# Status of Data Formats in Cryo-Electron Microscopy

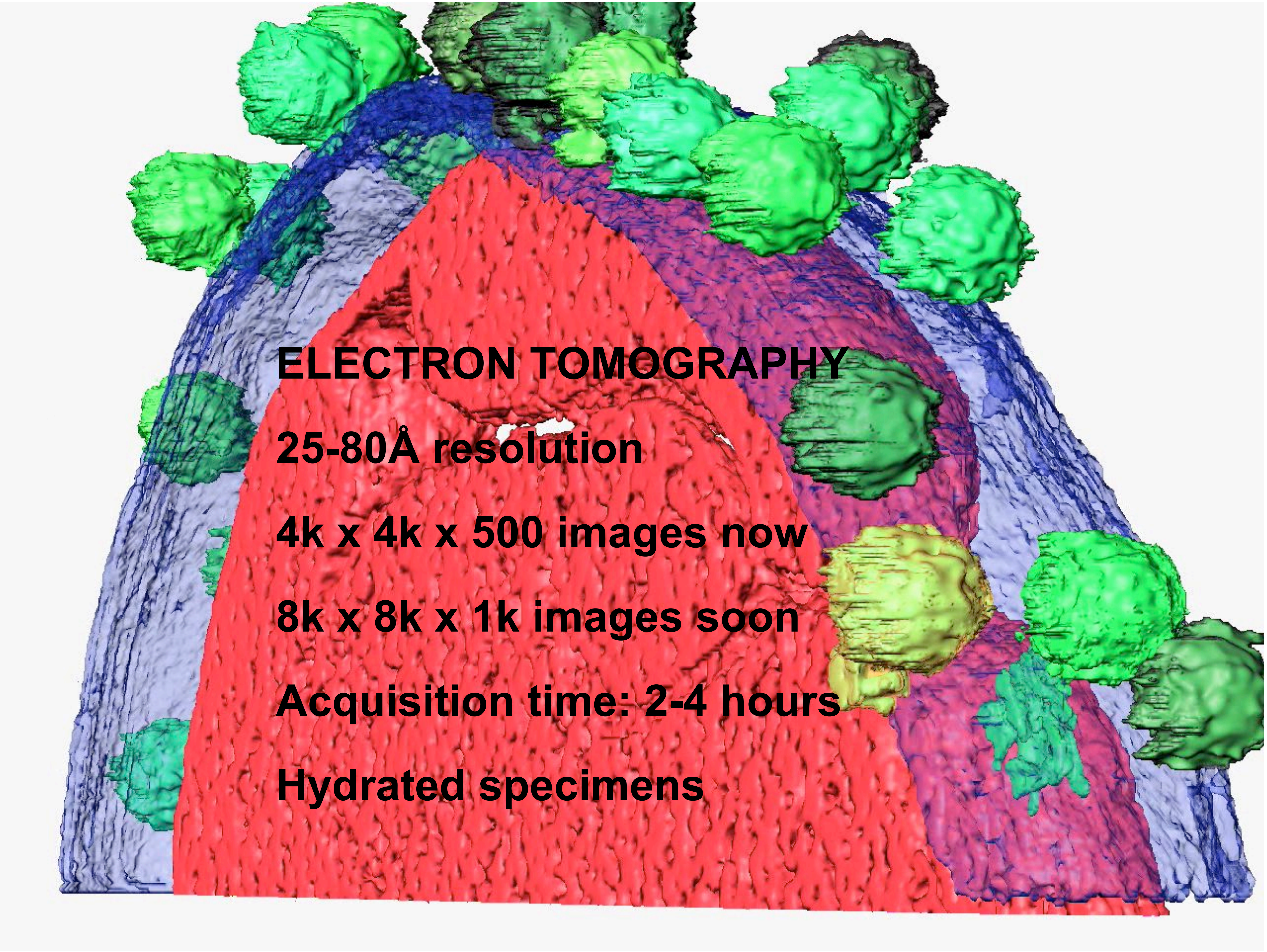
*Matthew Dougherty*

August 17, 2007

National Center for Macromolecular Imaging
Baylor College of Medicine
Houston, Texas

# What NCMI does

- Biomedical resource of NCRR/NIH
- EM technology development



# ELECTRON TOMOGRAPHY

25-80Å resolution

4k x 4k x 500 images now

8k x 8k x 1k images soon

Acquisition time: 2-4 hours

Hydrated specimens

## SINGLE PARTICLE RECONSTRUCTION

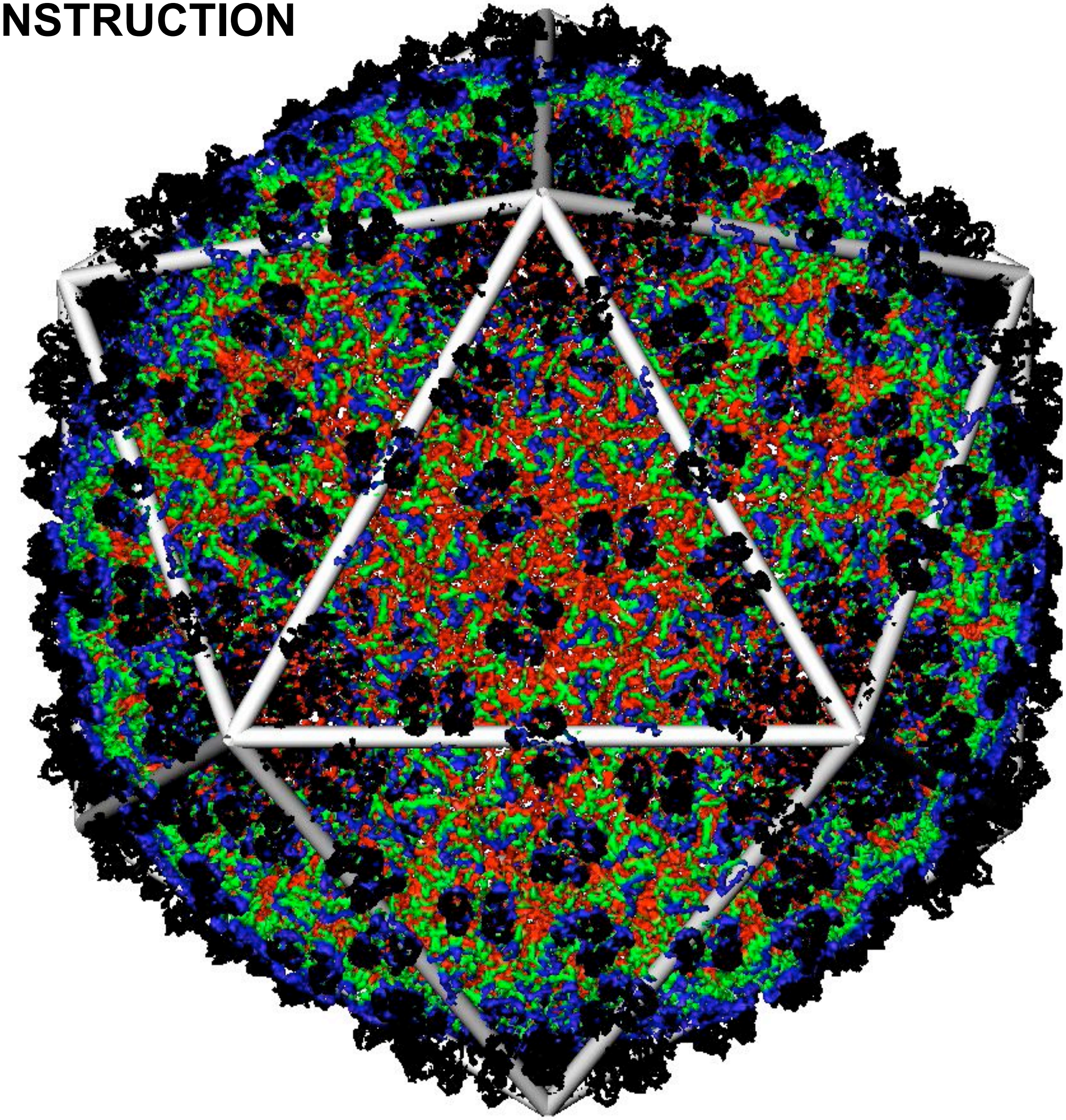
4.5Å resolution

1k x 1k x 1k images now

100k particles required

Ability to trace backbone

Hydrated specimens



# What NCMI does

- Biomedical resource of NCRR/NIH
- EM technology development
- EM service bureau
- Visualization and animation
- Cryo-EM task force (NCMI/EBI/PDB)

# Status of cryo-em image formats

- Three primary formats
- Three new formats
- Tied to s/w packages
- One header plus pixels
- Z stack of 2D images
- Single 3D image
- NO parallel IO
- NO compression
- NO XML
- NO user involvement
- NO extensibility
- NO regional data extraction
- NO multiscale

# Important non-EM image formats

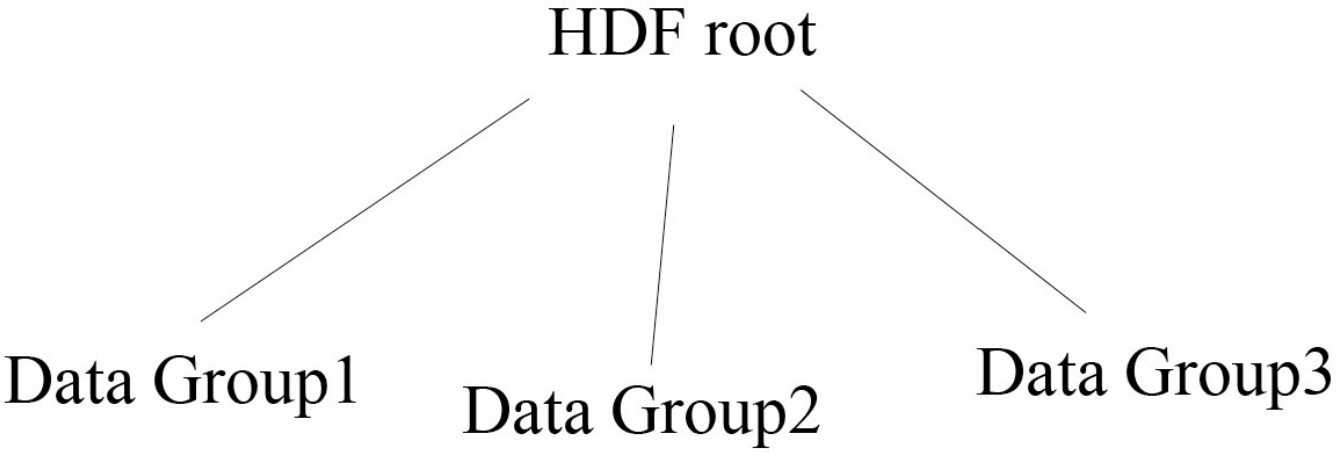
- DICOM
- OME
- TIFF
- imgCIF & NeXus
- Other scientific (astronomical, EOS)

- JPEG2000 part 10 (3D compression)
- XML (format builder)
- HDF (format builder)

# HDF

- Researching it in 2000, discussion in 2002
- Heterogenous data
- Sub-volume compression & extraction
- High performance
- Used by NeXus
- Open source
- Python
- Closest to a digital metric
- Resources & mission
- Development of the user community
- Encapsulation

# Encapsulation



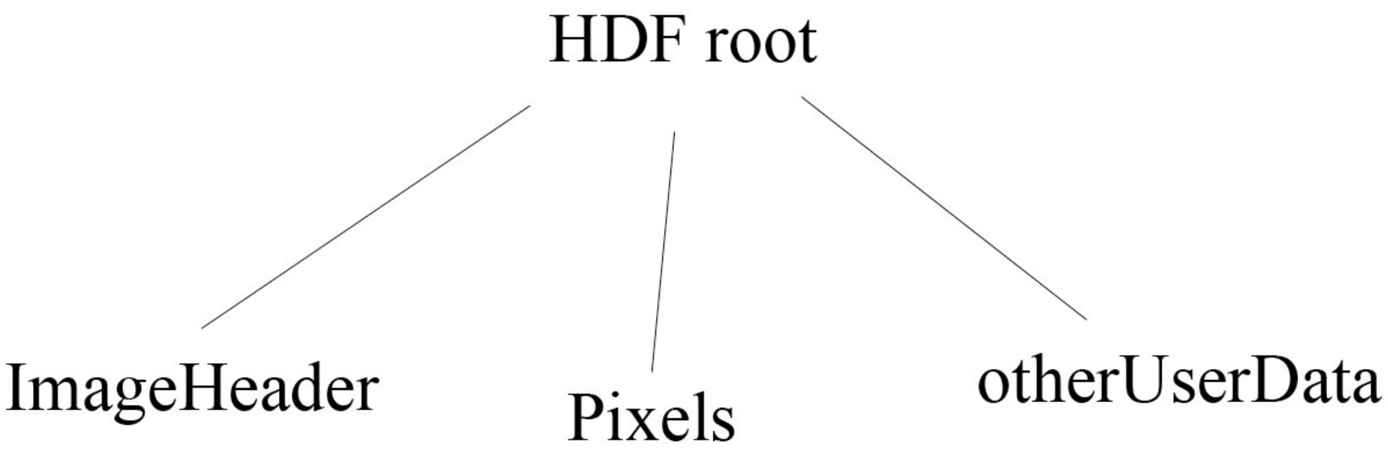
HDF root

Data Group1

Data Group2

Data Group3

# Encapsulation



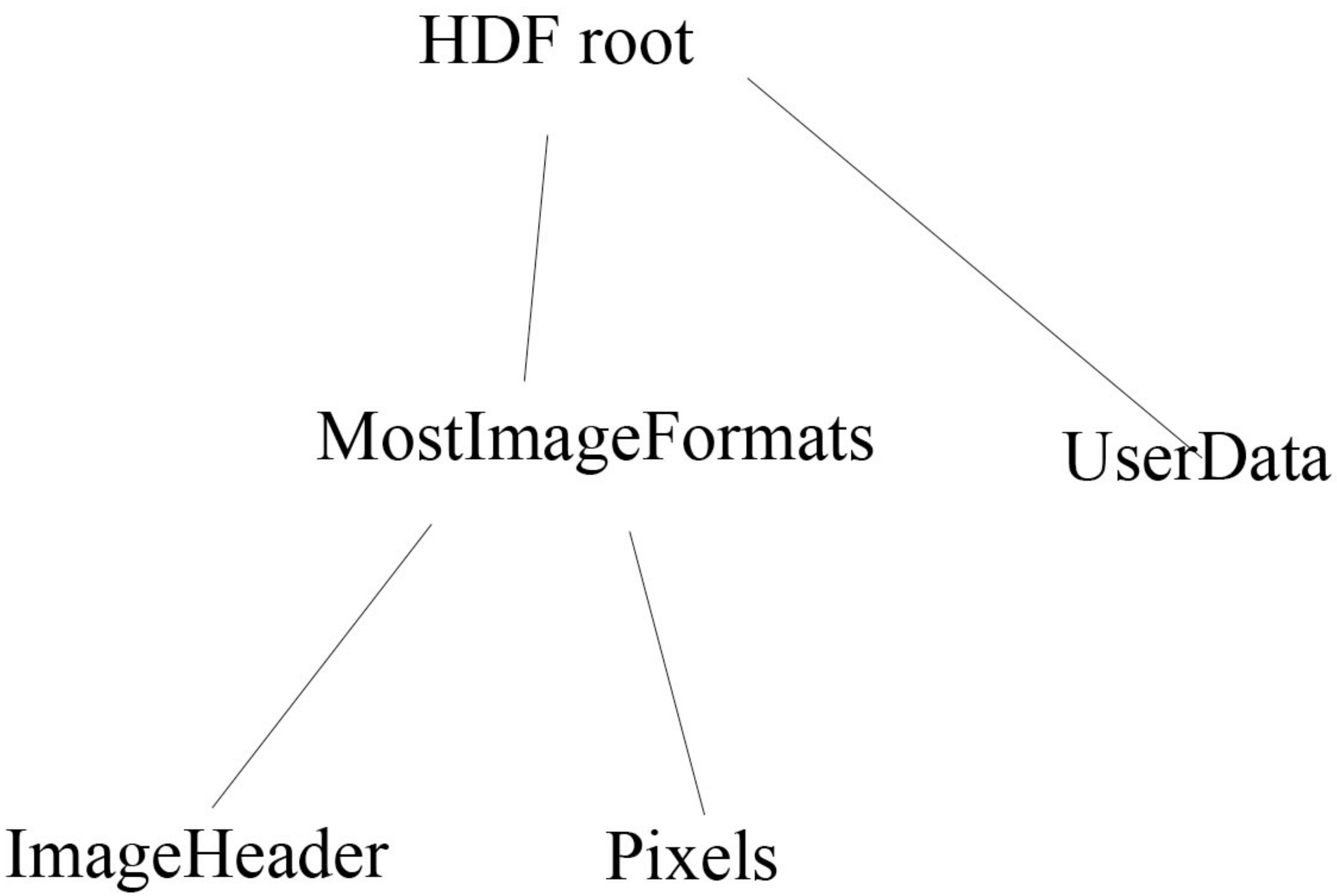
HDF root

ImageHeader

Pixels

otherUserData

# Encapsulation



HDF root

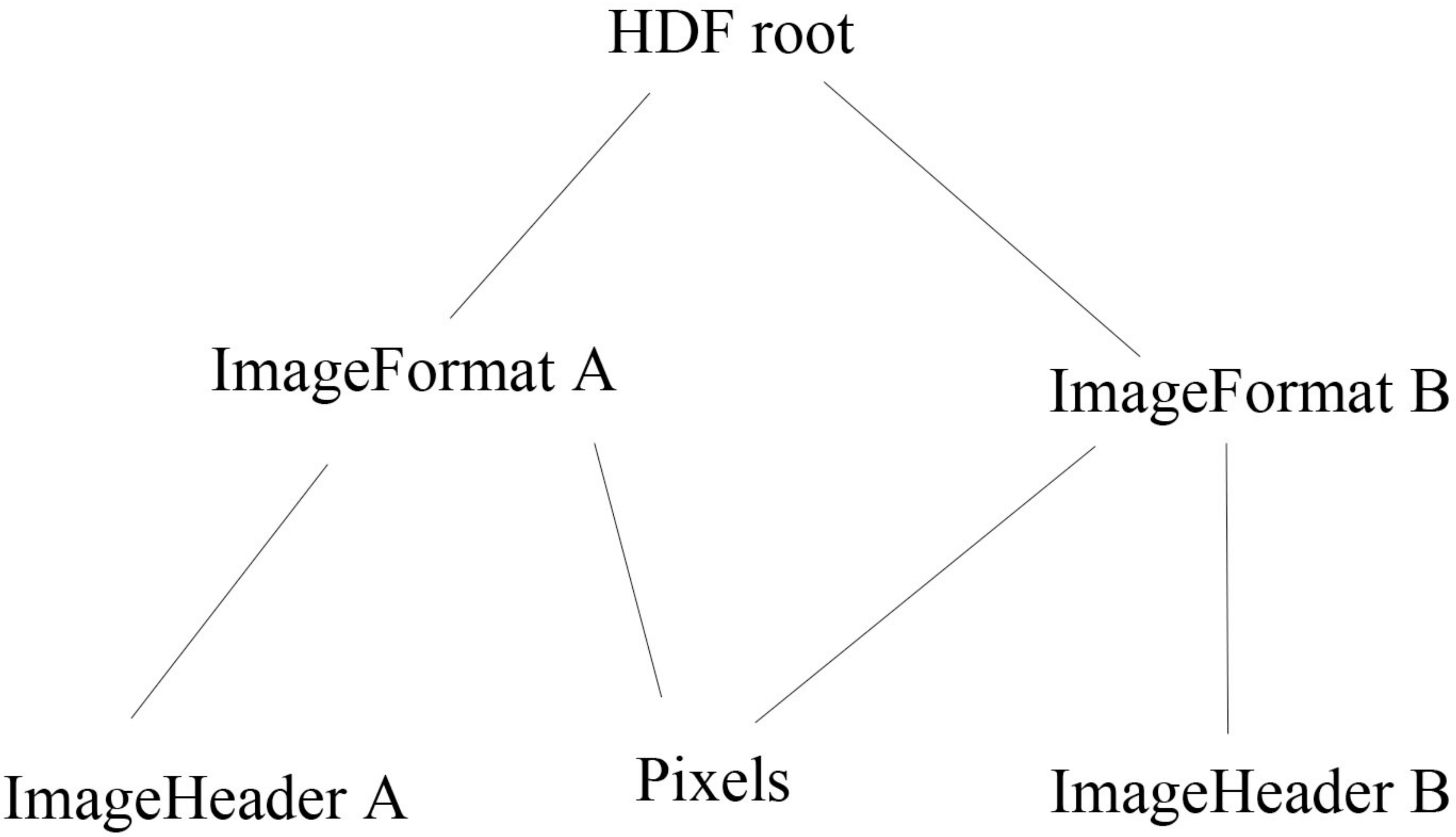
MostImageFormats

UserData

ImageHeader

Pixels

# Encapsulation



# What is a scientific image?

- N-dimensional grid
- Uniform grid of pixels
- Constant pixel model at each node
- Physical description and unit size

- Container(s): metadata & pixels

# What is it used for?

- Data acquisition
- Reconstruction
- Visualization & animation
- Segmentation & annotation
- Models
- Repositories
- Initial research & future research

# What is needed in a ‘scientific image’ data format?

- High performance
- Extensible
- Archival

# What is needed in a ‘scientific image’ data format?

- N-dimensional, multi-image
- Heterogeneous datasets
- User ability to attach user defined data
- Simple scientific image definition
- Integration with XML
- Multiple image headers
- Interactive multi-scale
- Regional compression
- Symmetry correction
- Segmentation and regional data analysis
- Version management of software and data formats
- Provenance
- Open source
- Documented
- Formal standards (NISO/DublinCore/METS)
- Somebody to manage & maintain it

# What can be done?

- Avoid Namespace collisions
  - Registry of root groups created by the research communities
  - Registry & archive of research data formats
  - Managed by HDF, championed by MEDSBIO
- Common image definition
  - Registry for pixel models created by the research communities
  - Involve the research, viz, storage, and archival communities
  - White paper and wiki
  - Organized and maintained by HDF, support by MEDSBIO & IUBS/TDWG-image
- Development of formal standards
  - NISO registration
  - Dublin Core changes
  - Modification to Metadata Encoding and Transmission Standard
  - Adherence to Open Archival Information System Reference Model
  - Lead by HDF
- Acquire, integrate, and disseminate Best Practices
  - wisdom from various communities
  - Teleconferencing, regular discussions
  - Intersection of datasets when possible & practical
  - Simplify
  - Auto-document, common s/w codeq
  - Developed and disseminated by HDF

# Summary

- There is unique opportunity and urgent need for a universal definition of a ‘scientific image’ that could serve most scientific communities.
- Such a definition would make the majority of scientific datasets compatible
- HDF5 is the most logical infrastructure to implement this definition
- MEDSBIO should be central in making this happen.