# Balancing Performance and Preservation Lessons learned with HDF5

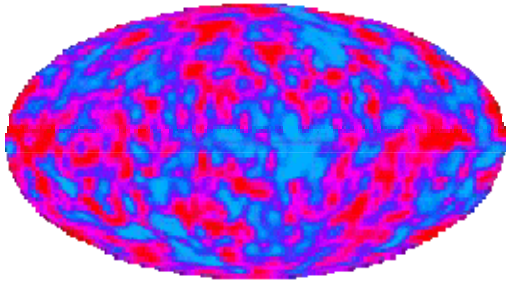
Mike Folk

The HDF Group

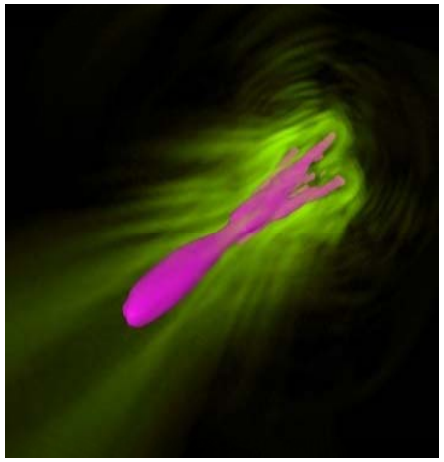
US DPIF Workshop
NIST, Gaithersburg, Maryland
March 29-31, 2010

# Data Challenges

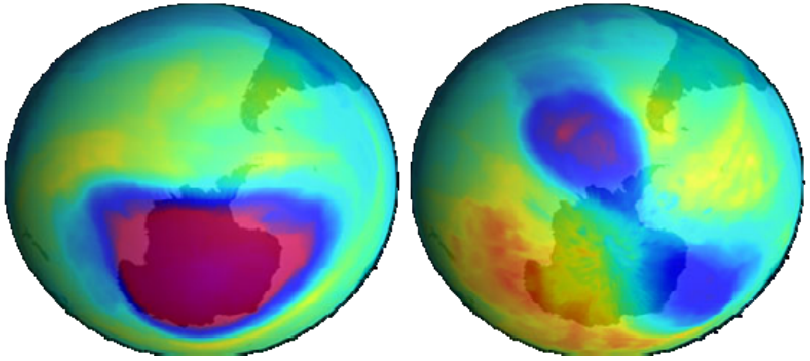
# Answering big questions …



**Matter and the universe**



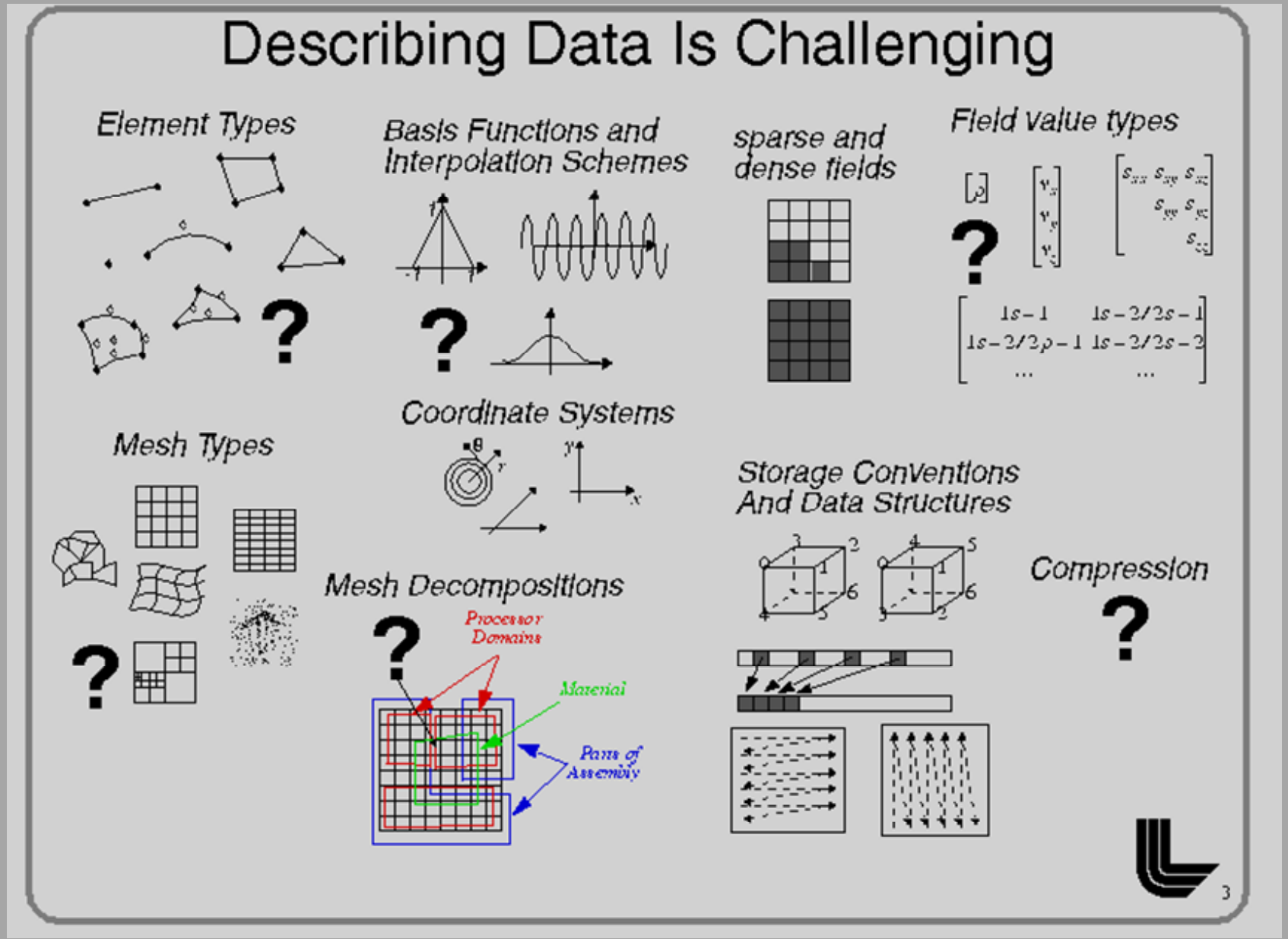
**Life and nature**



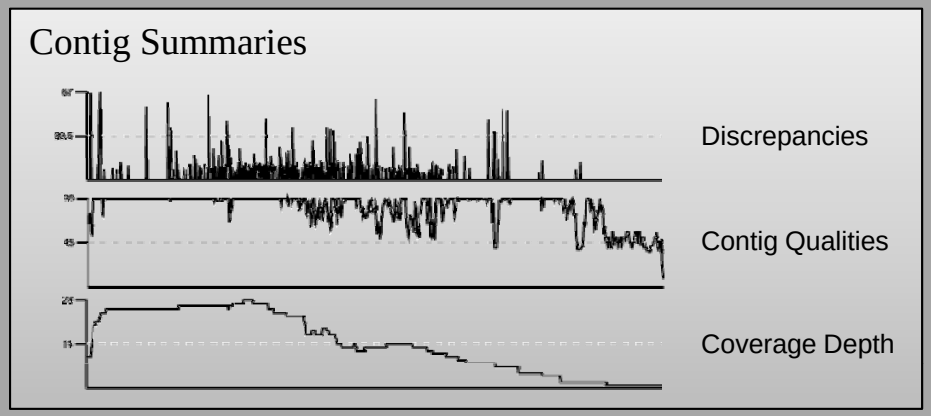

**Weather and climate**

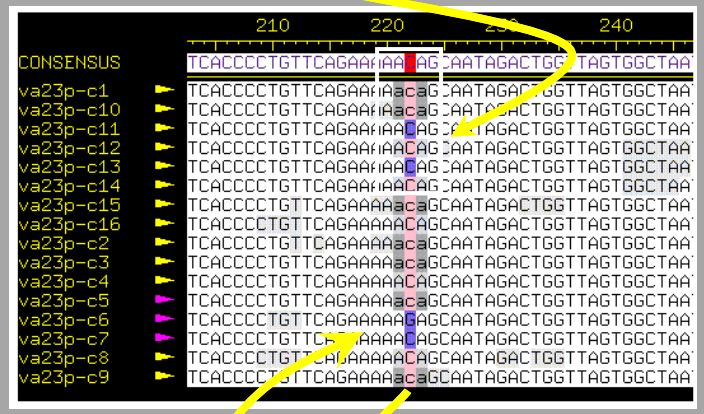
# … involves big data …

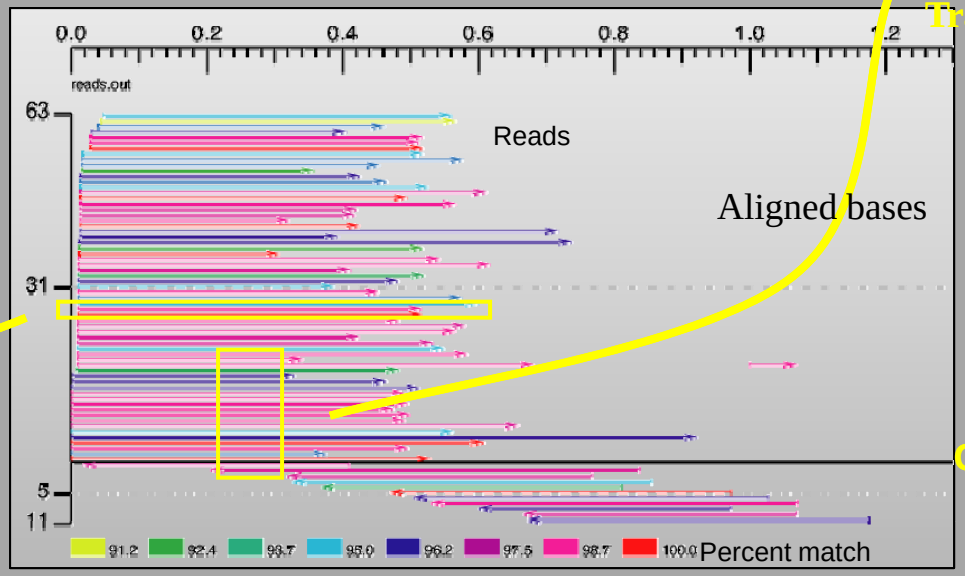
# … highly varied data …

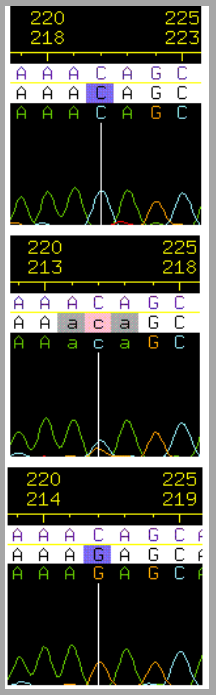


# … and complex relationships …

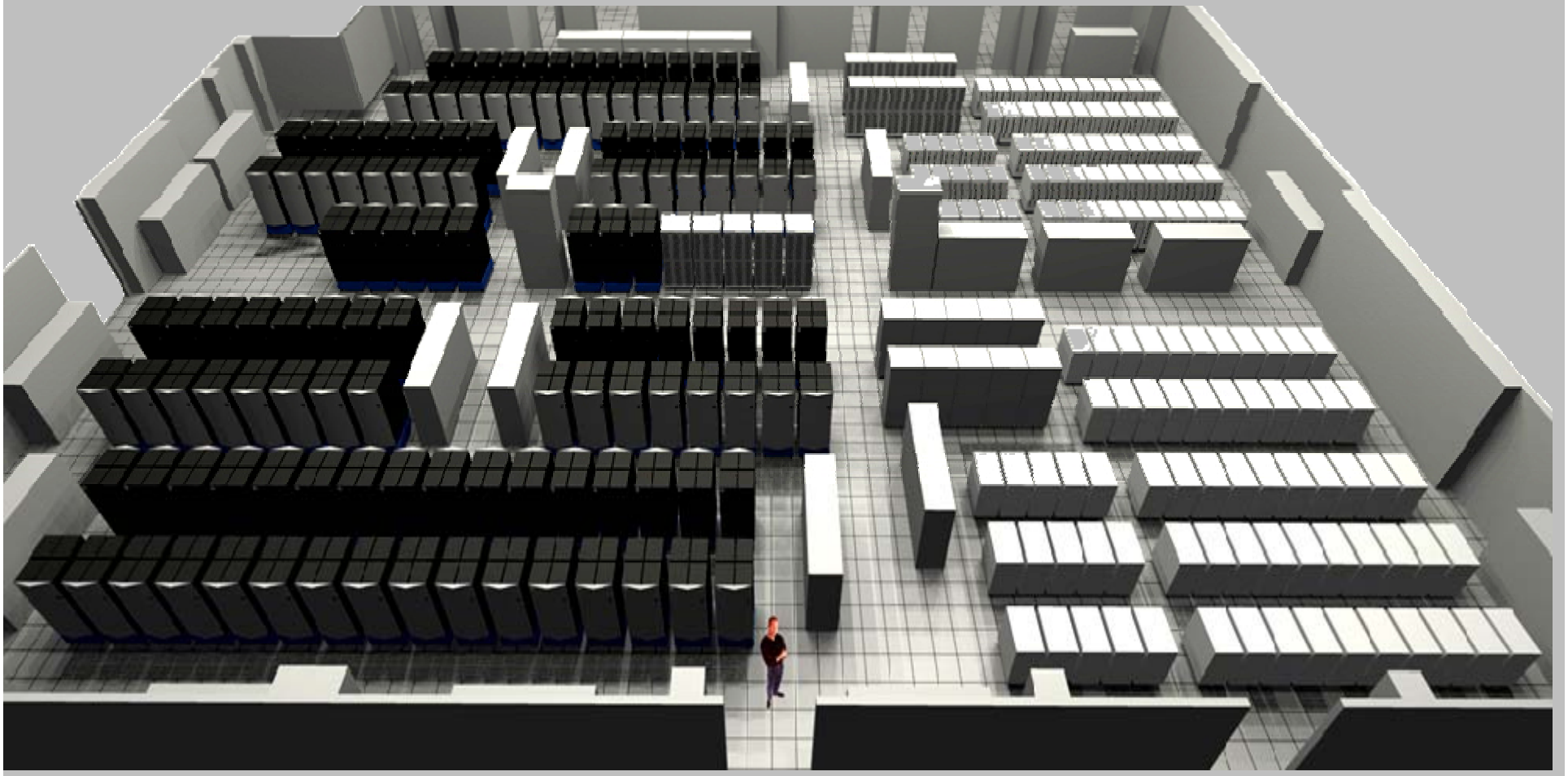



# … on big computers …



… and small computers …

# HDF

- At once, HDF serves as
  - a container for big data and varied data
  - a platform upon which to build data applications,
  - high performance middleware for capturing, storing, and accessing data

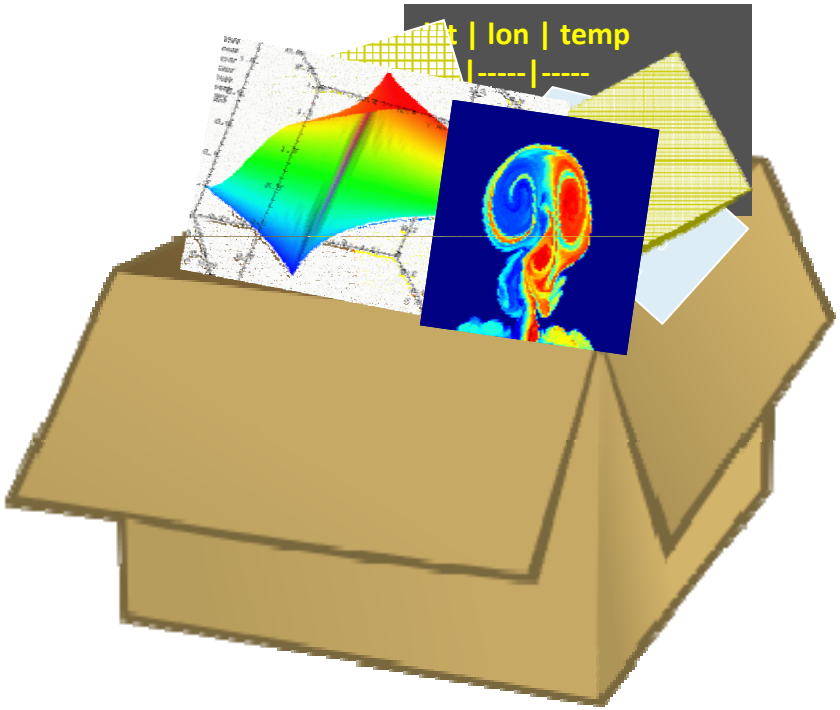
# HDF = Hierarchical Data Format

- HDF4 is the first HDF format
  - Originally called HDF
  - First release was 1988
  - Still supported by The HDF Group

- HDF5 is the second HDF format
  - First release was in 1998

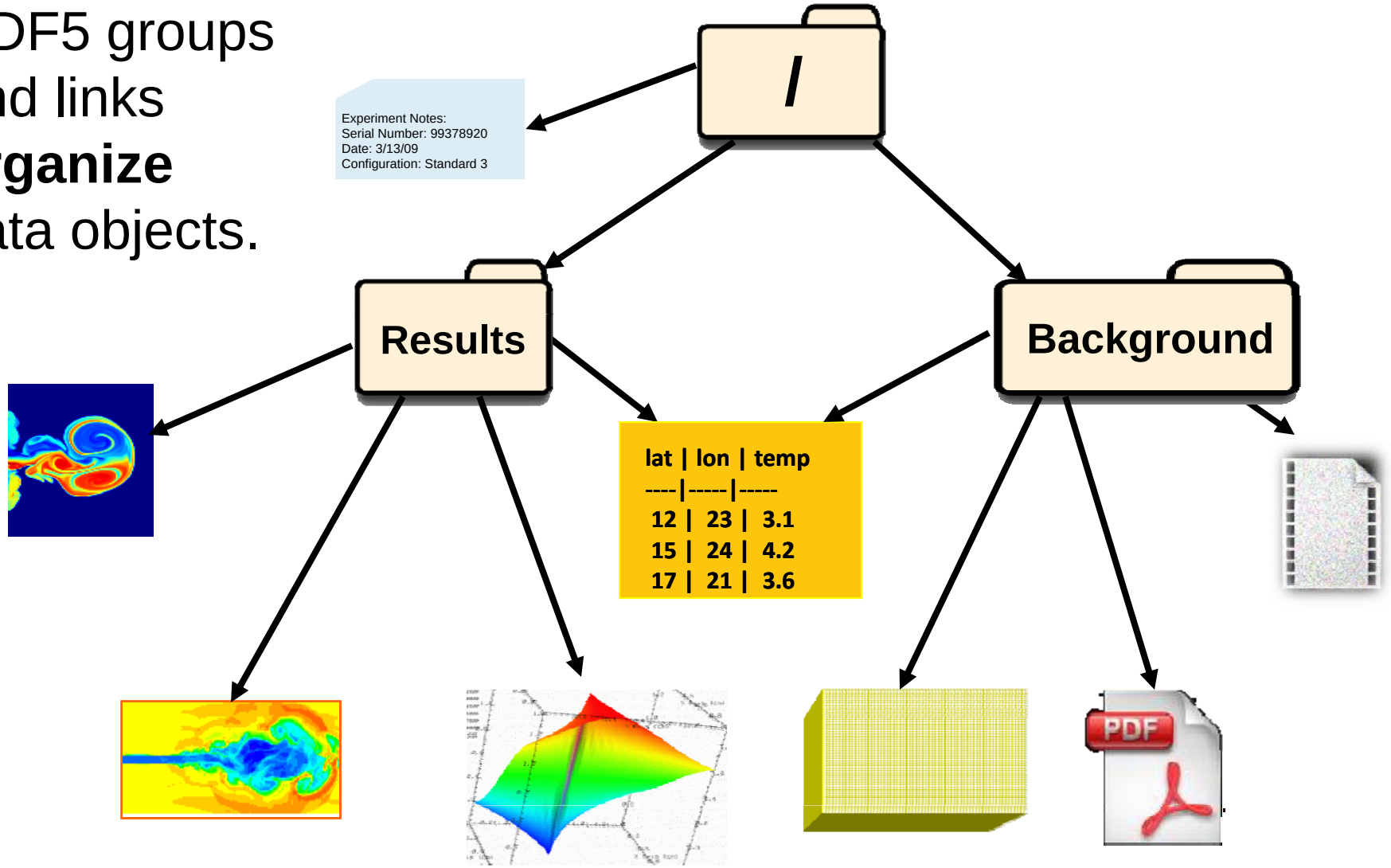
# HDF5 File

An HDF5 file is a **container** that holds data objects.



# Organizing data with HDF5

HDF5 groups and links **organize** data objects.

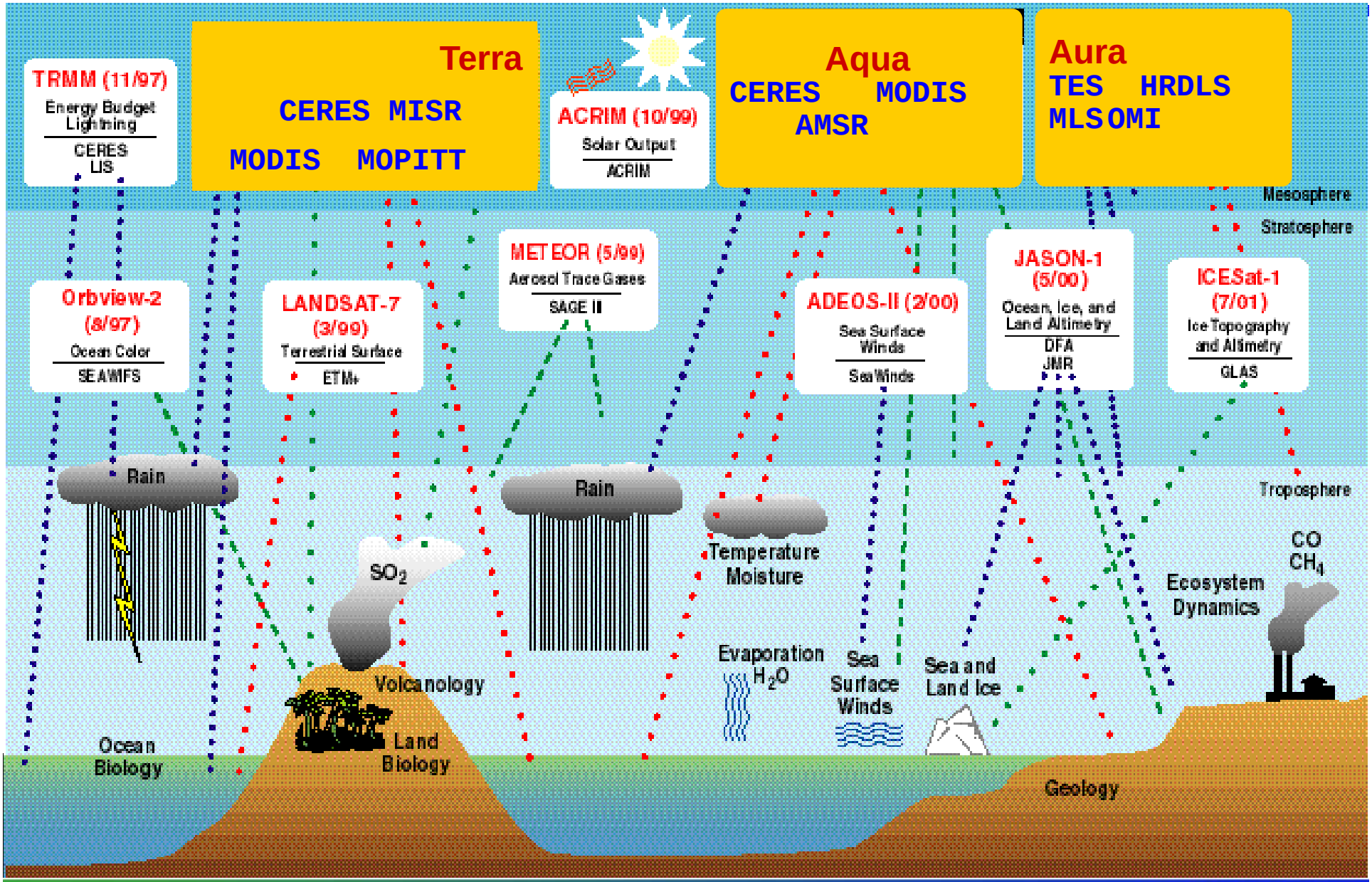


# HDF5 Technology Platform

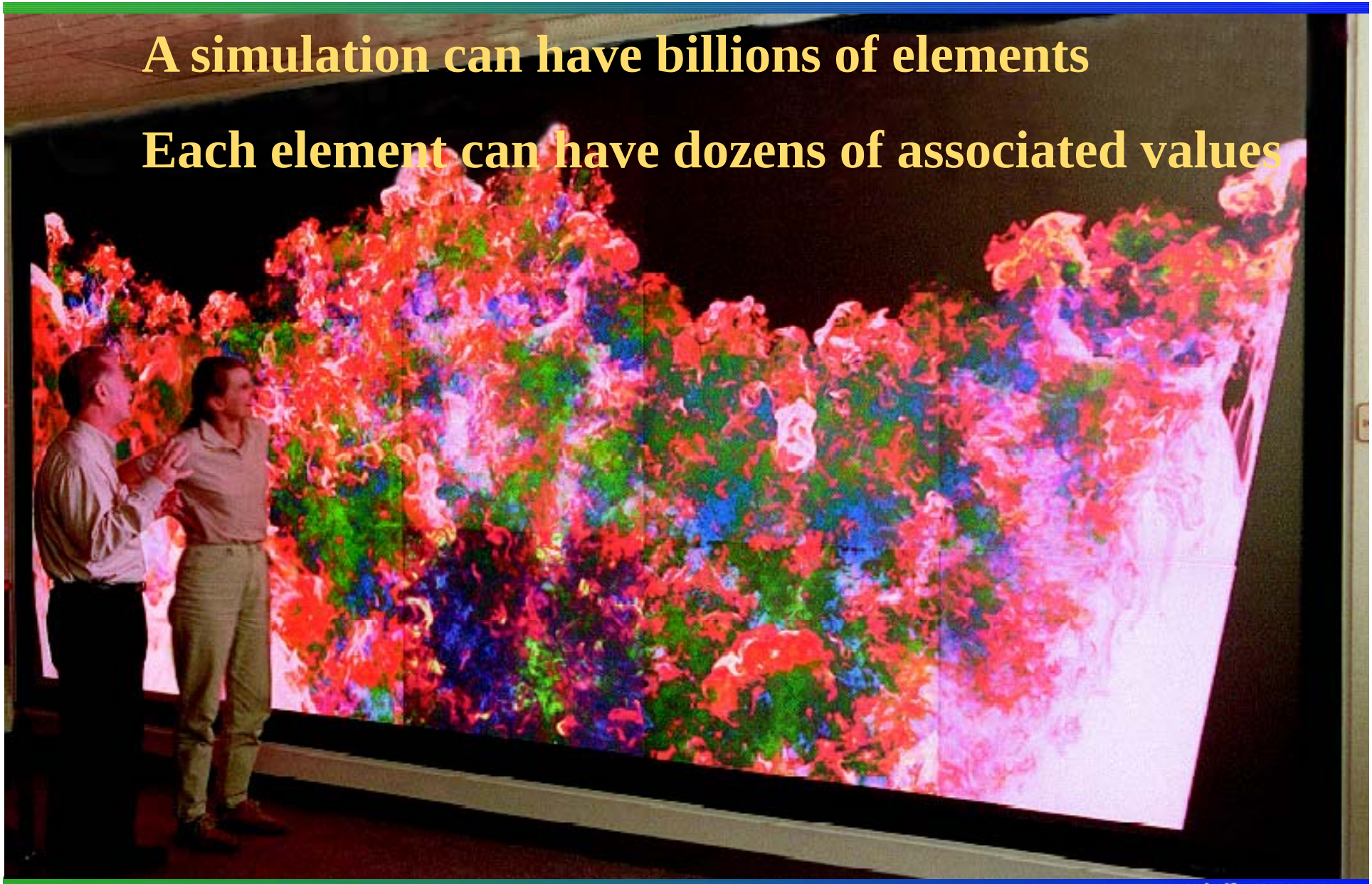
- HDF5 Software
  - Manage, analyze, view, query data
- HDF5 Data Model
  - Building blocks for data organization and storage
- HDF5 Binary File Format
  - Bit-level organization of HDF5 file

# Uses and users of HDF5

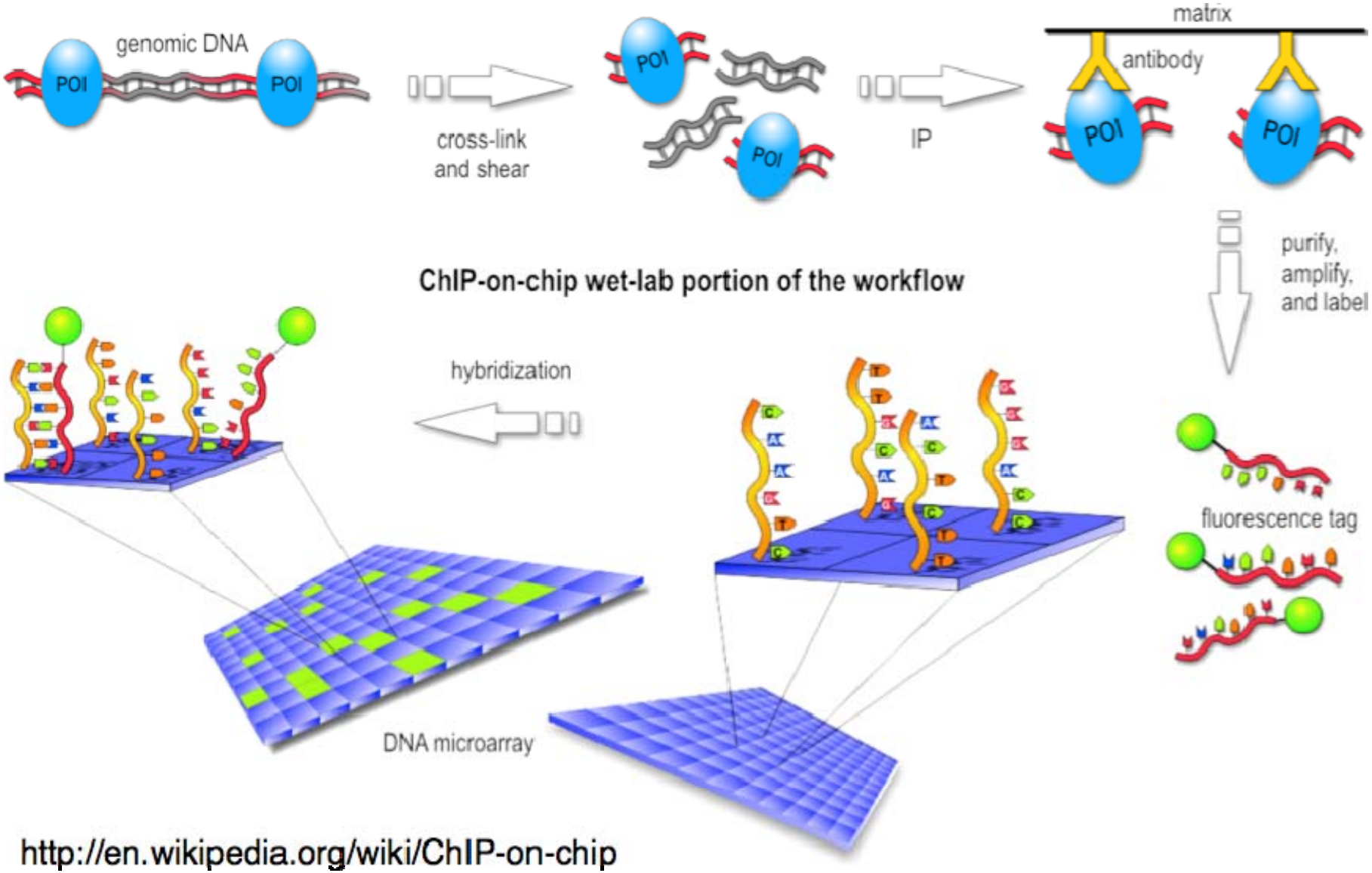
# Earth Science (Earth Observing System)



# Big simulations

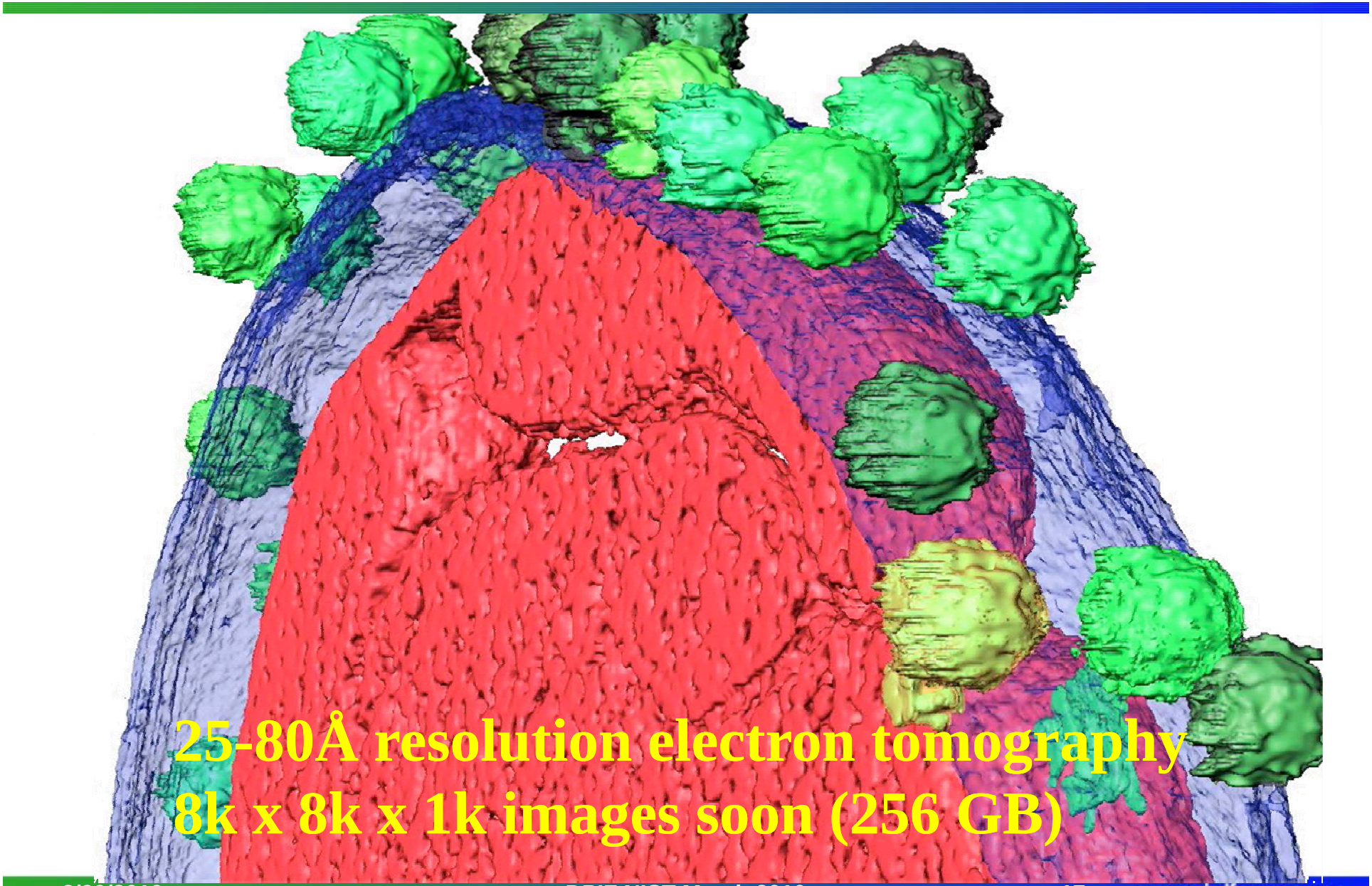


# Bioinformatics

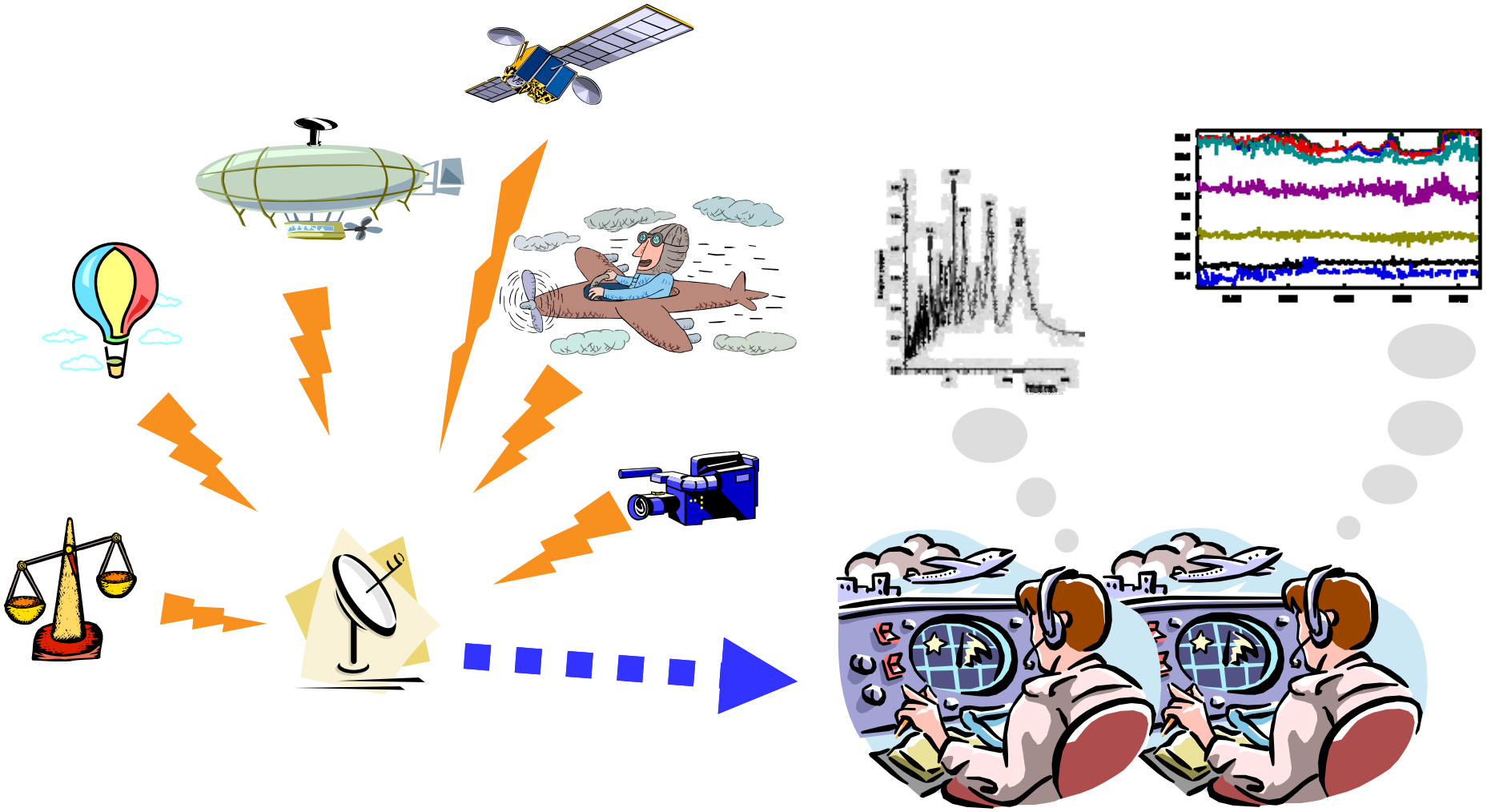



http://en.wikipedia.org/wiki/ChIP-on-chip

# Images


25-80Å resolution electron tomography
8k x 8k x 1k images soon (256 GB)

# Flight testing



# Vehicle testing

# Making movies


Spiderman 3

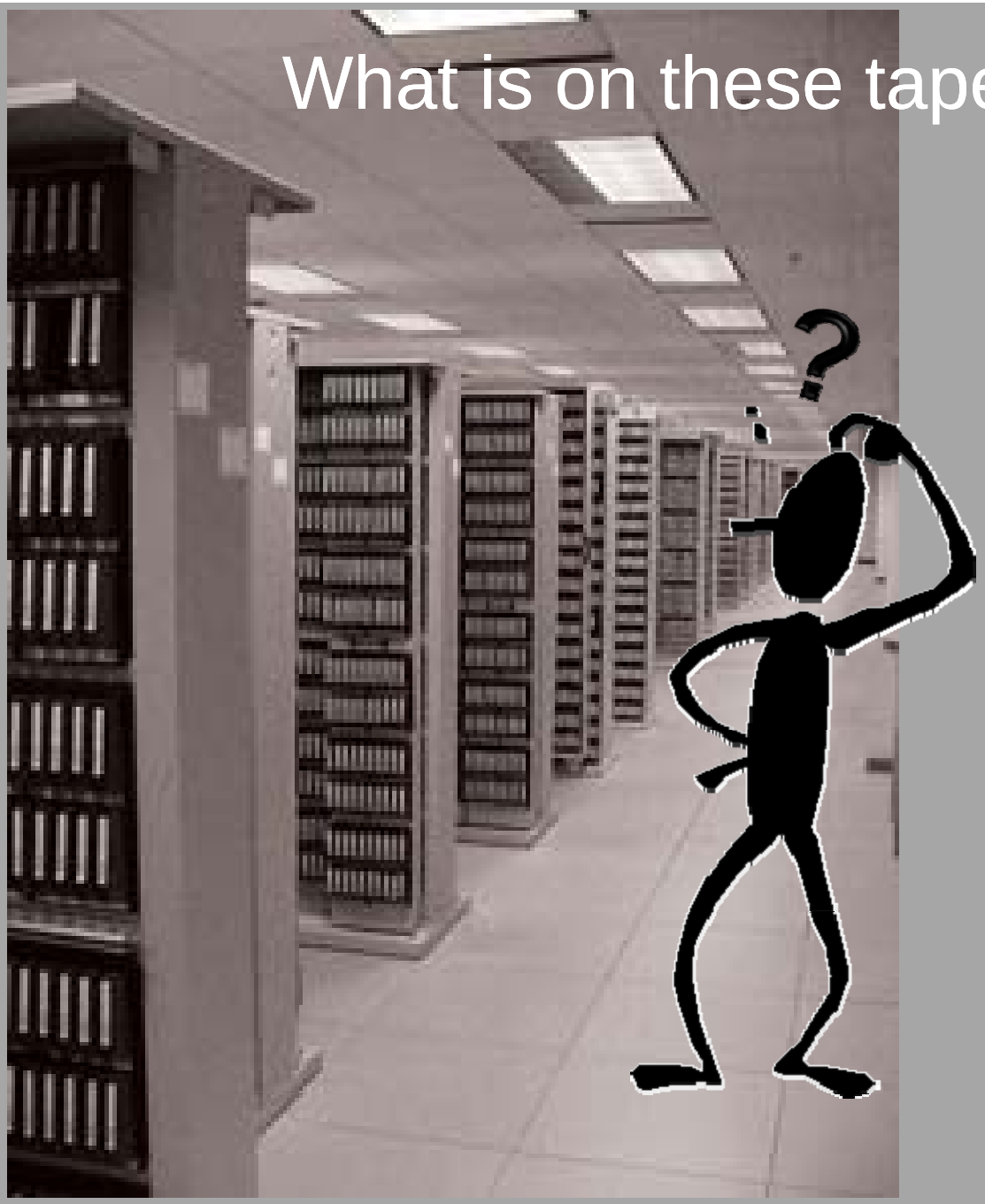
The Polar Express

# Target audience

- Applications facing big data challenges
- Academia, government, industry
- Hundreds of different of apps
- Millions of users world-wide

# Something is missing

# What is on these tapes?



# What about users in the future?

# HDF is .. *(revised)*

# HDF is.. *(revised)*

- A technology platform for addressing some of today's greatest data challenges
- **A set of features and practices to help preserve access to data for the long term**

# target audience ... (revised)

- Users today
  - Those who face challenges in organizing, accessing and integrating big, complex data.
- **Future users, and we don't know…**
  - **what data will be important to them**
  - **what they will do with the data once they get it**
  - **what knowledge and tools they will have for accessing and interpreting the data**

"What makes a good archive format?"
(1997, Folk)

"Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries"
(2002, Folk and Barkstrom)*

**And what can we do about it?**

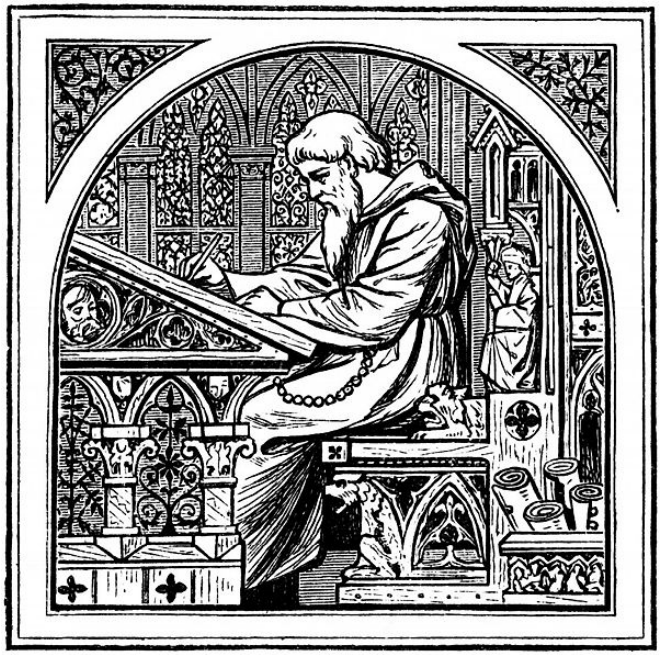
*http://www.hdfgroup.org/projects/nara/Sci_Formats_and_Archiving.pdf

# What Makes a Good Archive Format?

- Ease of Archival Storage
  - Compactness
  - Size
  - Ability to aggregate related objects.
- Ease of Archival Access
  - Raw I/O efficiency
  - Ease of subsetting
- Usability
  - Popularity
  - Availability of readers
  - Ability to embed data extraction software in the files
  - Ease of implementing readers
  - Simplicity
  - Ability to name file elements

# What Makes a Good Archive Format?

- Support for Data Scholarship
  - Provenance traceability
  - Rigorous definition
  - Self-describing
  - Referential extensibility
  - URN  embedding
  - Citability
- Support for Data Integrity
  - Source verification
  - File corruption detection & correction



# What Makes a Good Archive Format?

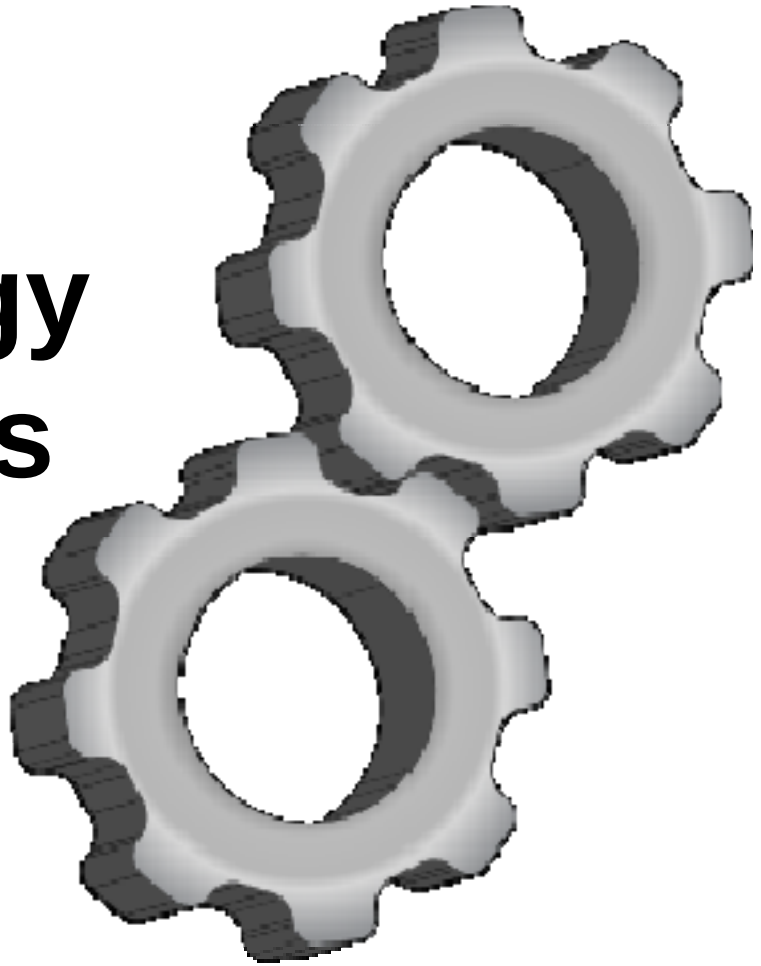
- Maintainability and Durability
  - Long-term institutional support
  - Suitability for a variety of storage technologies
  - Stability
  - Formal (BNF- or XML-like) description of format
  - Multi-language implementation of library software
  - Open Source software or equivalent

# HDF strategies for long-term preservation

**Technological**

**Institutional**

# Technology strategies



# A simple, durable but evolvable model and implementation

John of England signs Magna Carta

# Self-description

# Specification documentation

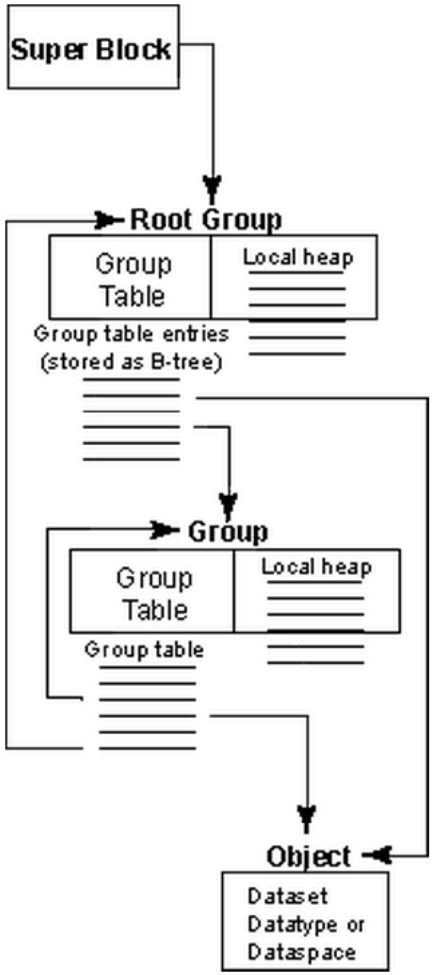
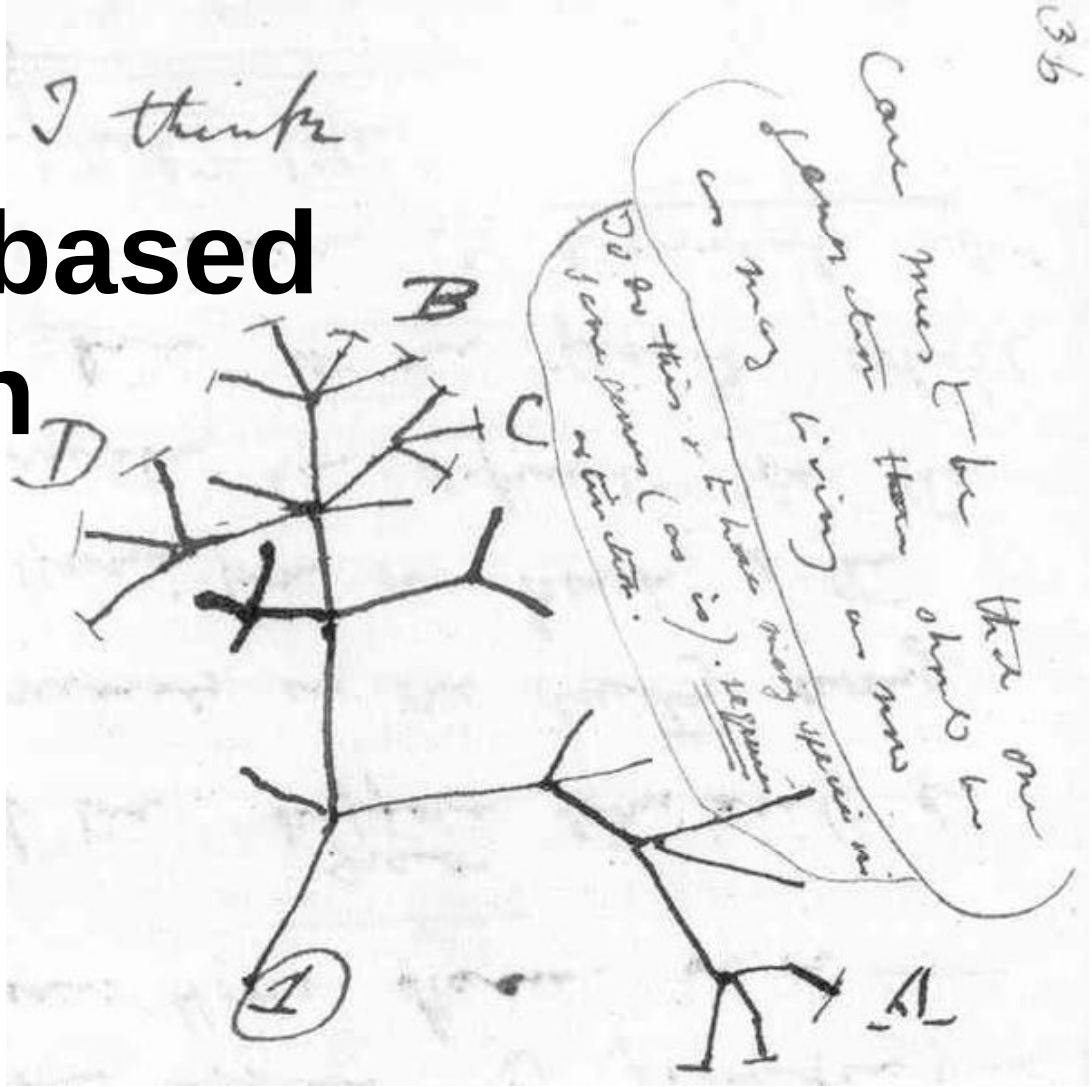



**Figure 1:** Relationships among the HDF5 root group, other groups, and objects

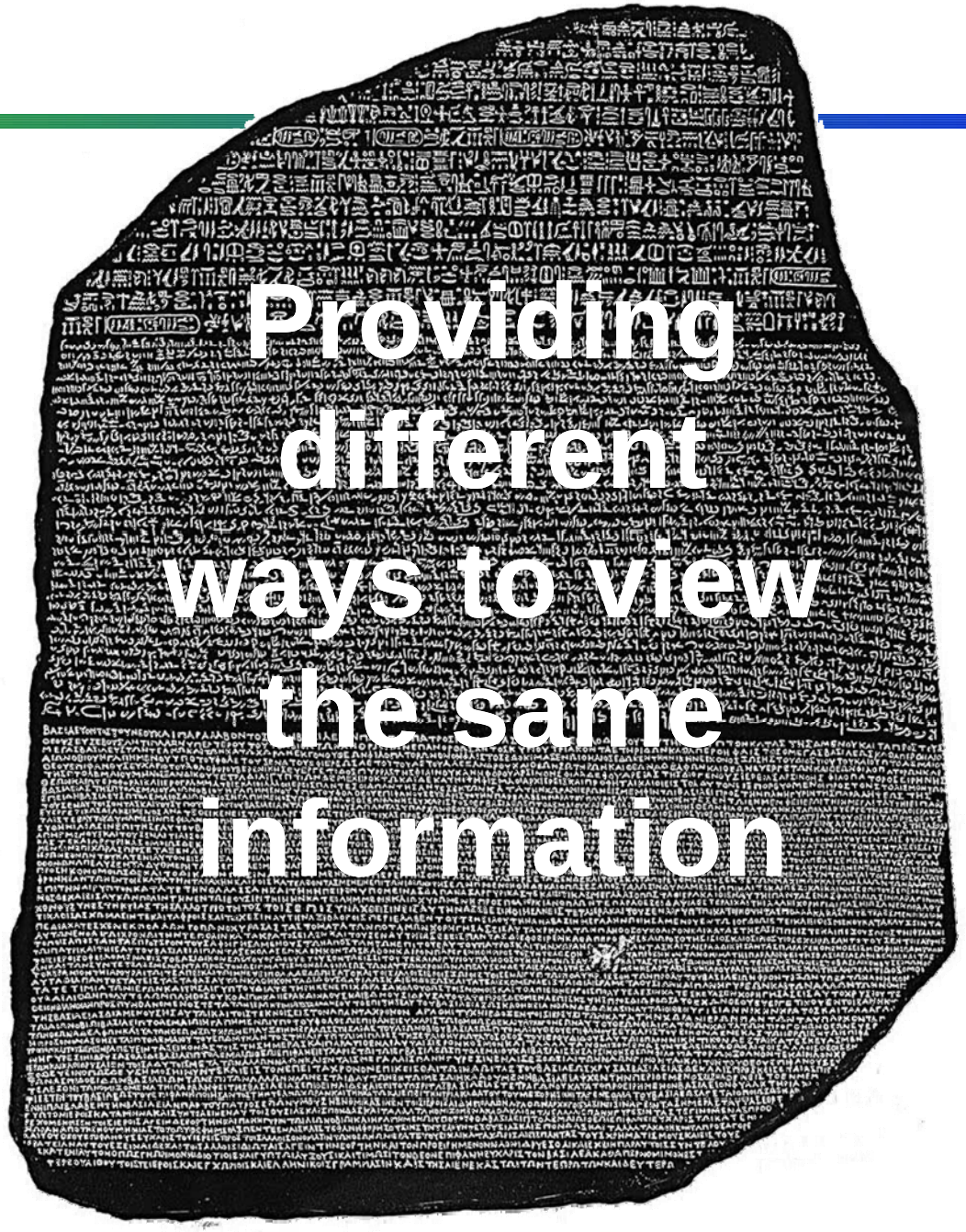
# Preservation-based evolution




Darwin's first evolutionary tree - 1837

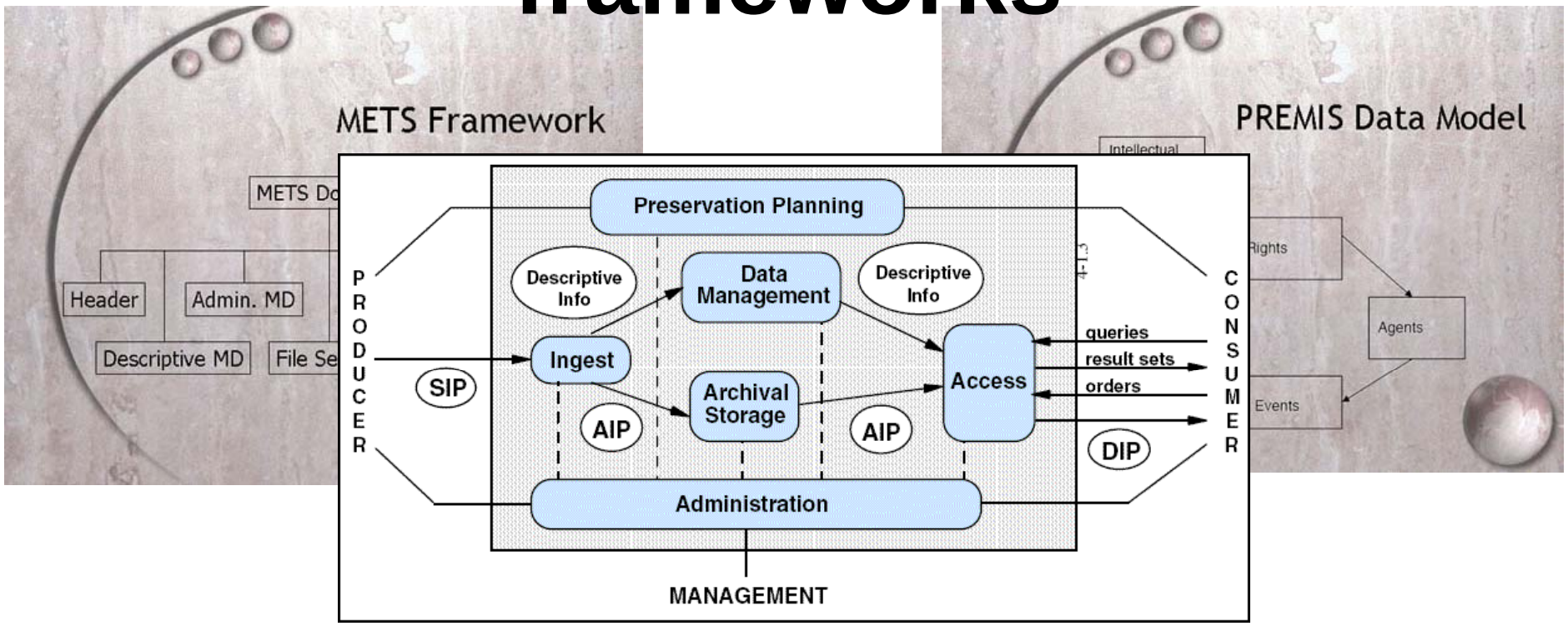
# Preservation-based evolution

a technology development strategy that allows the software and format to evolve, at the same time giving legacy applications a decent chance to meet their users' needs, and preserving access to all data.



# Providing different ways to view the same information

# Integration with preservation frameworks



# Institutional strategies

# Long-term institutional support

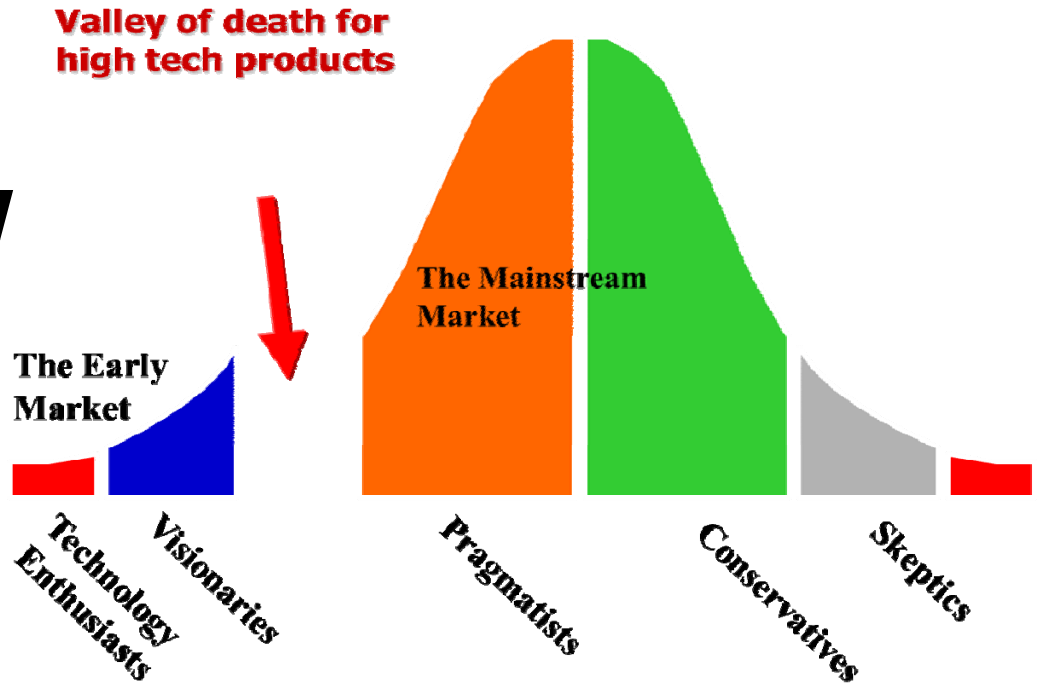
# A mission-driven business

# Human, financial, legal foundations for sustainability

# Open source

# One keeper of the format and software

# Cross the chasm to new users and applications



# Promoting standardization

# Summary

- Technical strategies
  - A simple, durable but evolvable model and implementation
  - Self-description
  - Specification documentation
  - Preservation-based evolution
  - Providing different ways to view the same information
  - Integration with preservation frameworks
- Institutional strategies
  - Long-term institutional support
  - A mission-driven business
  - Human, financial, legal foundations for sustainability
  - Open source
  - One keeper of the format and software
  - Cross the chasm to new users and applications
  - Promoting standardization

# HDF Group Mission

**To ensure long-term accessibility of HDF data through sustainable development and support of HDF technologies.**

# Thank you.