Big Data in Biomedical Imaging: Pathomics Analysis of Cancer

Tahsin Kurc

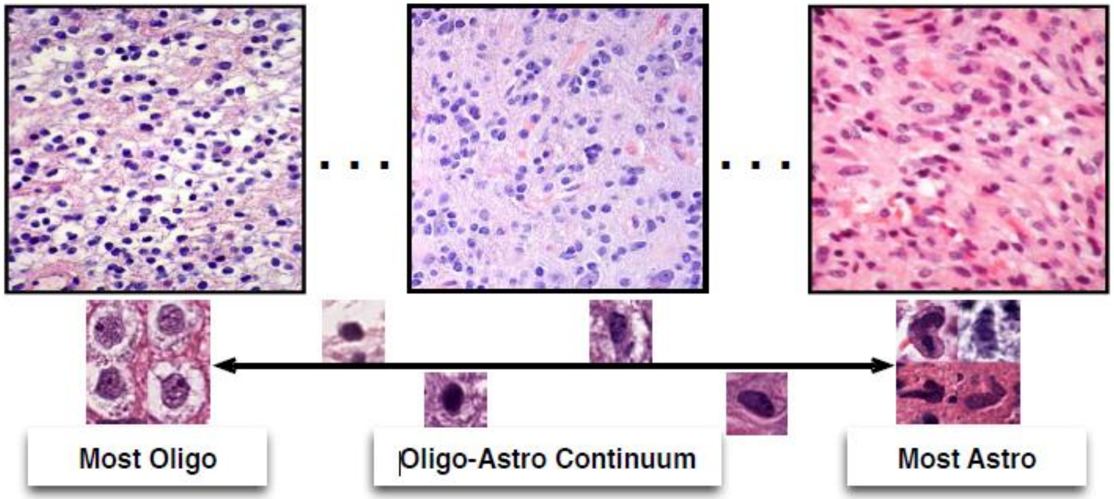
Biomedical Informatics Department
Stony Brook University

Team

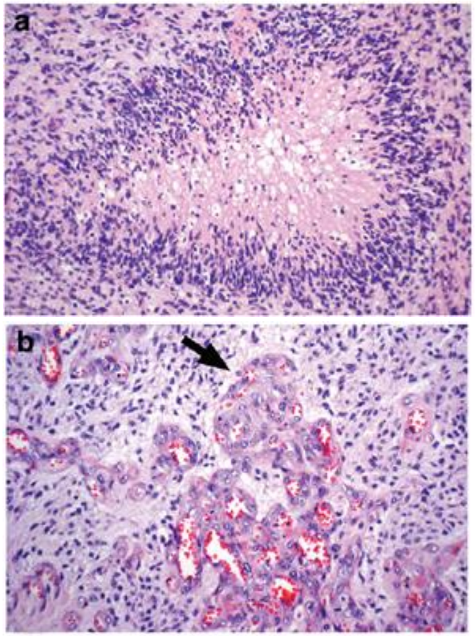
- Stony Brook:
  - Joel Saltz, Erich Bremer, Feiqiao Wang, Tammy Di Prima, Jonas Almeida, Mary Saltz, Le Hou, Vu Nyugen, Si Wen, Maozheng Zhao, Tianhao Zhao, Raj Gupta, Fusheng Wang, Furqan Baig, Yi Gao, Alina Jasniewski
- Emory University:
  - Ashish Sharma, Ganesh Iyer, Adam Marcus
- Oak Ridge National Laboratory:
  - Scott Klasky, Jeremy Logan, David Pugmire
- University of Brasilia, Brazil
  - George Teodoro, Luís F. R. Taveira, Alba C. M. A. Melo

Tissue Image Analysis

Different types of nuclei and cells



Regions of necrosis and angiogenesis

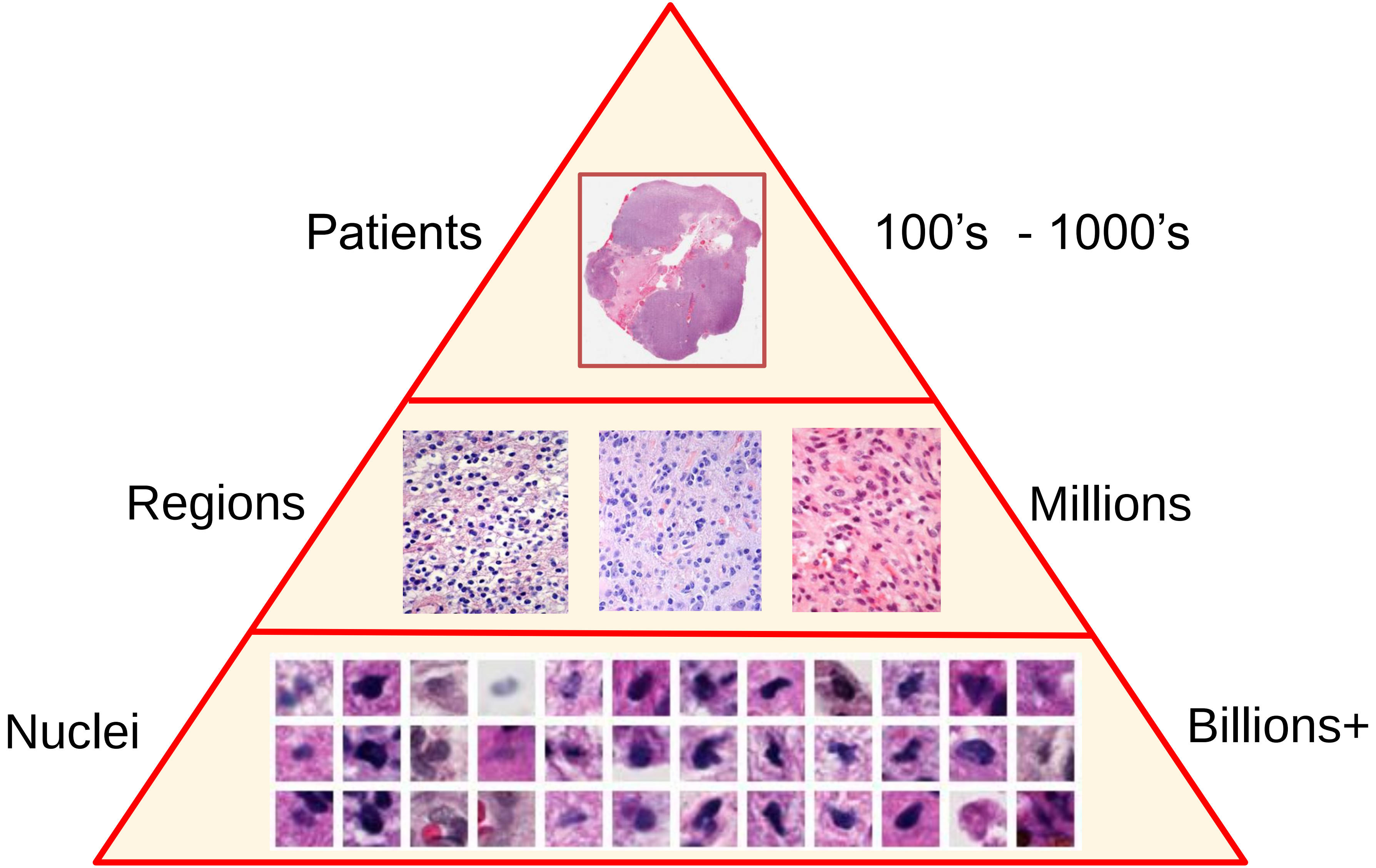


- Tumors are complex and heterogeneous
- Datasets of rich nuclear morphological information
  - Nuclear morphology
  - Maps of tumor infiltrating lymphocytes

Tissue Imaging

- Confluence of several technological advances is making slide scanning more practical
- Image scanning technology has progressed significantly in recent years
  - Sophisticated auto-focus mechanisms
  - Slide trays for batch scanning of slides
- Time required to scan a slide at high-resolution has reduced from multiple hours to several minutes
  - 200-300 slides / day
- Disk storage is getting cheaper

Big Data Challenges in Pathomics



Patients 100’s - 1000’s

Regions Millions

Nuclei Billions+

TCGA Microscopy Image Data

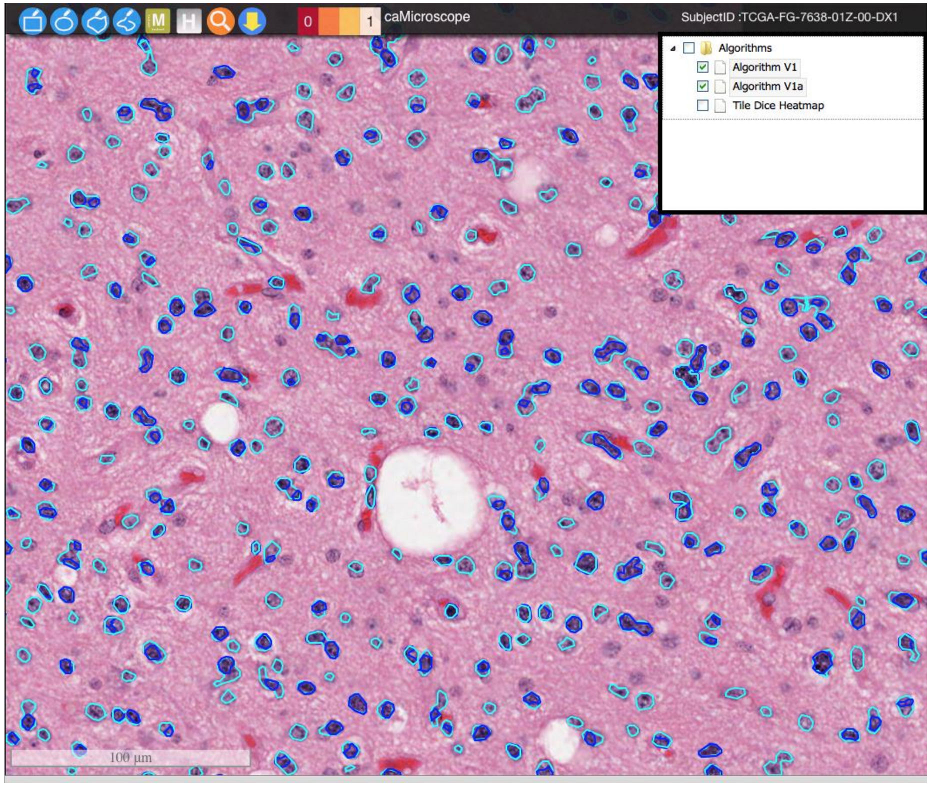
- Data from 11,000 subjects
- 30,000+ tissue slide images
- Image resolutions ranging from 4000x6000 pixels to 130,000x250,000 pixels
  - On average about 3.6 Billion pixels per image

Big Data: Analyzing Whole Slide Tissue Images

- Needs large main memory
  - 130,000x250,000x3 (RGB) = 90GB
- Takes from 30 minutes to 10-12 hours to process an image
  - A dataset with 100 images would require 8 days assuming 2 hours per image on average
- Hundreds of thousands to millions of objects in a whole slide tissue image

Analysis Sensitivity: Generating Robust Feature Sets

- Image analysis pipelines are sensitive to input parameters



□ Algorithms
☑ Algorithm V1
☑ Algorithm V1a
□ Tile Dice Heatmap

Generating Robust Feature Sets

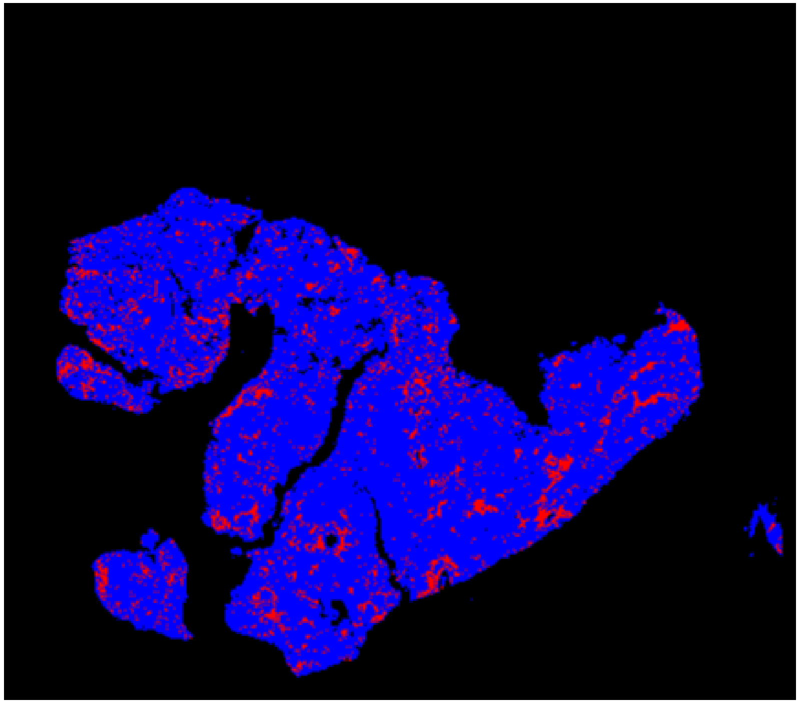
- Run multiple analyses
- Store, index, interact with results
- Computational comparison of results
  - Sensitivity analysis for algorithm evaluation and development
  - Parameter tuning
- Visual comparison of results
  - Curation

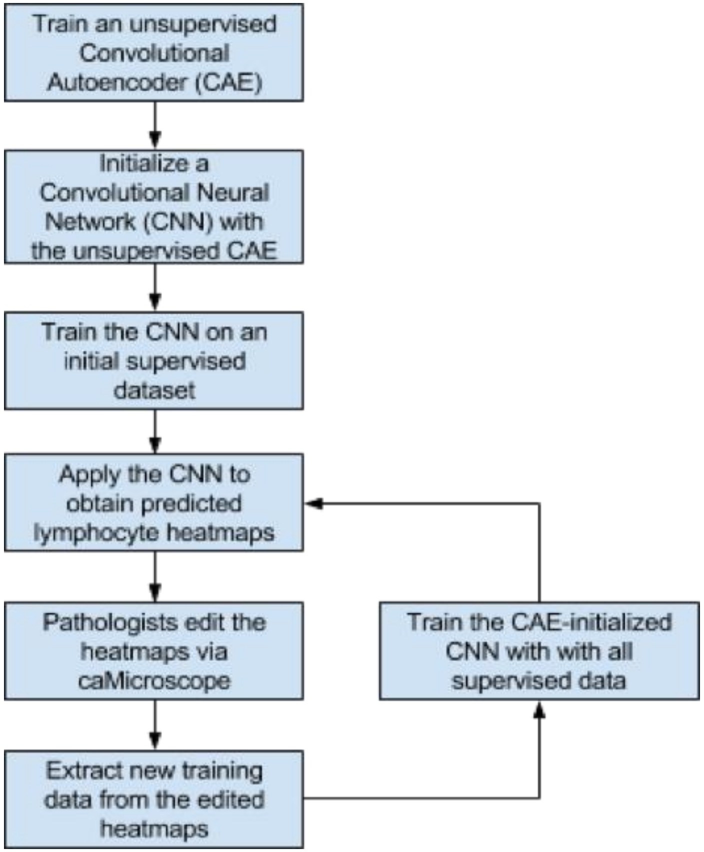
Feature Sets

- 40-70 features per object (nucleus)
- Analysis of 4000 images
  - About 2 Billion segmented objects
- Multi-analysis of 300+ images
  - 6-10 analyses per image
  - About 2 Billion segmented objects

Deep Learning

- Tumor infiltrating lymphocytes





Train an unsupervised Convolutional Autoencoder (CAE)

Initialize a Convolutional Neural Network (CNN) with the unsupervised CAE

Train the CNN on an initial supervised dataset

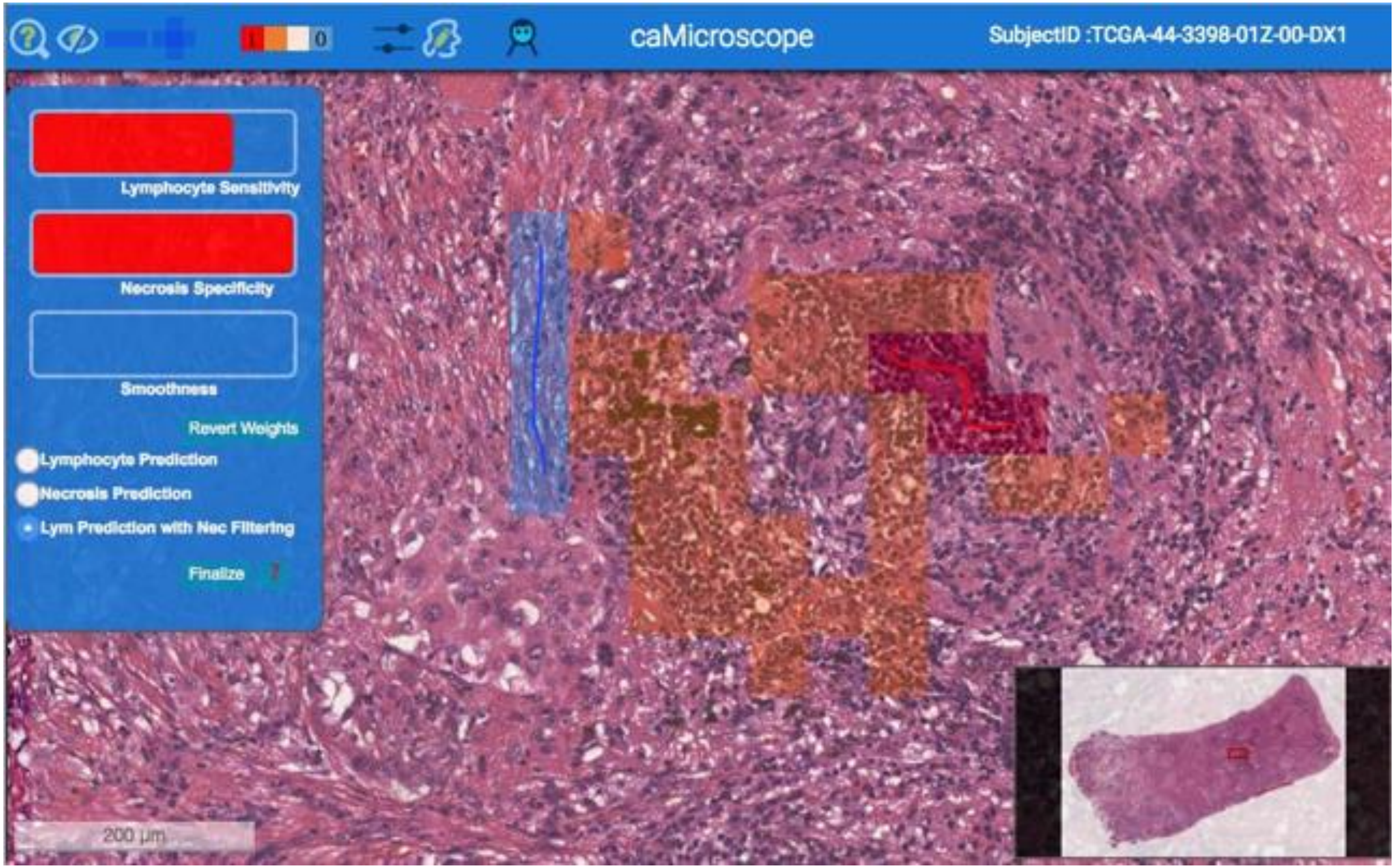
Apply the CNN to obtain predicted lymphocyte heatmaps

Pathologists edit the heatmaps via caMicroscope

Extract new training data from the edited heatmaps

Train the CAE-initialized CNN with with all supervised data

TIL Maps



caMicroscope

SubjectID :TCGA-44-3398-01Z-00-DX1

Lymphocyte Sensitivity

Necrosis Specificity

Smoothness

Revert Weights

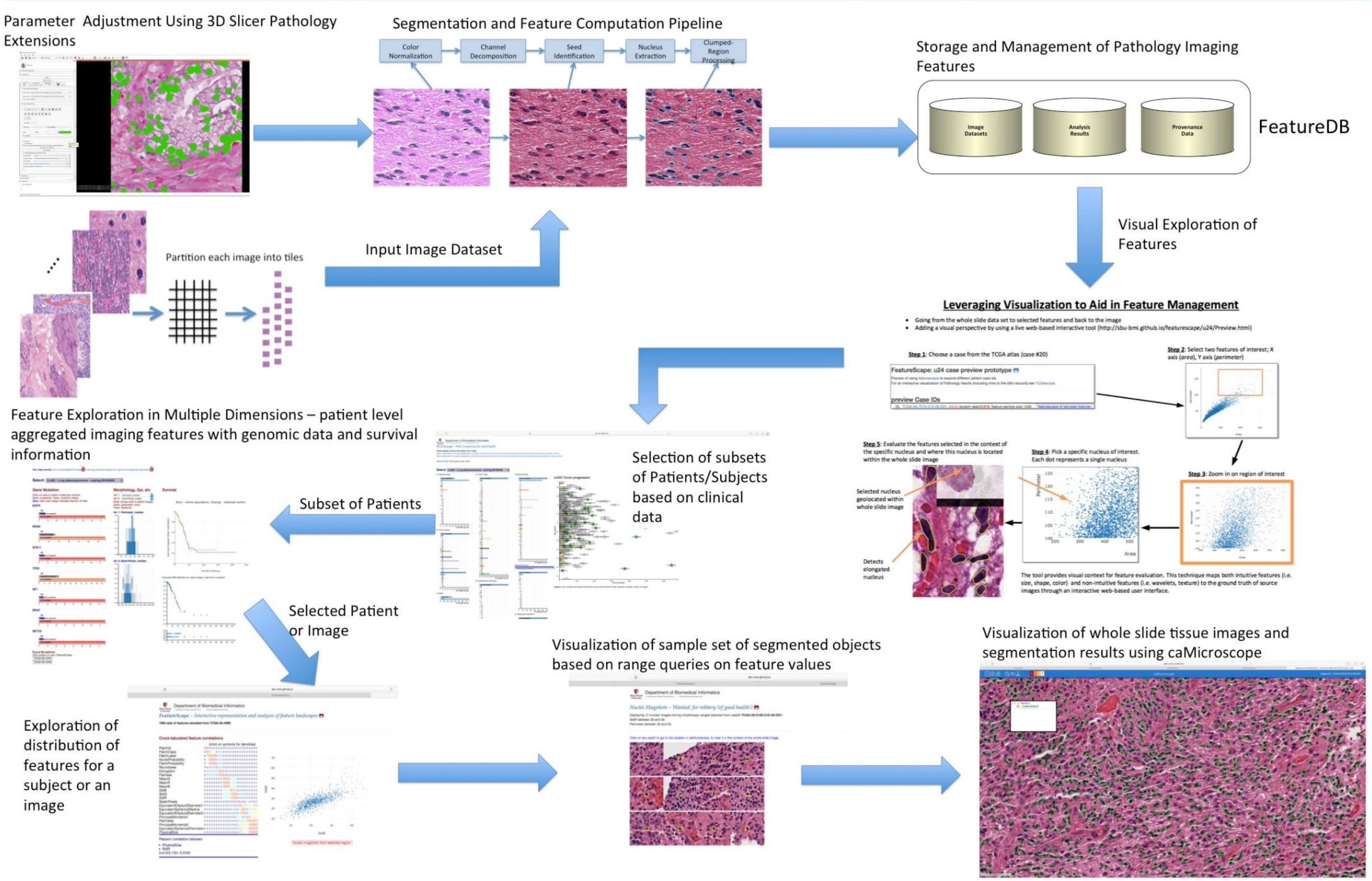
○ Lymphocyte Prediction

○ Necrosis Prediction

● Lym Prediction with Nec Filtering

Finalize

Pathomics Framework



Parameter Adjustment Using 3D Slicer Pathology Extensions

Segmentation and Feature Computation Pipeline

Storage and Management of Pathology Imaging Features

Partition each image into tiles

Input Image Dataset

Visual Exploration of Features

Leveraging Visualization to Aid in Feature Management

- Going from the whole slide data set to selected features and back to the image
- Adding a visual perspective by using a live web-based interactive tool (http://sbu-bmi.github.io/featurescape/u24/Preview.html)

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

FeatureScape: u24 case preview prototype

preview Case IDs

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

Selected nucleus geolocated within whole slide image
Detects elongated nucleus
Perimeter
125
120
115
110
105
100
200
300
400
500
Area

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.

Feature Exploration in Multiple Dimensions – patient level aggregated imaging features with genomic data and survival information

Selection of subsets of Patients/Subjects based on clinical data

Subset of Patients

Selected Patient or Image

Exploration of distribution of features for a subject or an image

Department of Biomedical Informatics

FeatureScape – Interactive representation and analysis of feature landscapes

Cross-tabulated feature correlations

Visualization of sample set of segmented objects based on range queries on feature values

Department of Biomedical Informatics

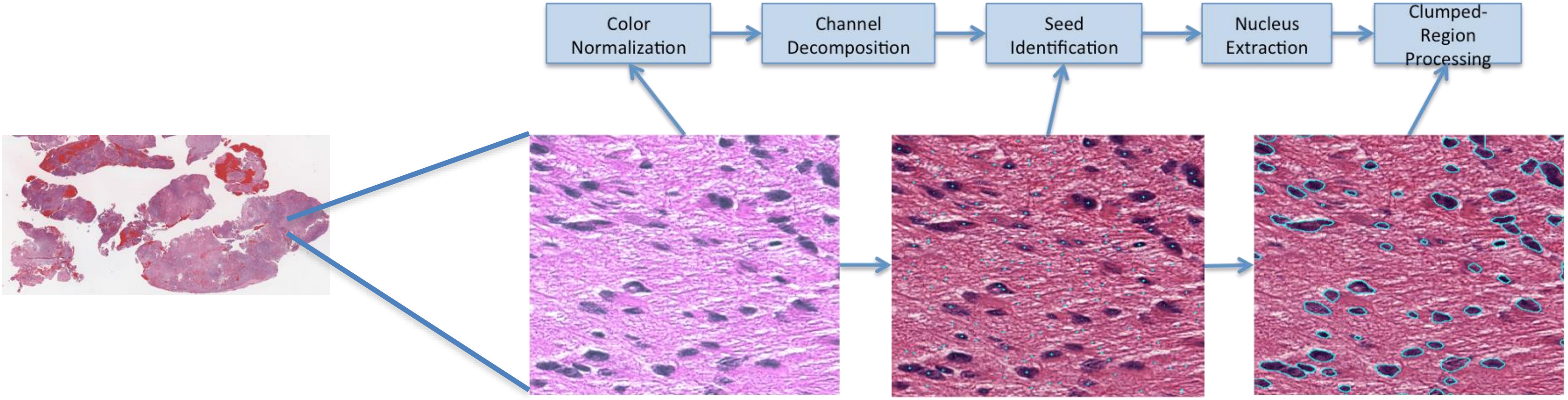
Nuclei Mugshots – Wanted: for robbery (of good health!)

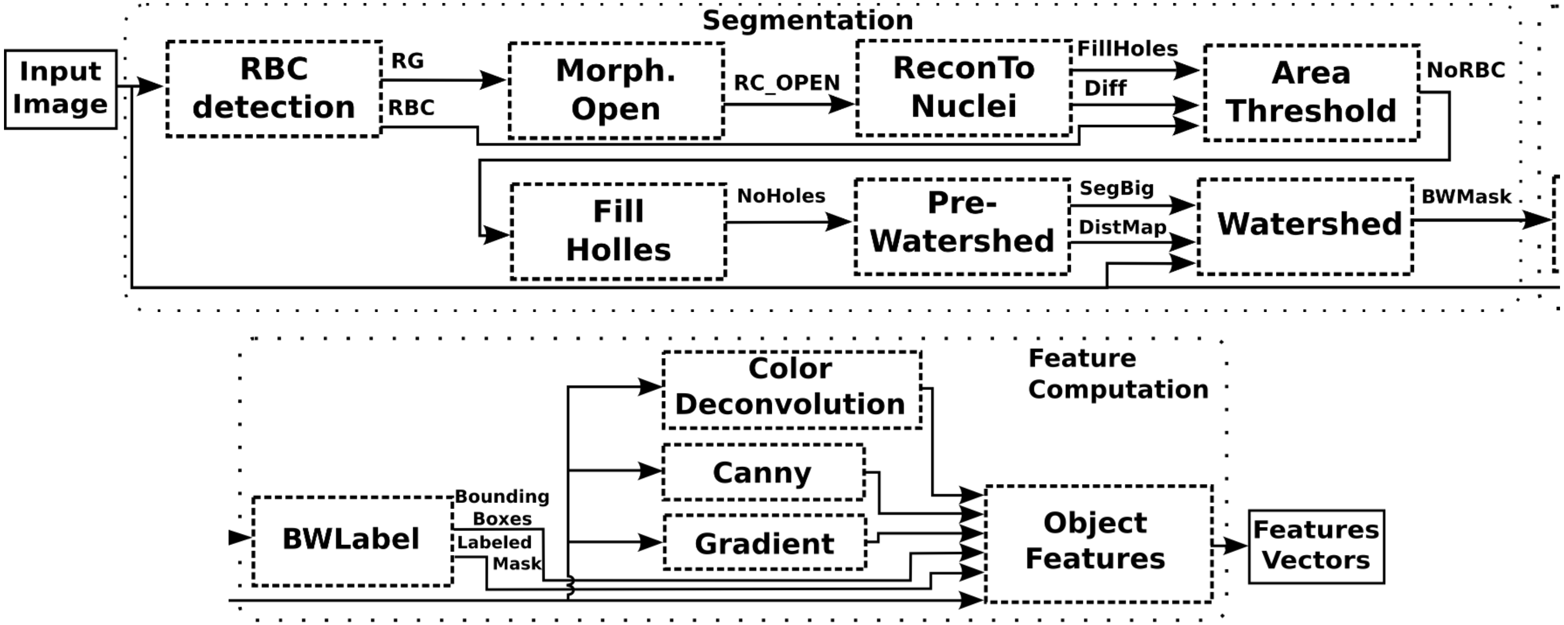
Visualization of whole slide tissue images and segmentation results using caMicroscope

Leveraging Hybrid High Performance Computing Systems

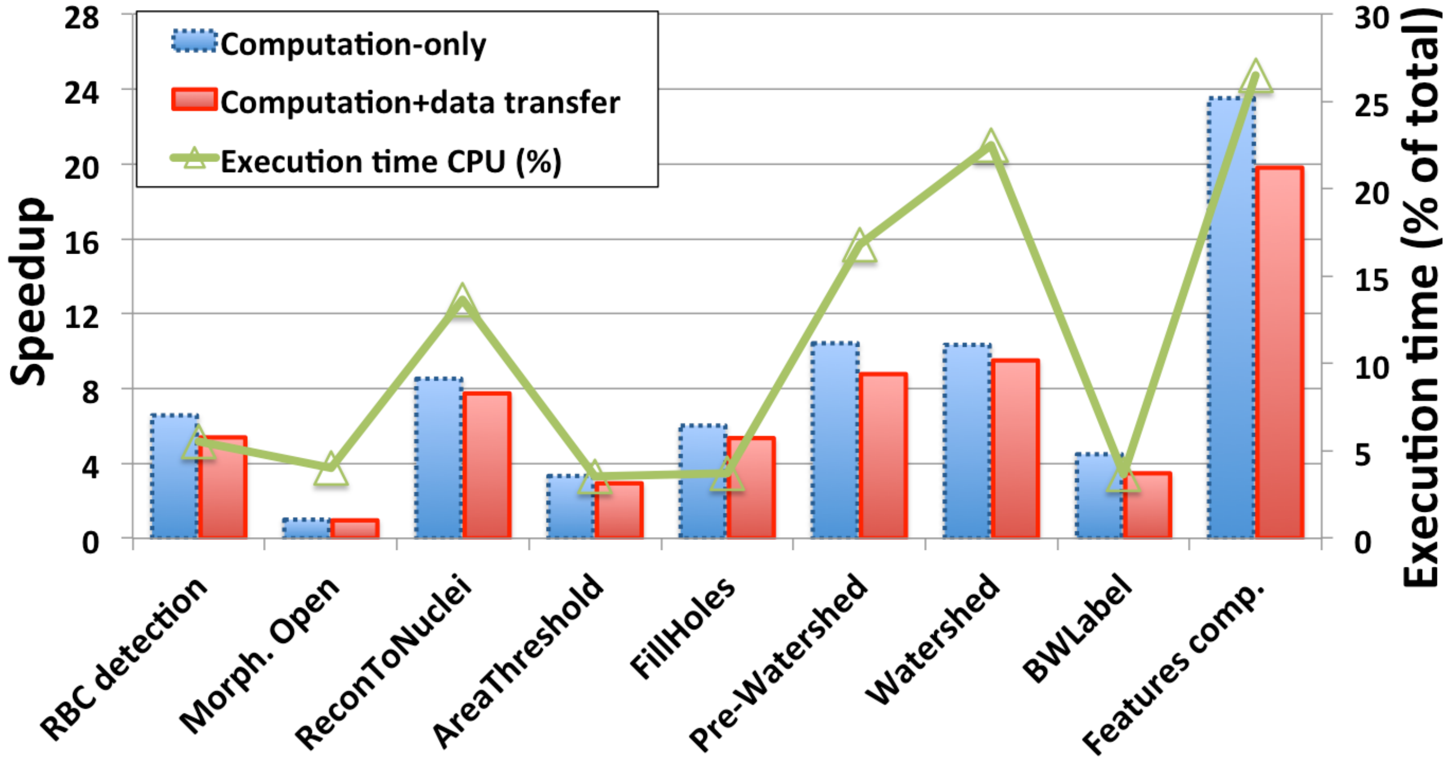
- Lots of nodes
- Multi-core CPUs, GPUs, and other hardware accelerators per node
- XSEDE, Supercomputing Centers

Nuclear Imaging Features



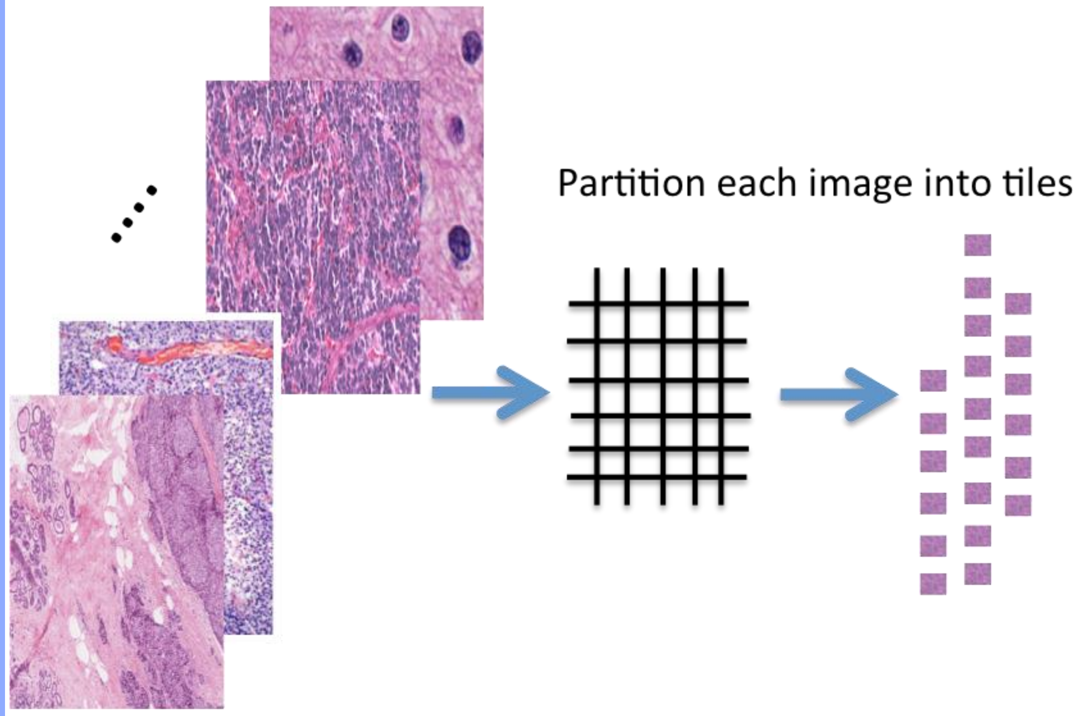


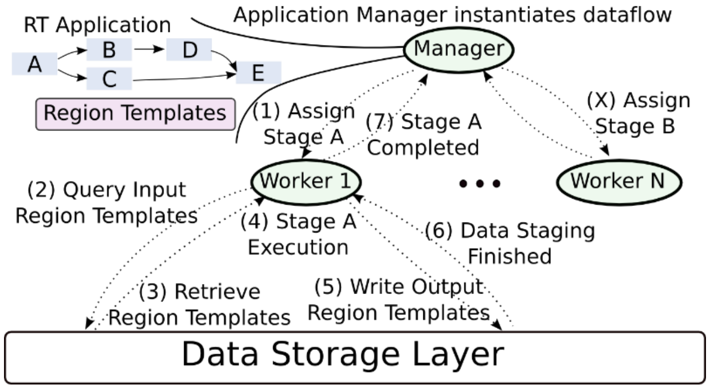
Inter-operation Performance Variations



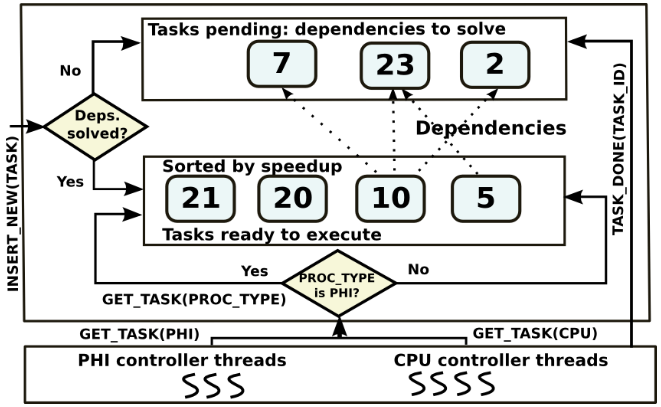
Experiments on Keeneland: equipped w/ M2090 GPUs

Analysis on HPC Platforms



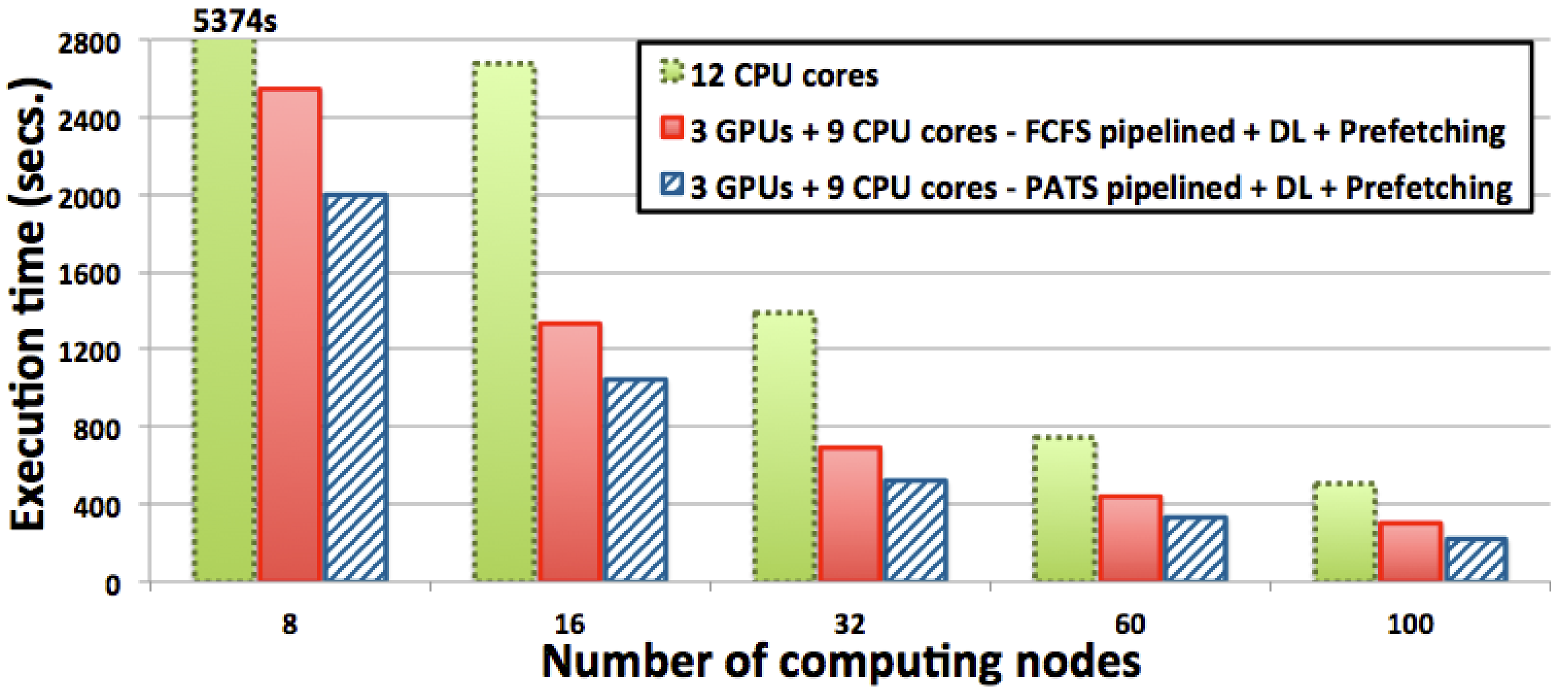


RT Application
A
B
C
D
E
Region Templates
Application Manager instantiates dataflow
Manager
(1) Assign Stage A
(7) Stage A Completed
(X) Assign Stage B
Worker 1
Worker N
(2) Query Input Region Templates
(4) Stage A Execution
(6) Data Staging Finished
(3) Retrieve Region Templates
(5) Write Output Region Templates
Data Storage Layer



- Thousands of CPU cores
- Hundreds of GPUs
- Large distributed memory

Coordinated Use of CPUs and GPUs

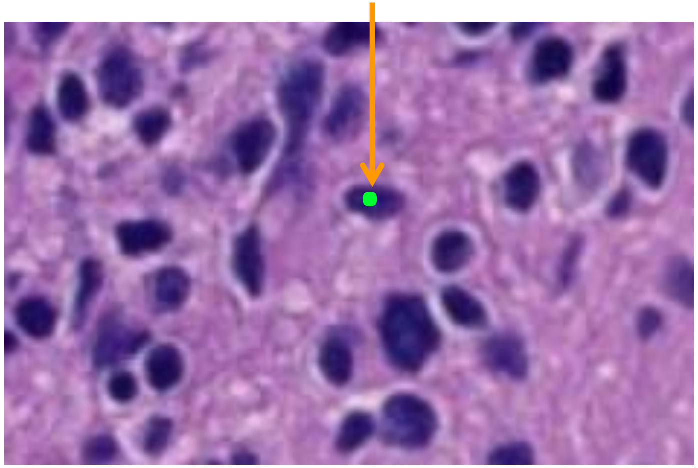


36,844 4Kx4K-pixel tiles from 340 whole slide tissue images.

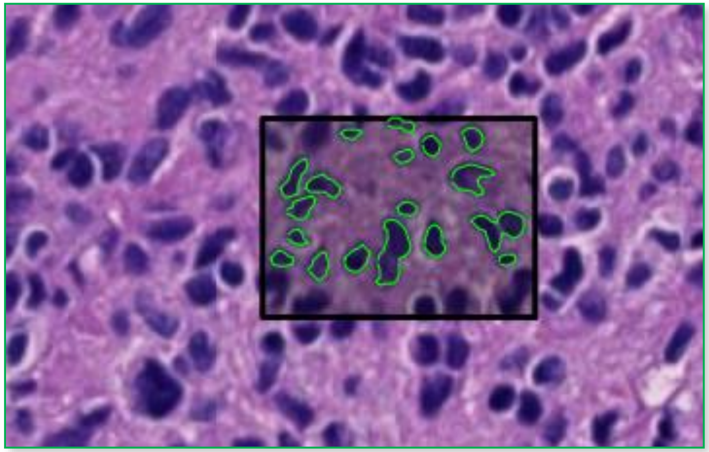
On 100 nodes, less than 4 minutes to process 36K tiles.

Support for Spatial Queries

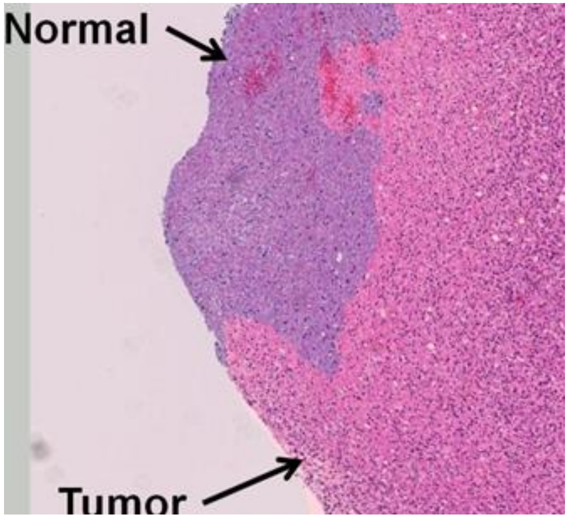
POINT query: human marked point inside a nucleus



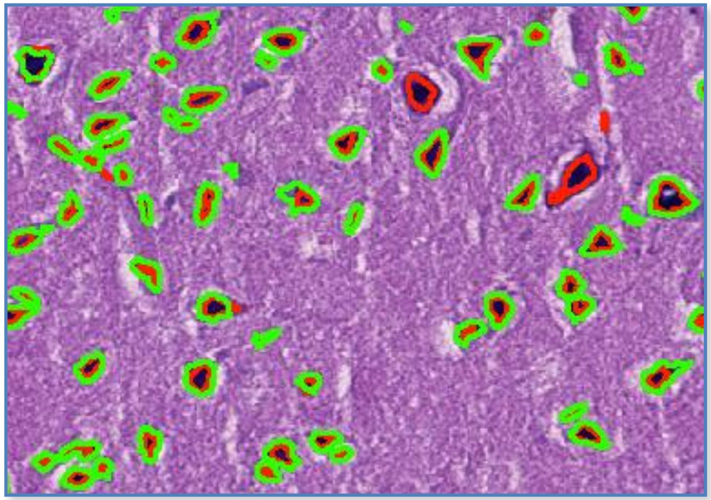
WINDOW query: return markups contained in a rectangle



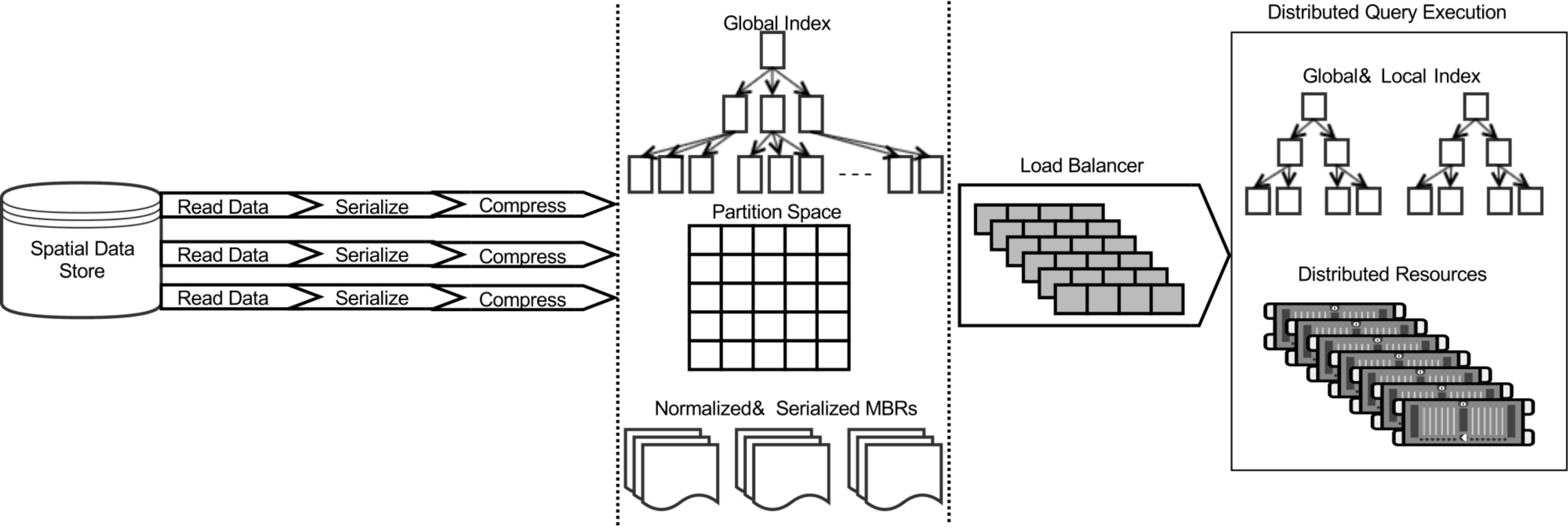
CONTAINMENT query: nuclear feature aggregation in tumor regions



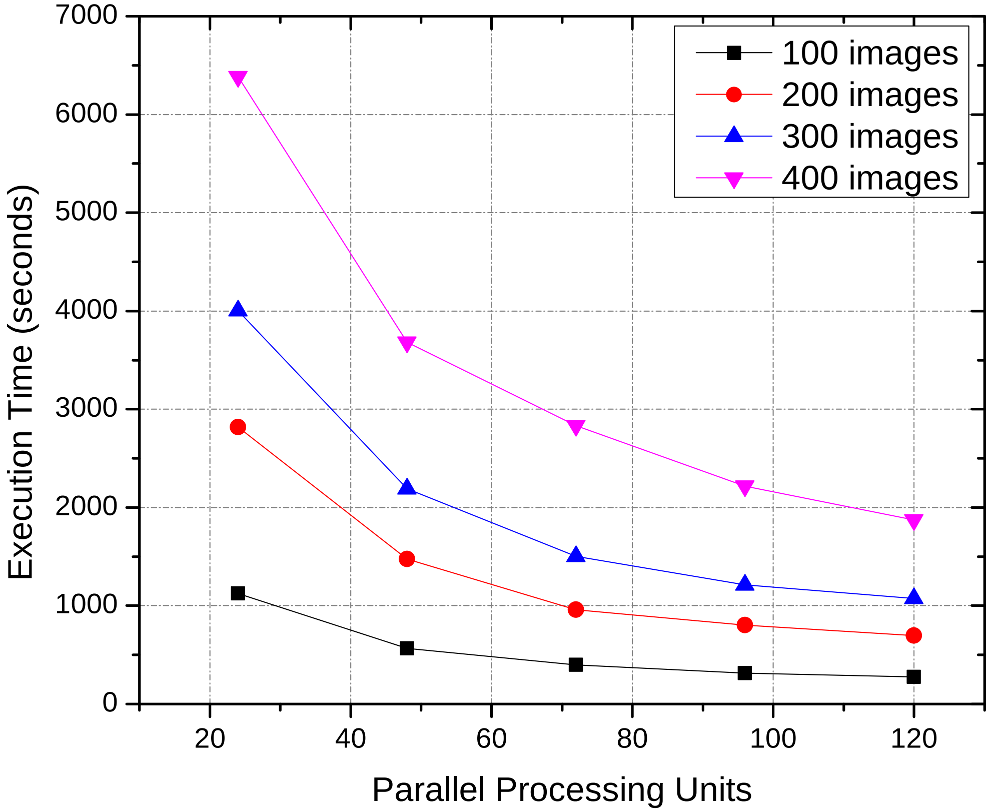
SPATIAL JOIN query: algorithm validation/comparison



SparkGIS


Spatial Data Store
Read Data
Serialize
Compress
Read Data
Serialize
Compress
Read Data
Serialize
Compress
Global Index
- - -
Partition Space
Normalized& Serialized MBRs
Load Balancer
Distributed Query Execution
Global& Local Index
Distributed Resources

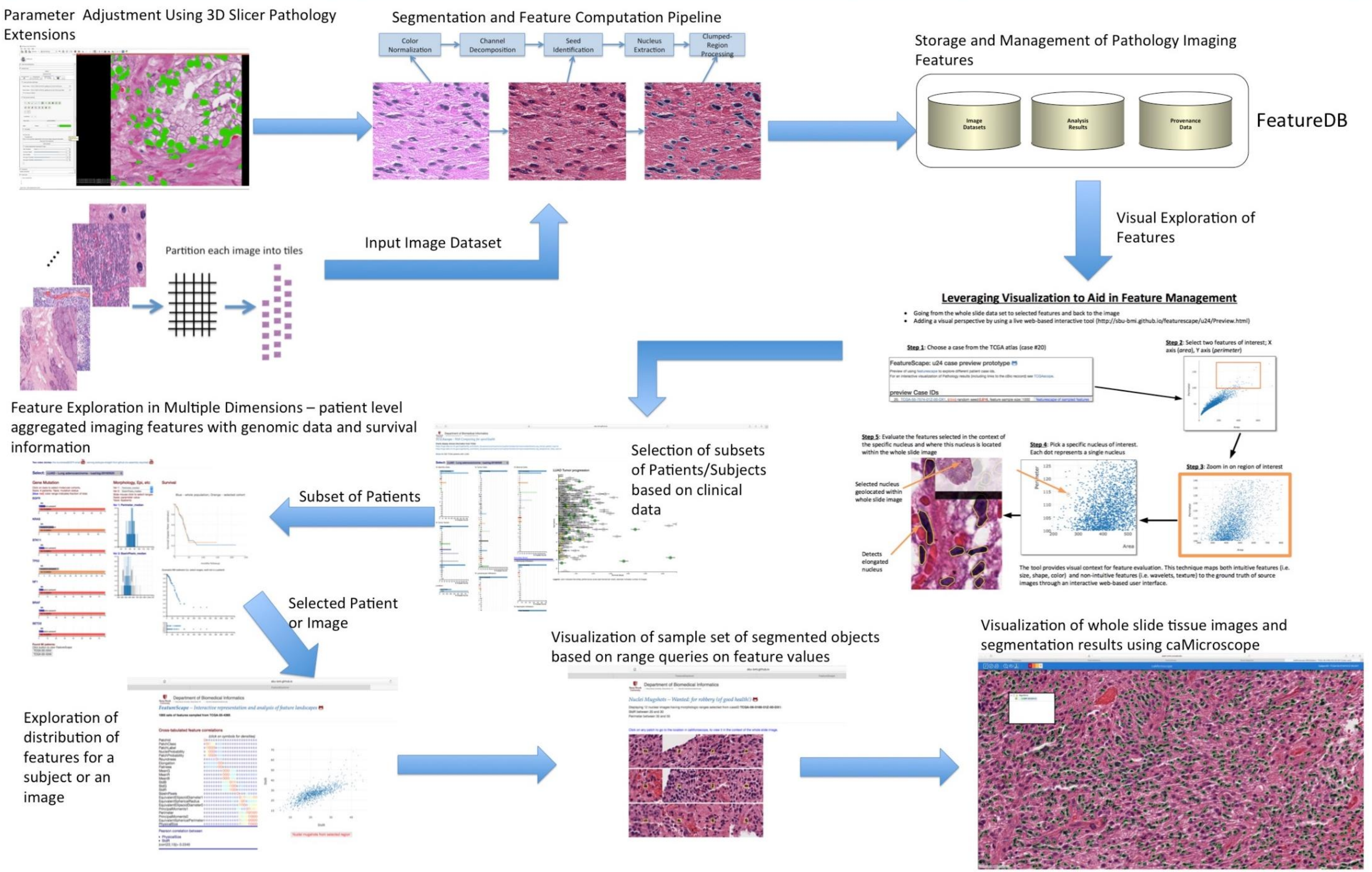
Heatmap Computations: Spatial Joins + Dice/Jaccard Metric



Going Forward

- Containers
  - Facilitates a modular design.
  - Self-contained, Isolated.
    - Software to support tissue image analysis needs to leverage a variety of existing libraries and tools.
  - Flexibility
    - Move computation to data
    - Move back and forth between Cloud platforms and local resources
- Scripting
  - JavaScript: Take advantage of web browsers
  - Python: Take advantage of large set of libraries

Pathomics Framework



Parameter Adjustment Using 3D Slicer Pathology Extensions

Segmentation and Feature Computation Pipeline

Storage and Management of Pathology Imaging Features

Partition each image into tiles

Input Image Dataset

Visual Exploration of Features

Leveraging Visualization to Aid in Feature Management

- Going from the whole slide data set to selected features and back to the image
- Adding a visual perspective by using a live web-based interactive tool (http://sbu-bmi.github.io/featurescape/u24/Preview.html)

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

Selected nucleus geolocated within whole slide image
Detects elongated nucleus

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.

Selection of subsets of Patients/Subjects based on clinical data

Subset of Patients

Feature Exploration in Multiple Dimensions – patient level aggregated imaging features with genomic data and survival information

Selected Patient or Image

Exploration of distribution of features for a subject or an image

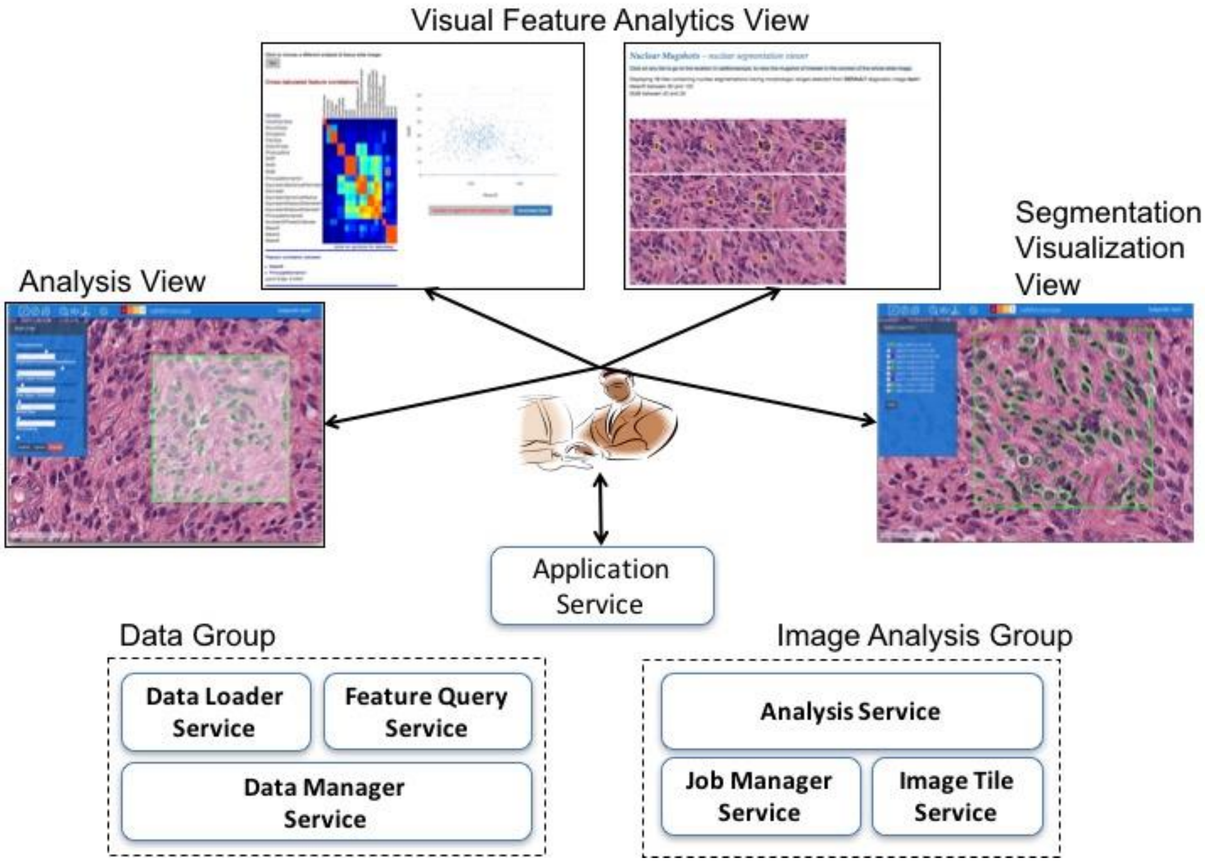
Department of Biomedical Informatics
FeatureScape – Interactive representation and analysis of feature landscapes
Cross-tabulated feature correlations

Visualization of sample set of segmented objects based on range queries on feature values

Department of Biomedical Informatics
Nuclei Mugshots – Wanted: for robbery (of good health!)

Visualization of whole slide tissue images and segmentation results using caMicroscope

Containerized Software for Tissue Image Analysis



Visual Feature Analytics View

Analysis View

Segmentation Visualization View

Application Service

Data Group

Data Loader Service

Feature Query Service

Data Manager Service

Image Analysis Group

Analysis Service

Job Manager Service

Image Tile Service

Containerized Software for Tissue Image Analysis

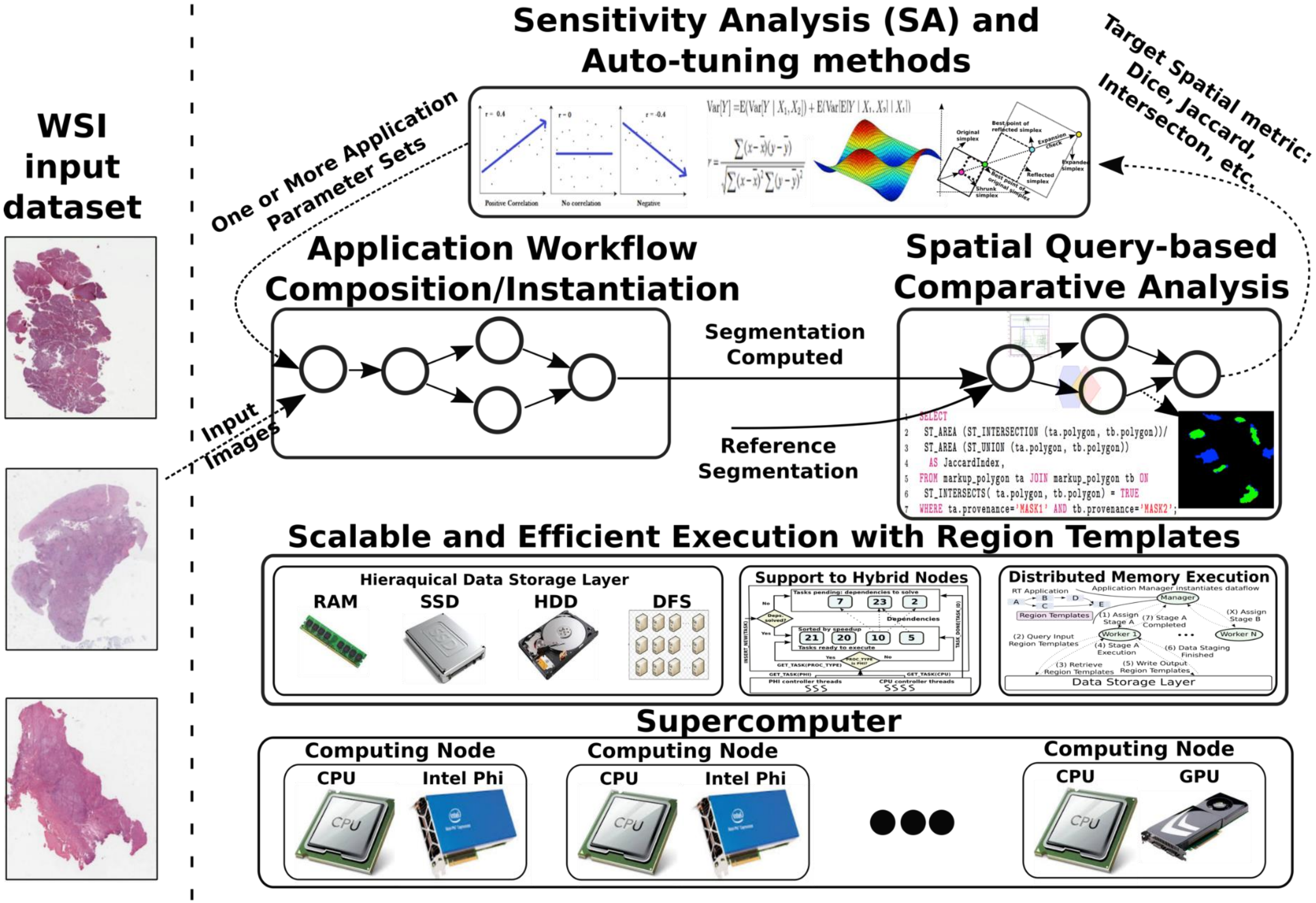
- ***The application service group*** is a single container that hosts a suite of Web applications to view images and interact with analysis results.
- ***The image analysis group*** is made up of three containers, which collectively execute image analysis requests.
  - Analysis service – hosts analysis pipeline
  - Job Manager service – tracks jobs
  - Image tile service – services image tiles for analysis
- ***The data management service group*** is implemented as a set of three containers for data loading, data management, and query processing.
  - Data loader service -- load image metadata and analysis results
  - Data manager service – manage and index image data, analysis results, features
  - Feature query service – query feature data for visualization and exploration

Software: https://github.com/SBU-BMI/quip_distro.git

Thank you.

This work was supported in part by 1U24CA180924-01A1 from the NCI and by R01LM011119-01 and R01LM009239 from the NLM and NCIP/Leidos 14X138

Support for Sensitivity Analysis



$$\mathrm{Var}[Y] = E(\mathrm{Var}[Y \mid X_1, X_2]) + E(\mathrm{Var}[E[Y \mid X_1, X_2] \mid X_1])$$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

```sql
SELECT
 ST_AREA (ST_INTERSECTION (ta.polygon, tb.polygon))/
 ST_AREA (ST_UNION (ta.polygon, tb.polygon))
  AS JaccardIndex,
FROM markup_polygon ta JOIN markup_polygon tb ON
 ST_INTERSECTS( ta.polygon, tb.polygon) = TRUE
WHERE ta.provenance='MASK1' AND tb.provenance='MASK2';
```