ESML, Subsetting, Mining Tools

MODIS Science Team Meeting
July 24, 2002

Sara Graves
Rahul Ramachandran
Information Technology and Systems Center (ITSC)
University of Alabama in Huntsville (UAH)
www.itsc.uah.edu

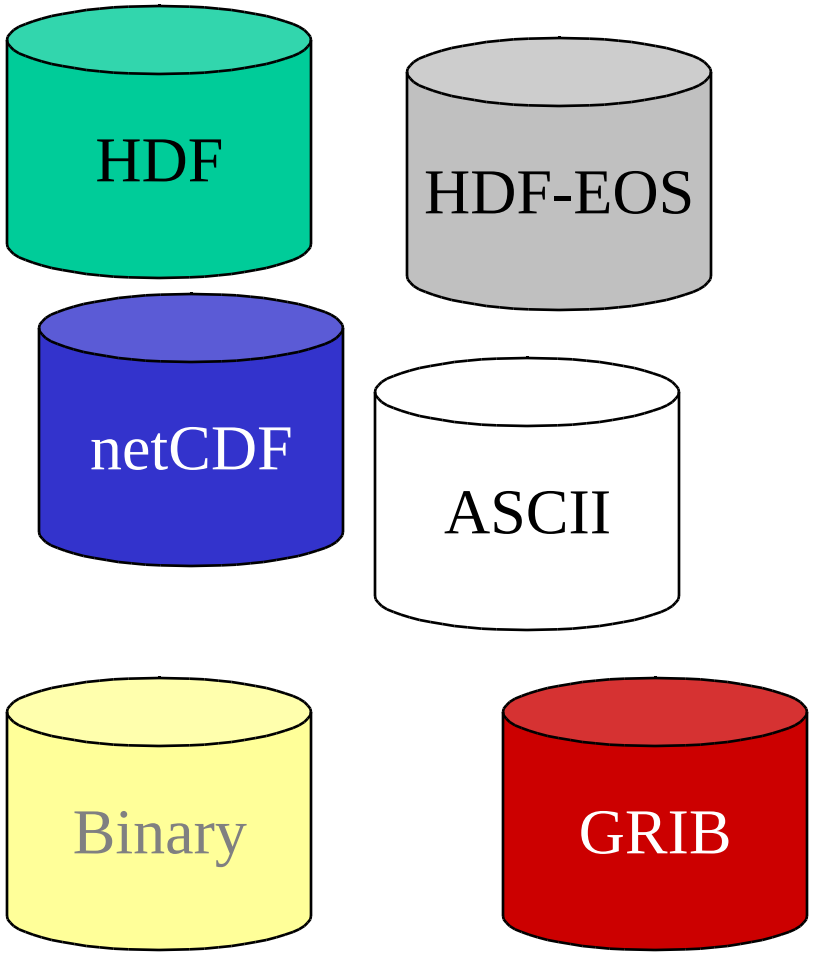
Tools Encompassing All Phases of Scientific Analysis

- Science Data Usability
  - Data/Application Interoperability
    - Earth Science Markup Language (ESML)
- Science Data Preprocessing
  - Subsetting
    - Various Subsetting Tools such as HEW
- Science Data Analysis
  - Data Mining
    - Algorithm Development and Mining (ADaM) System
- Mission/Project/Field Campaign Coordination
  - Electronic Collaboration

Science Data Usability

http://esml.itsc.uah.edu

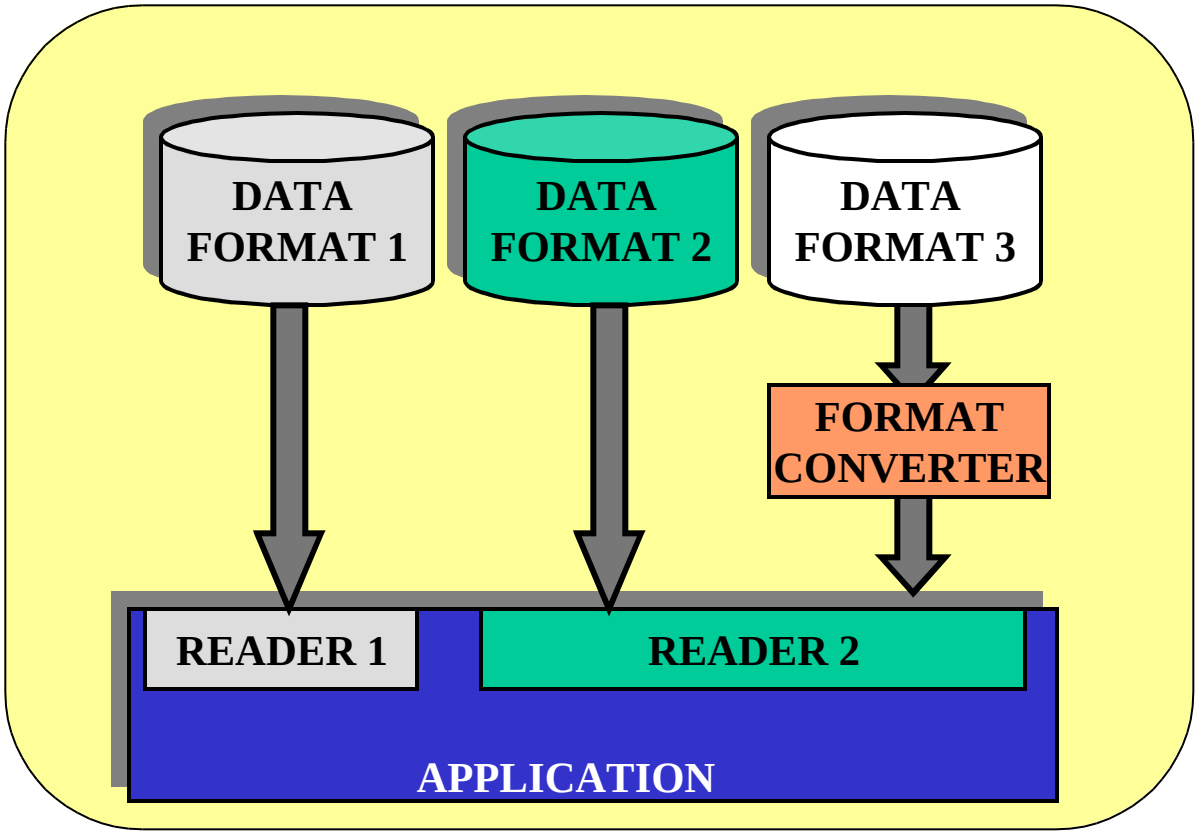
Earth Science Data Characteristics



- Different formats, types and structures (18 and counting for Atmospheric Science alone!)
- Different states of processing (raw, calibrated, derived, modeled or interpreted)
- Enormous volumes

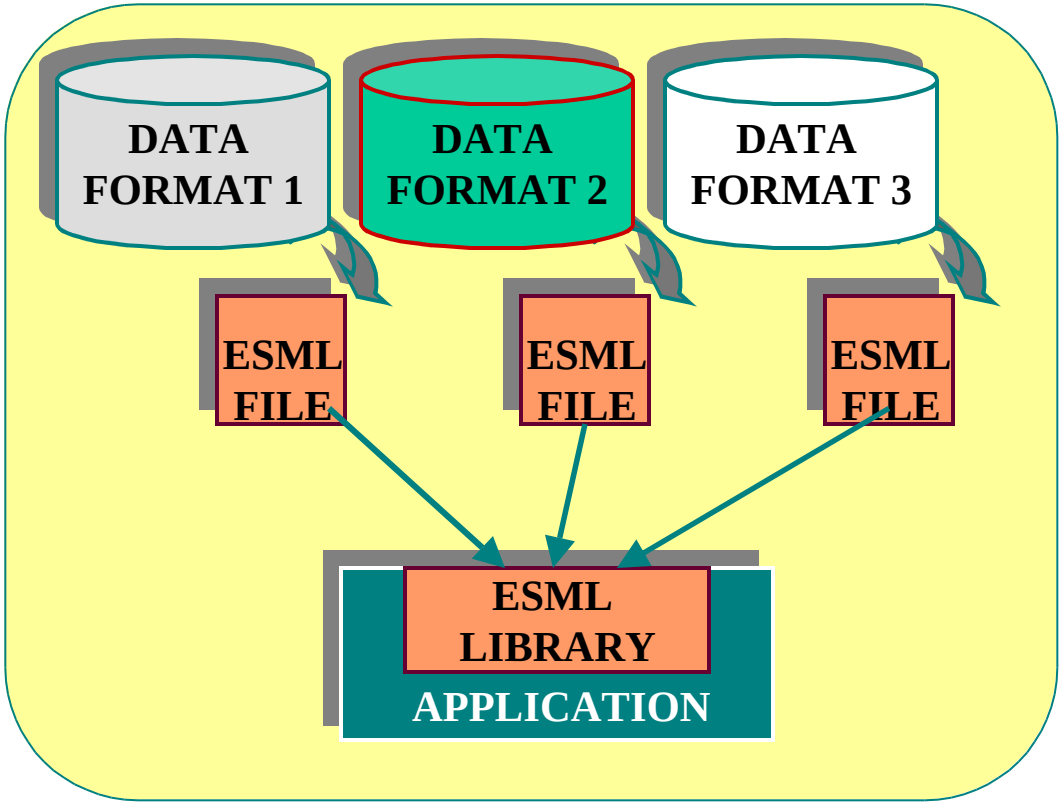
➢ **Heterogeneity leads to Data usability problem**

Data Usability Problem



- Requires specialized code for every format
  - Difficult to assimilate new data types
  - Makes applications tightly coupled to data
- One possible solution - enforce a Standard Data Format
  - Not practical, especially for legacy datasets

ESML Solution



- ESML (external metadata) files containing the structural description of the data format
- Applications utilize these descriptions to figure out how to read the data files resulting in data interoperability for applications

What is ESML?

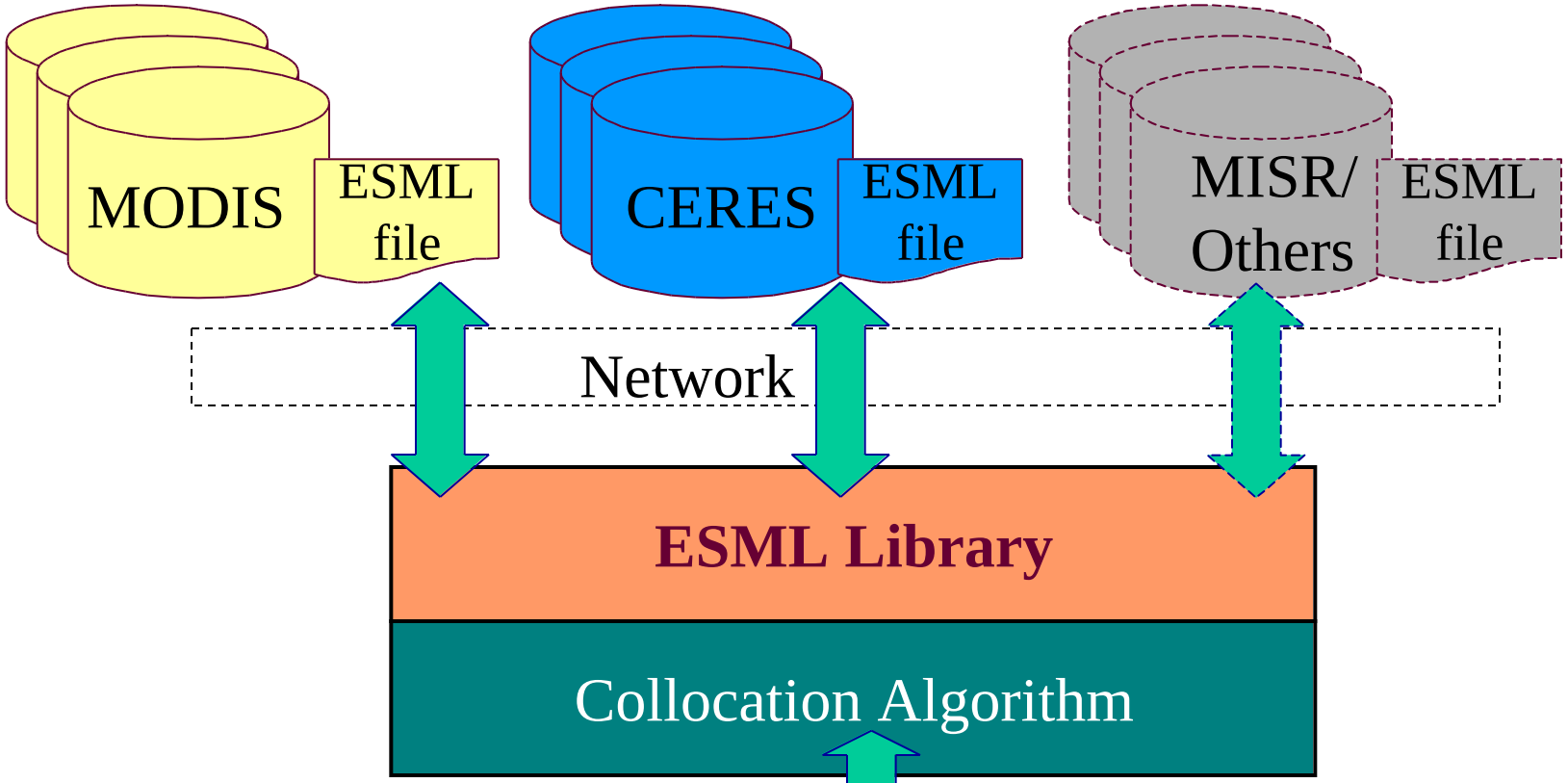
- It is a specialized markup language for Earth Science metadata based on XML
- It is a machine-readable and -interpretable representation of the structure and content of any data file, regardless of data format
- ESML description files contain external metadata that can be generated by either data producer or data consumer (at collection, data set, and/or granule level)
- ESML provides the benefits of a standard, self-describing data format (like HDF, HDF-EOS, netCDF, geoTIFF, …) without the cost of data conversion
- ESML is an Interchange Technology that allows data/application interoperability

ESML Tools/Products Available

http://esml.itsc.uah.edu

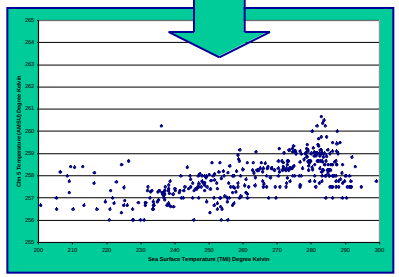
Tools/Products	Features
ESML Schema	•Defines ASCII, Binary (McIDAS), HDF-EOS, GRIB formats (more formats to follow) •Provides preprocessing, wild card, symbols and semantic tagging capabilities
ESML Library (C++/Java JNI)	•WINDOWS and LINUX •URL access •Handles ASCII, Binary (McIDAS), HDF-EOS, GRIB files •Handles preprocessing, wild card and symbols
ESML Editor	Ability to write and validate ESML descriptions
ESML Data Browser	Ability to browse data files using the ESML description files

MODIS/CERES Collocation Application



Purpose:

- *To study the relationship between shortwave flux and cloud/aerosol properties*
- *Important for climate change studies*



Analysis

Scientists can:

- *Select remote files across the network*
- *Select fields by modifying semantic tags in the ESML file*

Science Data Preprocessing

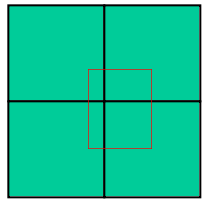
http://subset.org

Currently Available/Planned Subsetting Applications

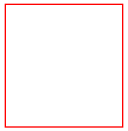
- HEW Subsetting
  - Complete System (available)
  - Subsetting Engine Only (available)
  - Subsetting Center (available)
  - SPOT - Subsettability Checker (available)
  - HEW Integration with ECS (in work)
  - Remote Subsetting Service (planned)
  - Subsetting as a Web Service (planned)
- Customized Subsetting
  - MODIS tools (available)
  - Coarse-grain SSM/I Subsetter (available)
- General Purpose Customizable Subsetting
  - Based on ADaM Data Mining Engine (available)
  - Subsetting Tool using ESML (in work)

Tools developed for MODIS Scientists

- MODIS – Land, Quality Assessment
  - •*modland* – subsetter for MODIS gridded data
  - • *stitcher* – pieces together 2 or 4 contiguous MODIS tiles

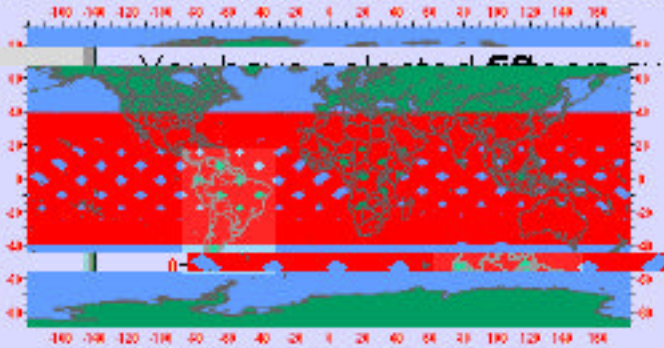


- MODIS – Atmosphere
  - •*modair* - specialized subsetter for MODIS swaths



Select Geotemporal Bounds

You have selected **fifteen** swaths for subsetting from each file.



- Red areas or dots indicate coverage area
- Drag an edge or corner of highlight to resize selection area
- Drag the middle of the highlight to move selection area
- For more precision, type in values in the boxes below

Help!
Home
Select directory
Select files
Select objects
Select bounds

Select the geotemporal bounds of your area of interest. Then click the "Next" button at the bottom of the screen. To reset the bounds to their default values, click the "Reset" button.

If you want to subset by spatial bounds, select the bounding rectangle of your area of interest

Top
39.04
17.000

Left
-180.00
-88.000

ddd ddddd or
ddd:mm:ss.ttt

-32.000
Right
180.00

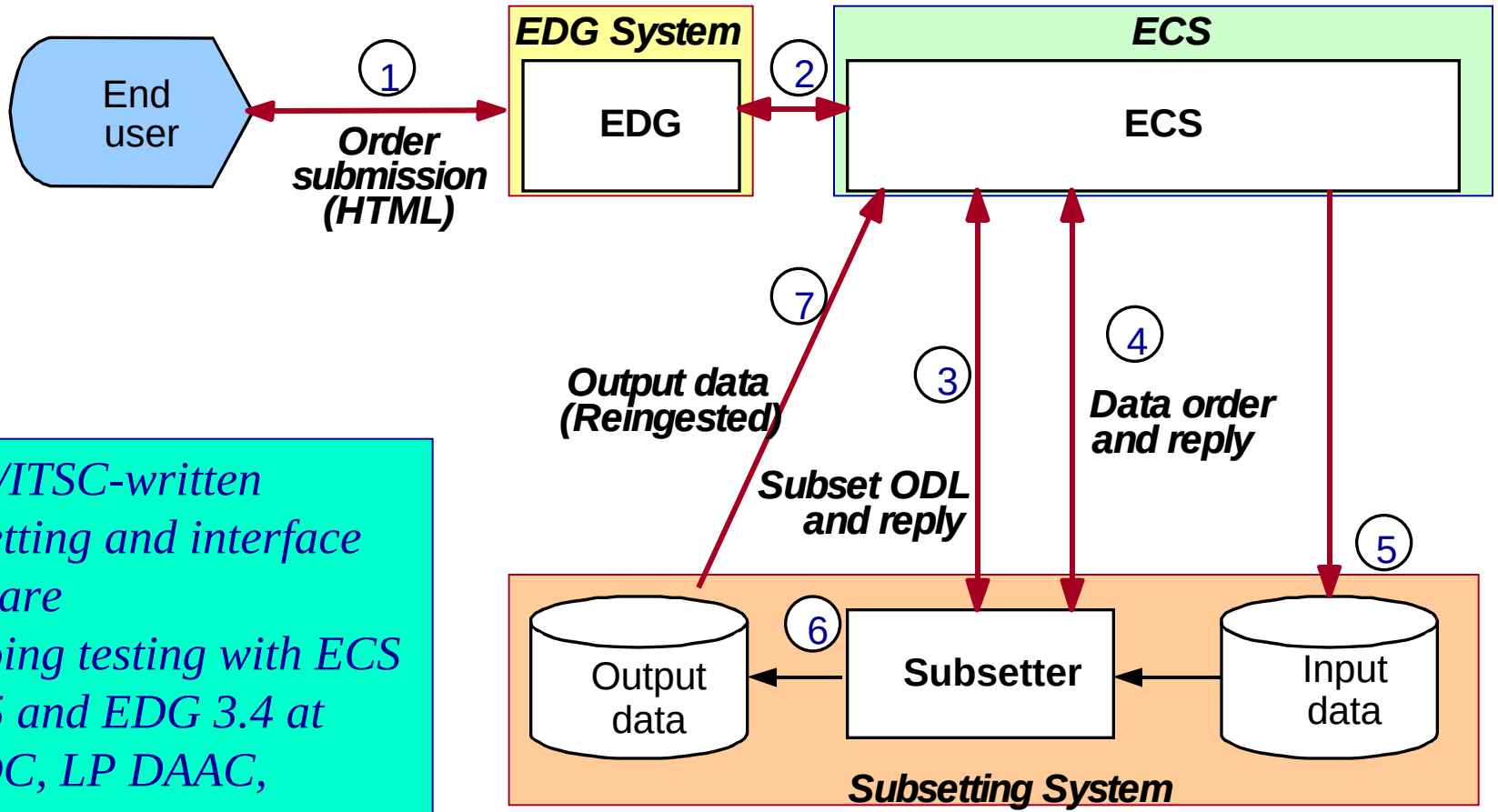
117.000

Bottom
-39.12

If you'd like to subset by time, select the date and time span

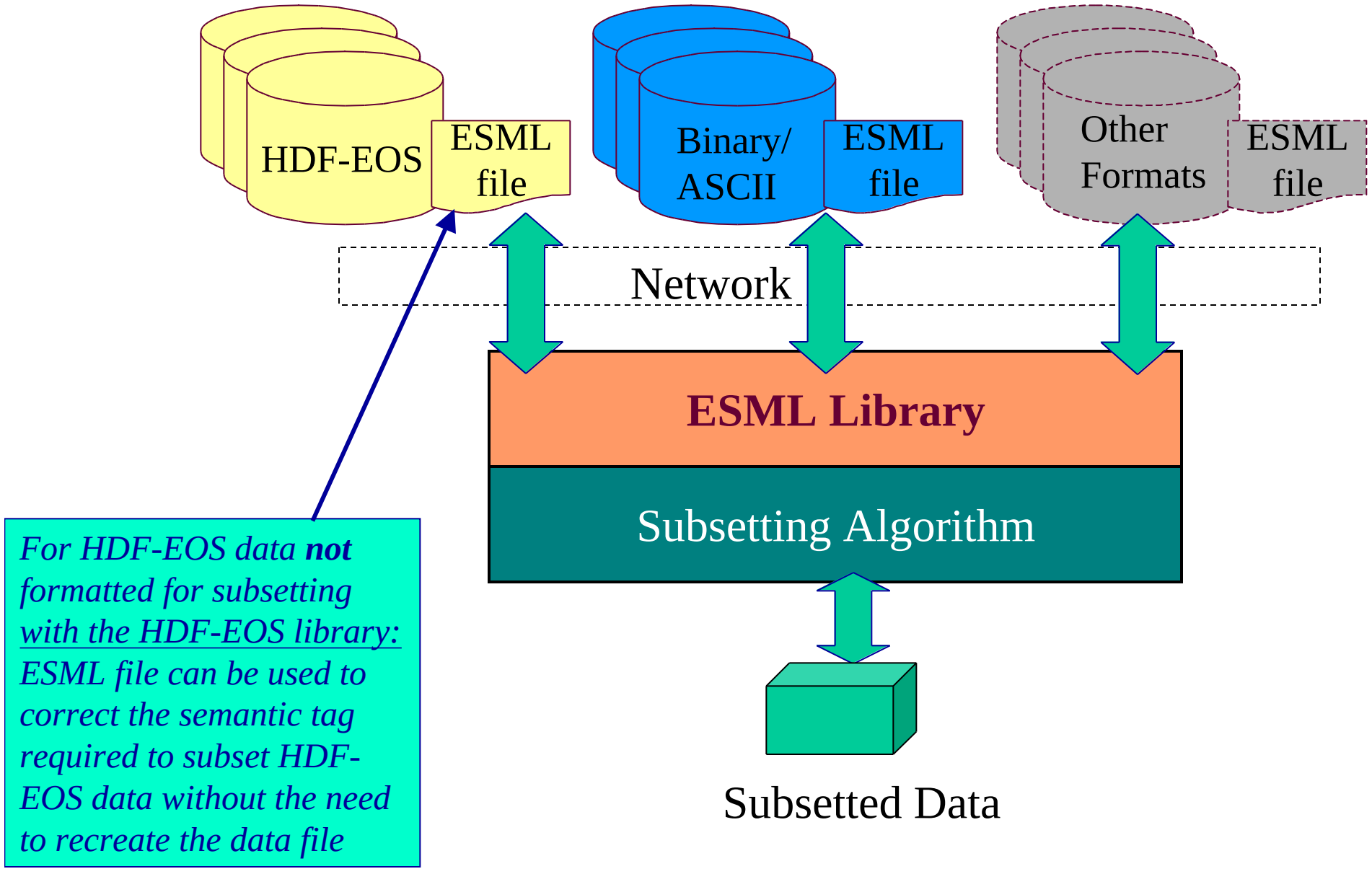
	yyyy-mm-dd	hh:mm:ss.ttttt		yyyy-mm-dd	hh:mm:ss.ttttt
From	2002-07-17	01:25:01.290100	thru	2002-07-18	00:30:42.287964
	2002-07-17	01:25:01.290100		2002-07-18	00:30:42.287964

HEW integration with ECS

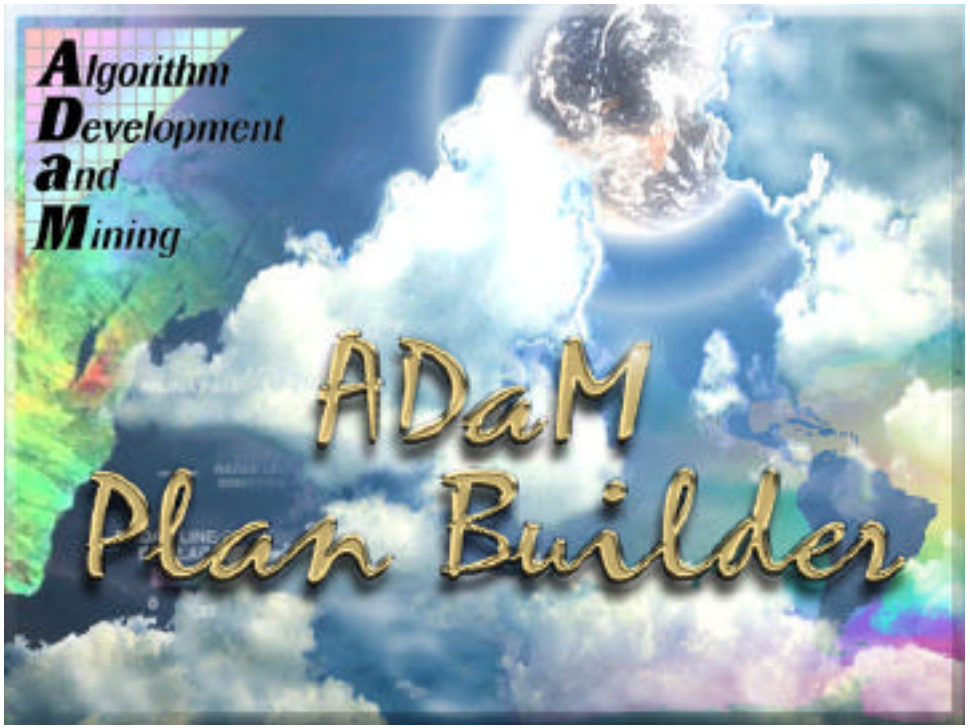


- *UAH/ITSC-written subsetting and interface software*
- *Ongoing testing with ECS 6a.05 and EDG 3.4 at NSIDC, LP DAAC, GDAAC*
- *Enhancements for DAACs may be made*

ESML enabled generic Subsetter



Science Data Analysis

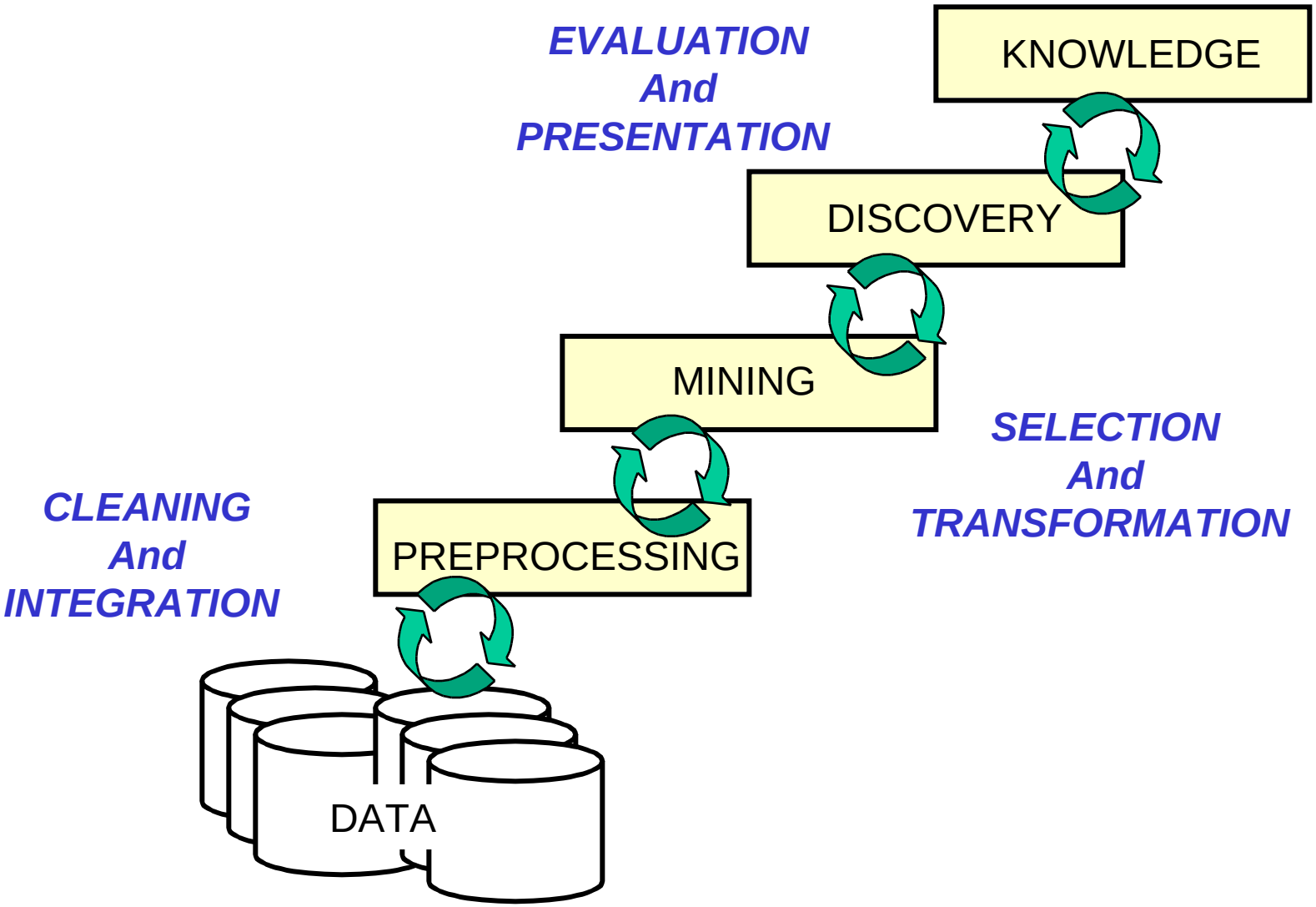


http://datamining.itsc.uah.edu

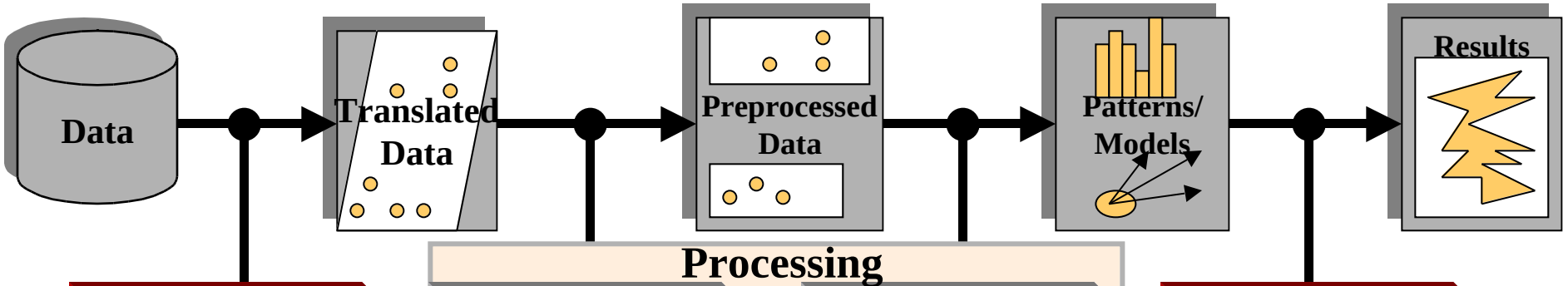
Data Mining

- Data Mining is the task of discovering interesting patterns/anomalies and extracting novel information from large amounts of data
- Data Mining is an interdisciplinary field drawing from areas such as statistics, machine learning, pattern recognition and others

Iterative Nature of the Data Mining Process



ADaM Engine Architecture



Input

HDF
HDF-EOS
GIF PIP-2
SSM/I Pathfinder
SSM/I TDR
SSM/I NESDIS Lvl 1B
SSM/I MSFC
  Brightness Temp
US Rain
Landsat
ASCII Grass
Vectors (ASCII Text)

Intergraph Raster
Others...

Processing

Preprocessing

Selection and Sampling
  Subsetting
  Subsampling
  Select by Value
  Coincidence Search
Grid Manipulation
  Grid Creation
  Bin Aggregate
  Bin Select
  Grid Aggregate
  Grid Select
  Find Holes
Image Processing
  Cropping
  Inversion
  Thresholding
Others...

Analysis

Clustering
  K Means
  Isodata
  Maximum
Pattern Recognition
  Bayes Classifier
  Min. Dist. Classifier
Image Analysis
  Boundary Detection
  Cooccurrence Matrix
  Dilation and Erosion
  Histogram
Operations
  Polygon
Circumscript
  Spatial Filtering
  Texture Operations
Genetic Algorithms
Neural Networks
Others...

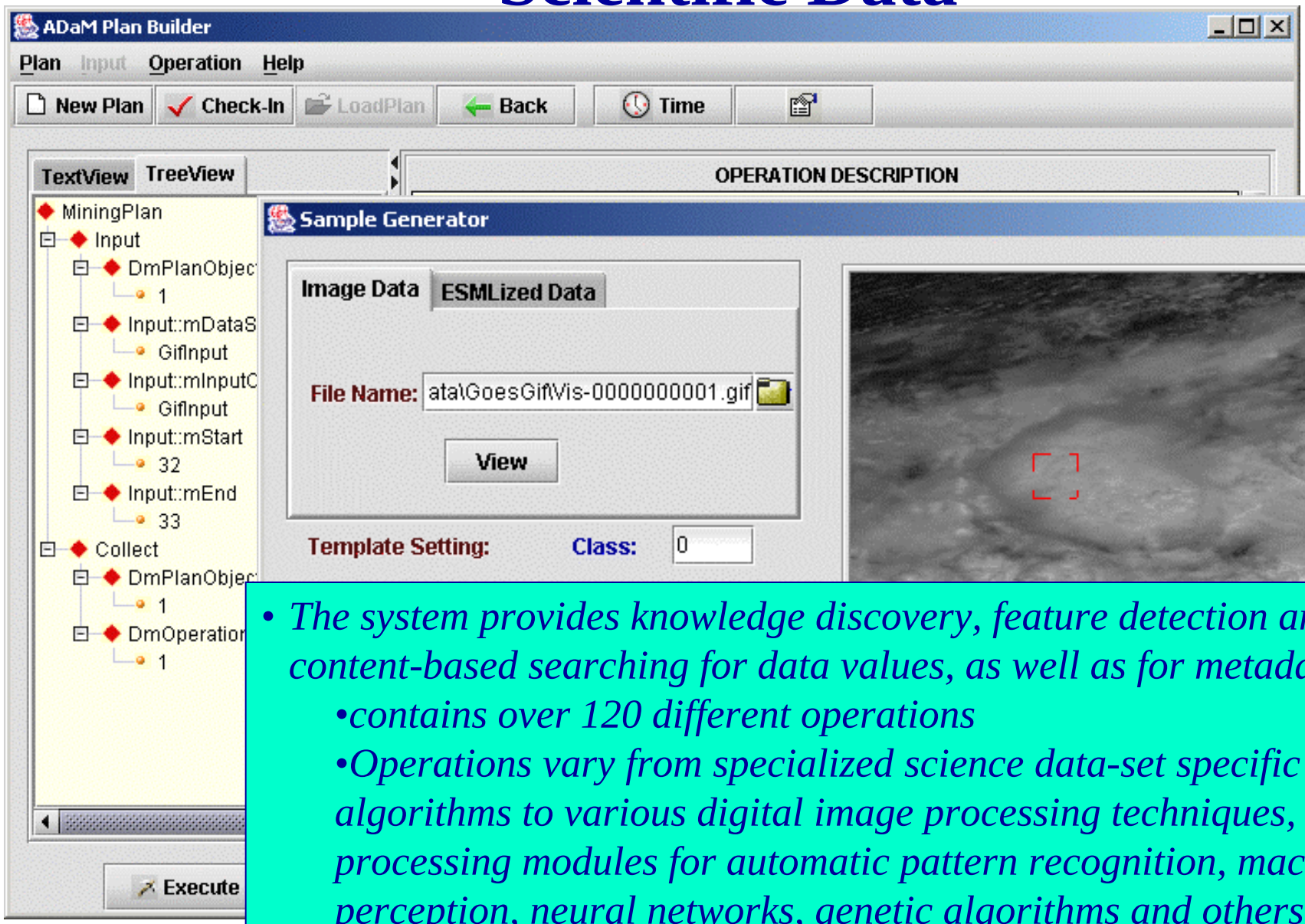
Output

GIF Images
HDF-EOS
HDF Raster Images
HDF SDS
Polygons (ASCII, DXF)
SSM/I MSFC
  Brightness Temp
TIFF Images
Others...

Reasons for Building a Data Mining Environment

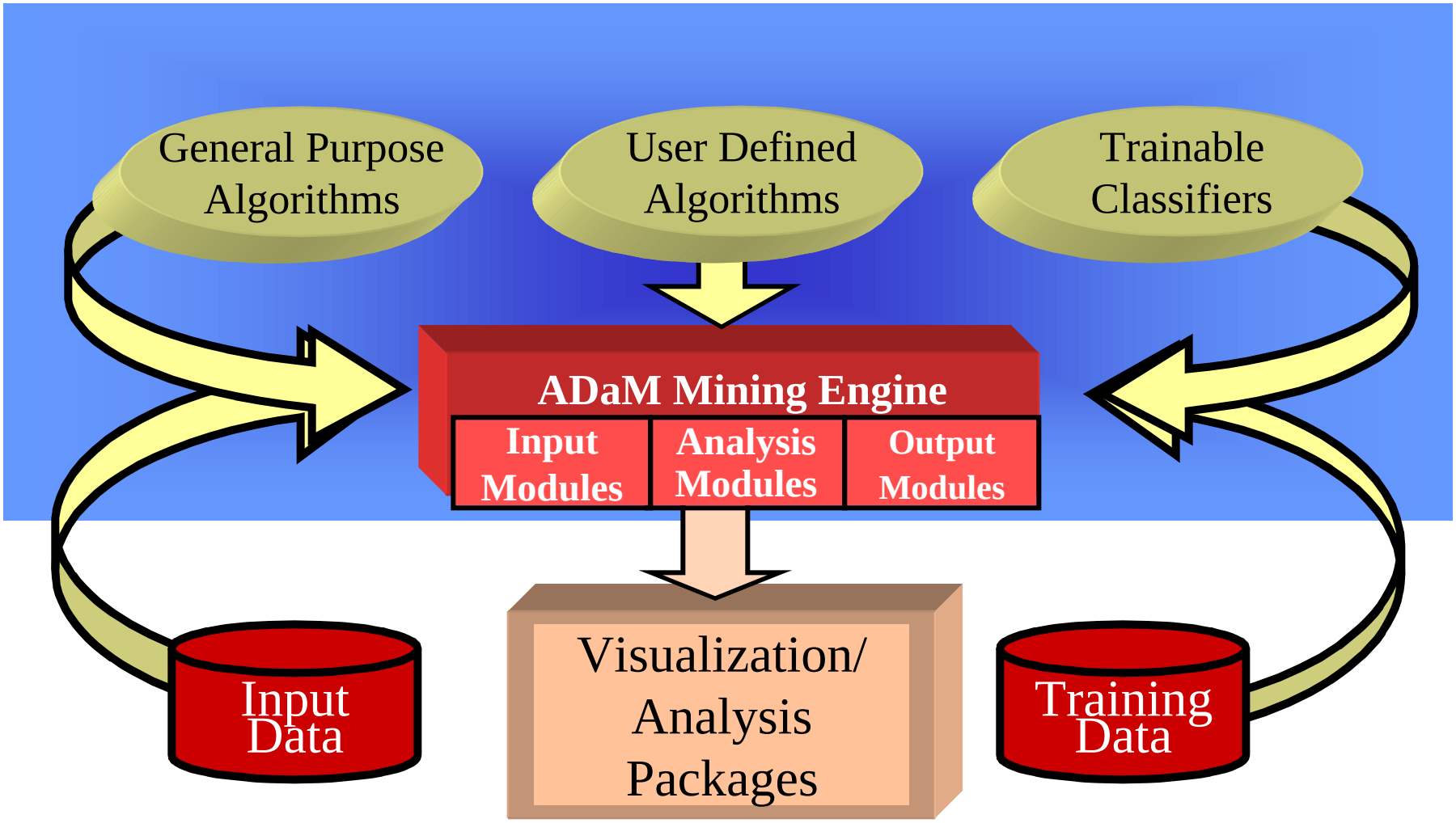
- Provide scientists with the capabilities to iterate
  - Allow the flexibility of creative scientific analysis
- Provide data mining benefits of
    - Automation of the analysis process
    - Reduction of data volume
- Provide a framework to allow a well defined structure for the entire analysis process
- Provide a suite of mining algorithms for creative analysis
- Provide capabilities to add “science algorithms” to the framework

ADaM : Mining Environment for Scientific Data



- *The system provides knowledge discovery, feature detection and content-based searching for data values, as well as for metadata.*
  - *contains over 120 different operations*
  - *Operations vary from specialized science data-set specific algorithms to various digital image processing techniques, processing modules for automatic pattern recognition, machine perception, neural networks, genetic algorithms and others*

Extensibility of ADaM



Reasons for using ADaM for Scientific Data Analysis

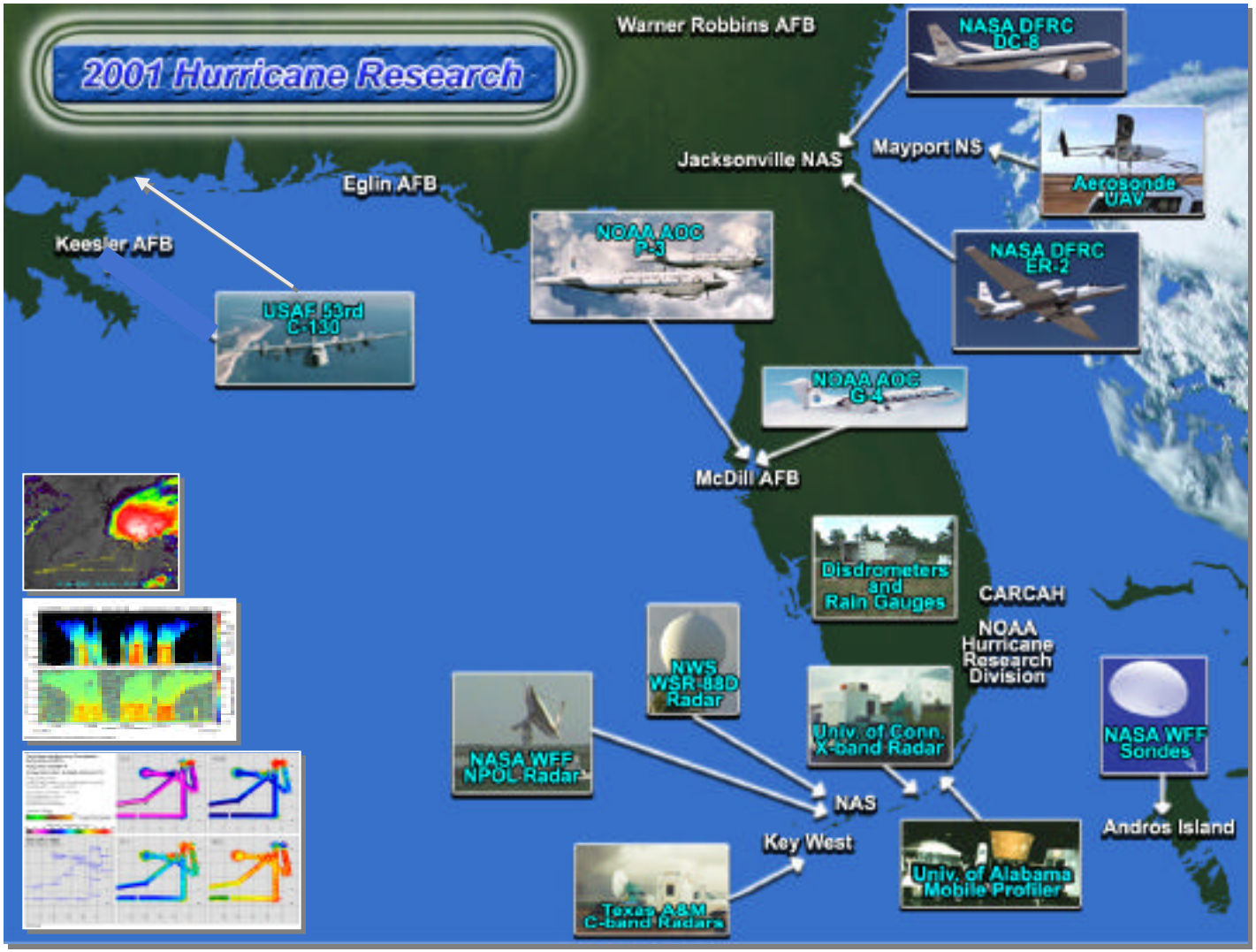
- Provide scientists with the capabilities to iterate
  - Allow the flexibility of creative scientific analysis
- Is a powerful tool for research and analysis given the volume of science data
- Extremely useful when manual examination of data is impossible
- Allows scientists to add problem specific algorithms to the ADaM toolkit
- Minimizes scientists' data handling to allow them to maximize research time
- Reduces "reinventing the wheel"

Mission/Project/Field Campaign Coordination

Electronic Collaboration

Strategic and Tactical Coordination

Technologies to coordinate complex projects



- Data acquisition and integration from multiple platforms, instruments and agencies for quick exploitation
- Intra-project communications before, during, and after CAMEX campaigns

CAMEX-4 Coordination: pre-flight

NASA managers may review status of aircraft, instruments, flight plans at various times throughout the mission

Coordination Clearinghouse

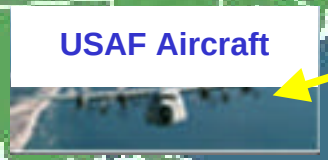
RDBMS

Web-based interface with customized information access for different user groups; rapid development, scalability and portability

Scalable, reliable data management

Experiment PI: Coordinates with all participants, posts plan of the day

Forecaster: Contacts local weather, forecast centers, weather support web sites to prepare and post daily morning weather briefing



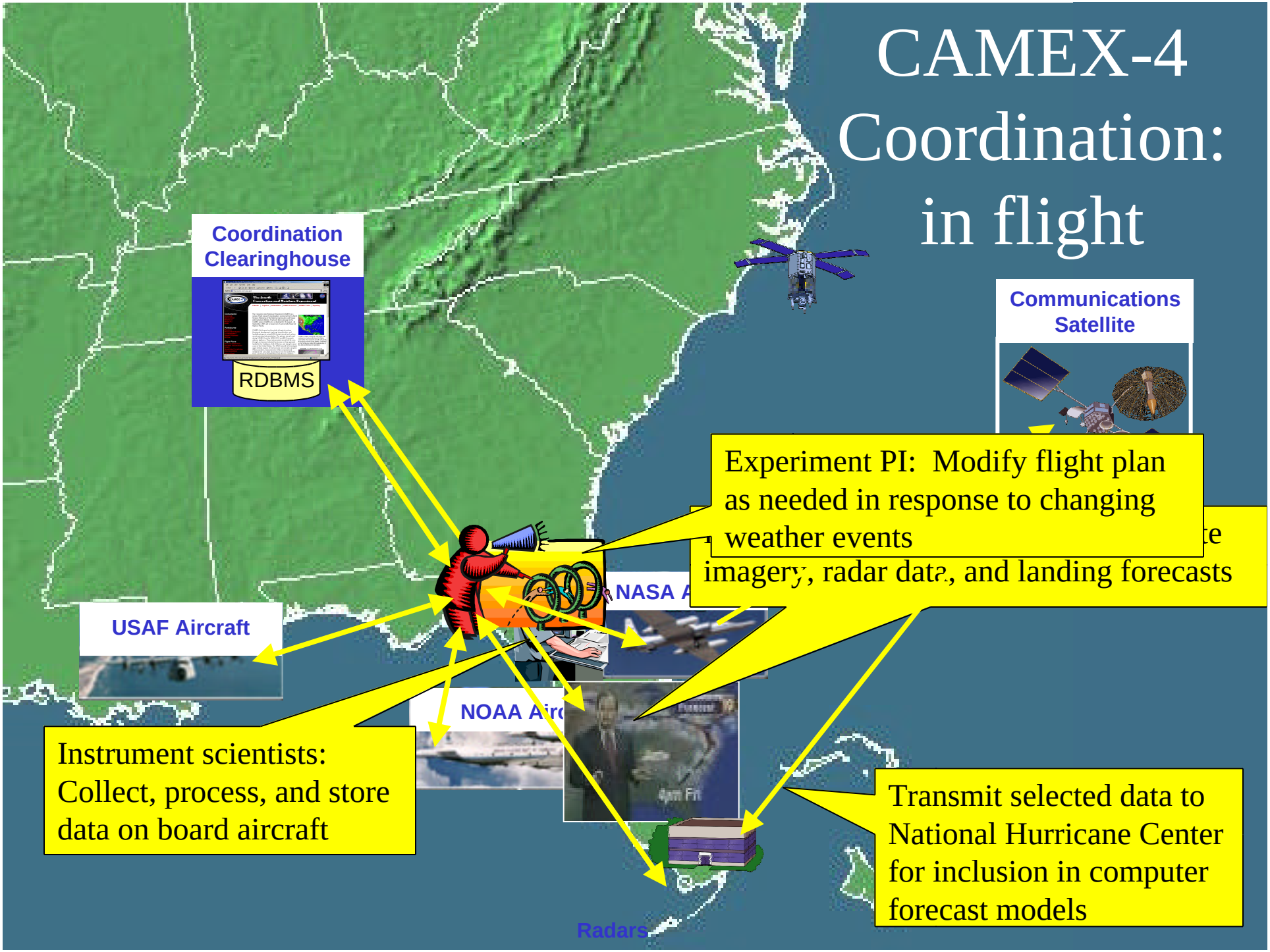
USAF Aircraft

Aircraft Crew: Perform aircraft maintenance and report status.

Preflight mission briefing and flight planning

Radars

CAMEX-4 Coordination: in flight



CAMEX-4 Coordination: post-flight

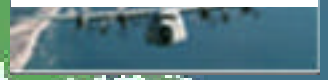
Coordination Clearinghouse

RDBMS

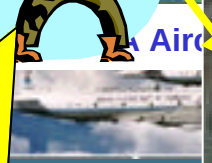
Mission and Instrument scientists: Post sortie and instrument reports and quicklook data

Forecaster: Prepare post-flight weather briefing

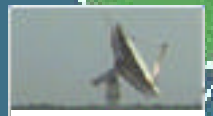
USAF Aircraft



NASA Aircraft



Aircraft Crew: Prepare aircraft and instruments for next flight and update status.



Radars