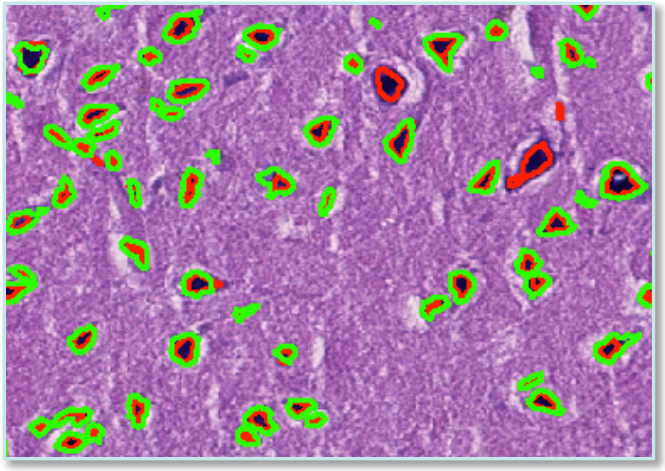
Biomedical Informatics Applications, Big Data, & Cloud Computing

Patrick Widener, PhD
Assistant Professor, Biomedical Engineering
Senior Research Scientist, Center for Comprehensive Informatics
Emory University
patrick.widener@emory.edu

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

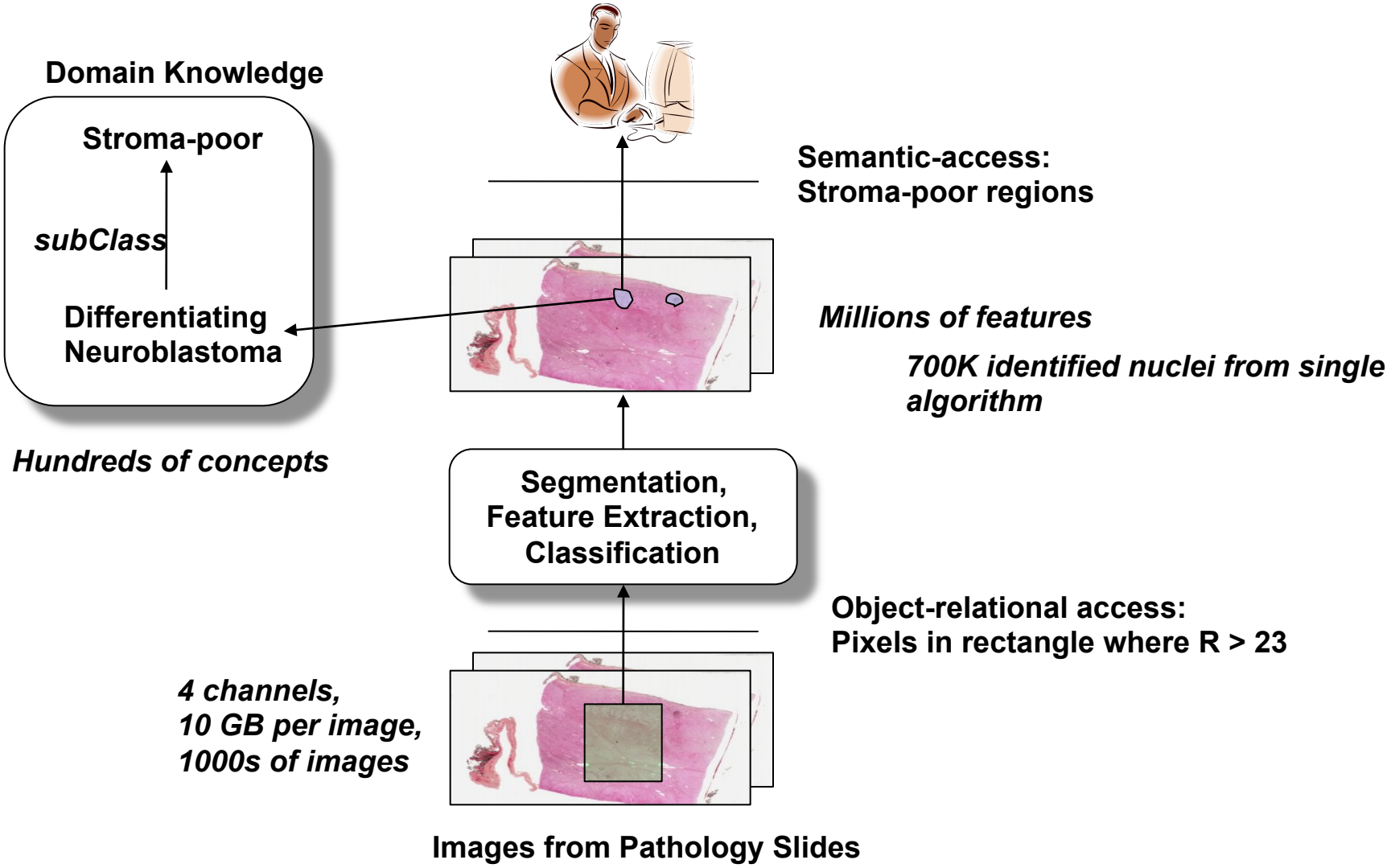
Pipelines to Carry out Feature Extraction and Classification in Brain Tumor Research



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%

Exascale feature extraction / analysis

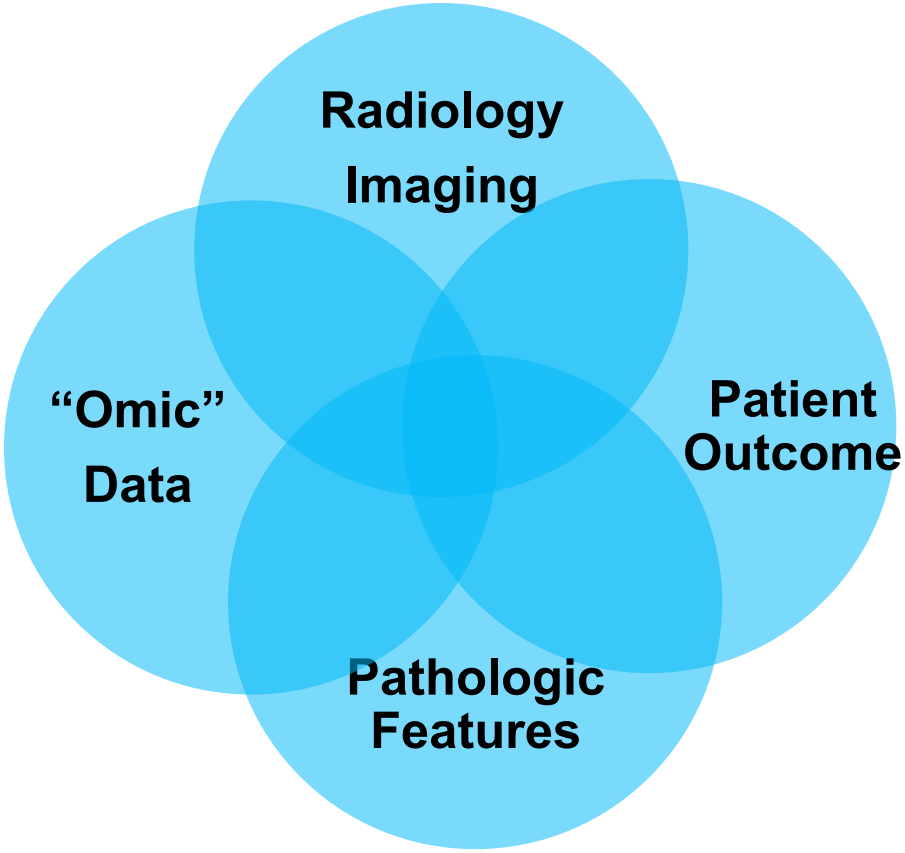


Pipeline for Whole Slide Feature Characterization

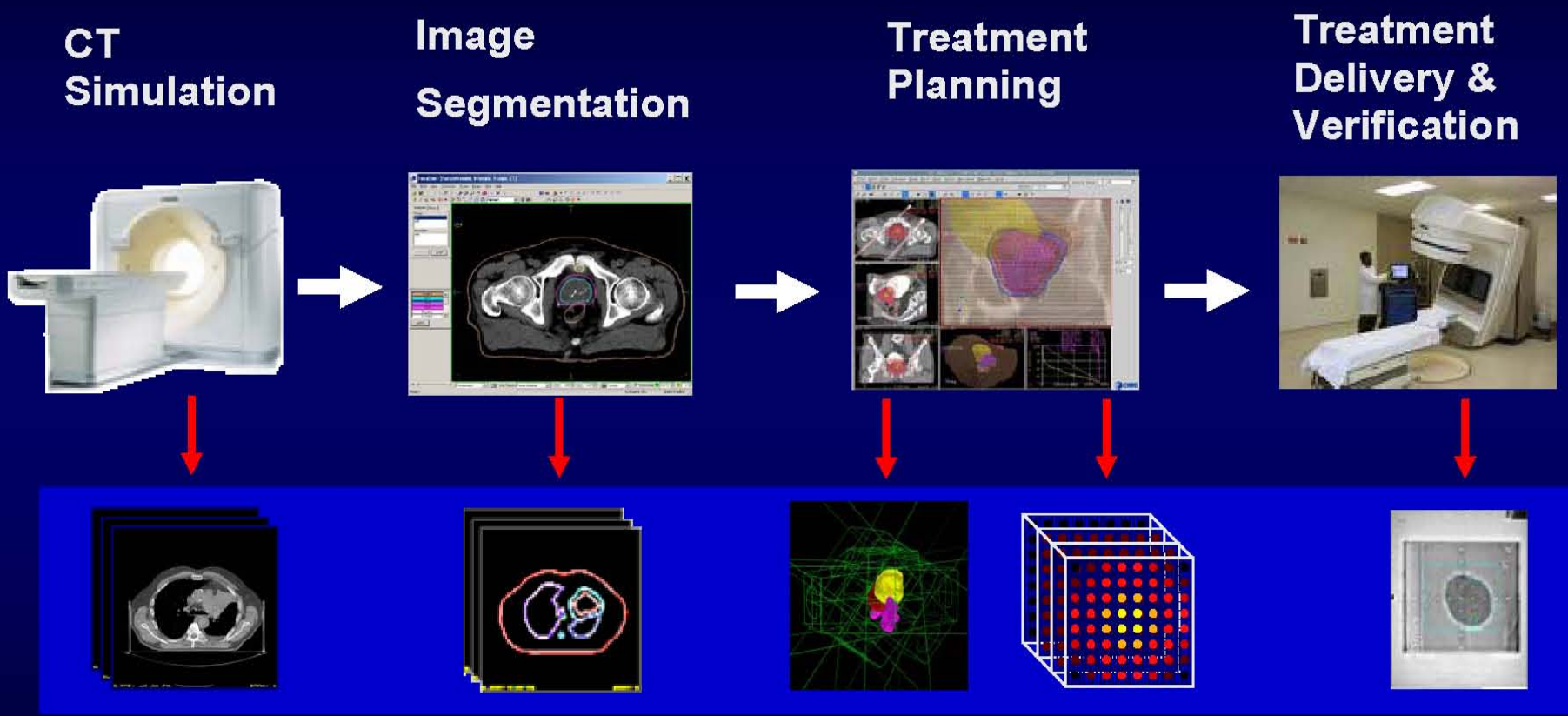
- 10^{10} pixels for each whole slide image
- 10 whole slide images per patient
- 10^{8} image features per whole slide image
- 10,000 brain tumor patients
- 10^{15} pixels
- 10^{13} features
- *Hundreds of algorithms*
- *Annotations and markups from dozens of humans*

Integration of heterogeneous multiscale information

- **Pathology, Radiology, "omics", Clinical information**
- **Reproducible characterization at gross level (Radiology) and fine level (Pathology)**
- **Integration with multiple types of "omic" information**
- **Exploit synergies to improve ability to forecast survival & response.**



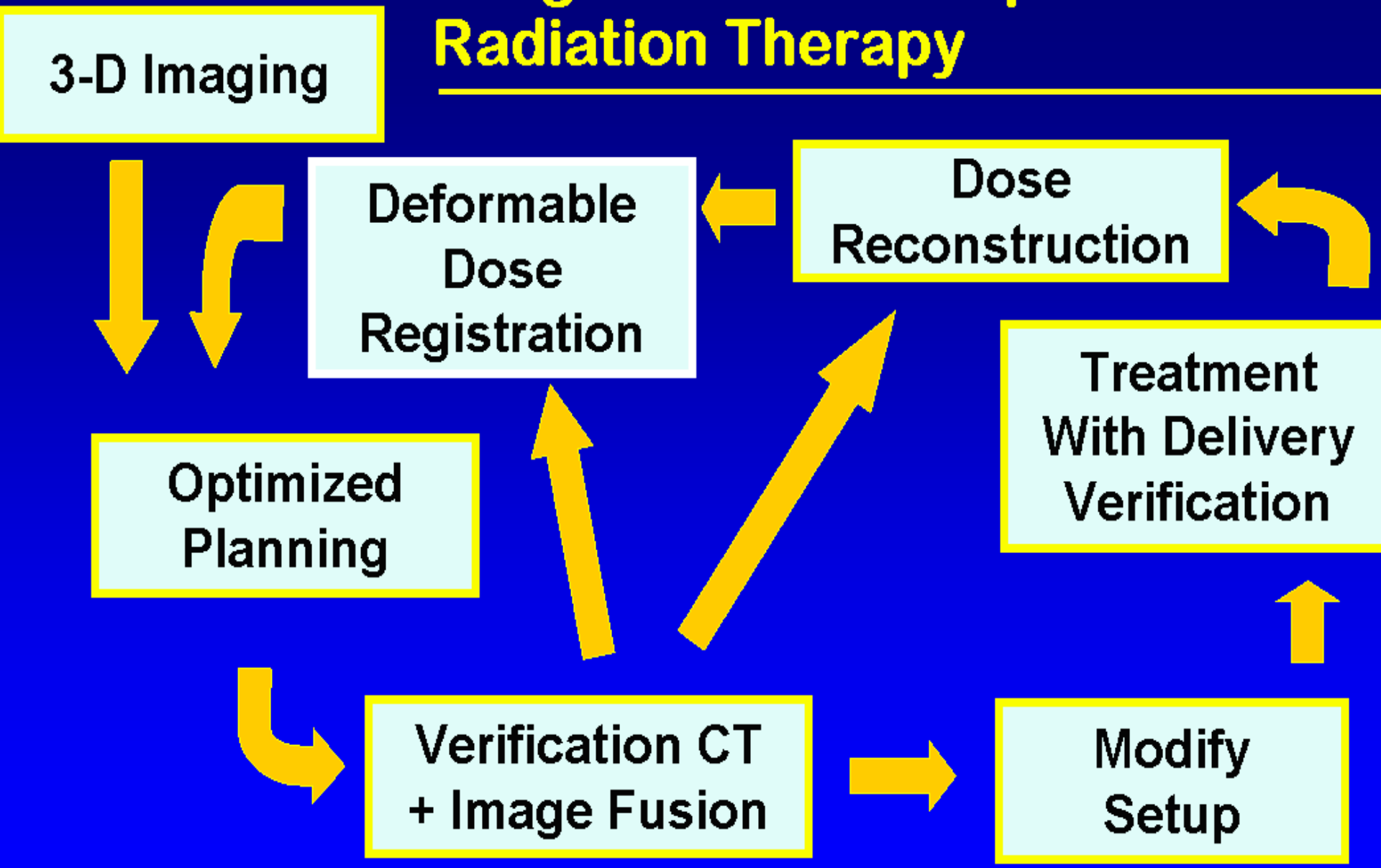
3D Conformal Radiation Therapy: Simplified Workflow and Data Objects



Slide from Walter Bosch

Typical Data Set per Patient ~ 100 MB

Image Guided - Adaptive Radiation Therapy



Analogous Feature Extraction and Classification Issues in Most Fields

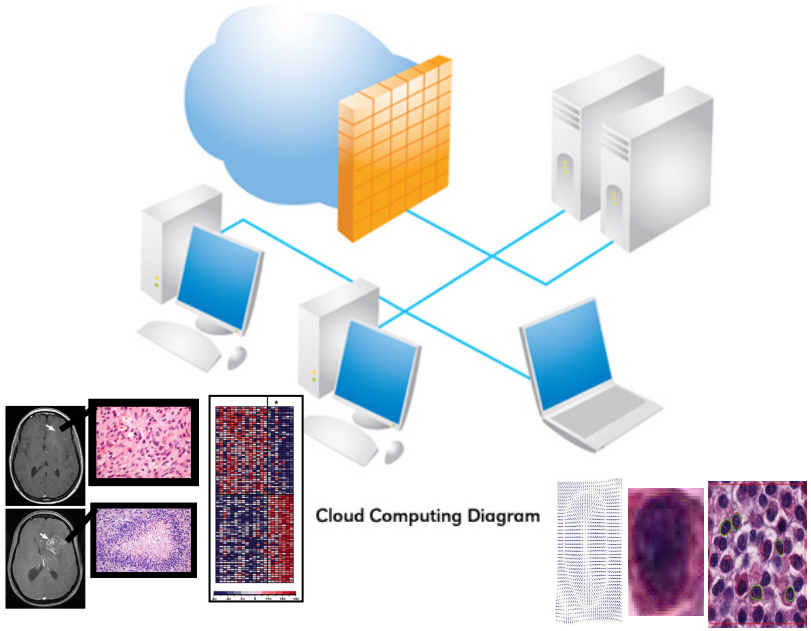
Astrophysics	*Which portions of a star's core are susceptible to implosion over time period [t1, t2] ?*	Compute streamlines on vector field *v* within grid points [(x1,y1)-(x2,y2)]
Material Science	*Is crystalline growth likely to occur within range [p1, p2] of pressure conditions ?*	Compute likelihood 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)

Data Science Research Challenges

- Coordination and management of algorithms and metadata needed to carry out high throughput feature extraction and classification
- Interactive on-demand user interactivity with exascale complex multi-algorithm analysis frameworks
- Computer assisted annotation and markup for very large datasets – development of the actual image analysis and machine learning algorithms
- 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

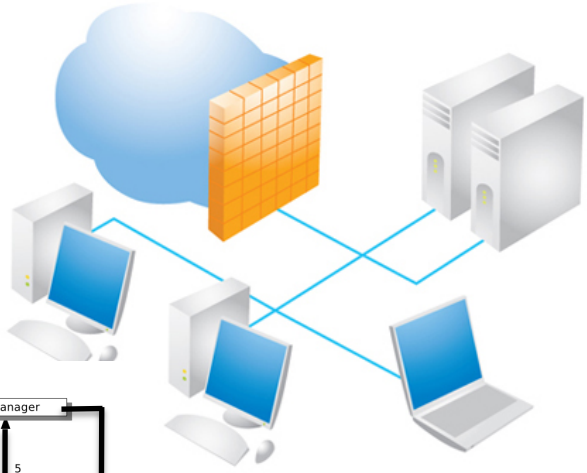
Biomedical informatics research challenges

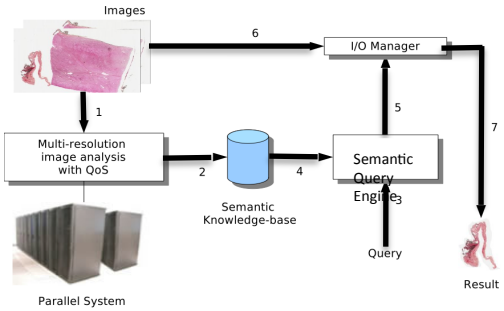
- Integration of multiple data sources with conflicting metadata and conflicting data
- Efficient methods for semantic query involving complex multi-scale features associated with peta-/exascale ensembles of highly annotated images
- Systems to support combinations of structured and irregular accesses to exascale datasets



Correlative analysis

Image segmentation, classification





Distributed, federated query support

Tactical research issues

- Data privacy/security
- Seeding petascale data into the cloud
- Data to computation or computation to data
- Access to specialized analysis engines

GT / Emory Biomedical Research Cloud

- Jointly developed system software/middleware stack
- Leverage existing work in virtualization, power management
- Joint hardware, networking investments