HDF efforts to improve data preservation

Mike Folk
The HDF Group

ESIP Federation Summer Meeting 2009

HDF = Hierarchical Data Format

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

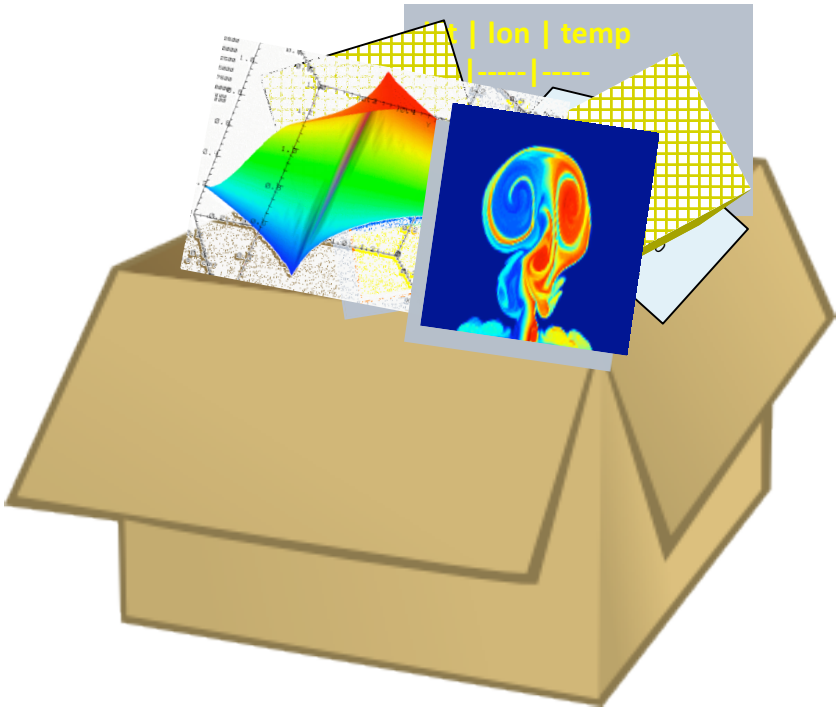
- HDF4 is the first HDF format
  - Originally called HDF
  - First release was 1988

HDF5 Technology Platform

- HDF5 Data Model
  - Defines the “building blocks” for data organization and specification
  - Files, Groups, Links, Datasets, Attributes, Datatypes
- HDF5 Software
  - Tools
  - HDF5 Library
    - C, Fortran, C++ Language Interface
- HDF5 Binary File Format
  - Bit-level organization of HDF5 file
  - Defined by HDF5 File Format Specification

HDF5 File

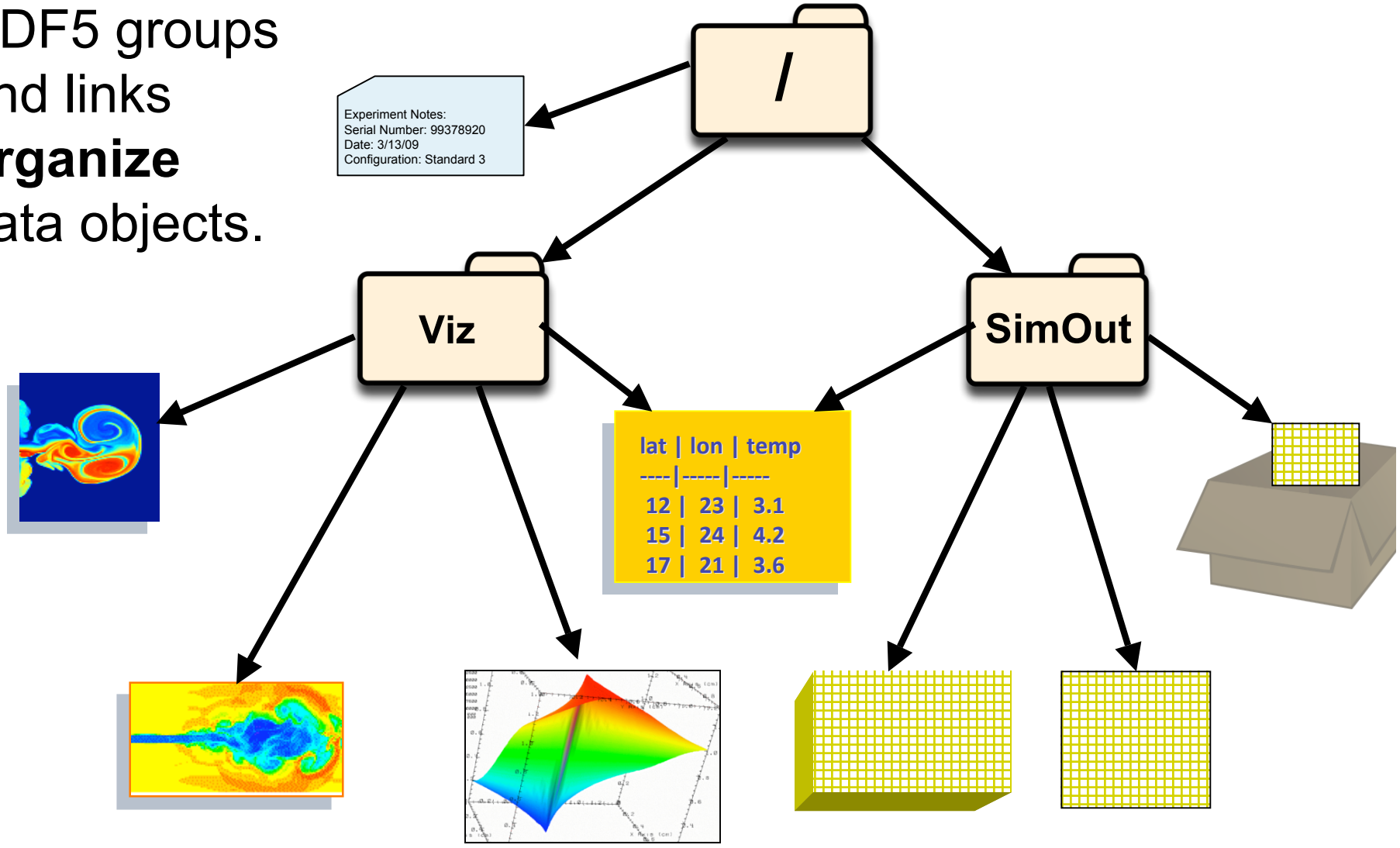
An HDF5 file is a **container** that holds data objects.



An HDF5 file is not necessarily a file on a filesystem.

HDF5 Groups and Links

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



Two Questions

Al Fleig (1990)

Ben Kobler (1996)

"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)*

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

WHAT MAKES A GOOD ARCHIVE FORMAT? (2002)

- Ease of Archival Storage
  - Compactness
  - Size
  - Ability to aggregate many objects in a single file.
- 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? (2002)

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

WHAT MAKES A GOOD ARCHIVE FORMAT? (2002)

- 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

Technological approaches

Incorporate preservation-friendly technologies

- HDF5 Archival Information Package (AIP) to archive HDF EOS2 data and metadata
  - A NOAA Scientific Data Stewardship project
  - A collaboration with Ruth Duerr (NSIDC)
- HDF5 AIP in METS
  - Define HDF5 AIP
  - Tool to create and edit HDF5 AIP
  - Collaboration with iRODS

Create alternate views of the data

- netCDF harmonies
  - netcdf API for HDF4 (1992)
  - Access to certain EOS data through netCDF-4 API
  - netCDF-4
- FITS harmonization (1998)
- Simplified access to data via independent mapping of hdf4 data with XML

Define durable but evolvable model, features, format

- Simple, comprehensive, extensible data model
- Scalable size, complexity
- "User block"

"Evolve with compatibility"

- Develop guidelines to keep evolution under control
- Develop process to ensure cross-generational compatibility

Organizational and social strategies

HDF Group Mission

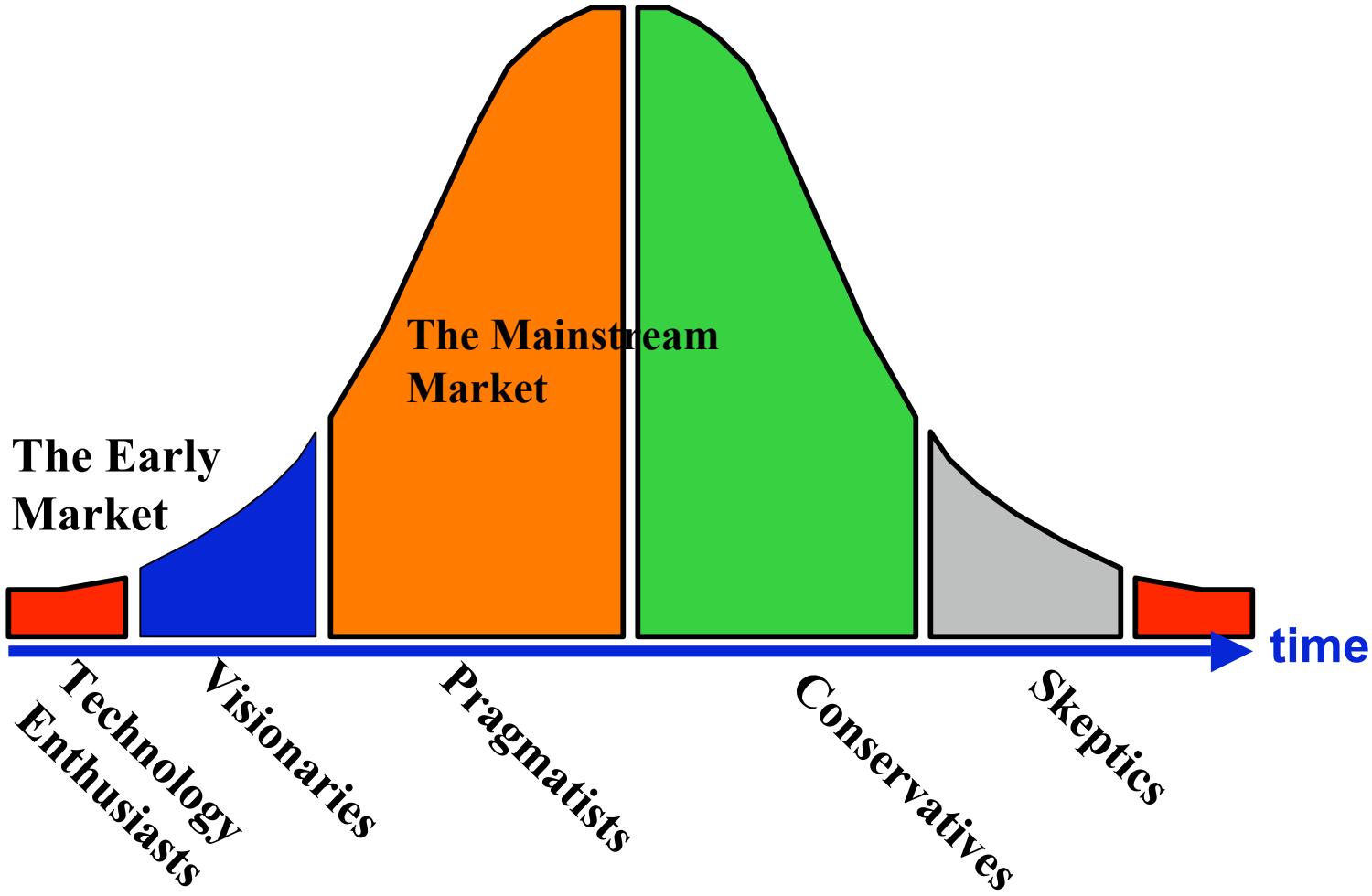
To ensure long-term accessibility of HDF data through sustainable development and support of HDF technologies.

HDF Group Challenges

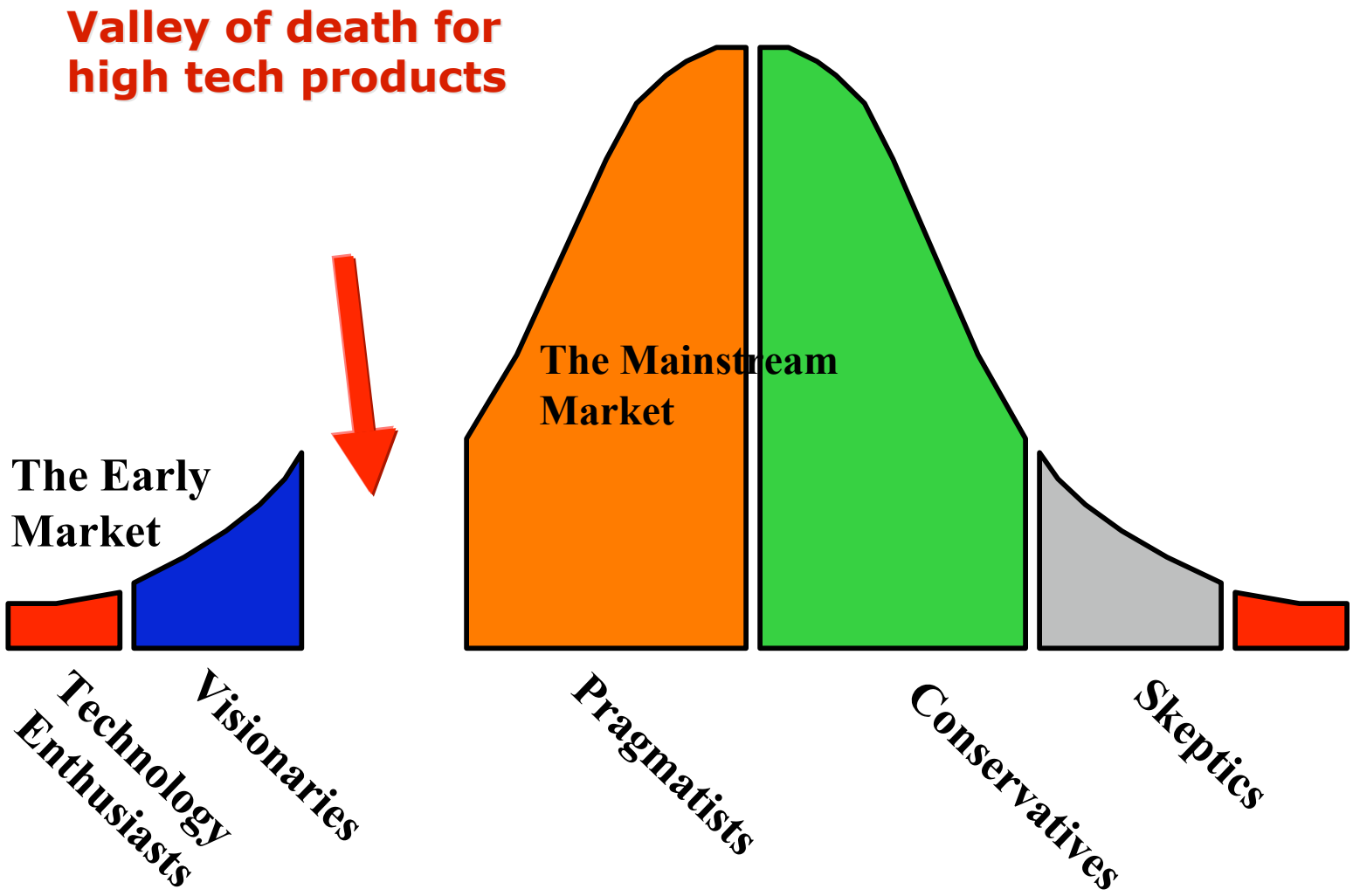
- Establish a mindset that values long-term preservation mission
- Find a business model that is sustainable
- Develop sustainability assets: funding, knowledge, people

Cross the Chasm
From innovators/visionaries
To pragmatists

Technology Markets



Crossing the chasm



Broaden the base - increase the need for long-term support

- Diverse application domains
- Diverse institutional types (govt, commercial, academic)
- Vendors
- More large, stable institutional use and support
- Whole product support
  - Tools
  - technology interoperability (iRODS, opendap, XML, RDBMS, MATLAB, ...)

Promote standard usage within domains

- Examples
  - HDF-EOS
  - CGNS
  - HDF Time history (Aerospace)
  - NeXuS
  - BioHDF
- components
  - Unified data model
  - API and implementation (preferably multi-language)
  - Tools
  - Lots of data

Steal ideas from successes, such as netCDF, PDF, FITS, TIFF, PDS

Thank you.