caBIG™ cancer Biomedical Informatics Grid™ an initiative of the National Cancer Institute

Squeezing Information from Data at Exascale

Joel Saltz

Emory University

Georgia Tech

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
- Much can be done at “data staging time”

NATIONAL CANCER INSTITUTE

caBIG cancer Biomedical Informatics Grid

Overview

- Integrative biomedical informatics analysis
  - feature sets obtained from Pathology and Radiology studies
- *This is the same CS problem as what we have seen in Oil Reservoir/Seismic analyses, astrophysics and in Computational Fluid Dynamics*
- Techniques, tools and methodologies for derivation, management and analysis of feature sets
- Ideas for how to move to exascale

Examples

Astrophysics	*Which portions of a star's core are susceptible to implosion over time period [t1, t2] ?*	Compute streamlines on vector field *v* within grid points [(x1,y1)-(x2,y2)]
Material Science	*Is crystalline growth likely to occur within range [p1, p2] of pressure conditions ?*	Compute likelihood of local cyclic relationships among nanoparticles within a frame
Cancer studies	*Which regions of the tumor are undergoing active angiogenesis in response to hypoxia ?*	Determine image regions where (blood vessel density > 20) and (nuclei and necrotic region are within 50 microns of each other)

Neuro-
imaging

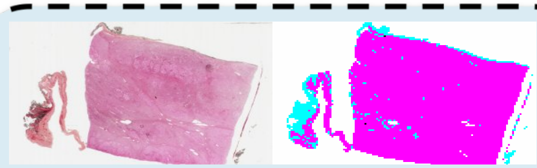
Typical data analysis scenario

Transformation of raw image data

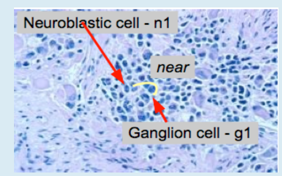
- Normalization: illumination.
- Spatial Alignment: displacements
- Stitching: seamless image mosaic
- Warping: standard template / canonical atlas
- …

Analysis

- Pixel-based computing
- Color decomposition
- Correcting for non uniform staining



computing
- Segmentation
- Feature extraction, classification



- Annotation of data
- Semantic querying
- Image mining

Data volume decreases; Data complexity & domain specificity increase

INTEGRATIVE BIOMEDICAL INFORMATICS ANALYSIS

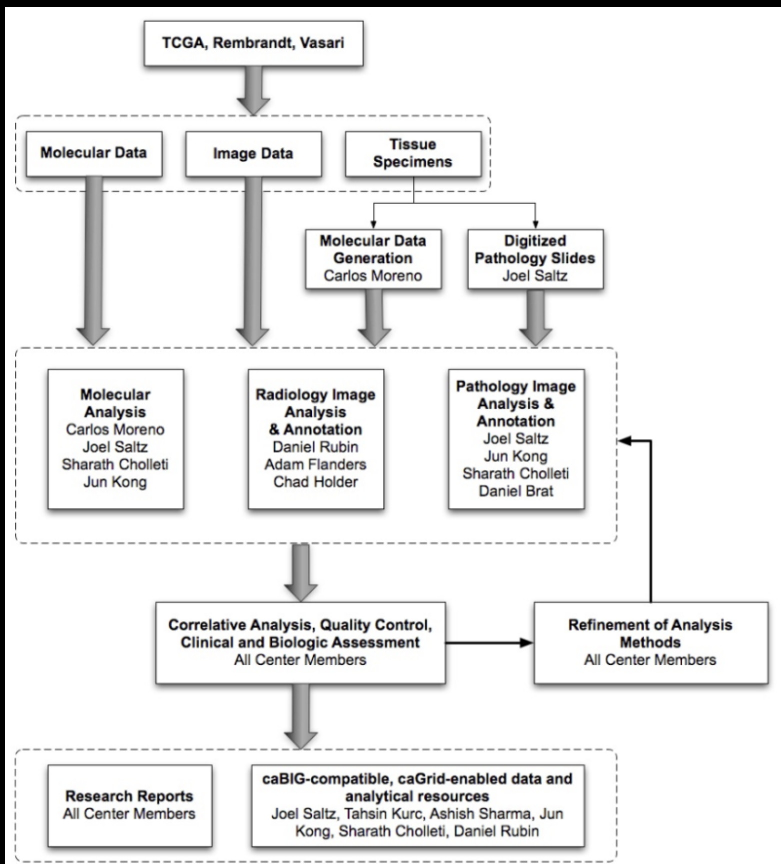
Reproducible anatomic/functional characterization at gross level (Radiology) and fine level (Pathology)

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

In Silico Center – Application Driven Computer Science (with National Cancer Institute flavor)

In Silico Center for Brain Tumor Research



EMORY UNIVERSITY

Specific Aims:

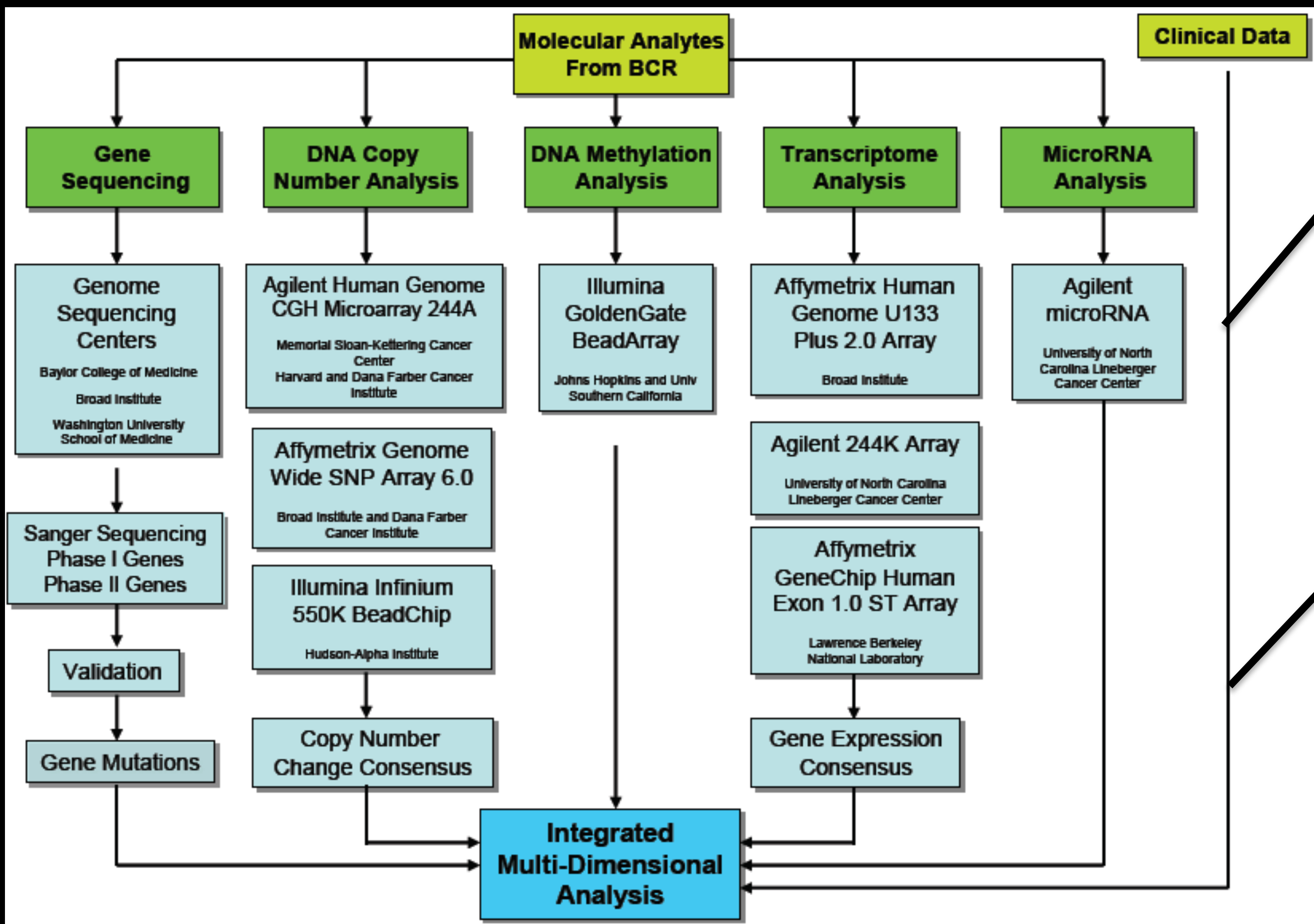
1. Influence of necrosis/ hypoxia on gene expression and genetic classification.

2. Molecular correlates of high resolution nuclear morphometry.

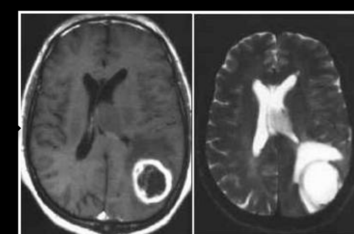
3. Gene expression profiles that predict glioma progression.

4. Molecular correlates of MRI enhancement patterns.

TCGA Research Network



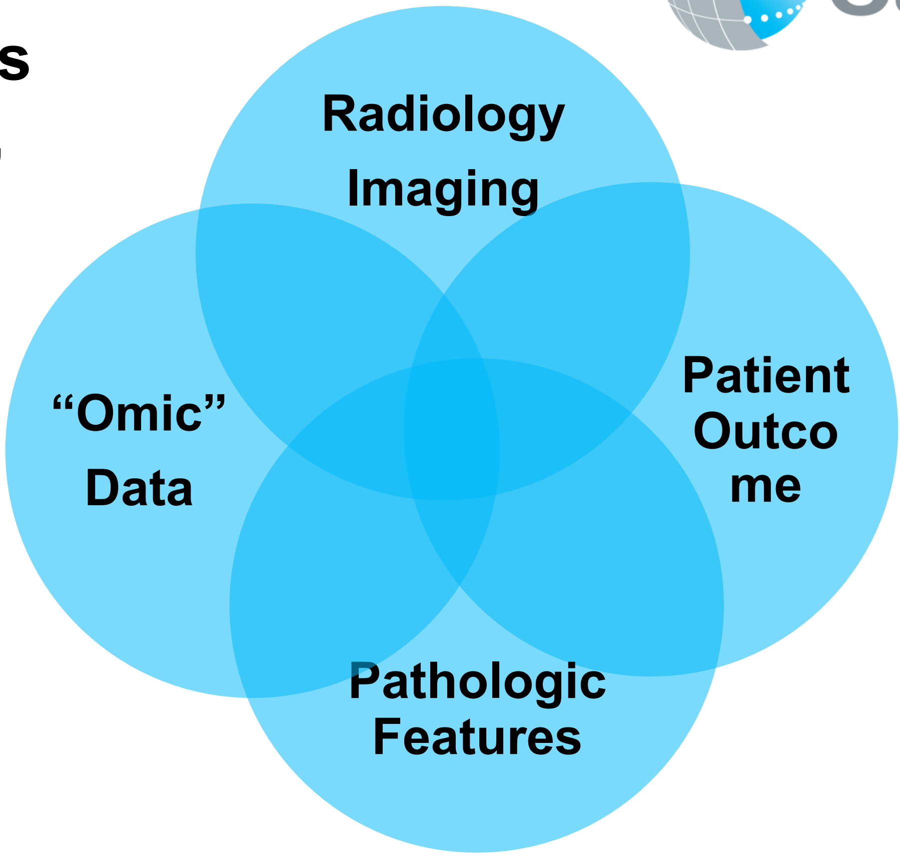
Digital Pathology

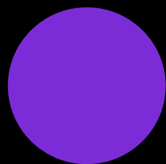


Neuroimaging

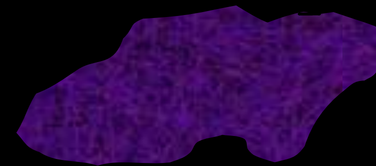
Integration of heterogeneous multiscale information

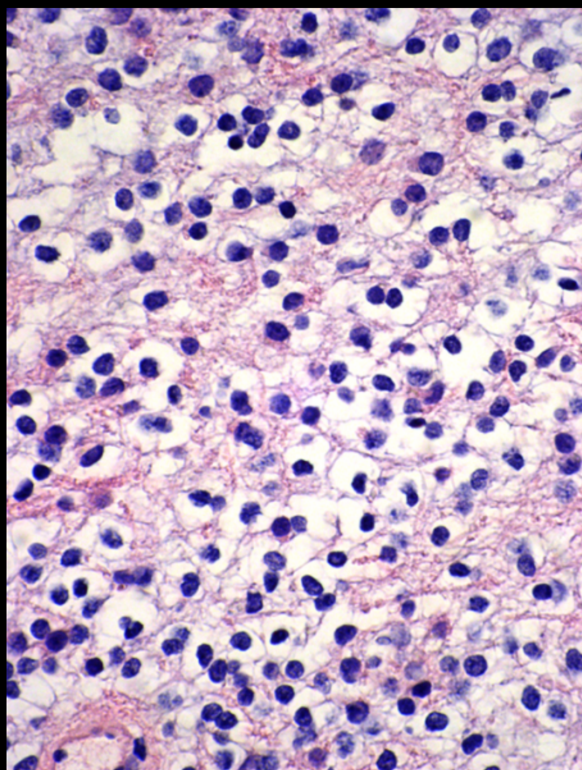
- **Coordinated initiatives Pathology, Radiology, "omics"**
- **Exploit synergies between all initiatives to improve ability to forecast survival & response.**

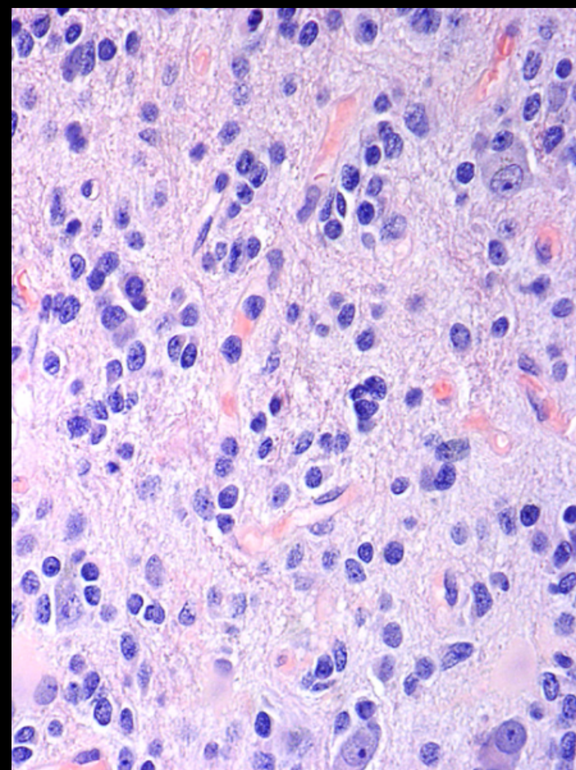


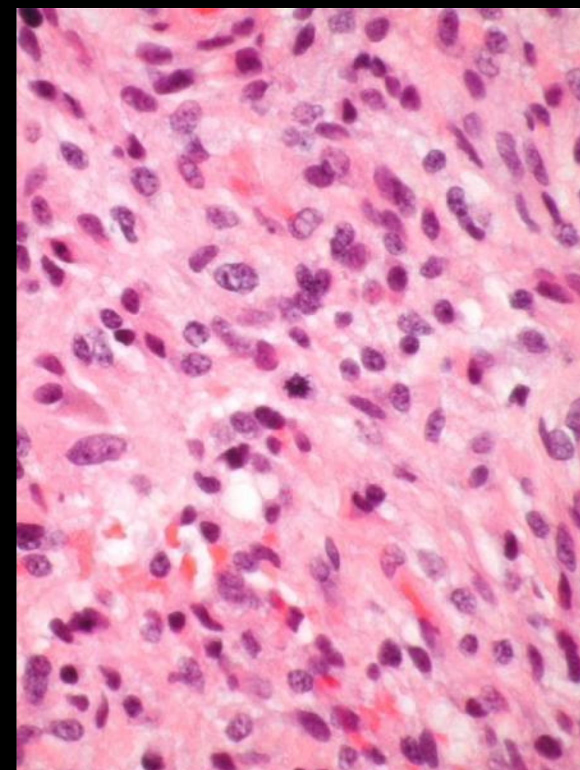


Nuclear Qualities





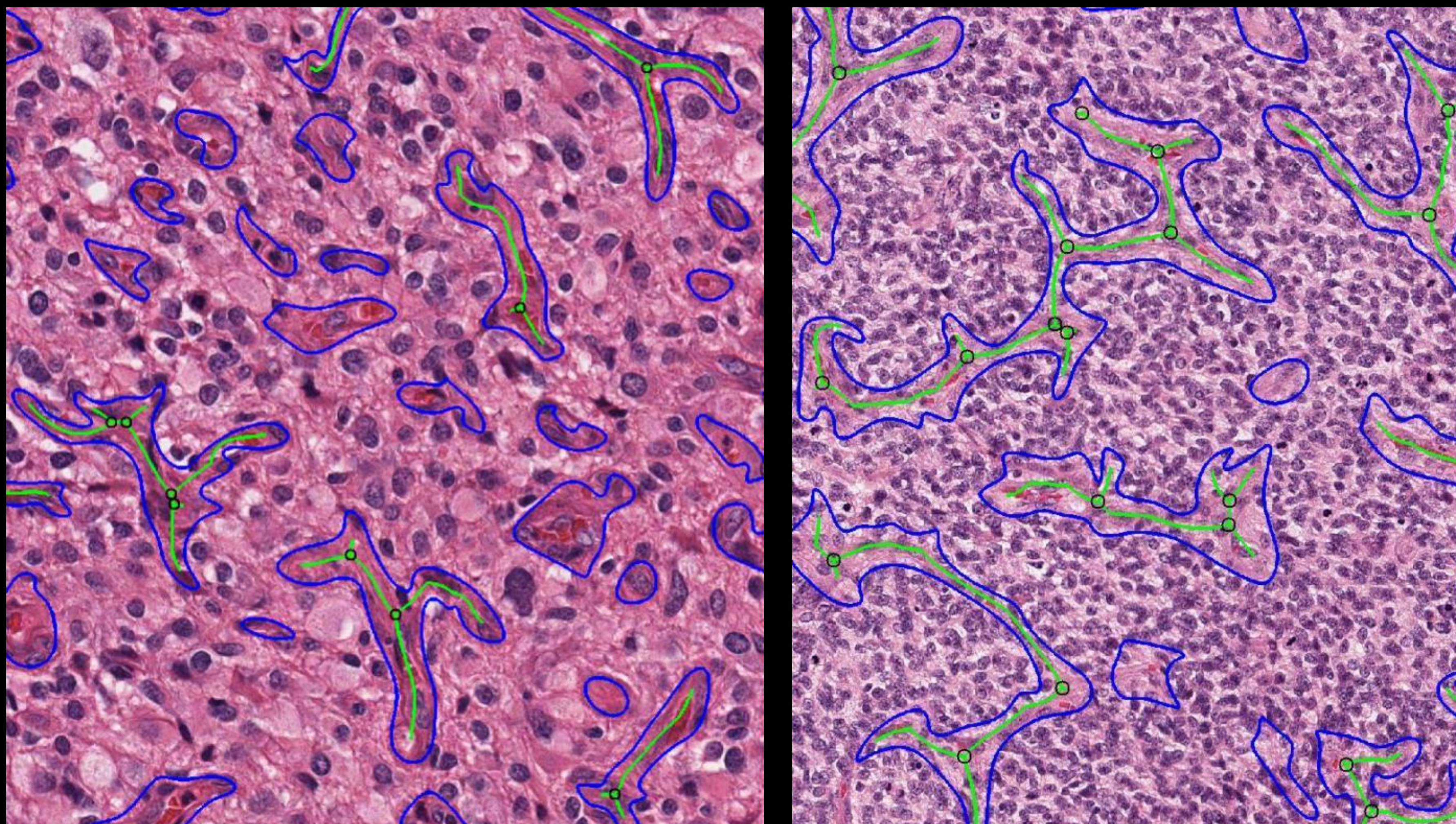




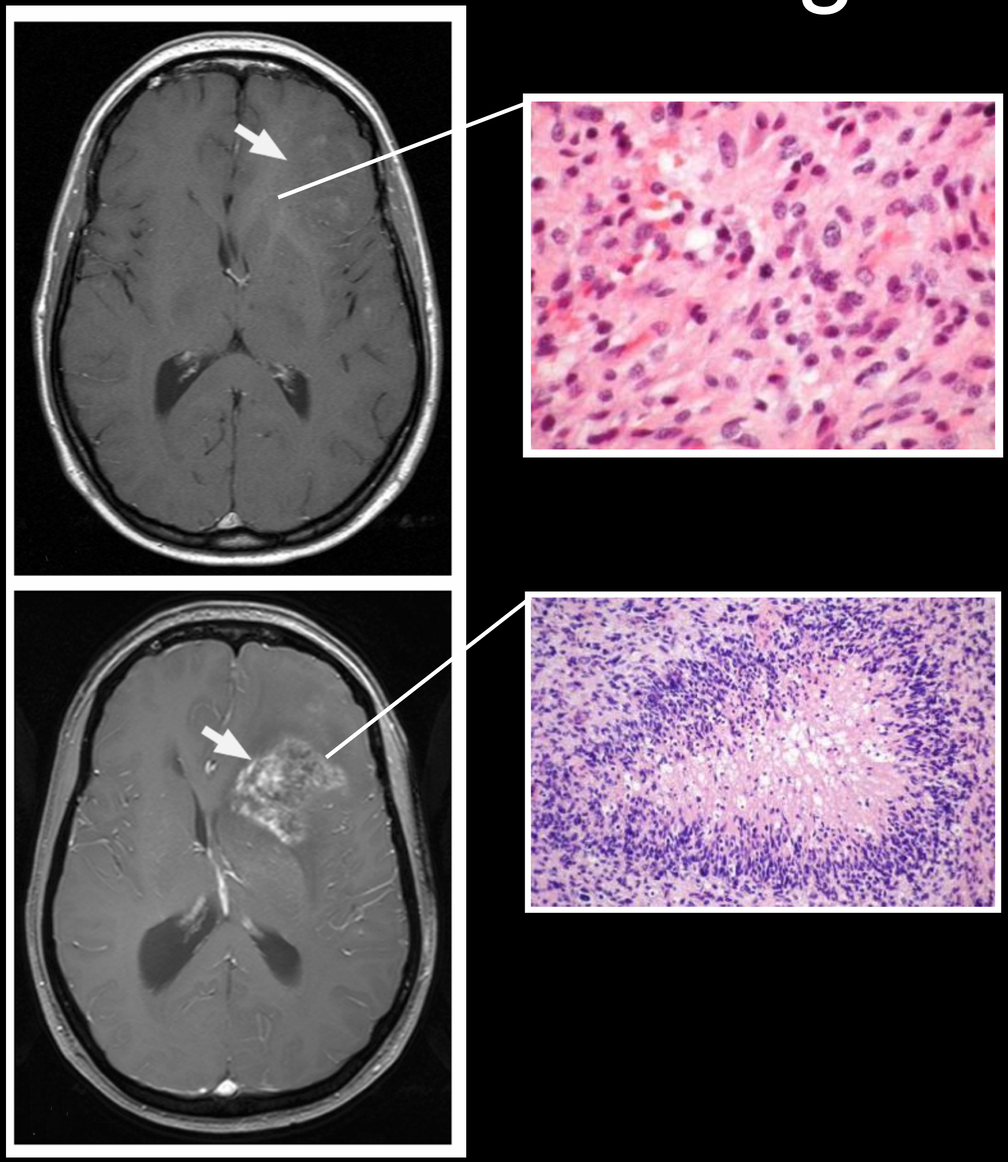
Oligodendroglioma **Astrocytoma**

Vessel Characterization

- Bifurcation detection



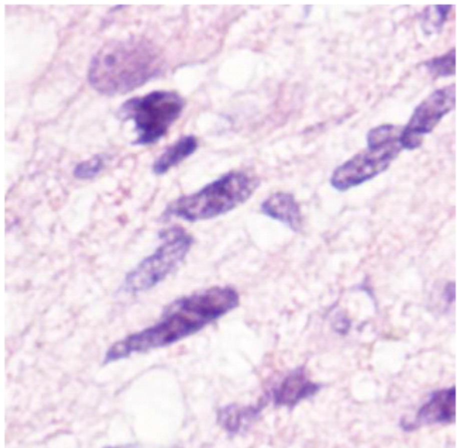
Progression to GBM



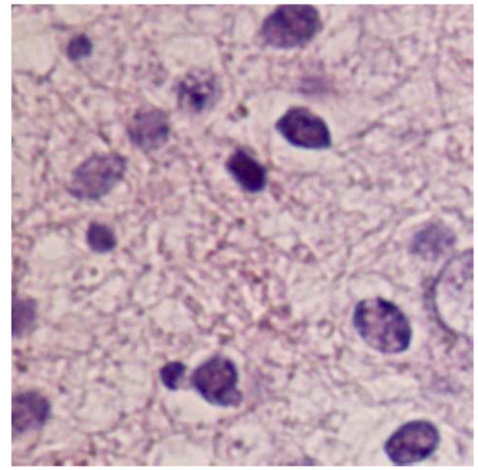
Anaplastic Astrocytoma (WHO grade III)

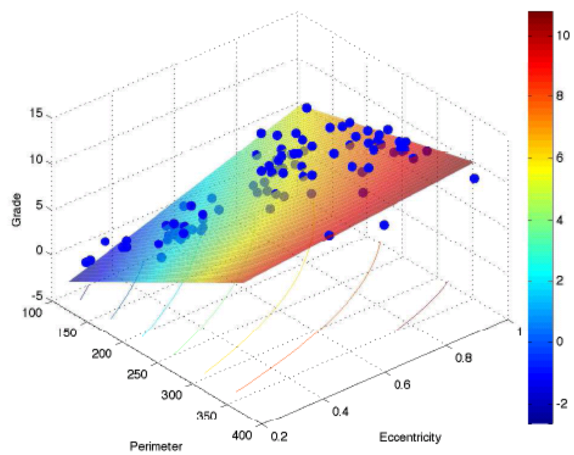
Glioblastoma (WHO grade IV)

Astrocytoma vs Oligodendroglima
Overlap in genetics, gene expression, histology



(a)



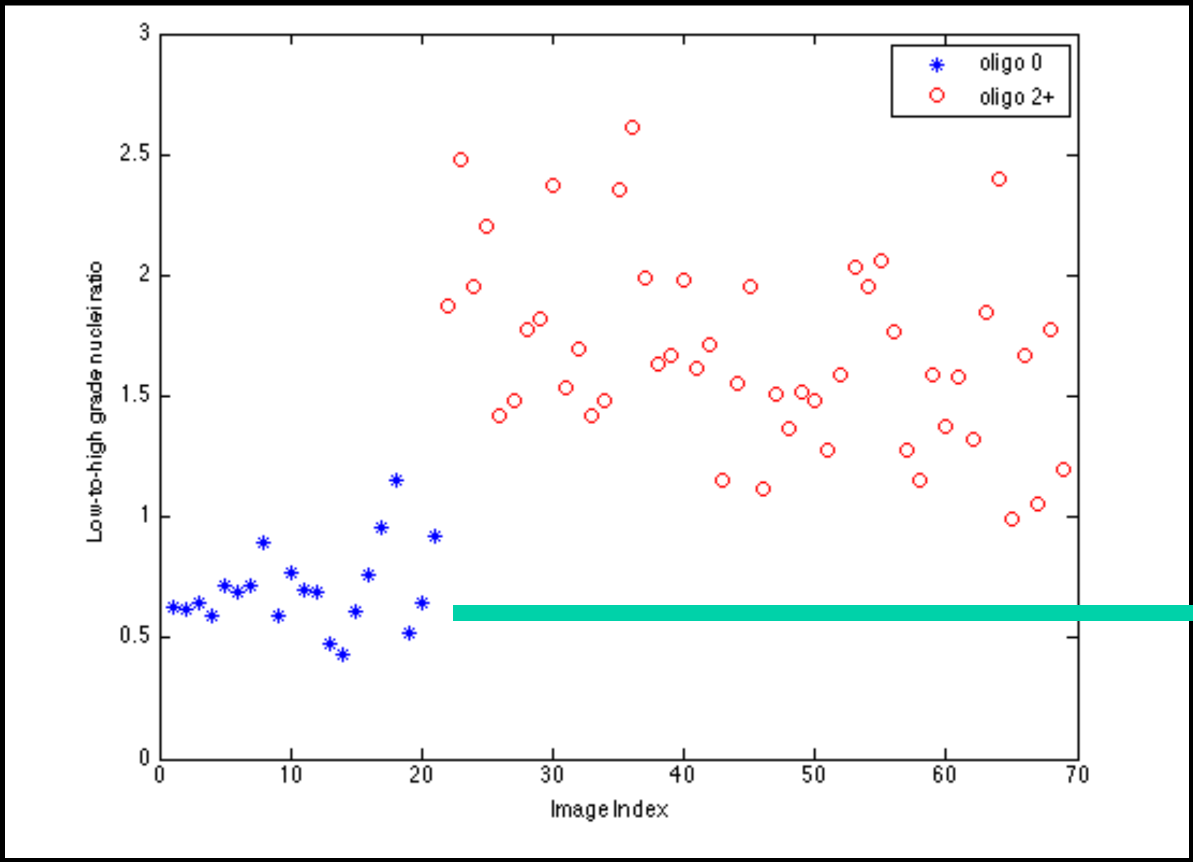


Astrocytoma vs Oligodendroglima

- Assess nuclear size (area and perimeter), shape (eccentricity, circularity major axis, minor axis, Fourier shape descriptor and extent ratio), intensity (average, maximum, minimum, standard error) and texture (entropy, energy, skewness and kurtosis).

Machine-based Classification of TCGA GBMs (J Kong)

Whole slide scans from 14 TCGA GBMS (69 slides)
7 purely astrocytic in morphology; 7 with 2+ oligo component
399,233 nuclei analyzed for astro/oligo features
Cases were categorized based on ratio of oligo/astro cells

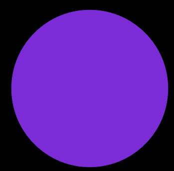


TCGA Gene Expression Query: *c-Met* overexpression

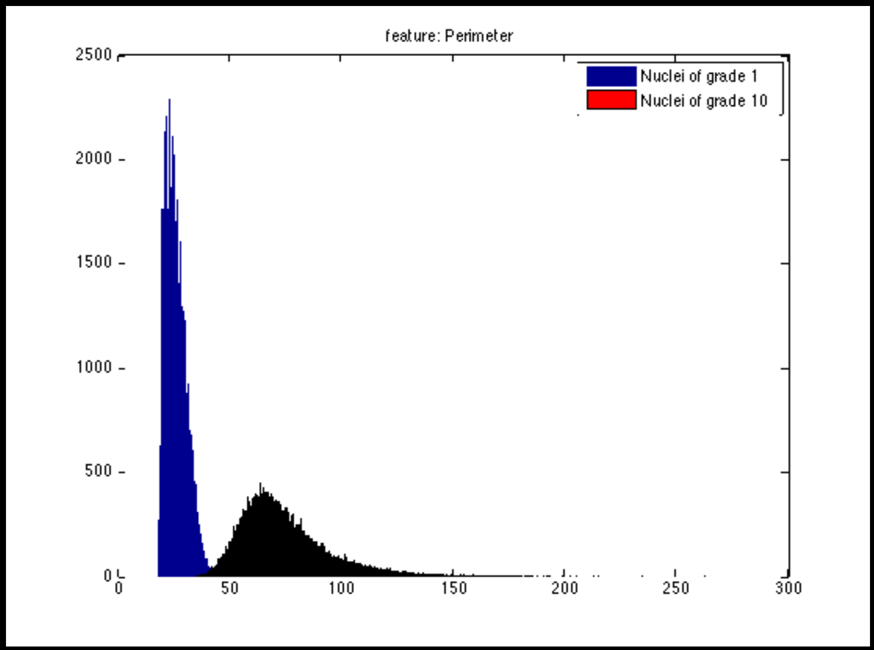
Classification Performance

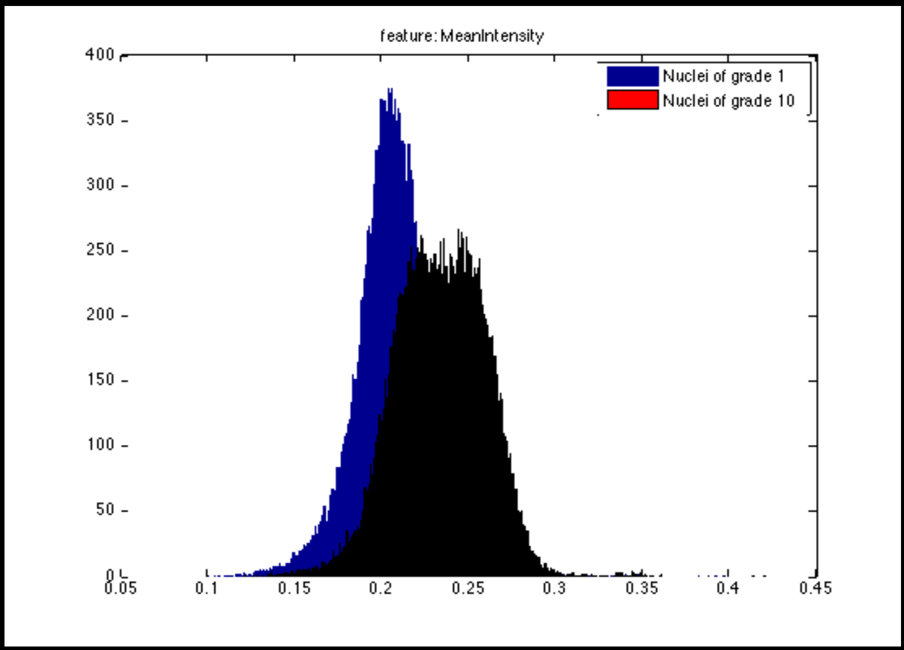
SFFS + 10% Filtering + 100 runs

	Neoplastic Astrocyte	Neoplastic Oligodendrocyte	Reactive Endothelial	Reactive Astrocyte	Junk
Neoplastic Astrocyte	91.89%	1.82%	2.88%	2.25%	1.16%
Neoplastic Oligodendrocyte	1.53%	95.60%	1.10%	0.14%	1.62%
Reactive Endothelial	4.87%	0.53%	88.96%	2.18%	3.47%
Reactive Astrocyte	5.37%	1.54%	6.21%	85.62%	1.27%
Junk	2.86%	1.34%	5.24%	0.64%	89.93%



Nuclear Qualities



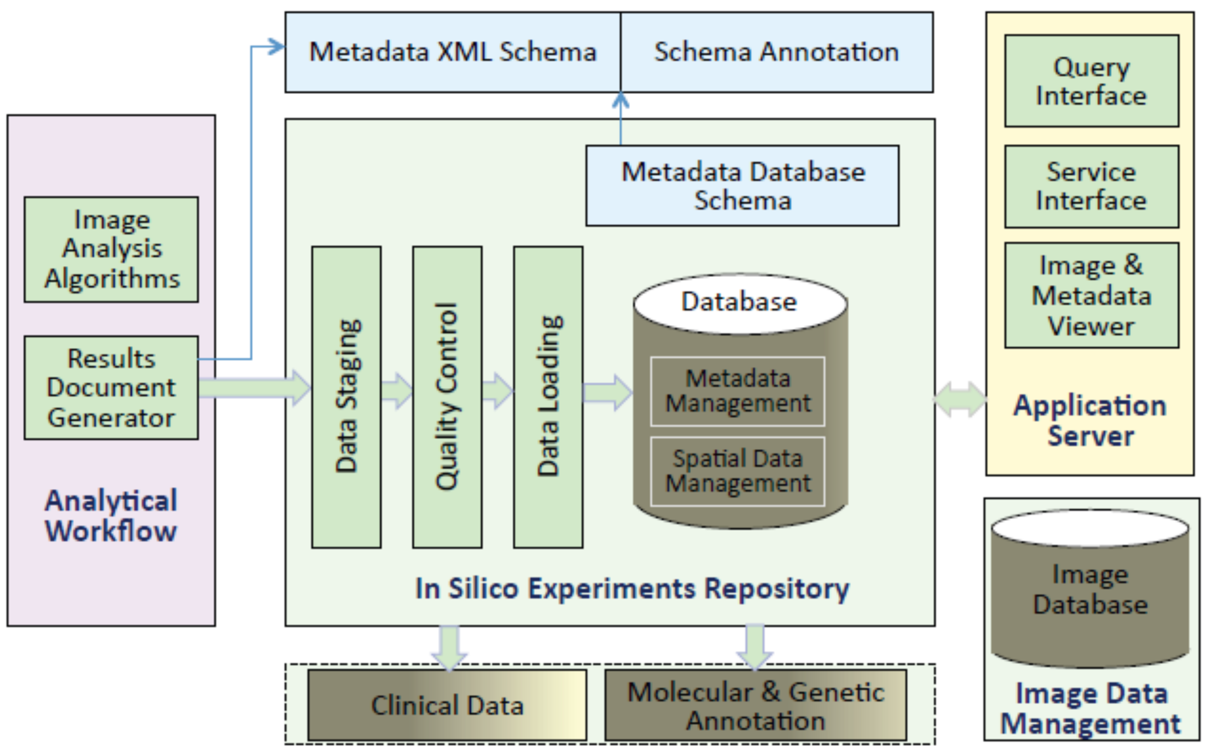


Which features carry most prognostic significance?
Which features correlate with genetic alterations?

Pipeline for 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*

Feature Management and Query Framework

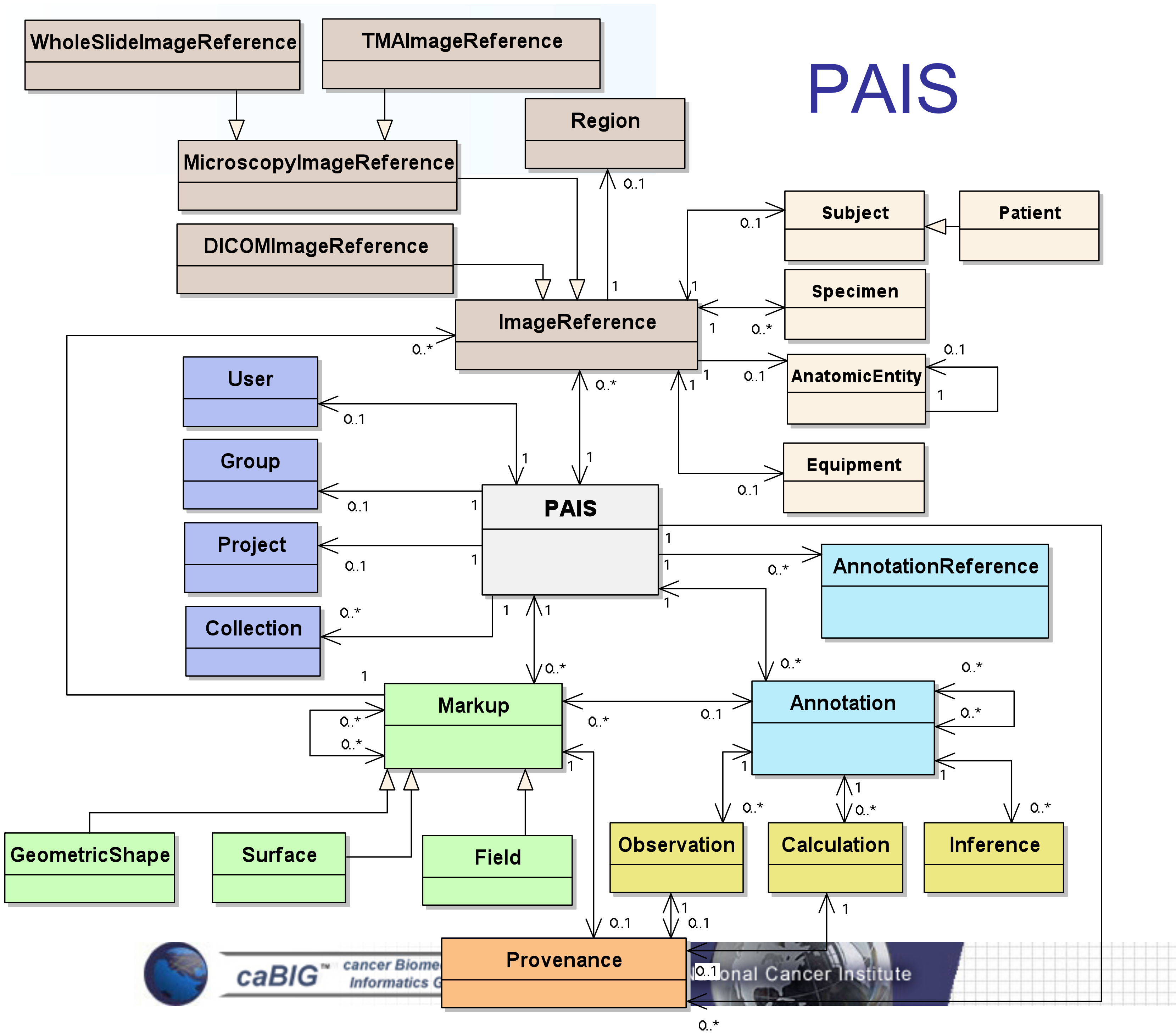


Data Models to Represent Feature Sets and Experimental Metadata

PAIS |pās| : Pathology Analytical Imaging Standards

- Provide semantically enabled data model to support pathology analytical imaging
- Data objects, comprehensive data types, and flexible relationships
- Reuse existing standards
- ***Data models (in general) likely route to integrating staging, immediate on line analyses and full scale analyses***
- ***Semantic models/annotations***
- ***Semantic directed runtime compilation that embedded various partitioners (work with Kennedy, Fox)***

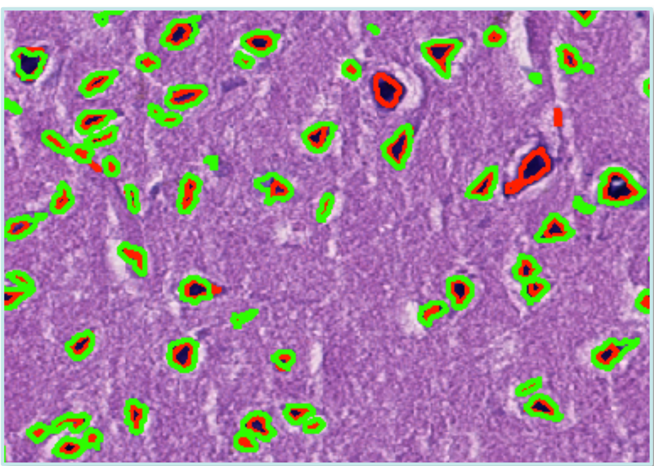
PAIS



WholeSlideImageReference
TMAImageReference
Region
MicroscopyImageReference
DICOMImageReference
ImageReference
Subject
Patient
Specimen
AnatomicEntity
Equipment
User
Group
Project
Collection
PAIS
AnnotationReference
Markup
Annotation
GeometricShape
Surface
Field
Observation
Calculation
Inference
Provenance

Compute Intersection Ratio and Distance Between Markups from Two Segmentation Algorithms

```
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 TCGA Query: Mean Feature Vector and Feature Covariance

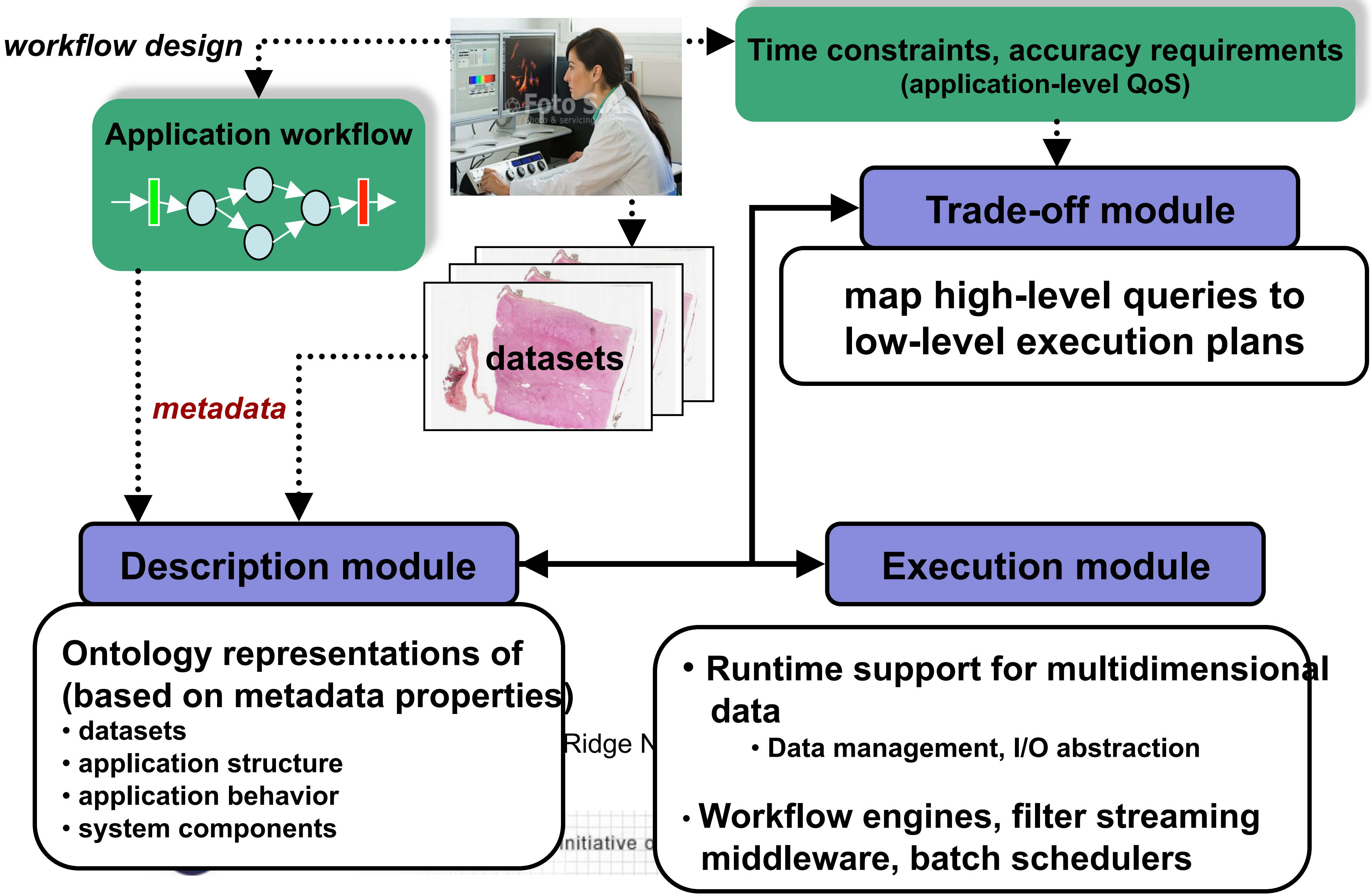
- Mean feature vector for each slide and tumor subtype

```
SELECT AVG(area), AVG(sum_canny_pixel), AVG(mean_canny_pixel)
FROM   pais.calculation_flat c, tctga.patient_charateristic pc, pais.patient p
WHERE  p.patientid = pc.patient_id AND p.pais_uid = c.pais_uid
GROUP BY c.pais_uid, pc.subtype;
```

- Covariance between features

```
SELECT
      COVARIANCE(PERIMETER, AREA) AS COV_PERIMETER_AREA,
      COVARIANCE(PERIMETER, ECCENTRICITY) AS COV_PERIMETER_ECCENTRICITY
FROM  pais.calculation_flat
WHERE PAIS_UID ='TCGA-06-0152-01Z-00-DX7_20x_20x_NS-MORPH_1';
```

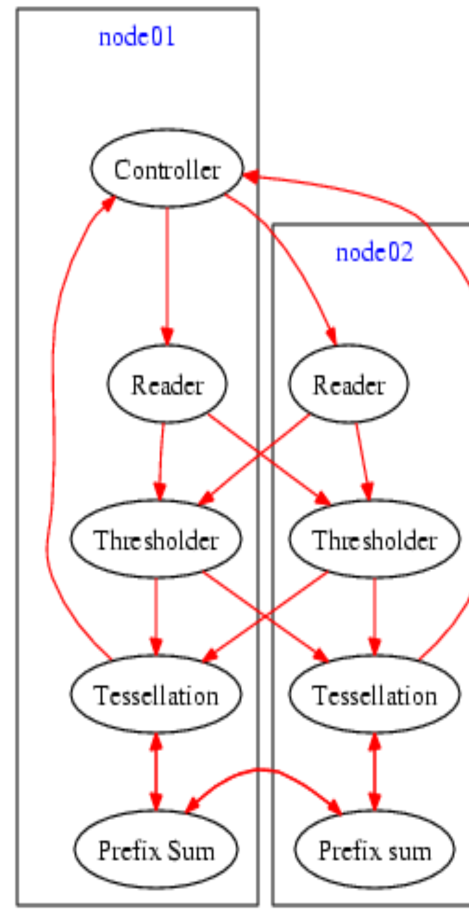
Analysis framework architecture



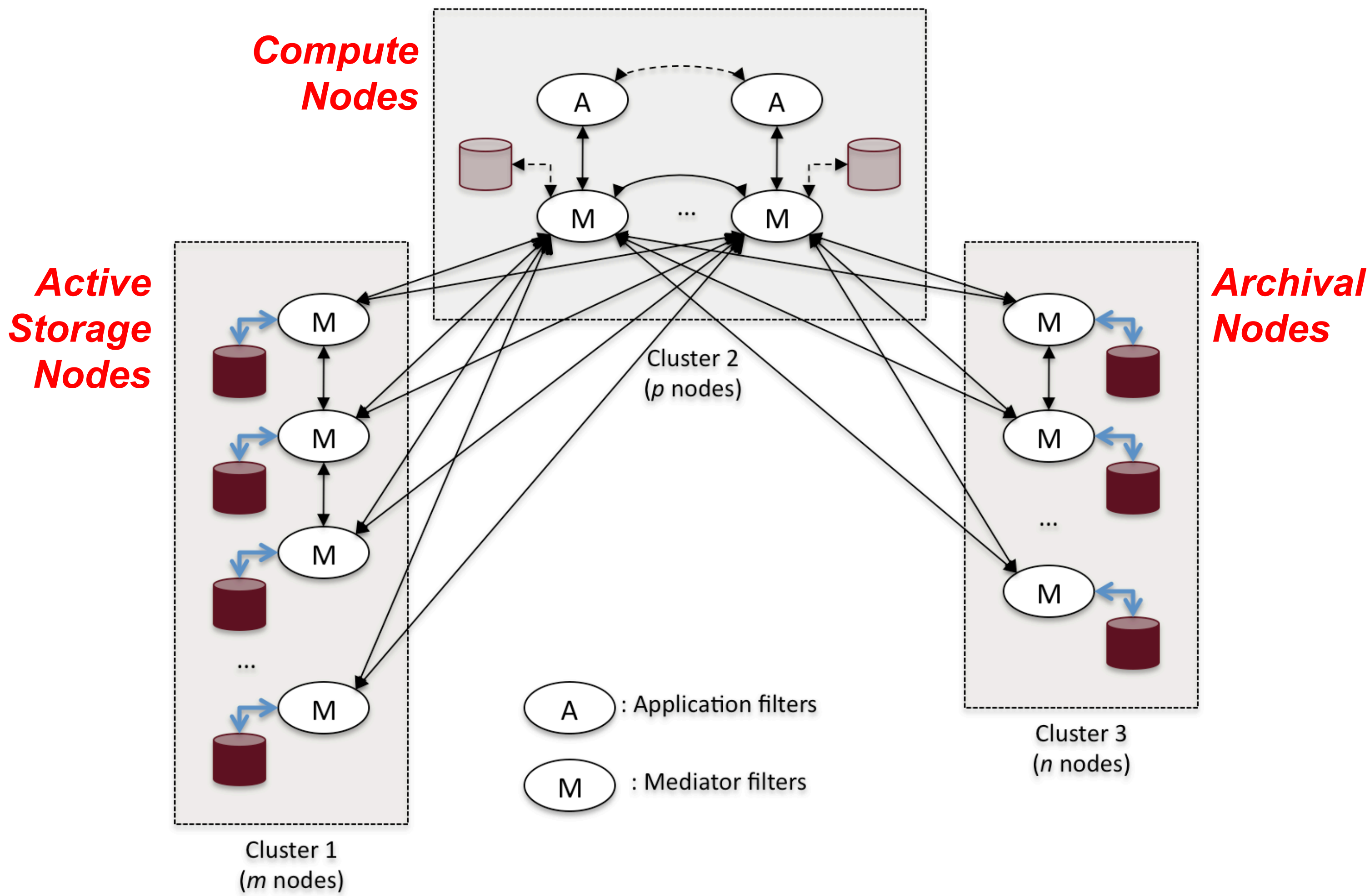
Execution Module: Runtime support for multidimensional data

OCVM

- Customize for specific domains
  - Out-of-core Virtual Microscope
- Out-of-core data?
  - Data stored as a collection of ***chunks***
  - Chunk: unit of data management (disk I/O, indexing and compression)
- Data model
  - Data ***spatially partitioned*** into chunks
  - Chunks distributed across nodes in a shared-nothing environment
- Semi-streaming programming model
  - Leverages lightweight filter-streaming, buffer management by streaming middleware (e.g., DataCutter, IBM System S)

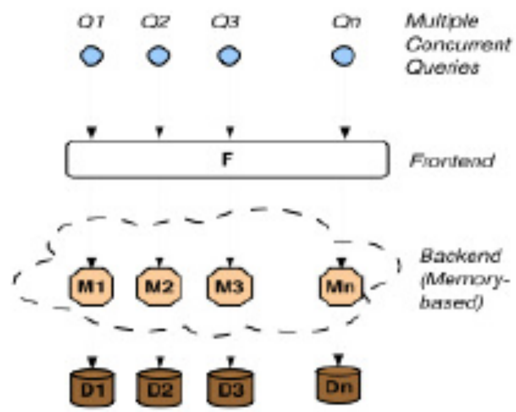


Mediators: I/O abstraction layer

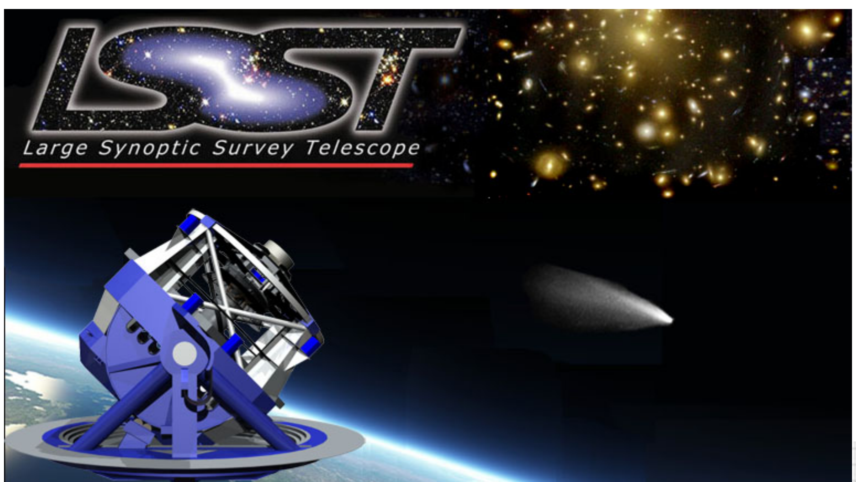


In Transit Processing using DataCutter Spatial Crossmatch

- Mapping to atlas and 3-D reconstruction frequently rely on spatial crossmatch
- We have studied spatial crossmatch with LLNL initially in an astronomy context
- Large Synoptic Survey Telescope (LSST) -- 3.2 Gigapixel camera that captures field of view every 15 seconds
- Catalog roughly 50 billion objects in 10 years
- Netezza (active disk) implementation vs two DataCutter based distributed mySQL implementations
- Benchmarked on Netezza and small (16 node) cluster



(c) Configuration 3



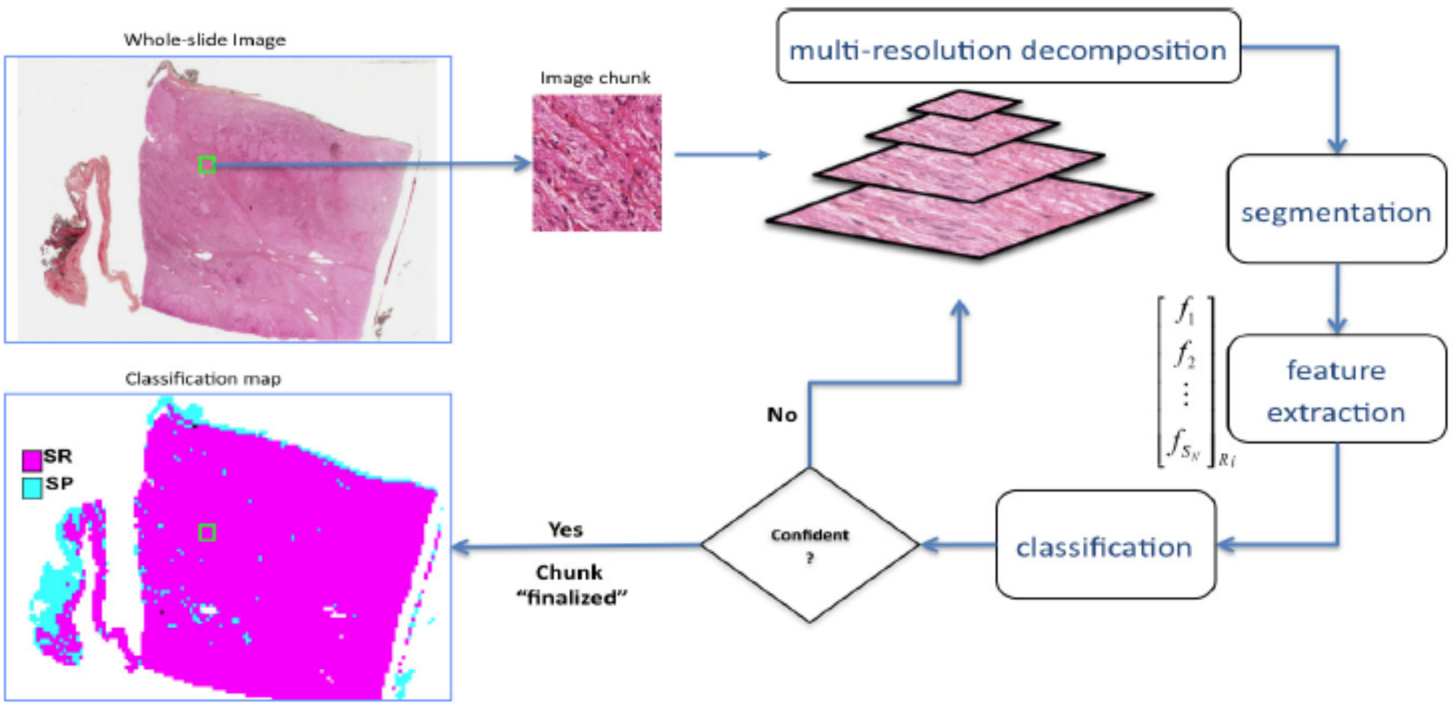
+ Multiple Concurrent Queries (high throughput)
+ Joins executed at backend
+ Memory-based storage
+ Non-transactional storage engin
+ User-controlled data partitioning

- Replica consistency

Semantic Workflows (Wings)
Collaborative Work with Yolanda Gil, Mary Hall

- **A systematic strategy for composing application components into workflows**
- **Search for the most appropriate implementation of both components and workflows**
- **Component optimization**
  - **Select among implementation *variants* of the same computation**
  - **Derive integer values of optimization *parameters***
  - **Only search promising code variants and a restricted parameter space**
- **Workflow optimization**
  - **Knowledge-rich representation of workflow properties**

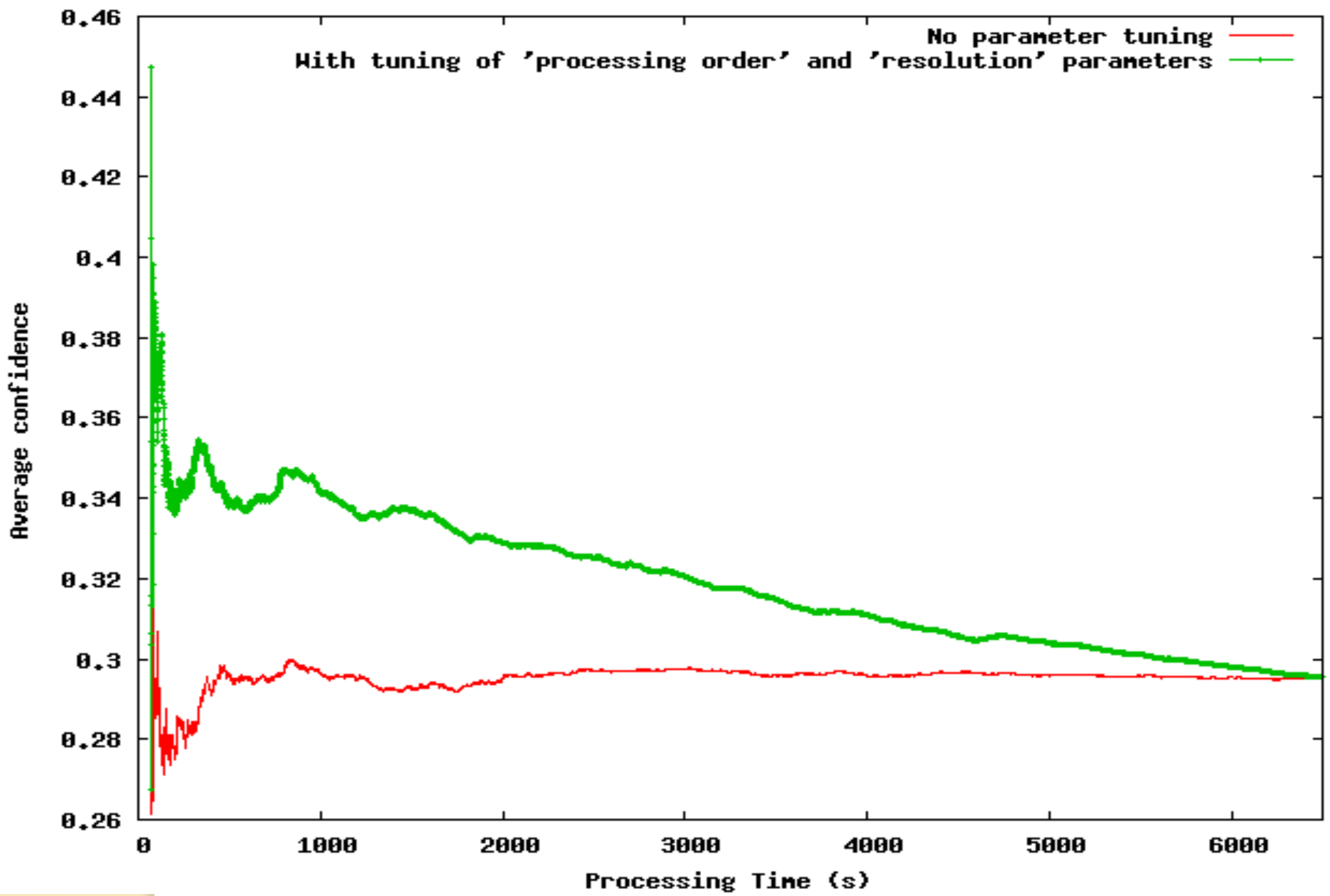
Adaptivity



Whole-slide Image
Image chunk
multi-resolution decomposition
segmentation
feature extraction
$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{S_N} \end{bmatrix}_{Ri}$
classification
Confident ?
No
Yes
Chunk "finalized"
Classification map
SR
SP

Time-constrained Classification: Sample Result

Query: "Maximize average classification confidence within time *t*"

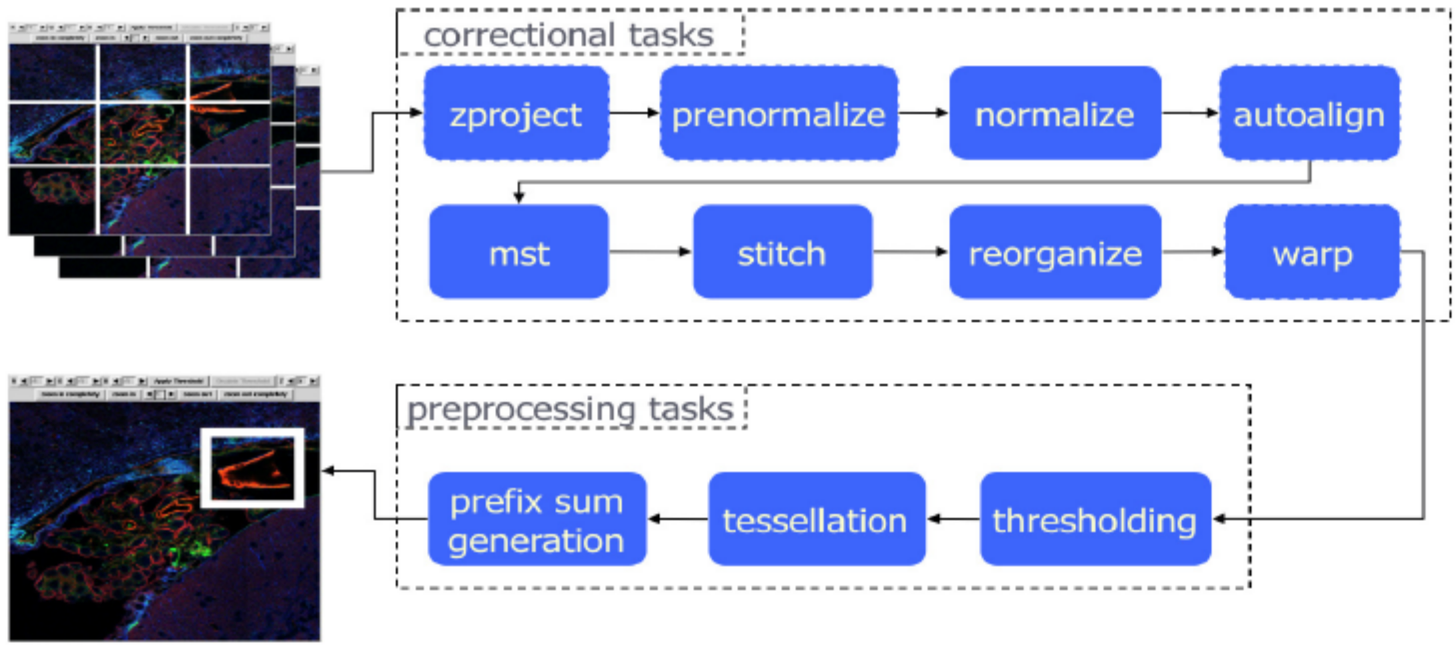


- **32 node cluster**
- **2.4 GHz AMD Opteron dual-processor**
- **8 GB of memory/node**
- **2x250GB local disks**
- **Disk I/O: 55 MB/sec**

Heuristics determine more favorable chunks at an earlier point of time

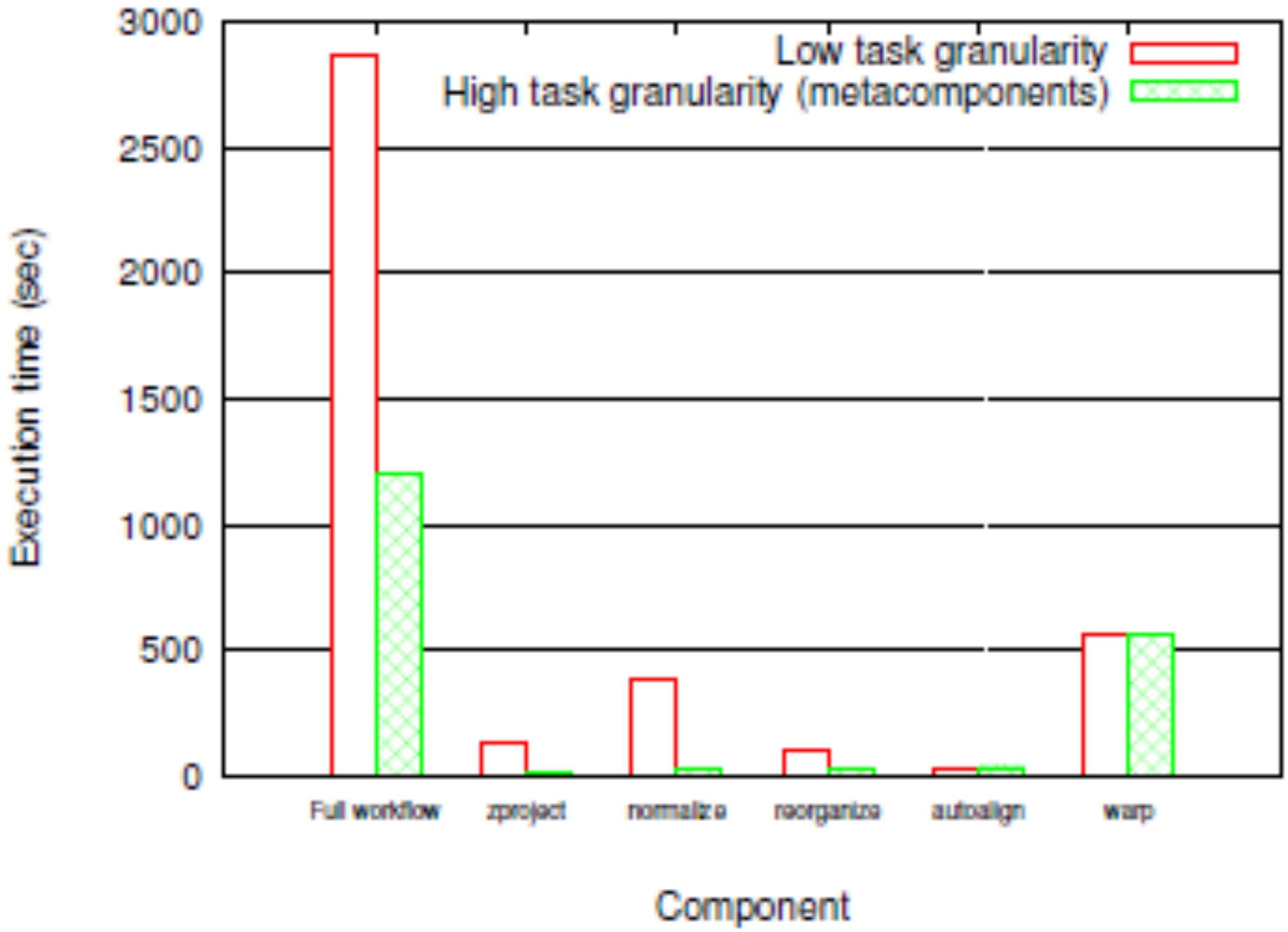
- **Tune 'order of execution' of chunks and 'data resolution' parameter per chunk**

Multiple Granularity Workflows
Map Images into Atlas, Measure Gene Expression



Fuse components into metacomponents
Tasks associated with metacomponent managed by execution module
Pegasus, DataCutter, Condor used to support multiple grained workflow

Performance Impact of Combined Coarse and Fine Grained Workflows



Data Science Research Challenges Driven by In Silico Discovery Research

- Data integration that targets multiple data sources with conflicting metadata and conflicting data
- Efficient methods for semantic query that targets questions involving complex multi-scale features associated with petascale and exascale ensembles of highly annotated images
- Computer assisted annotation and markup for very large datasets
- Systems to support combinations of structured and irregular accesses to exascale datasets

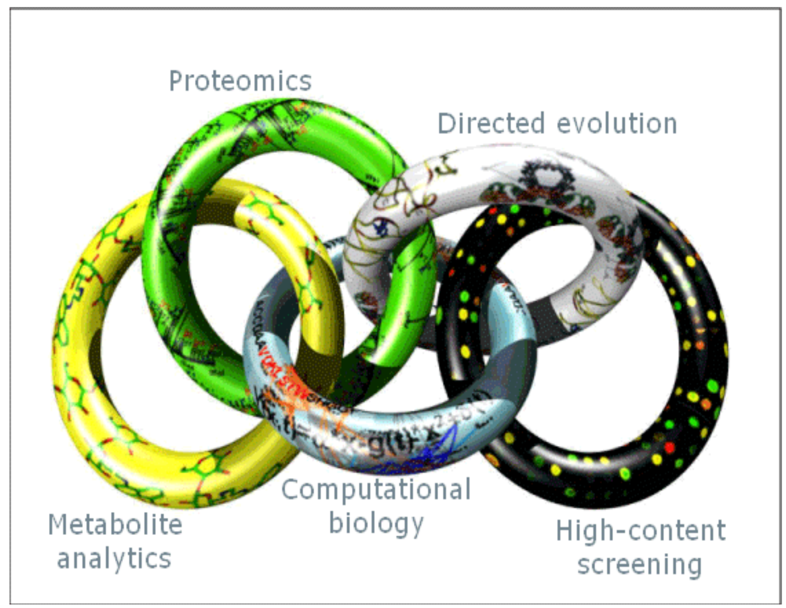
Data Science Research Challenges

- Structural and semantic metadata management: how to manage tradeoff between flexibility and curation
- Data and semantic modeling infrastructures and policies able to scale to handle distributed systems with an aggregate of 10*9 or more data models/concepts
- Three dimensional (time dependent) reconstruction, feature detection and annotation of 3-D microscopy imagery
- Workflow infrastructure for large scale data intensive computations

Final Data Science Challenge: Large Dataset Size

- Basic small mouse is 10 cm^3
- 1 μ resolution – very roughly 10^{13} bytes/mouse
- Molecular data (spatial location) multiply by 10^2
- Vary genetic composition, environmental manipulation, systematic mechanisms for varying genetic expression; multiply by 10^3

Total: 10^{18} bytes per big science animal experiment



Thanks to:

- **Tahsin Kurc, Vijay Kumar**
- **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)**

caBIG™ cancer Biomedical Informatics Grid™ an initiative of the National Cancer Institute

Thanks!