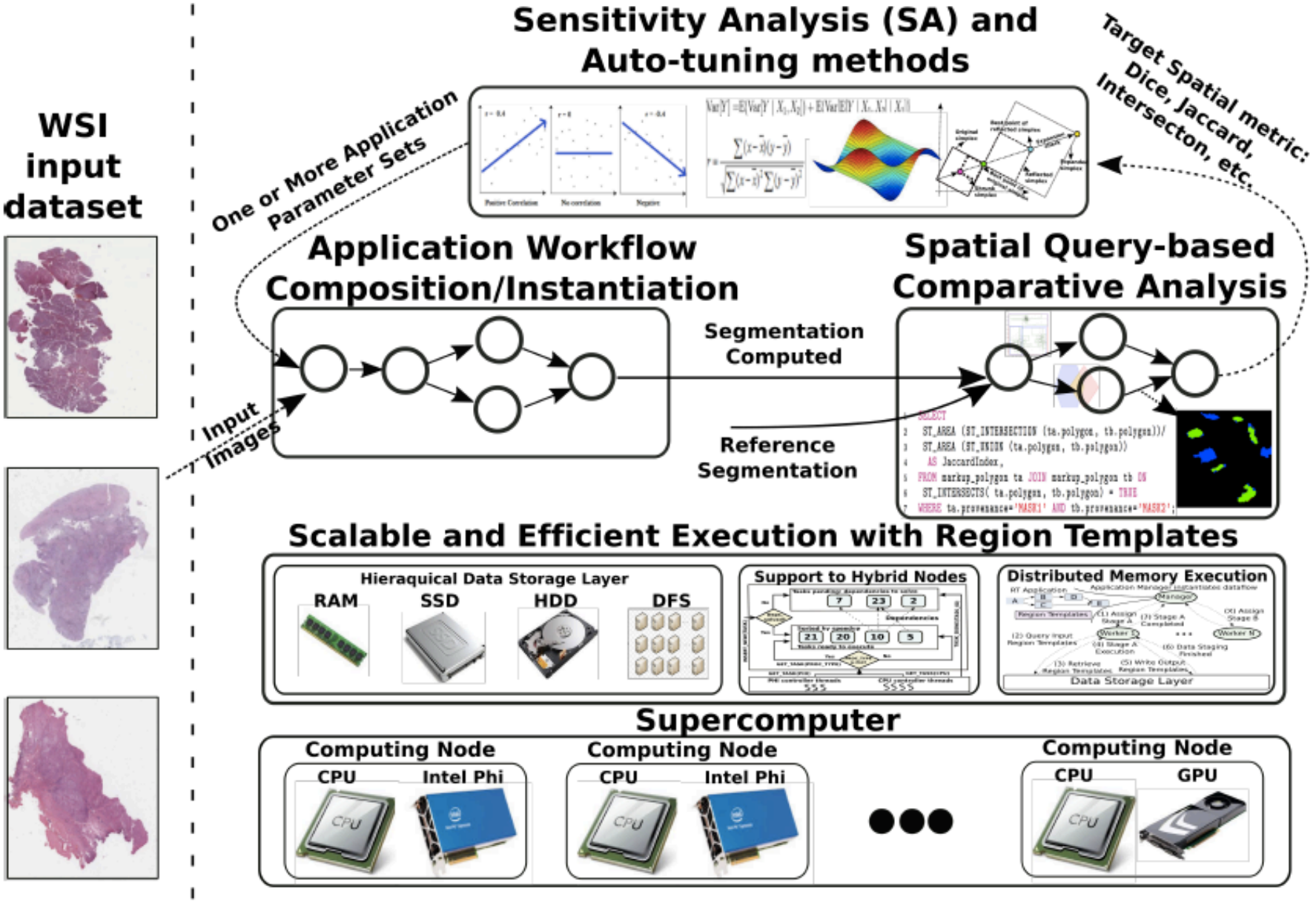


Things that Need to be Done with Spatio Temporal Data

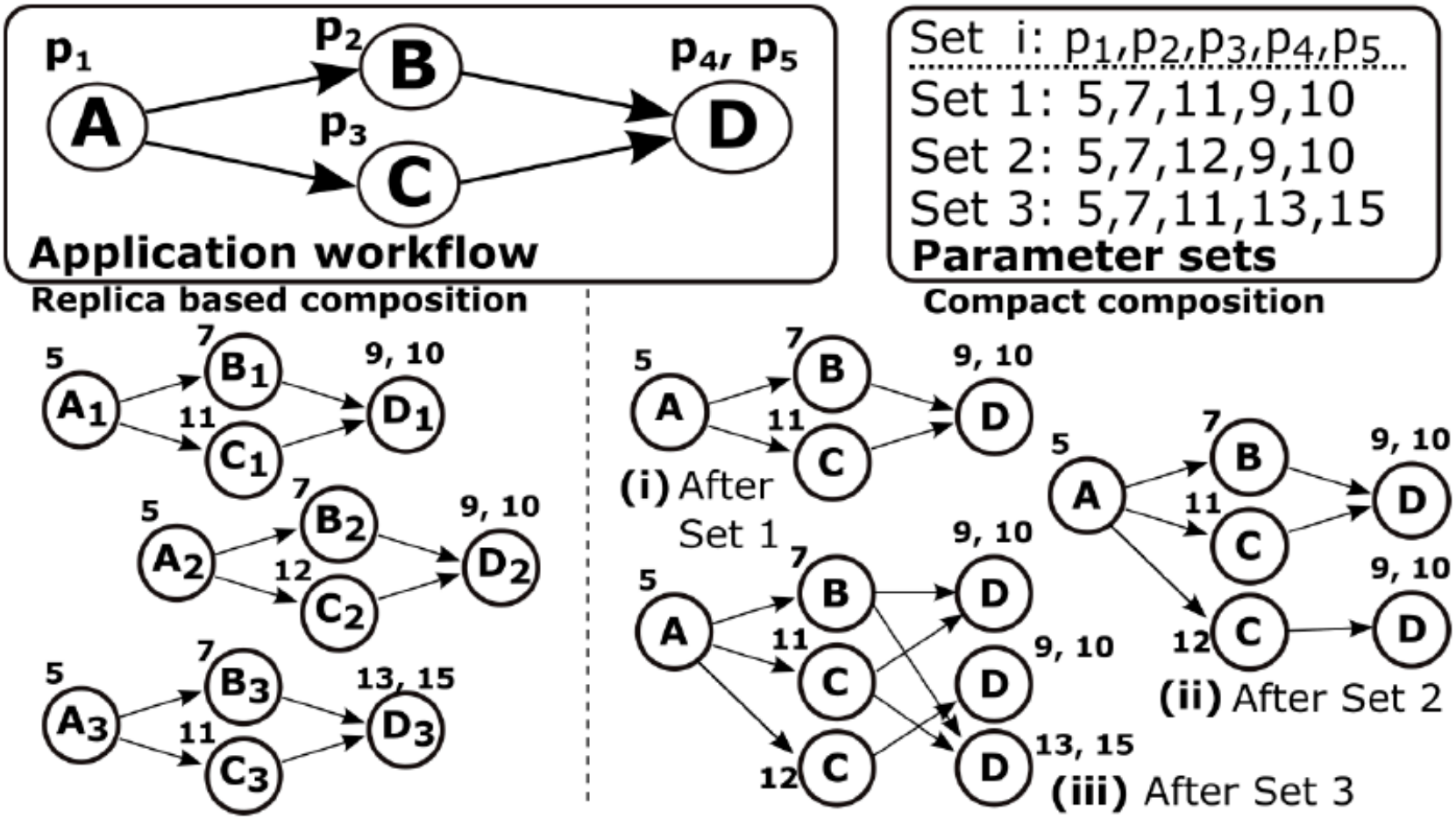
- **Generation of Features**
- **Sanity Checking and Data Cleaning**
- Qualitative Exploration
- Descriptive Statistics
- Classification
- Identification of Interesting Phenomena
- Prediction
- Control
- Save Data for Later (Compression)

Auto-tuning and feature extraction

- Goal – correctly segment trillions of objects (nuclei)
- Adjust algorithm parameters
- Autotuning – finds parameters that best match ground truth in an image patch
- Region template runtime support to optimize generation and management of multi-parameter algorithm results
- Eliminates redundant computation, manages locality
- Active Harmony – Jeff Hollingsworth!!
- Collaboration – George Teodoro, Tahsin Kurc

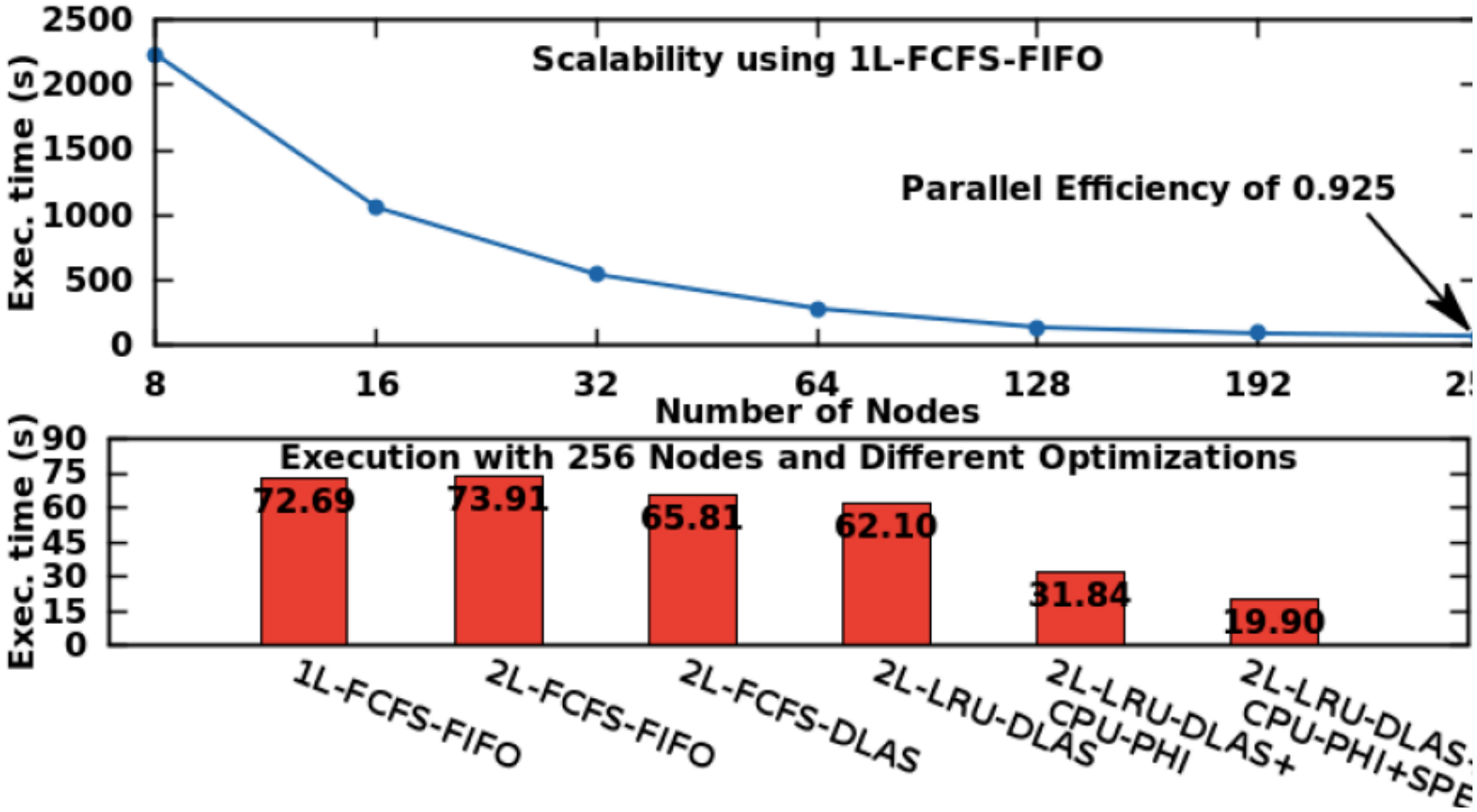

WSI input dataset
Sensitivity Analysis (SA) and Auto-tuning methods
One or More Application Parameter Sets
Target Spatial metric: Dice, Jaccard, Intersecton, etc.
Application Workflow Composition/Instantiation
Spatial Query-based Comparative Analysis
Segmentation Computed
Input Images
Reference Segmentation
Scalable and Efficient Execution with Region Templates
Hieraquical Data Storage Layer
RAM
SSD
HDD
DFS
Support to Hybrid Nodes
Distributed Memory Execution
Data Storage Layer
Supercomputer
Computing Node
CPU
Intel Phi
Computing Node
CPU
Intel Phi
Computing Node
CPU
GPU

E=Eliminate Duplicate Compuations

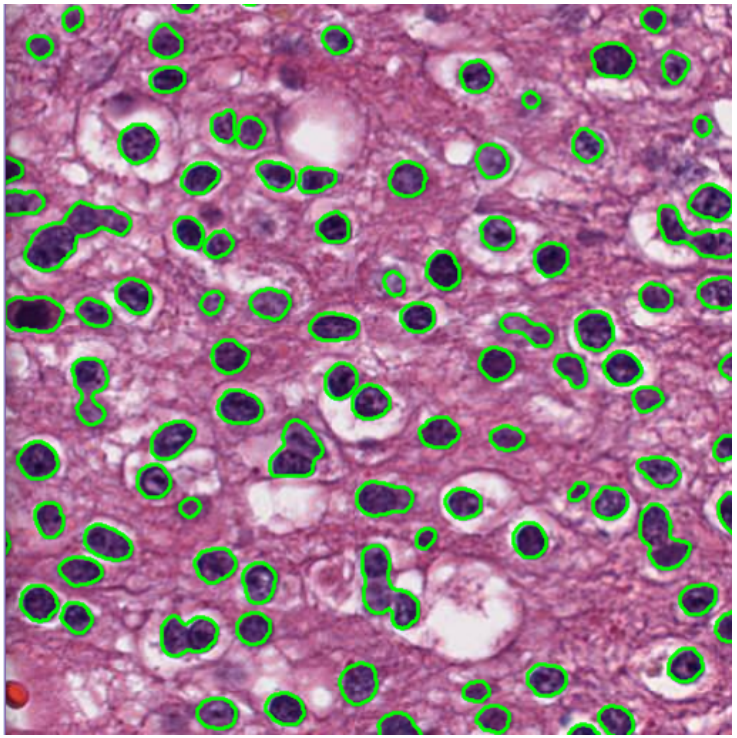


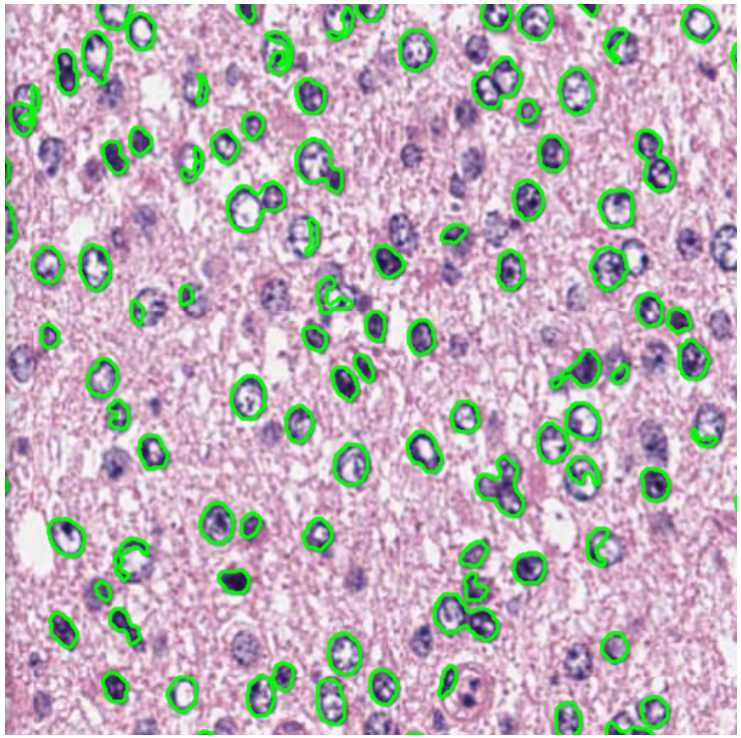
Performance Optimization

256 nodes of Stampede. Each node of the cluster has a dual socket Intel Xeon E5-2680 processors, an Intel Xeon Phi SE10P co-processor and 32GB RAM.The nodes are inter-connected via Mellanox FDR Infiniband switches.



Machine Learning and Quality Critiquing





	Good	Bad
Test as Good	2916	33
Test as Bad	28	2094

SVM Approach

Things that Need to be Done with Spatio Temporal Data

- Generation of Features
- Sanity Checking and Data Cleaning
- **Qualitative Exploration**
- **Descriptive Statistics**
- Classification
- Identification of Interesting Phenomena
- Prediction
- Control
- Save Data for Later (Compression)

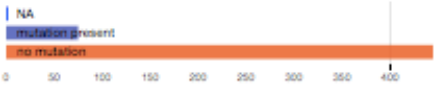
Feature Explorer - Integrated Pathomics Features, Outcomes and "omics" – TCGA NSCLC Adeno Carcinoma Patients

Gene Mutation

Click on bars to select molecular cohorts,
Xaxis: # patients; Yaxis: mutation status
[blue-red] color range indicates fraction of total.

EGFR

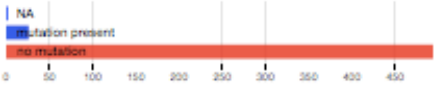
KRAS



STK11_LKB1

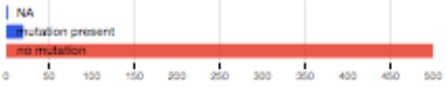
TP53

NF1



BRAF

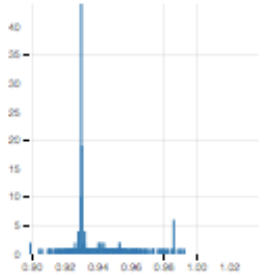
SETD2



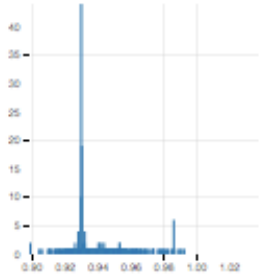
Morphology, Epi, etc

Var 1: Roundness_median
Var 2: StdR_median
Slide mouse click to select ranges
Xaxis: parameter value
Yaxis: #patients

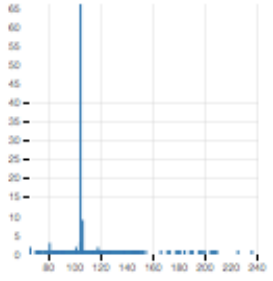
Var 1: Roundness_median



Var 1 Zoom: Roundness_median

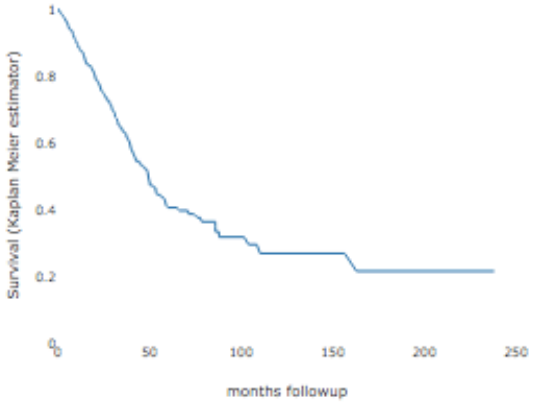


Var 2: StdR_median

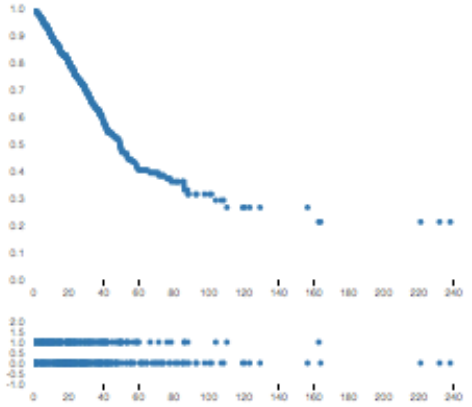


Survival

Blue - whole population; Orange - selected cohort



Zoomable KM estimator (i.e. select ranges, each dot is a patient)



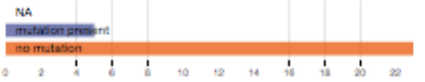
Feature Explorer - Integrated Pathomics Features, Outcomes and "omics" – TCGA NSCLC Adeno Carcinoma Patients

Gene Mutation

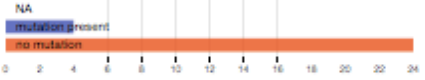
Click on bars to select molecular cohorts,
Xaxis: # patients; Yaxis: mutation status
[blue-red] color range indicates fraction of total.

EGFR

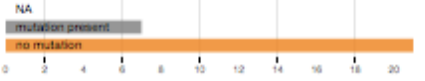
KRAS



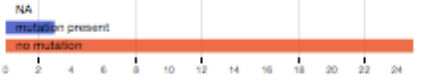
STK11_LKB1



TP53

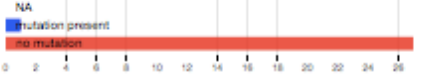


NF1



BRAF

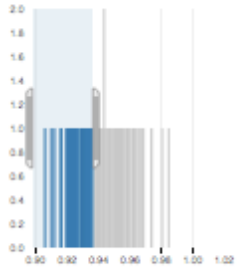
SETD2



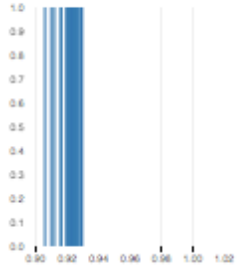
Morphology, Epi, etc

Var 1: Roundness_median
Var 2: StdR_median
Slide mouse click to select ranges
Xaxis: parameter value
Yaxis: #patients

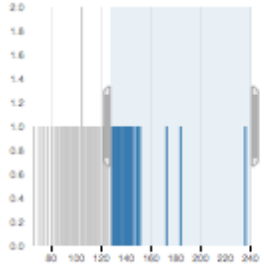
Var 1: Roundness_median



Var 1 Zoom: Roundness_median

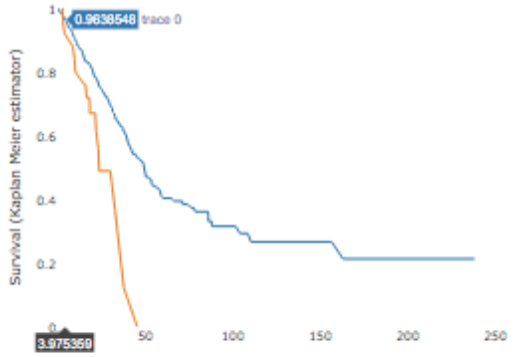


Var 2: StdR_median

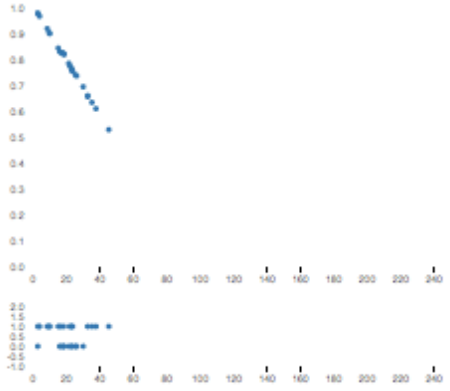


Survival

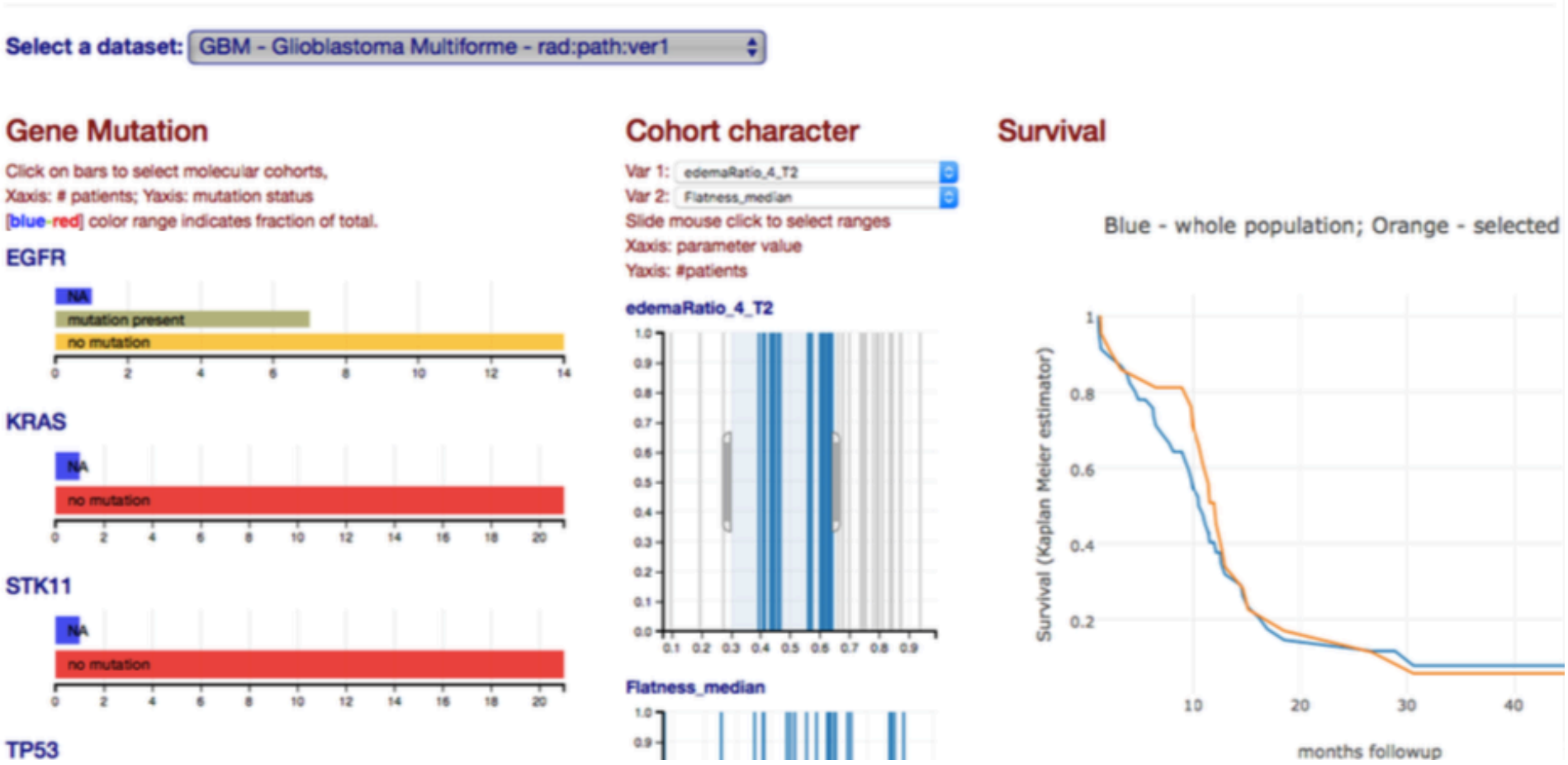
Blue - whole population; Orange - selected cohort



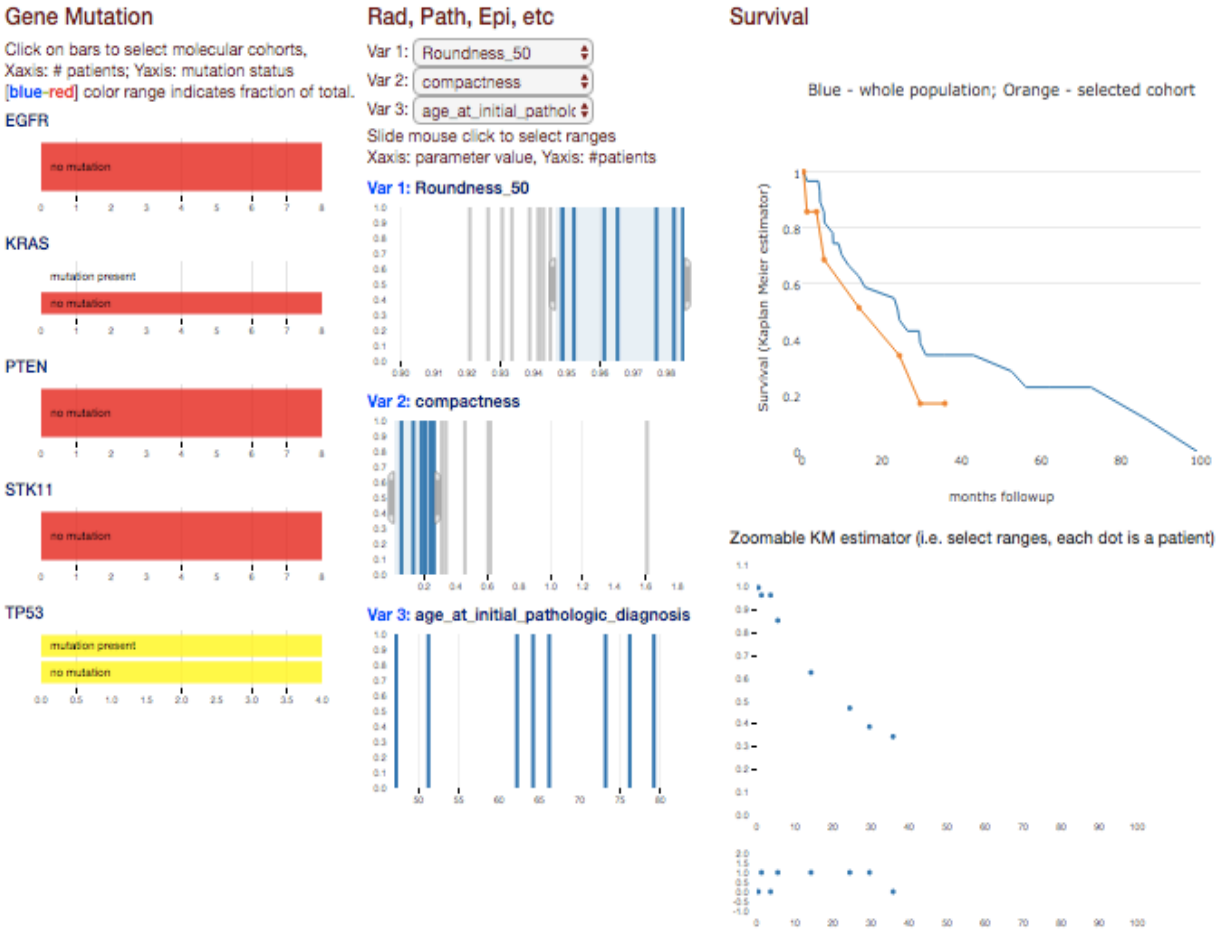
Zoomable KM estimator (i.e. select ranges, each dot is a patient)



Collaboration with MGH – Feature Explorer – Radiology Brain MR/Pathology Features



Collaboration with SBU Radiology – TCGA NSCLC Adeno Carcinoma
Integrative Radiology, Pathology, "omics", outcome



Mary Saltz, Mark Schweitzer SBU Radiology

Things that Need to be Done with Spatio Temporal Data

- Generation of Features
- Sanity Checking and Data Cleaning
- Qualitative Exploration
- Descriptive Statistics
- **Classification**
- **Identification of Interesting Phenomena**
- Prediction
- Control
- Save Data for Later (Compression)

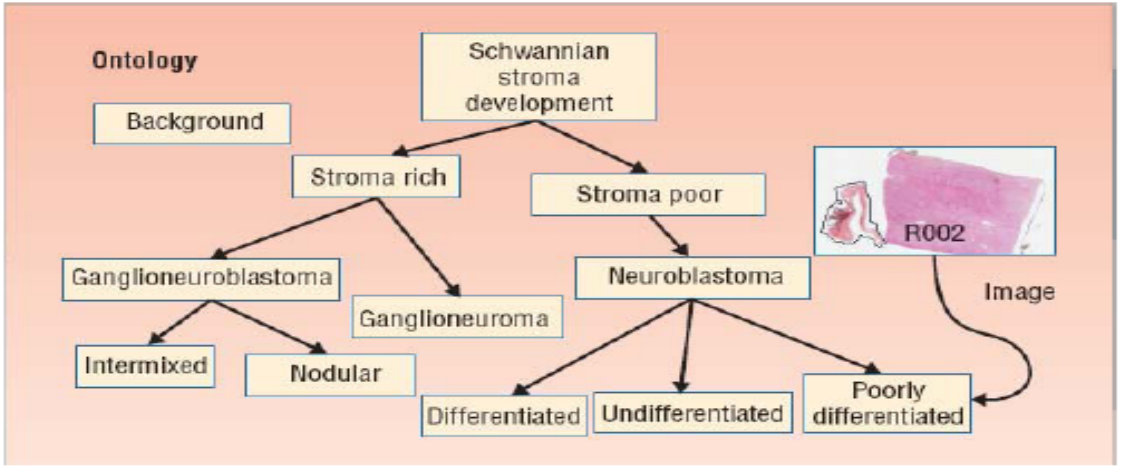
Classification

- Automated or semi-automated identification of tissue or cell type
- Variety of machine learning and deep learning methods
- Classification of Neuroblastoma
- Classification of Gliomas
- Quantification of lymphocyte infiltration

Classification and Characterization of Heterogeneity

BISTI/NIBIB Center for Grid Enabled Image Analysis - P20 EB000591, PI Saltz

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

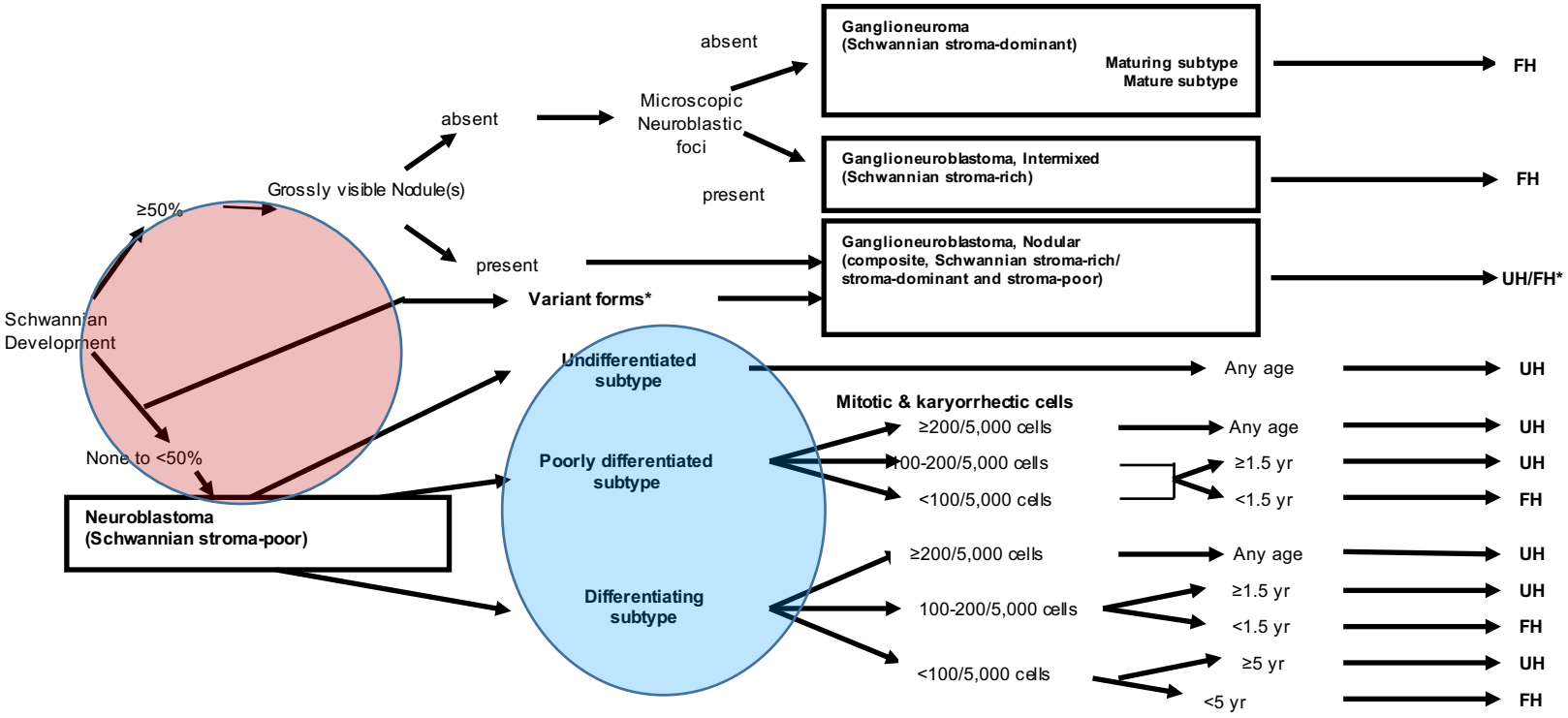


Morphological characterization of tissue used for prognosis

Hiro Shimada, Metin Gurcan, Jun Kong, Lee Cooper Joel Saltz

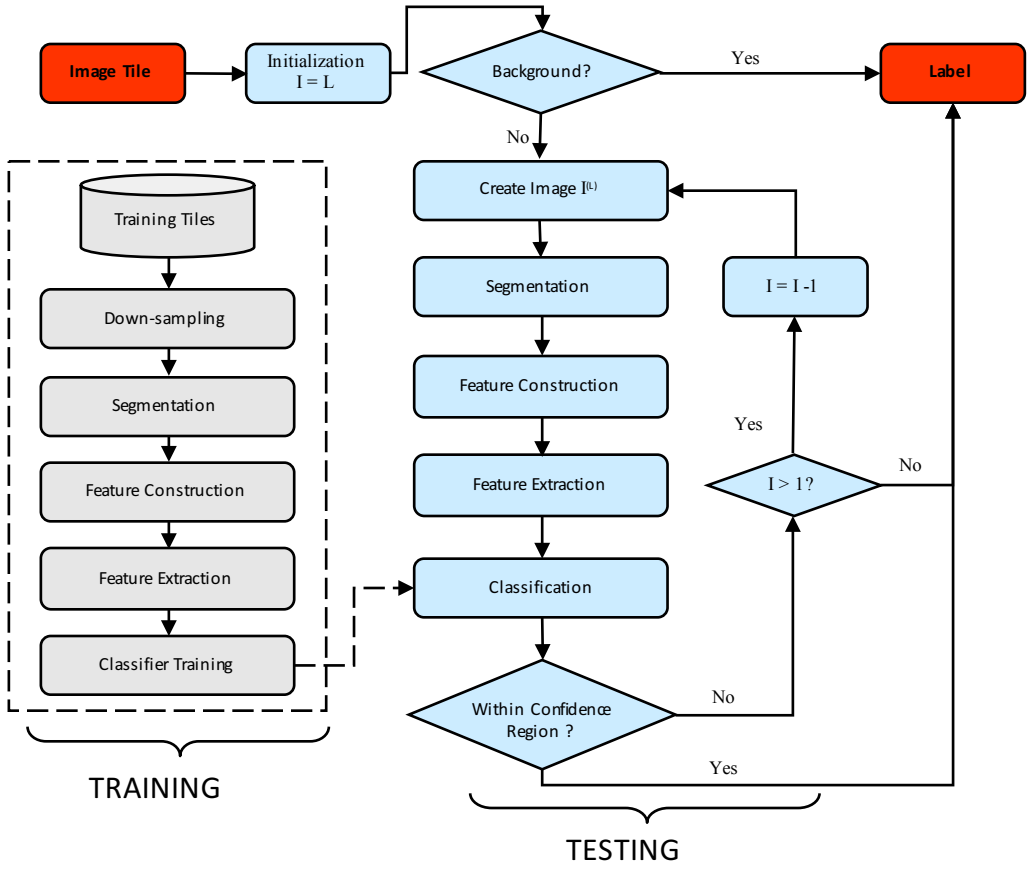
Gurcan, Shamada, Kong, Saltz

Neuroblastoma Classification



FH: favorable histology **UH:** unfavorable histology
CANCER 2003; 98:2274-81

Multi-Scale Machine Learning Based Shimada Classification System



- 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)

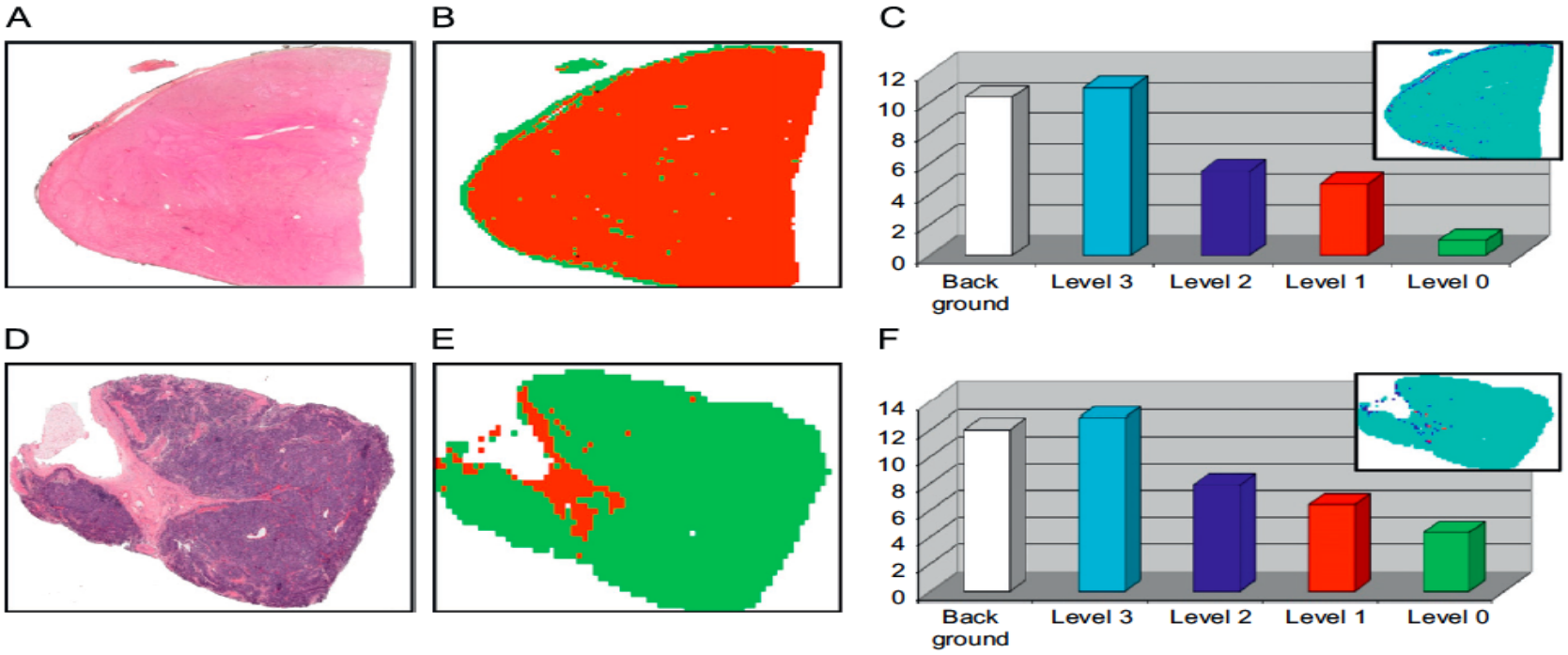


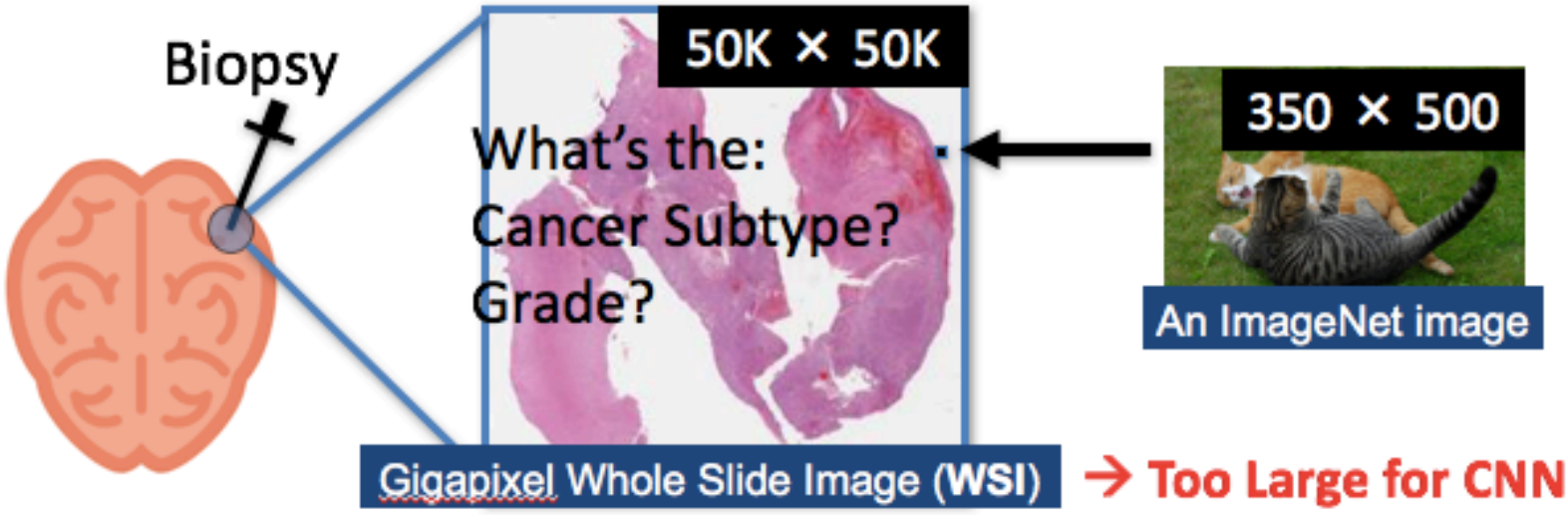
Fig. 8. Sample classification results after processing a whole-slide NB image. (a) and (d) are the H&E stained NB slides associated with stroma-rich and stroma-poor by an expert pathologist. (b) and (e) are the classification maps identified by the computerized system where the red color corresponds to stroma-rich regions and the green color corresponds to stroma-poor regions. (c) and (f) are the corresponding decision level statistics that show in log-scale the number of image tiles classified at a certain resolution level. In the resolution level map on the upper right, cyan color represents the lowest resolution and green color represents the highest resolution, respectively.

Brain Tumor Classification – CVPR 2016

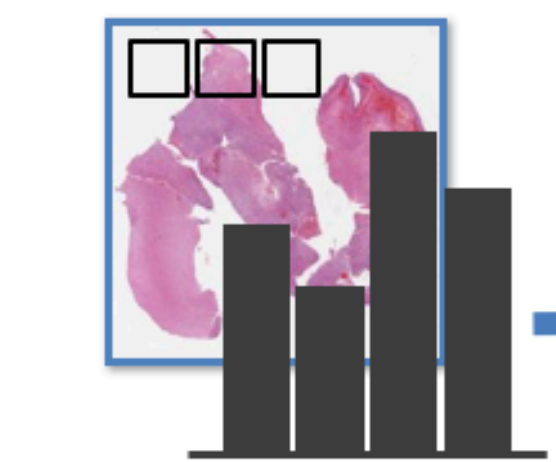
Contributions

- Automatic discriminative patch identification for patch-CNN training
- A robust, general method to combine patch-level predictions

An important application: cancer classification



Combining Information from Patches



Histogram of Patch-level Classes

Advantage:

- The prediction of every patch counts
  → robustness & generalization

SVM or Logistic Regression

Engineering details

- CNN architecture: AlexNet and VGG16
- Patch size: 500x500, Multiple scale
- Dataset: TCGA [gdc-portal.nci.nih.gov]
- Number of Patches: 1000 per WSI

Brain Tumor Classification Results

Glioma is
- The most common brain cancer
- The leading cause of cancer-related deaths in people under age 20

Methods	Accuracy
VGG16 features + BoW + SVM	0.667
Patch-CNN + Voting	0.710
Patch-CNN + Max-pooling	0.710
Our method	**0.771**
Pathologists' Agreement [M. Gupta 2015] (on a similar dataset)	0.7-0.8

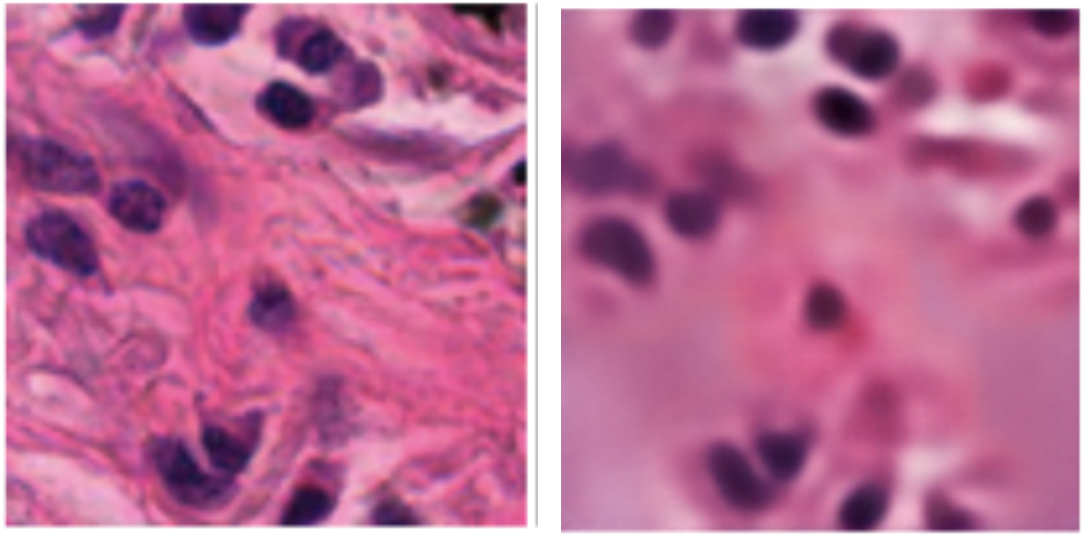
Confusion Matrix: OA is very hard even for pathologists	GBM	OD	OA	DA	AA	AO
Glioblastoma, Grade IV (GBM)	214		2		1	
Oligodendroglioma, Grade II (OD)	1	47	22	2		1
Oligoastrocytoma, Grade II & III (OA)	1	18	40	8	3	1
Diffuse Astrocytoma, Grade II (DA)	3	9	6	20		1
Anaplastic Astrocytoma, Grade III (AA)	3	2	3	3	4	
Anaplastic Oligodendroglioma, Grade III (AO)	2	2	3			1

Le Hou, Dimitris Samaras, Tahsin Kurc, Yi Gao, Liz Vanner, James Davis, Joel Saltz

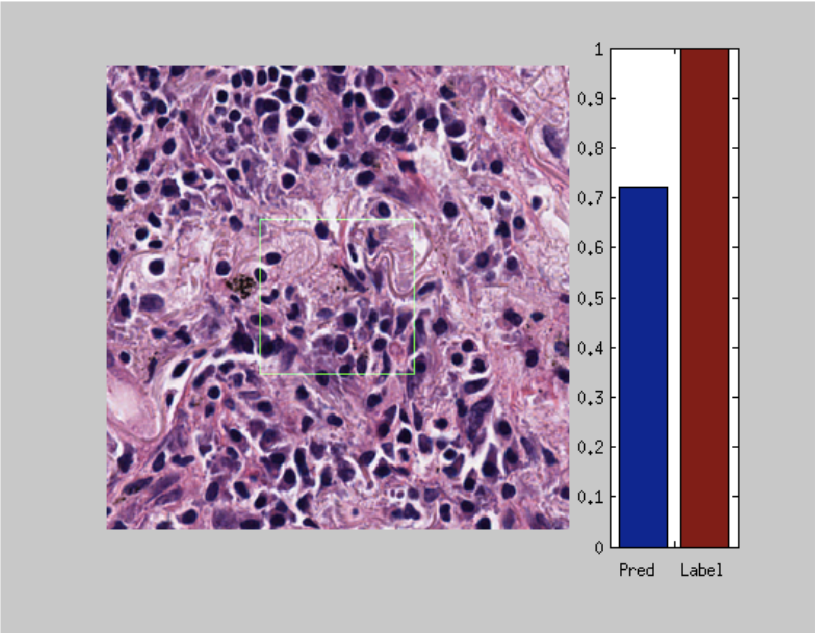
Tumor Infiltrating Lymphocyte quantification

- Convolutional neural network to classify lymphocyte infiltration in tissue patches
- Convolutional neural network and random forest to classify individual segmented nuclei
- Extensive collection of ground truth
- Joint work with Emory and TCGA PanCanAtlas Immune group

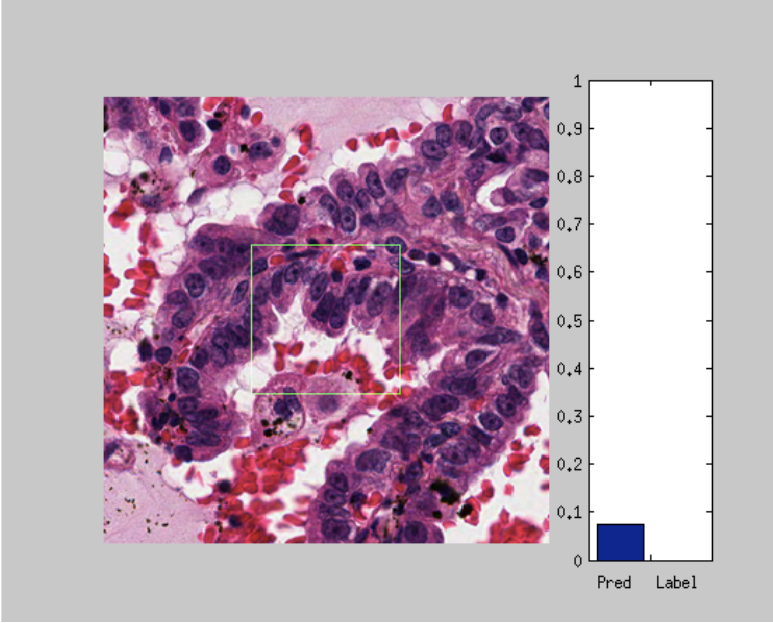
Unsupervised Autoencoder – 100 feature dimensions



Lymphocyte identification



Lymphocytes Infiltration



No Lymphocyte Infiltration

Receiver Operating Characteristic – Area Under Curve – 95%

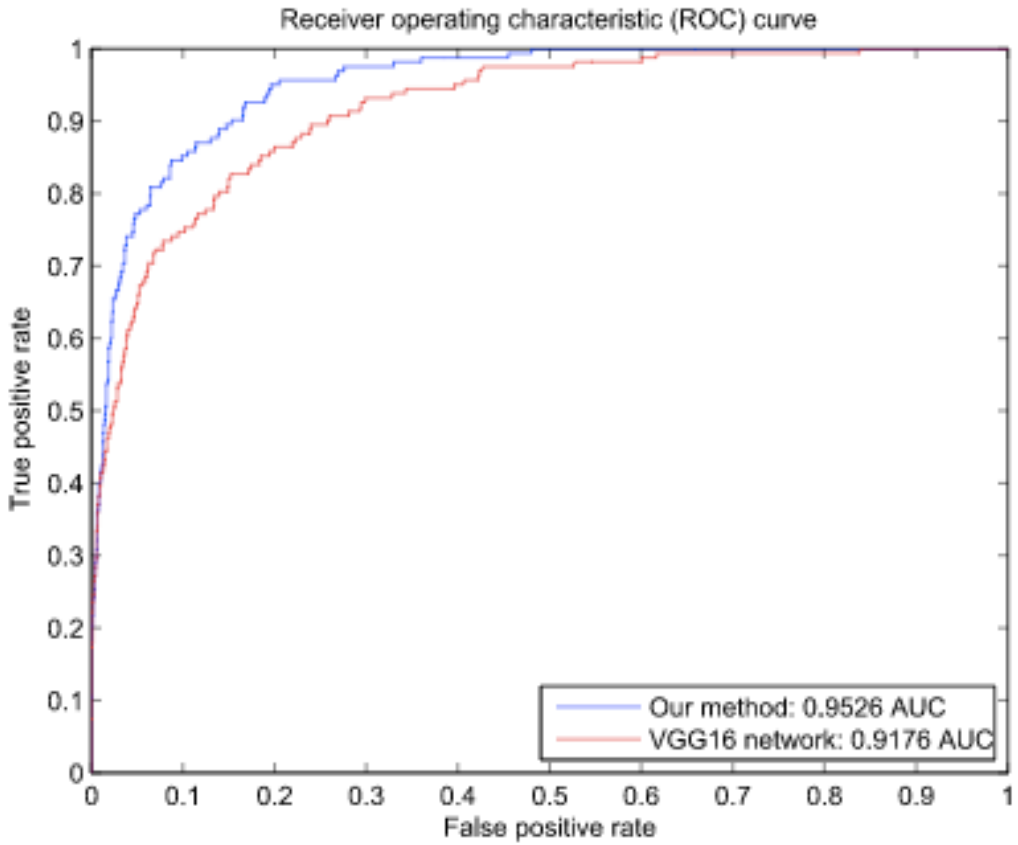
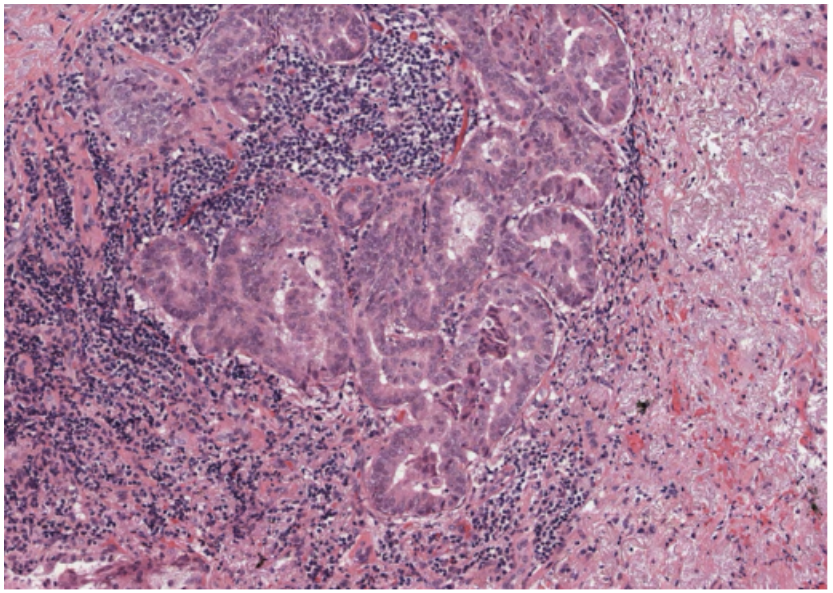
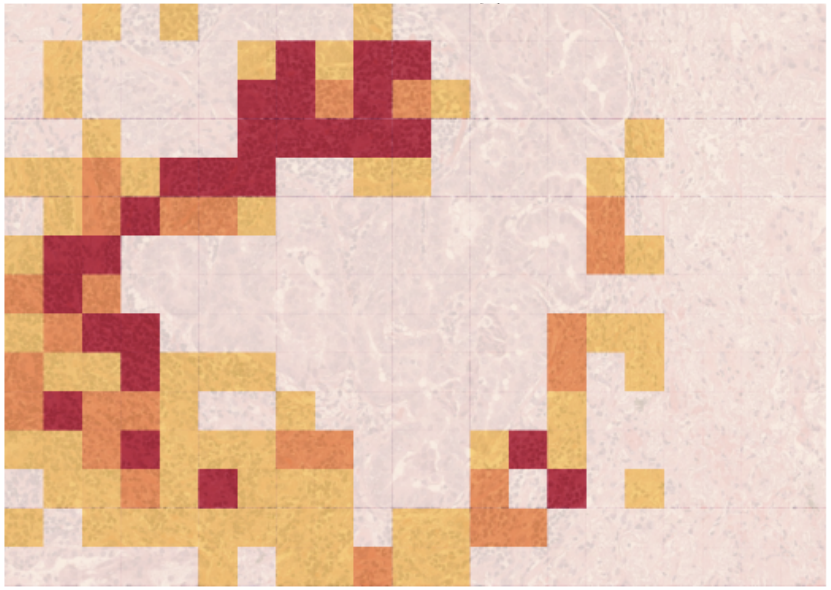


Figure 3 ROC curve

Lymphocyte Classification Heat Map

Trained with 22.2K image patches
Pathologist corrects and edits





Commonalities

- Provided quick but pretty deep dive into aspects of spatio temporal data analytics
- Requirements, methods and I think core infrastructure can be shared between disparate application classes
- These application classes are definitely data but spatio-temporal aspects are HPC community context friendly
- Most of this holds for analysis of scientific program generated data – ORNL Klasky collaborations

ITCR Team

Stony Brook University

Joel Saltz

Tahsin Kurc

Yi Gao

Allen Tannenbaum

Erich Bremer

Jonas Almeida

Alina Jasniewski

Fusheng Wang

Tammy DiPrima

Andrew White

Le Hou

Furqan Baig

Mary Saltz

Emory University

Ashish Sharma

Adam Marcus

Oak Ridge National Laboratory

Scott Klasky

Dave Pugmire

Jeremy Logan

Yale University

Michael Krauthammer

Harvard University

Rick Cummings

Funding – Thanks!

- This work was supported in part by U24CA180924-01, NCIP/Leidos 14X138 and HHSN261200800001E from the NCI; R01LM011119-01 and R01LM009239 from the NLM
- This research used resources provided by the National Science Foundation XSEDE Science Gateways program under grant TG-ASC130023 and the Keeneland Computing Facility at the Georgia Institute of Technology, which is supported by the NSF under Contract OCI-0910735.

Thanks!