UAH Information Technology and Systems Center

The
Data Mining and Data Usability Challenge

Sara J. Graves, Ph.D.

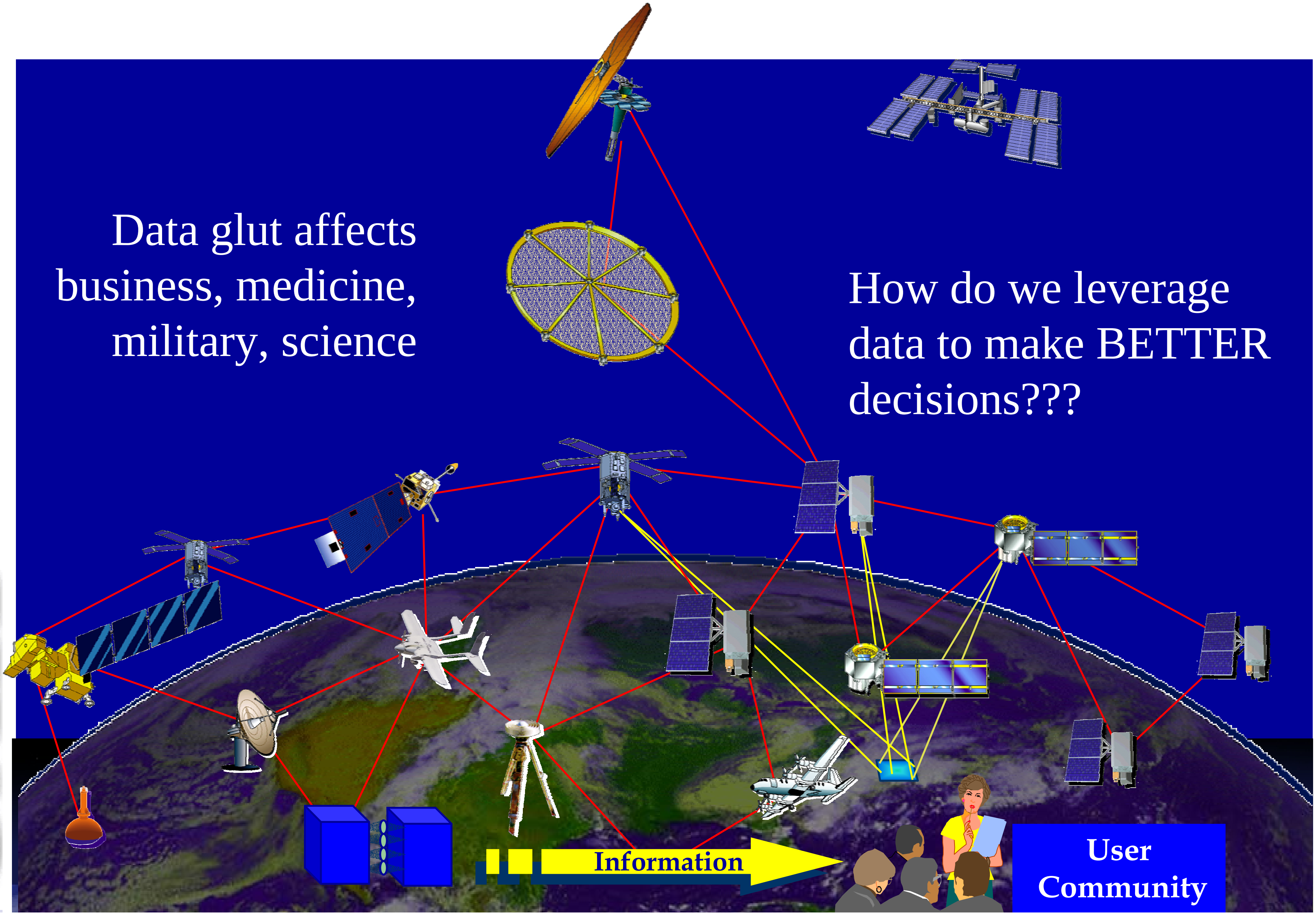
Director, Information Technology and Systems Center
University Professor, Computer Science Department
University of Alabama in Huntsville
Director, Information Technology Research Center
National Space Science and Technology Center
256-824-6064
sgraves@itsc.uah.edu

http://www.itsc.uah.edu

"…drowning in data but starving for knowledge"

Data glut affects business, medicine, military, science

How do we leverage data to make BETTER decisions???



UAH Information Technology and Systems Center

National Research Council Report

2002

Assessment of the Usefulness and Availability of NASA's Earth and Space Science Mission Data

Task Group on the Usefulness and Availability of NASA's Space Mission Data

Space Studies Board
Division on Engineering and Physical Sciences
Board on Earth Sciences and Resources
Division on Earth and Life Studies

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C.

UAH Information Technology and Systems Center

NASA Workshop Report

2004

NASA EOS Science Working Group on Data

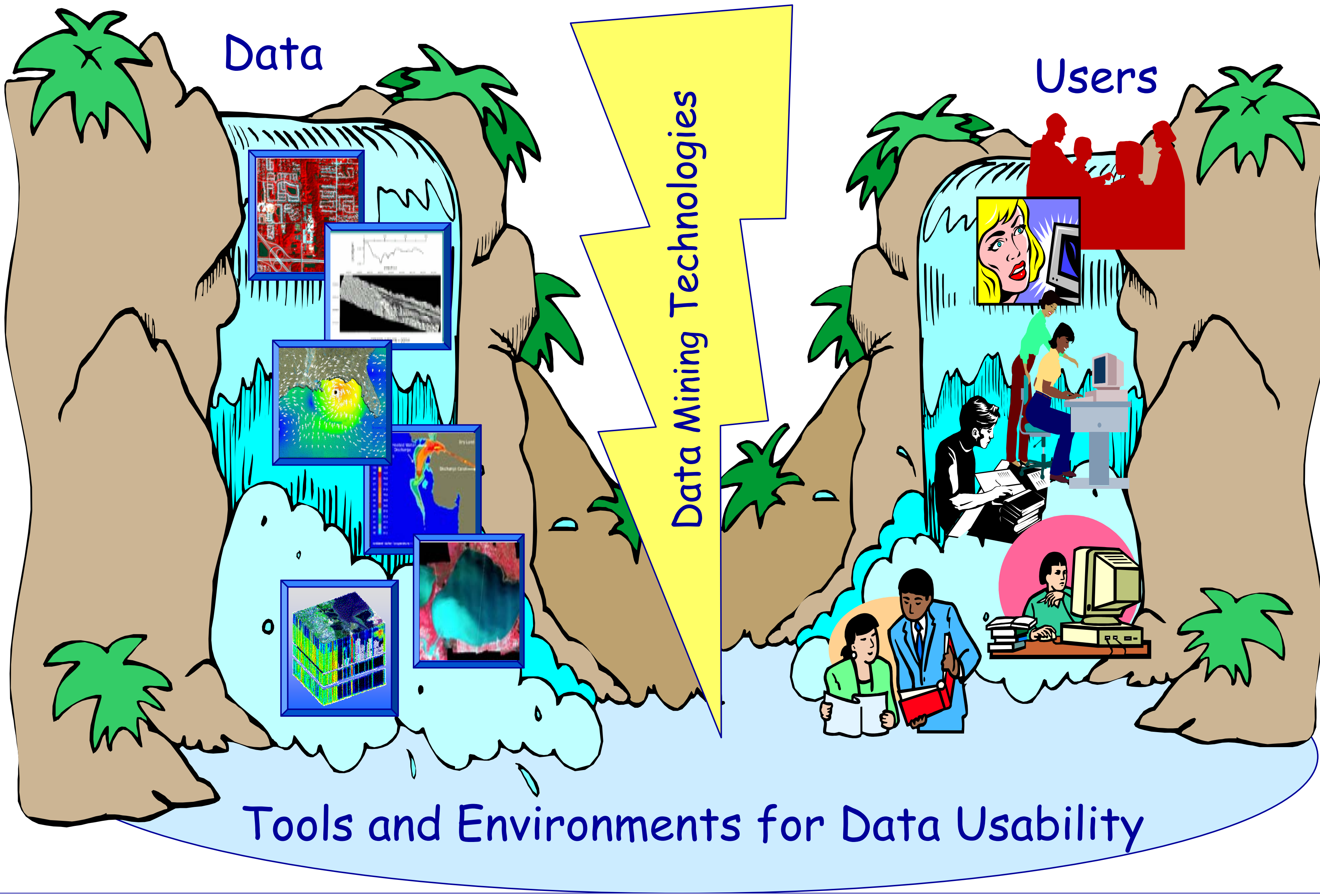
Data Access and Usability Workshop

Report

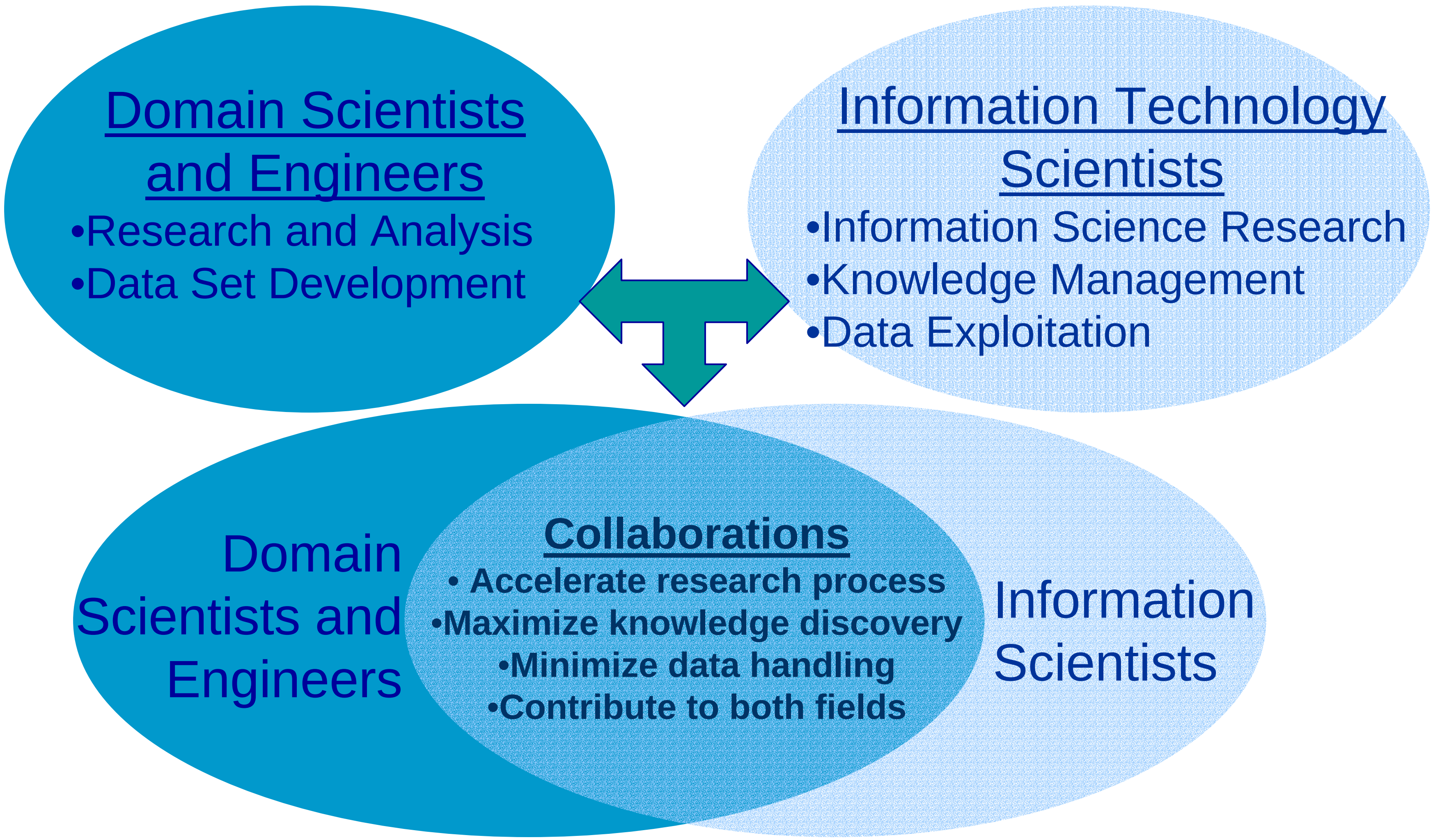
February 26, 2004

Contents

Challenge: **Increase usability of data and technologies to address the diverse needs of the flood of users.**



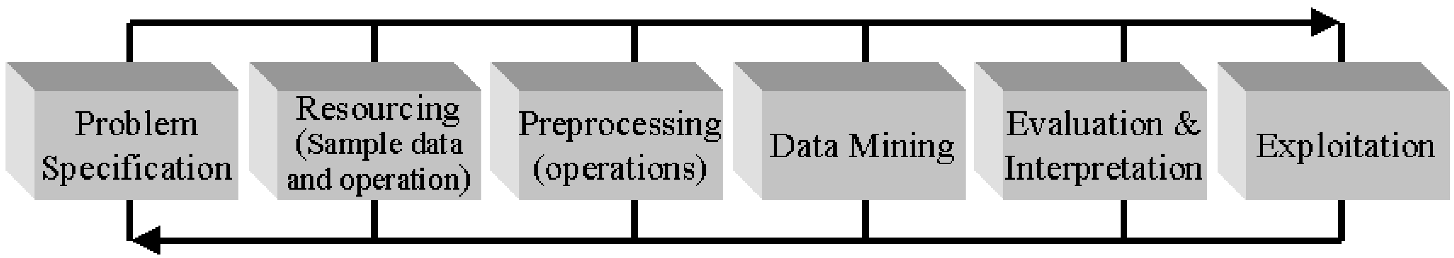
Data Usability Success Builds on the Integration of Various User Domains and Information Technology



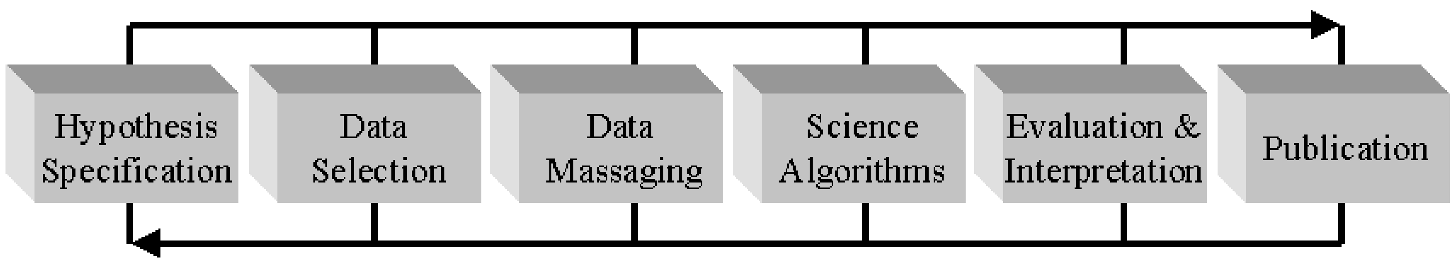
Scientific Analysis	Data Mining
• Harnesses human analysis capabilities – Highly creative • Based on theory and hypothesis formulation – Physical basis is normally used for algorithms • Drawing insights about the underlying phenomena • Rapidly widening gap between data collection capabilities and the ability to analyze data • Potential of vast amounts of data to be unused	• Provides automation of the analysis process • Can be used for dimensionality reduction when manual examination of data is impossible • Can have limitations – May not utilize domain knowledge – May be difficult to prove validity of the results • There may not be a physical basis • Should be viewed as complimentary tool and not a replacement for scientific analysis

Similarity between Data Mining and Scientific Analysis Processes

Mining Process



Scientific Analysis Process



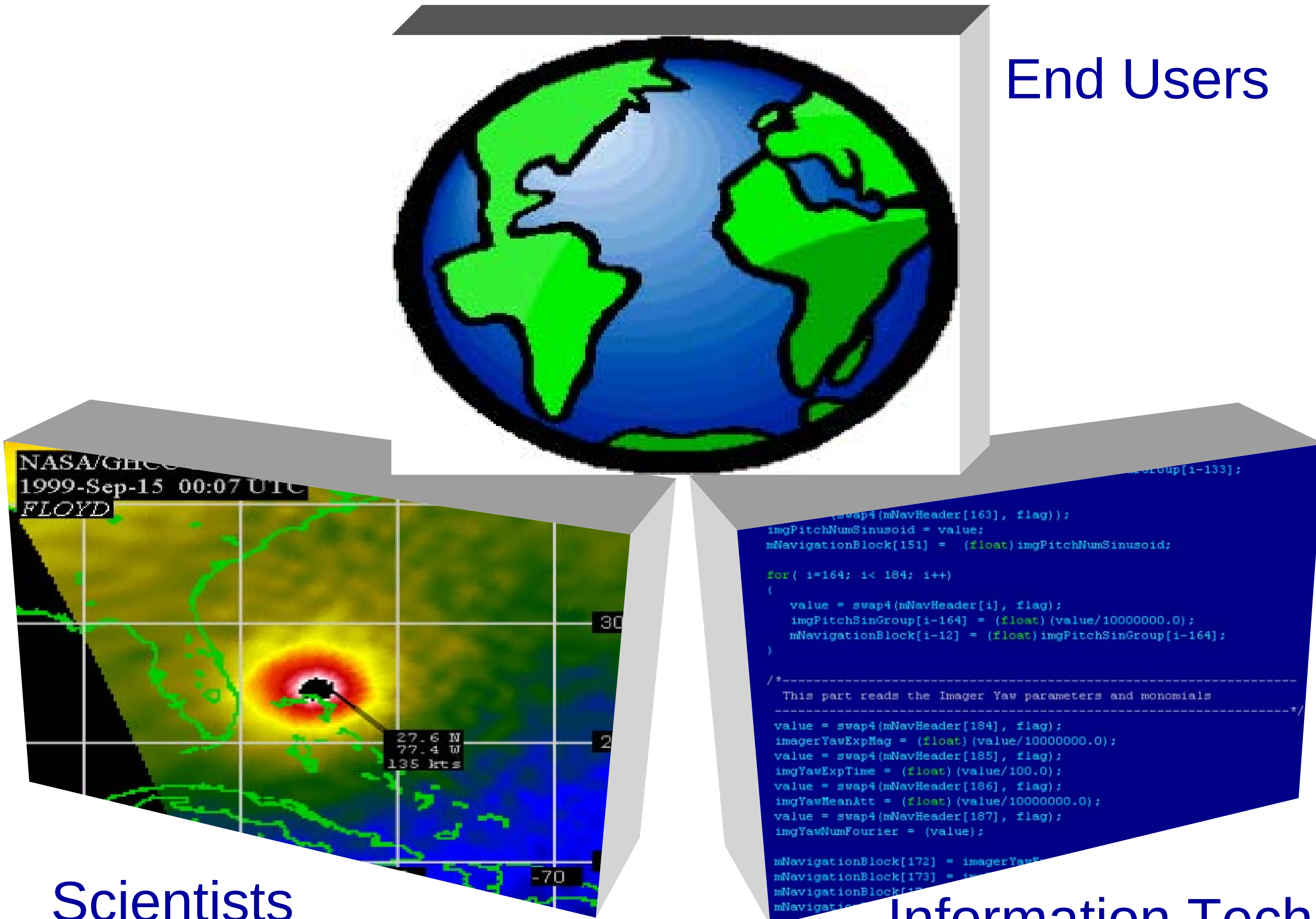
Characteristics of Science Data

- Varied kinds of data
  - Raster images
    - With structure and geometry
    - Multispectral
  - Time series and sequence data
  - Numerical model outputs
- Multiple resolutions/multiple scales
- Variability of data formats
- Granularity of data
- Includes spatial and temporal dimensions
- Physical basis/domain knowledge needed before applying algorithms
- Typically requires domain-specific algorithms

Reasons for Mining Science Data

- Powerful tool for research and analysis given the volume of science data
- Necessity when manual examination of data is impossible
- Can allow scientists to refine/add more layers to the knowledge bases
- Can minimize scientists' data handling to allow them to maximize research time
- Can reduce "reinventing the wheel"
- Can fully exploit reusable knowledge bases for different problems
- Can be integrated into a Next Generation Information System to provide additional services such as:
  - Custom Order Processing
  - Subsetting/Formatting/Gridding ….
  - Event/Relationship Searching

Key Collaborators



End Users

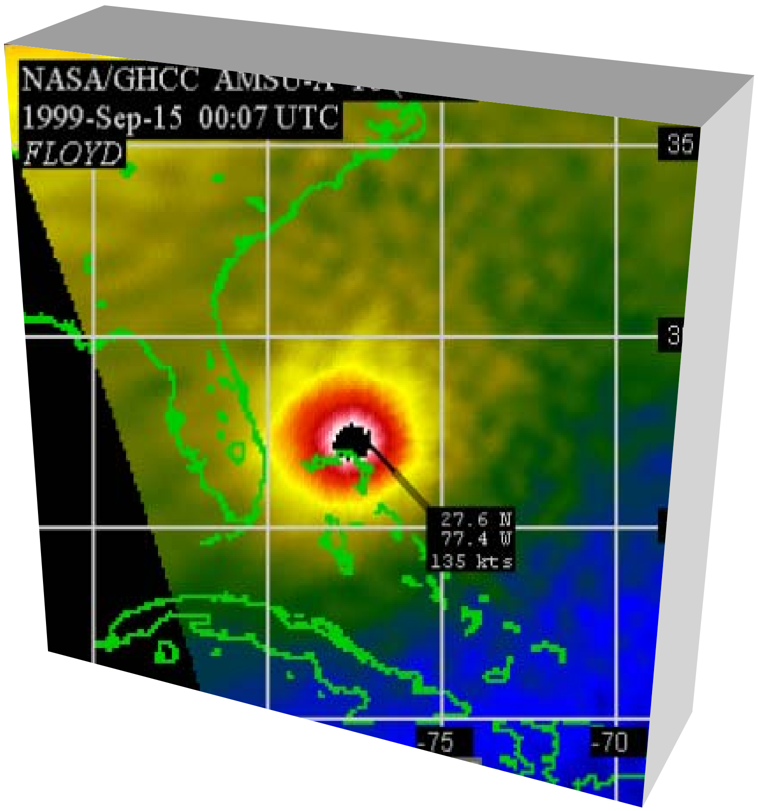
Scientists

Information Technology Specialists

Information Technology and Systems Center

UAH

Scientist's Perspective

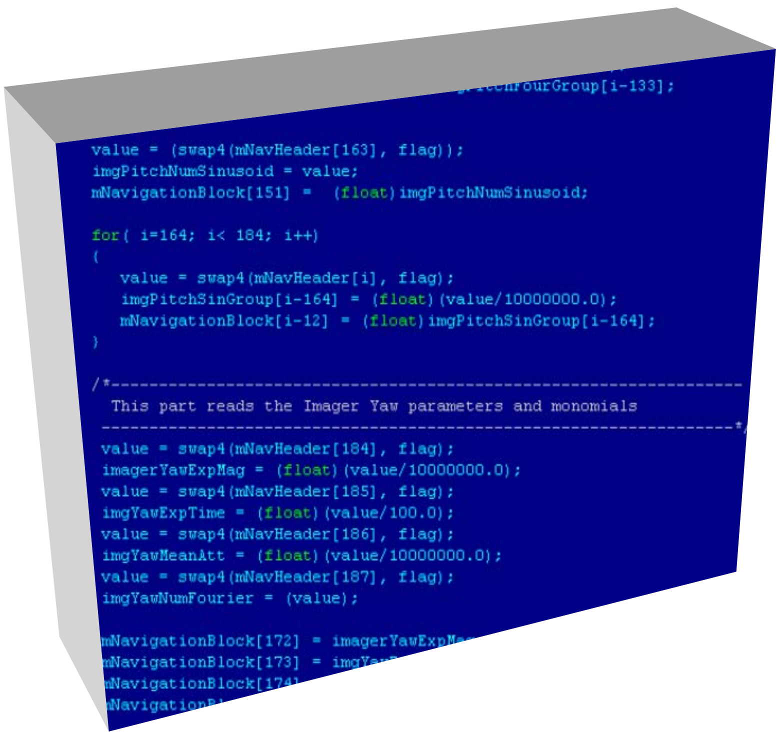


- Define the experiment
- Create reusable "Knowledge Base"
- Iterate over experiment to refine the knowledge base
- Minimize data handling/Maximize research
- Add more "layers" to the knowledge base
- Allow different levels of knowledge discovery:
  - Shallow knowledge
  - Hidden
  - Deep

End User's Perspective

- End users can be:
  - Students
  - Public
  - Decision makers
  - Other Scientists
- Access to data
- Access to knowledge base
- End products

Information Technology Specialist's Perspective



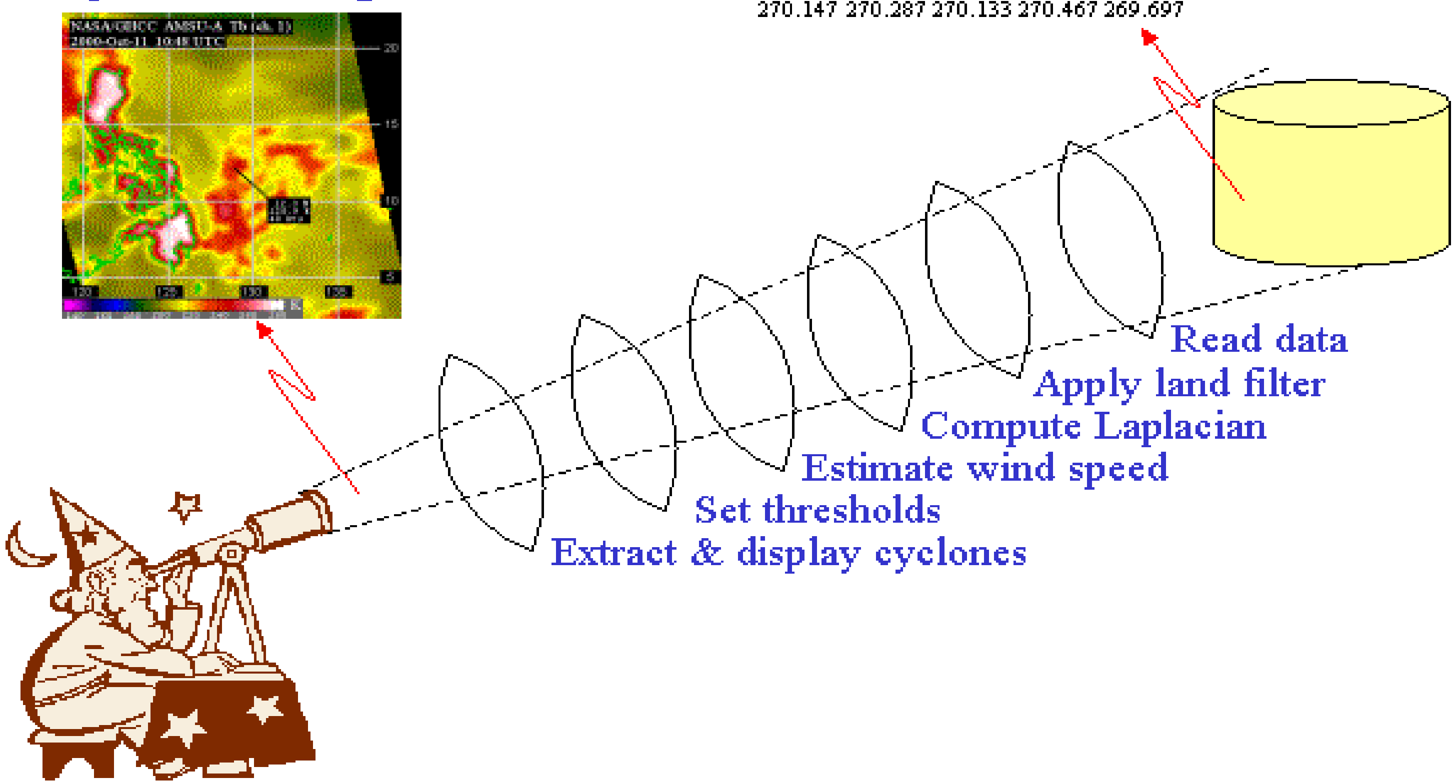
- Exploit IT to benefit science analysis
- Handle data processing for scientists
- Provide a dynamic Information System
- Provide services to the user community
  - Subsetting
  - Formatting
  - Gridding
  - Analysis tools
  - Visualization tools
- Provide knowledge access to all

Enhancing Data Usability: Focusing on finding information in data

Data

270.421	270.600	270.366	270.797	269.976
270.422	270.606	270.347	270.787	269.971
270.359	270.591	270.327	270.755	269.940
270.339	270.571	270.307	270.700	269.915
270.315	270.445	270.287	270.653	269.889
270.268	270.352	270.264	270.568	269.840
270.255	270.305	270.238	270.535	269.780
270.205	270.252	270.212	270.517	269.774
270.172	270.224	270.185	270.497	269.739
270.141	270.276	270.158	270.494	269.628
270.147	270.287	270.133	270.467	269.697

Cyclone Image



Mining Collaborations with Science and Engineering Domains

Problem formulation:

- Domain experts and mining experts collaborate to formulate problem

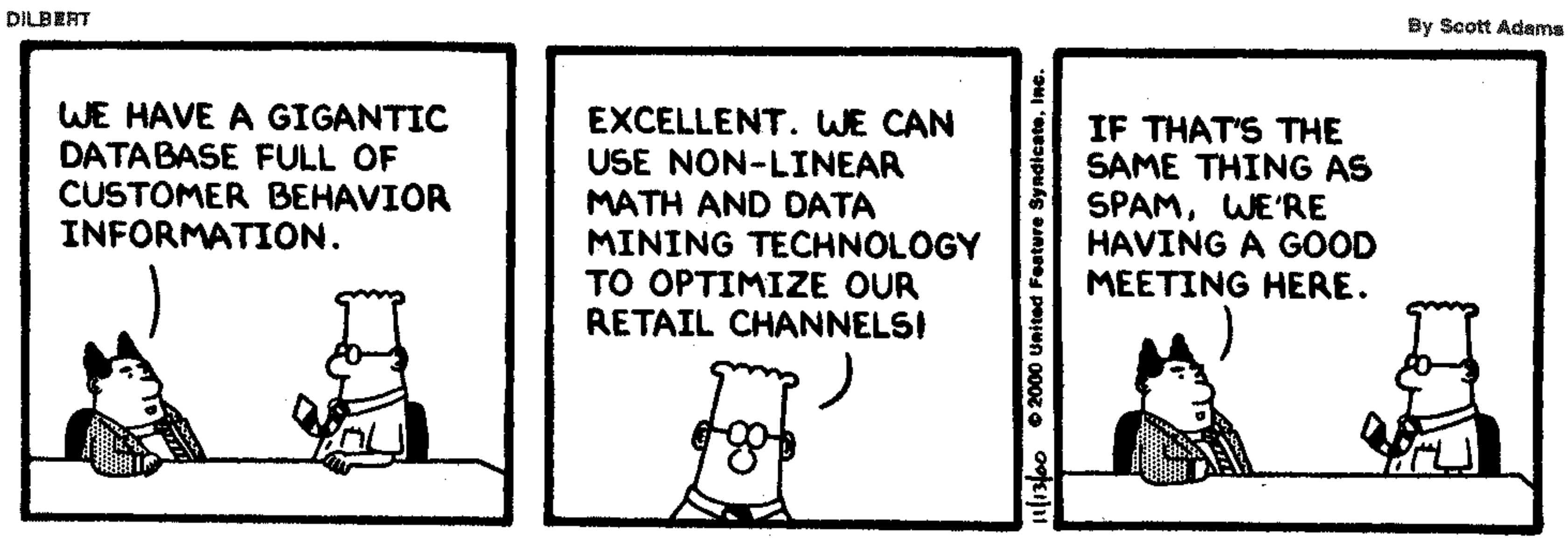
Problem solving - one approach:

- Domain experts provide data for analysis
- Mining experts solve specific problems using mining tools

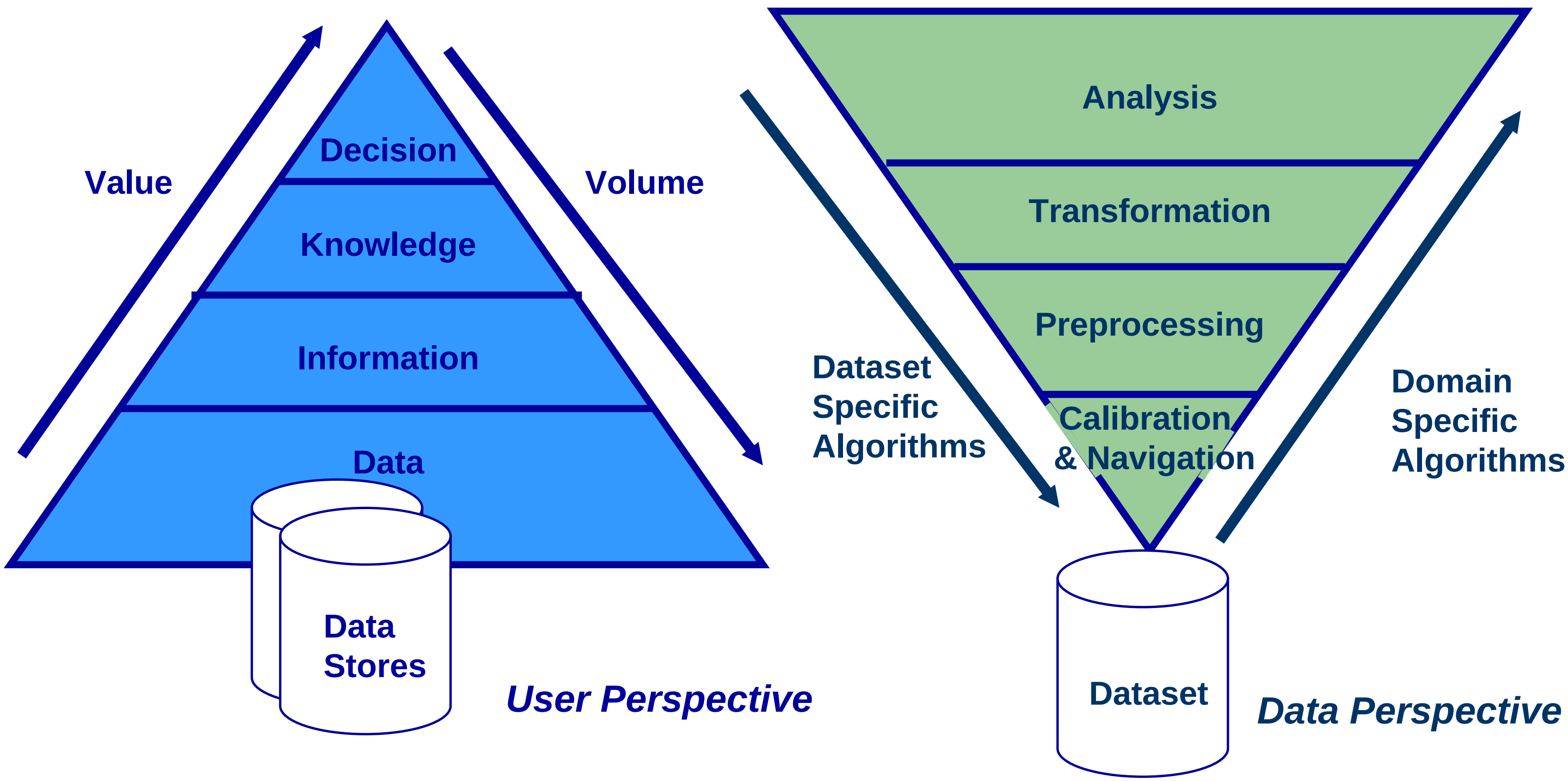
Problem solving - alternate approach:

- Mining experts consult with domain experts on running mining tools
- Integrated team works closely to identify mining algorithms, configure and train mining tools, and analyze results

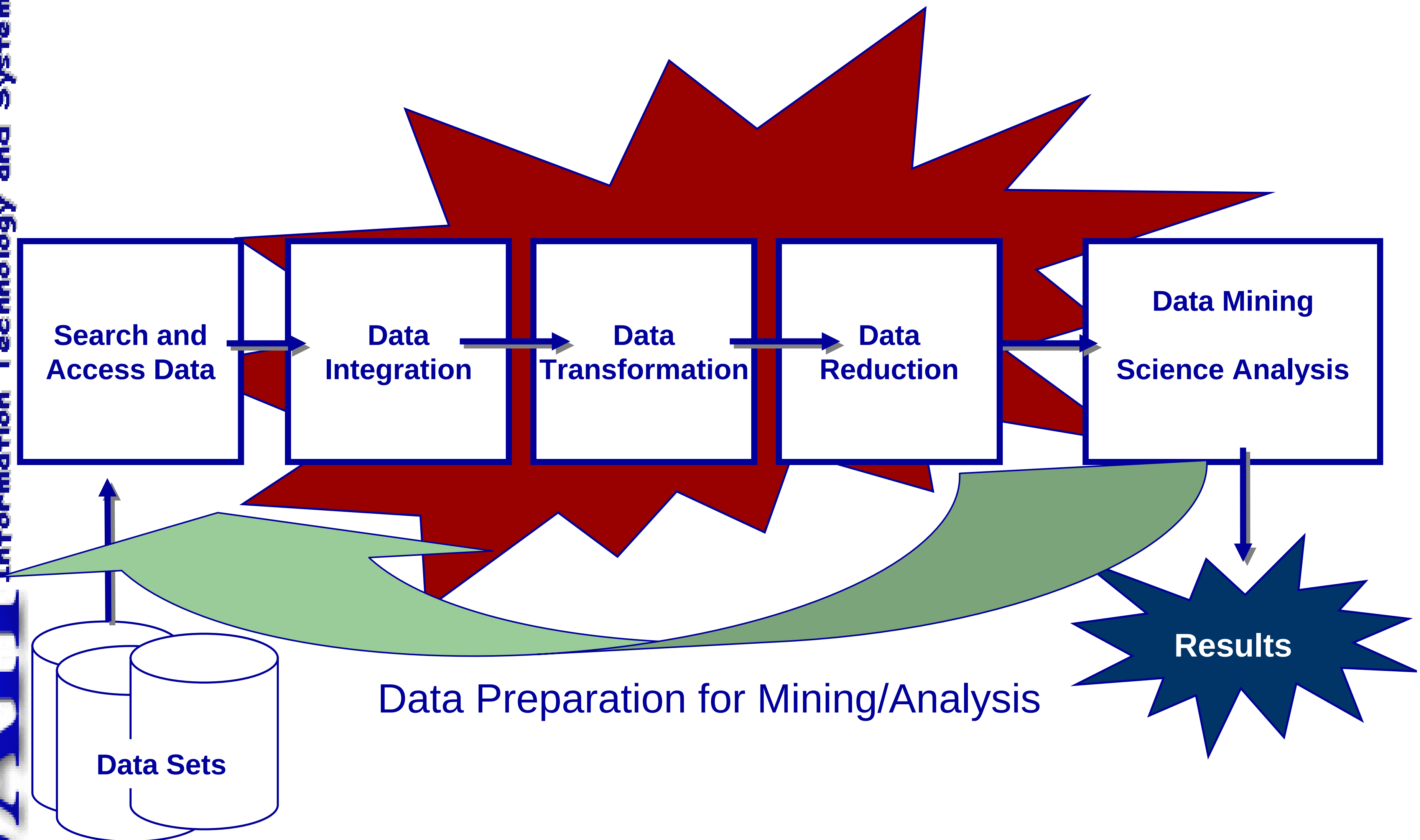
Data Misunderstanding ?



User Perspective and Data Perspective of the Data Mining Process



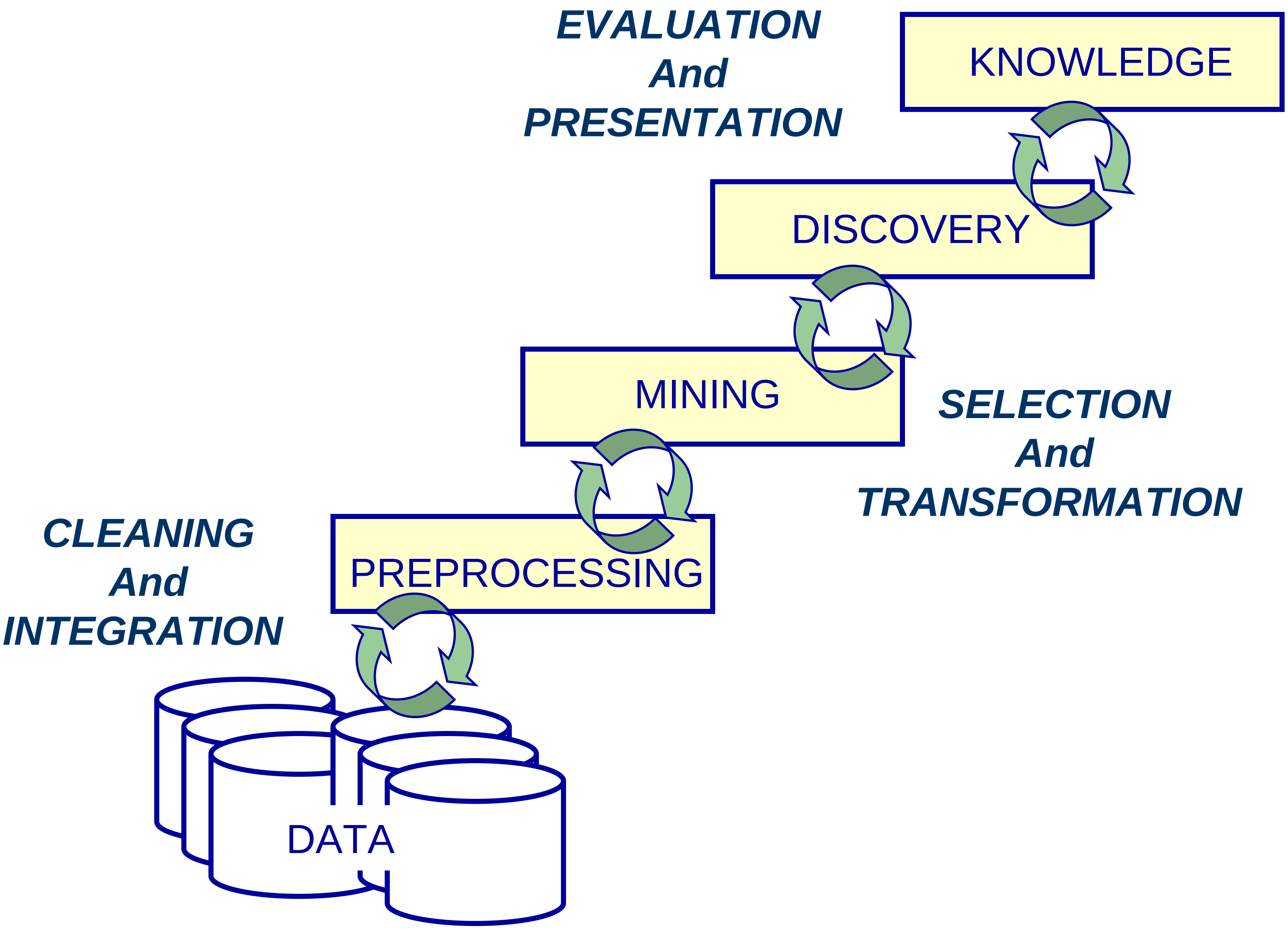
Data Challenge



Typical Data Preparation Operations

- Data Cleaning
  - Clean data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
  - Fairly well handled
- Data Integration
  - Integration of multiple data files
- Data Transformations
  - Normalization and aggregation
- Data Reduction
  - Obtain a reduced representation of the data set, which produces the same analytical results

Iterative Nature of the Data Mining Process



Issues

- Mining
  - Feature extraction
  - Finding anomalies in the data
  - Understandability of the derived model
  - Utilizing domain knowledge effectively
- Scientific Data Mining Environment
  - Ease of use
  - Adaptable to new science questions
  - Plug in favorite analysis tools

Reasons for a Data Mining Environment

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

Mining Environments

Multiple Configurations

- Complete System (Client and Engine)
- Mining Engine (User provides its own client)
- Application Specific Mining Systems
- Operations Tool Kit
- Stand Alone Mining Algorithms

Distributed/Federated Mining

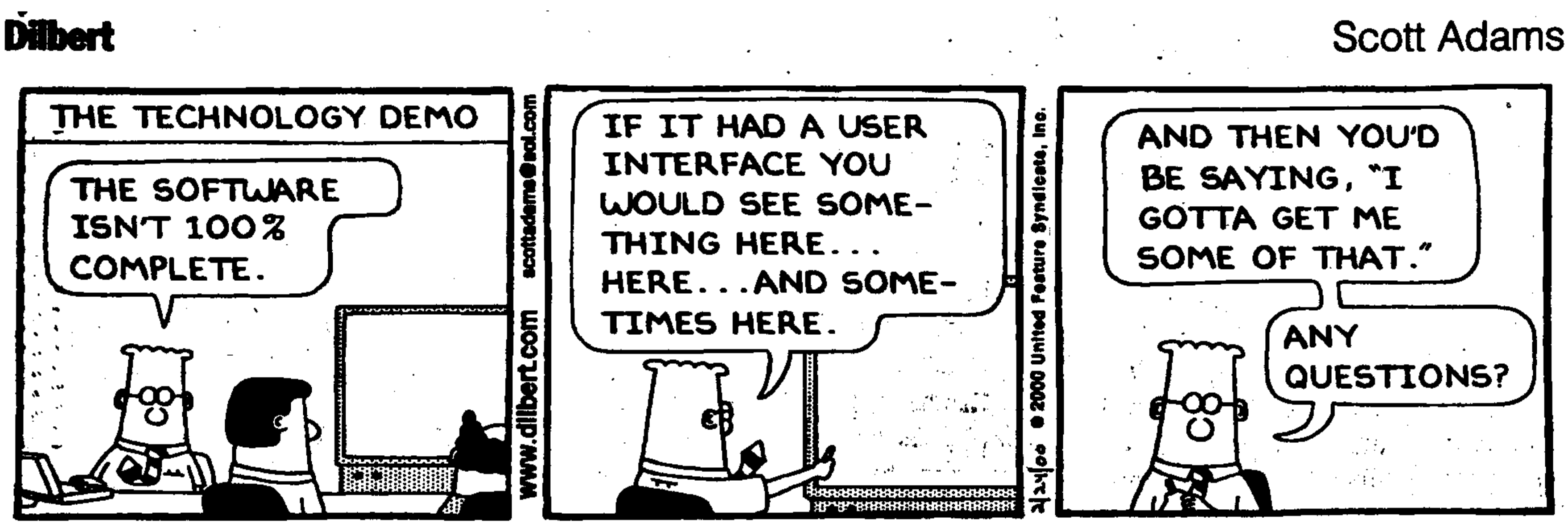
- Distributed services
- Distributed data
- Chaining using Interchange Technologies

On-board Mining

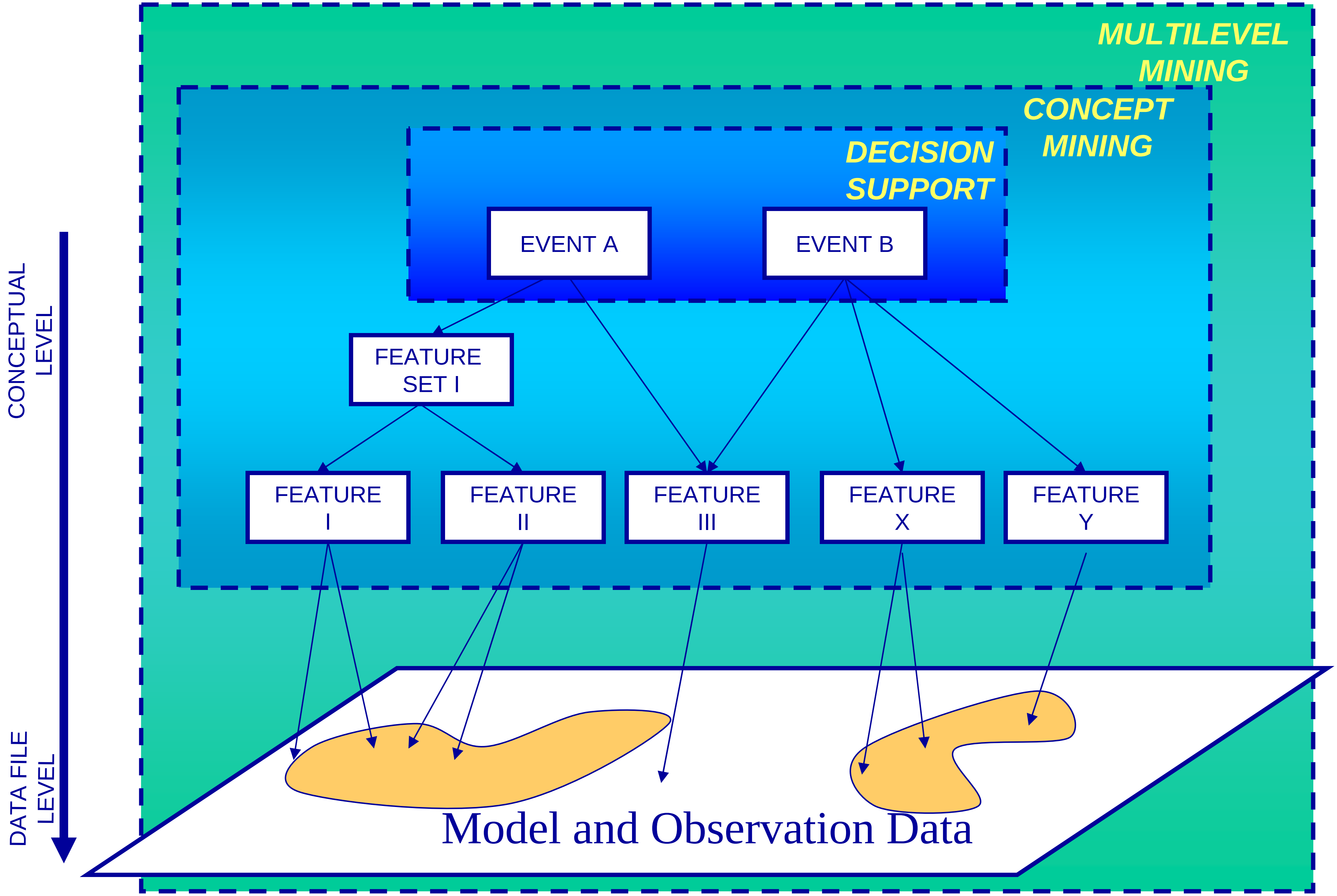
- Real time and distributed mining
- Processing environment constraints

UAH Information Technology and Systems Center

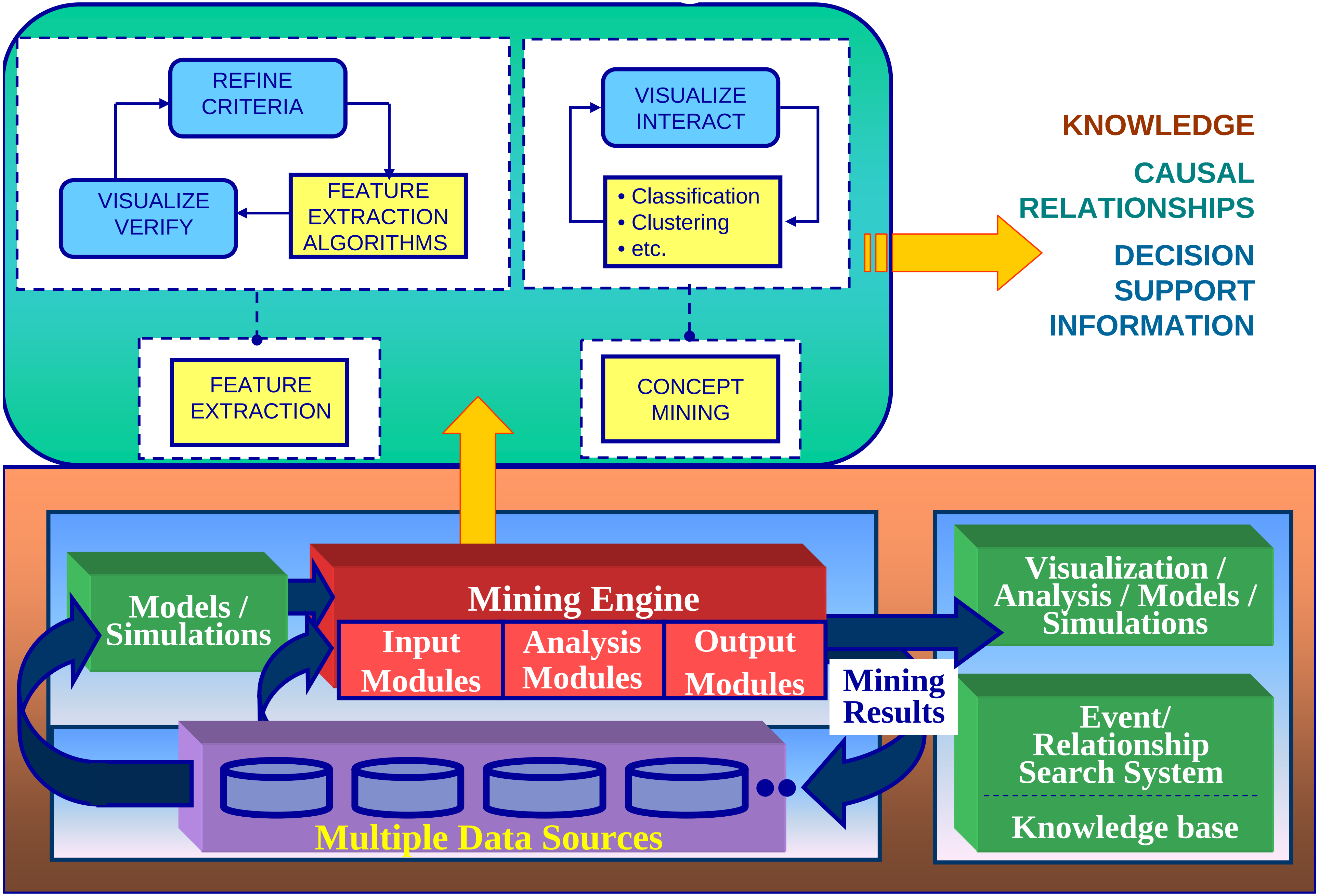
Providing User Interfaces ?



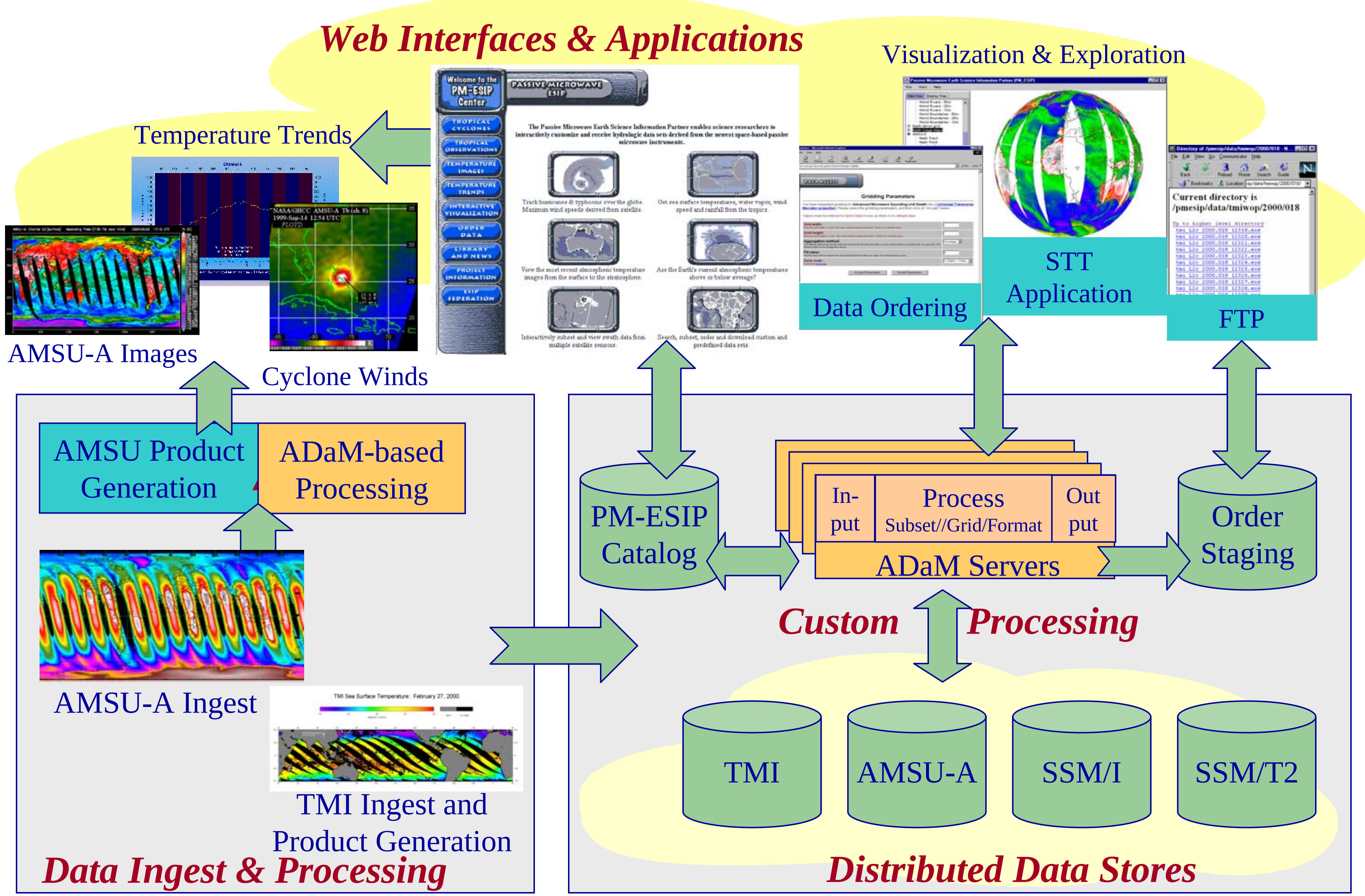
Concept Hierarchy for Data Mining and Fusion



Multilevel Mining

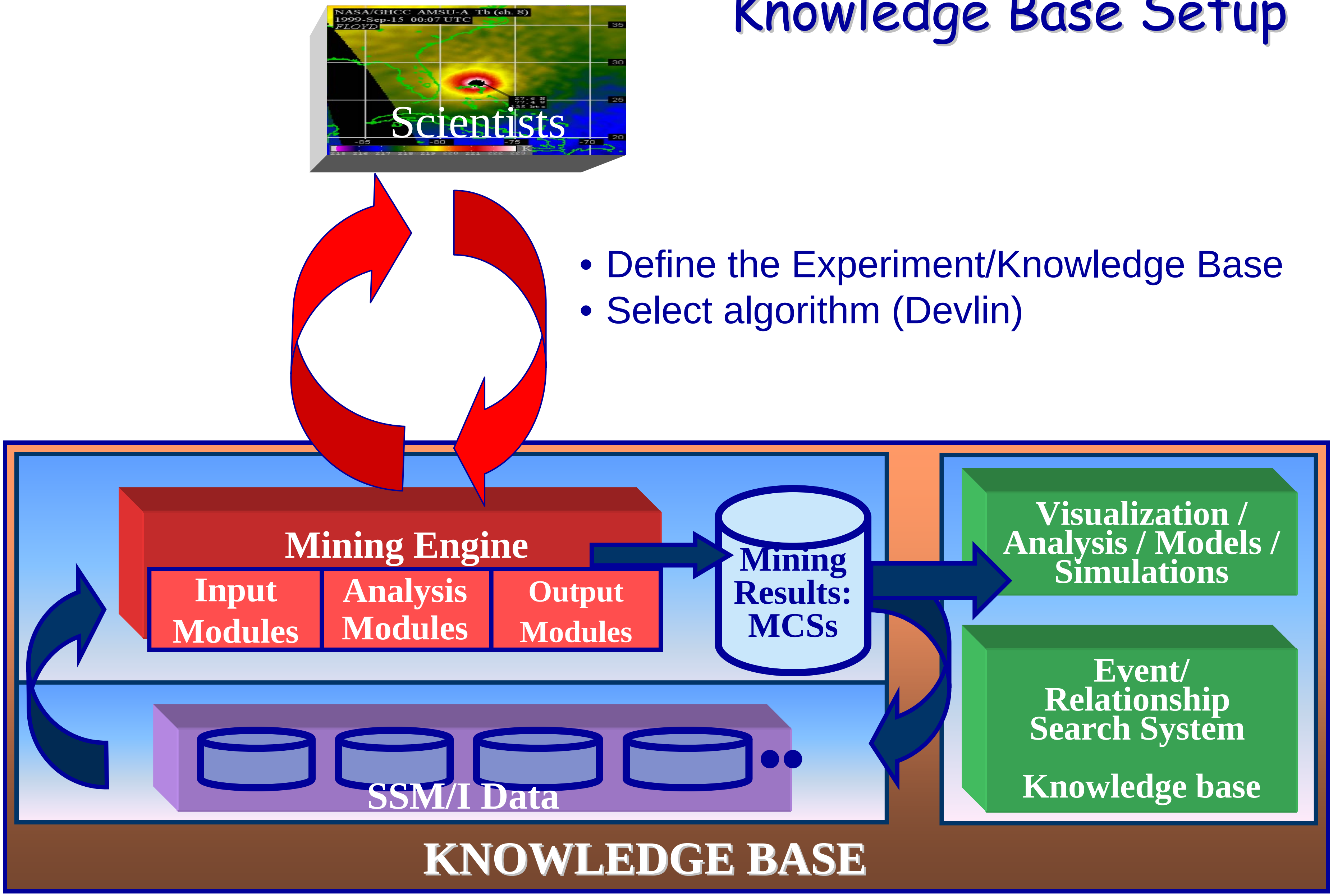


Multiple Mining Environments: Passive Microwave ESIP Information System

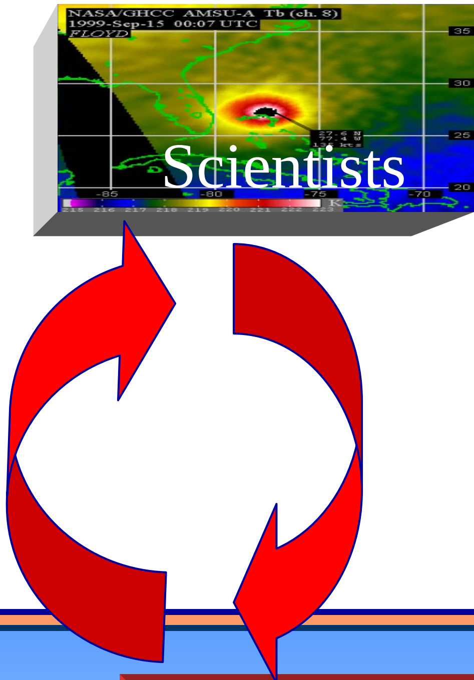


Mesoscale Convective System (MCS) Detection: Knowledge Base Setup

- Define the Experiment/Knowledge Base
- Select algorithm (Devlin)



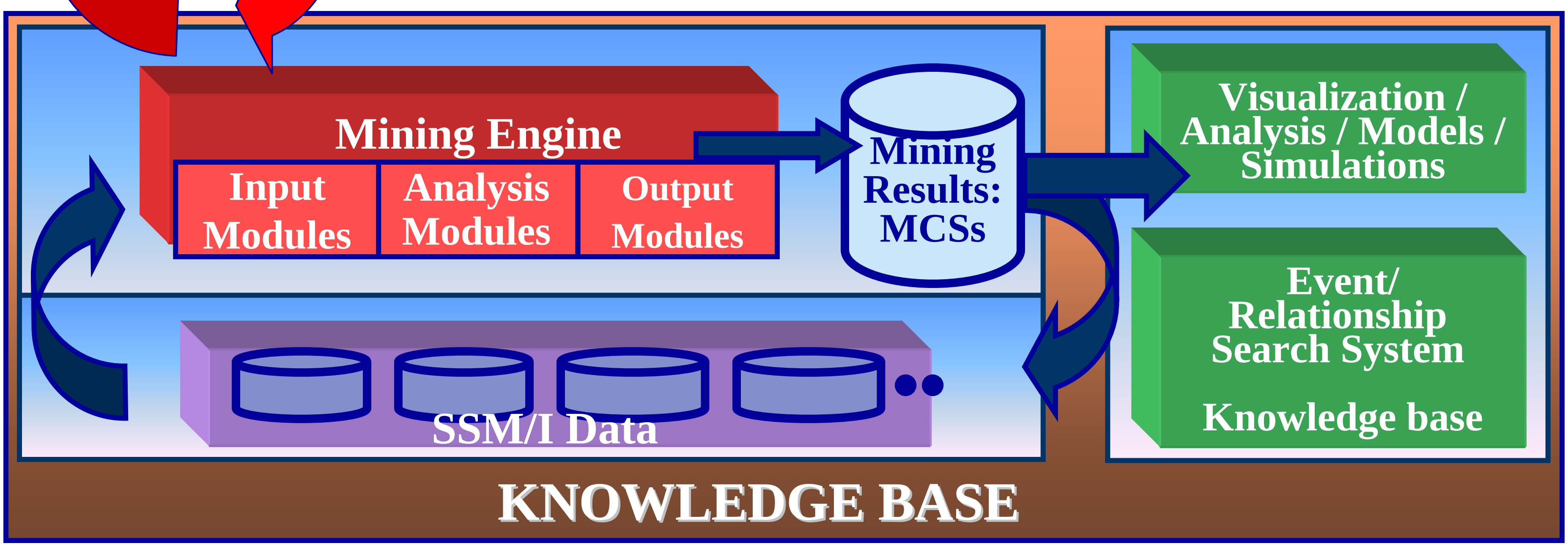
MCS Detection: Research Analysis



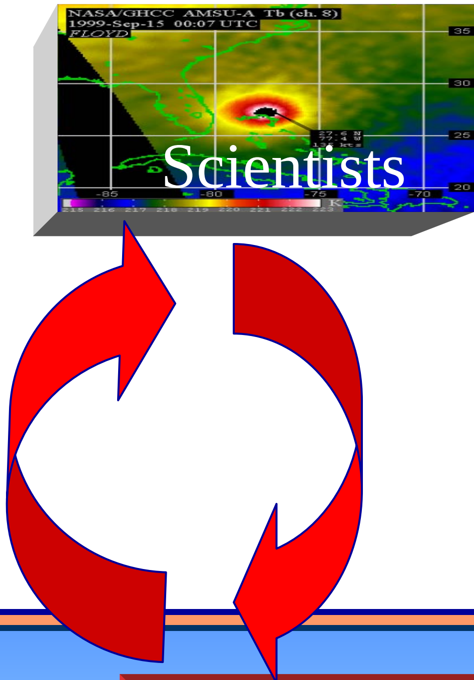
Analysis:

- Find MCS's over river basins in Middle East
- Data Sets
  - MCSs
  - River basin data set
  - Political boundaries

- **Allow scientists to pose questions and get "results"**
- **Allow easy visualization**
- **Maximize knowledge discovery / minimize data handling**
- **Scientists can refine their knowledge repository**
- **Answer the science questions**

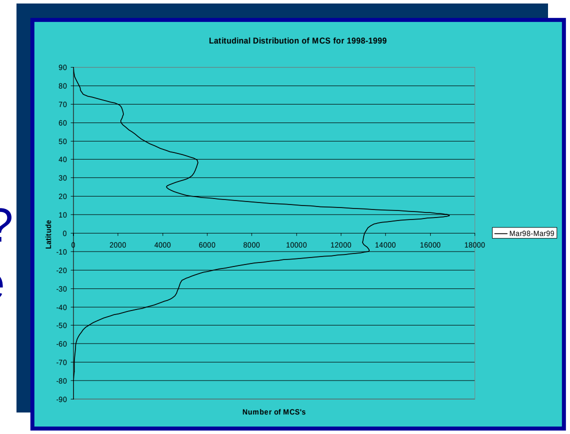


MCS Detection: Knowledge Reuse

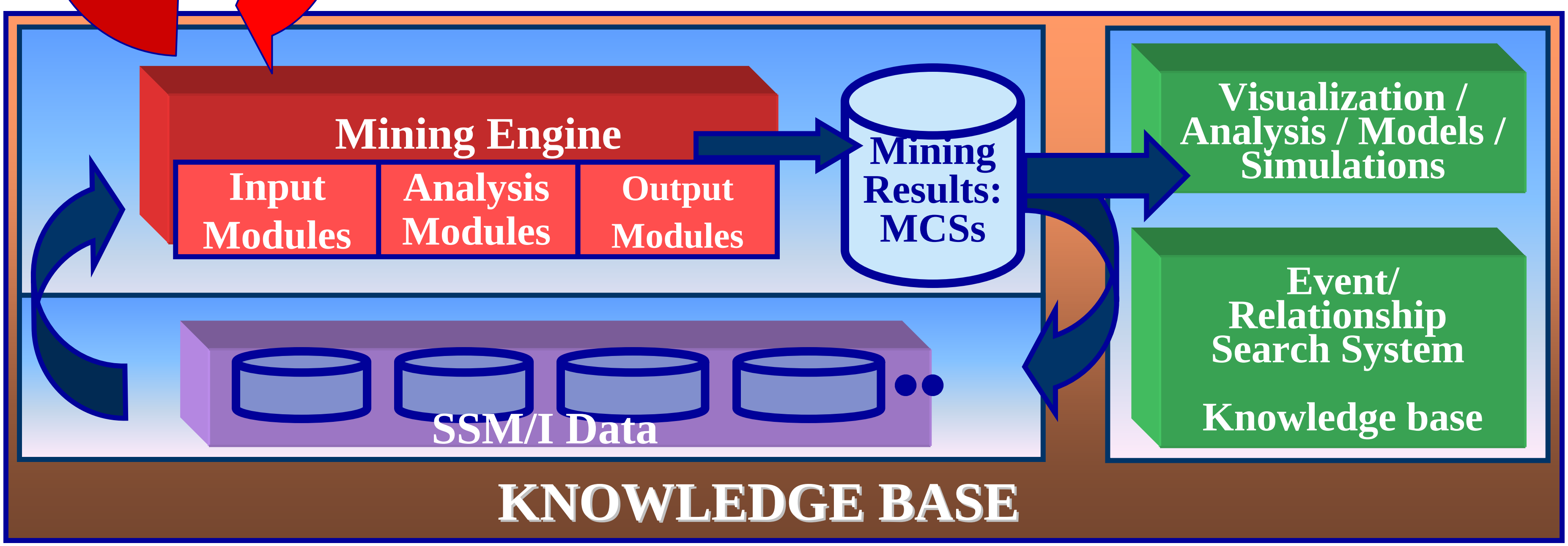


Climatological Study

- What is the latitudinal distribution of MCSs?
- Which continent has more MCSs?
- What is the size distribution of the MCSs for JUN-JUL-AUG?
- What is the relationship between the number of MCSs and their intensities?
- Do results vary for El-Nino years?



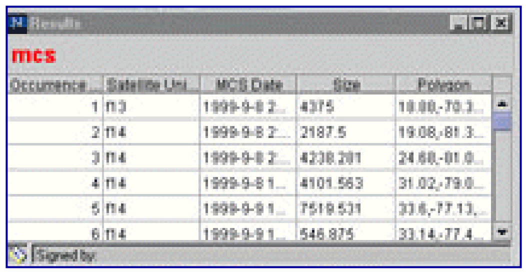
Knowledge Reuse



MCS Detection: Services Provided

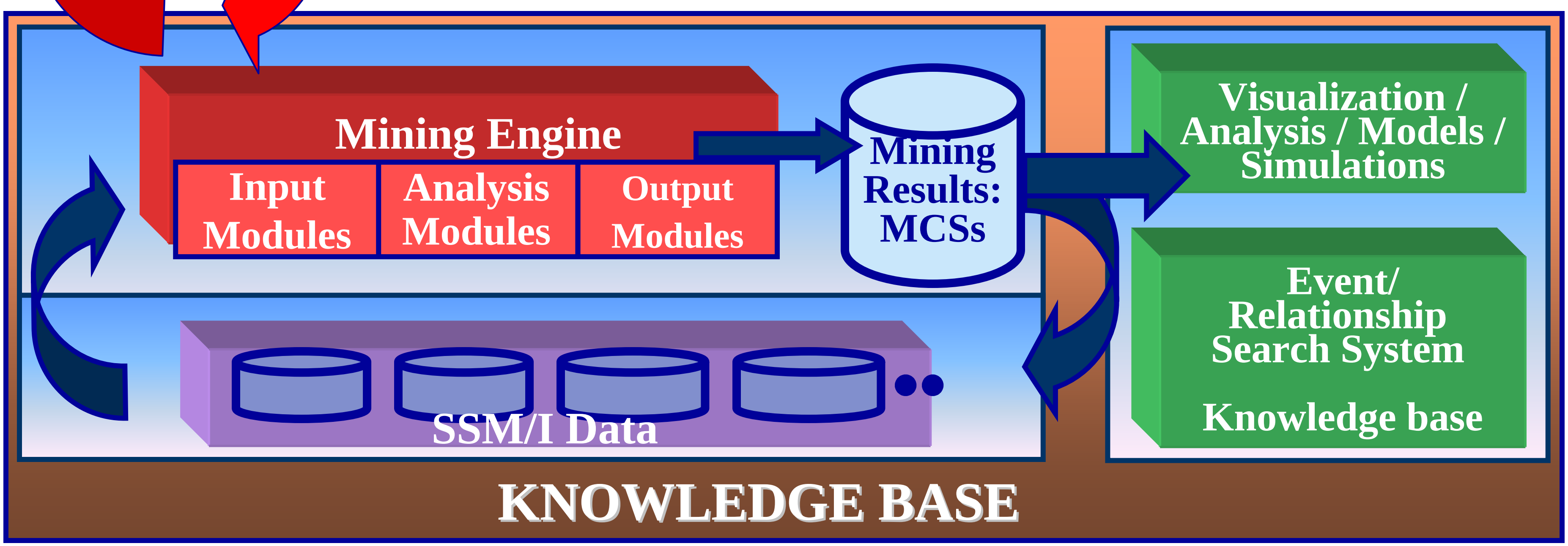
Provide services such as:

- Ability to search for phenomena based on spatial/temporal parameters
- Order specific data files

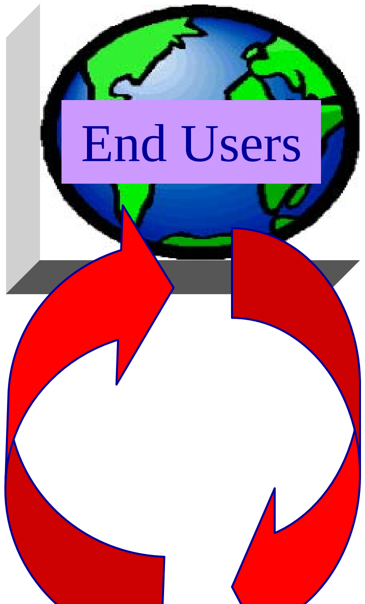


Can provide additional services:

- **Custom order processing**
- **Products on demand**

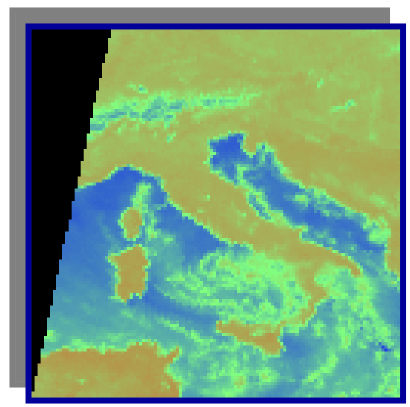


MCS Detection: Product Generation



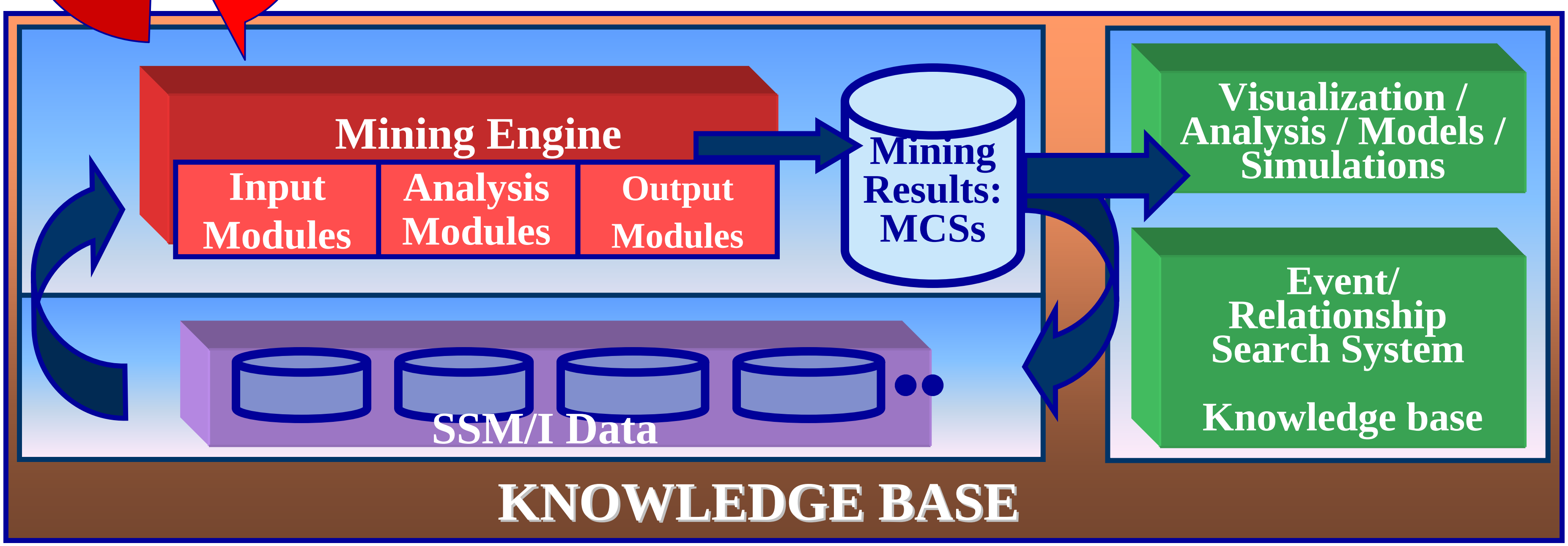
End Products:

- Subset over Italy for flood forecasting



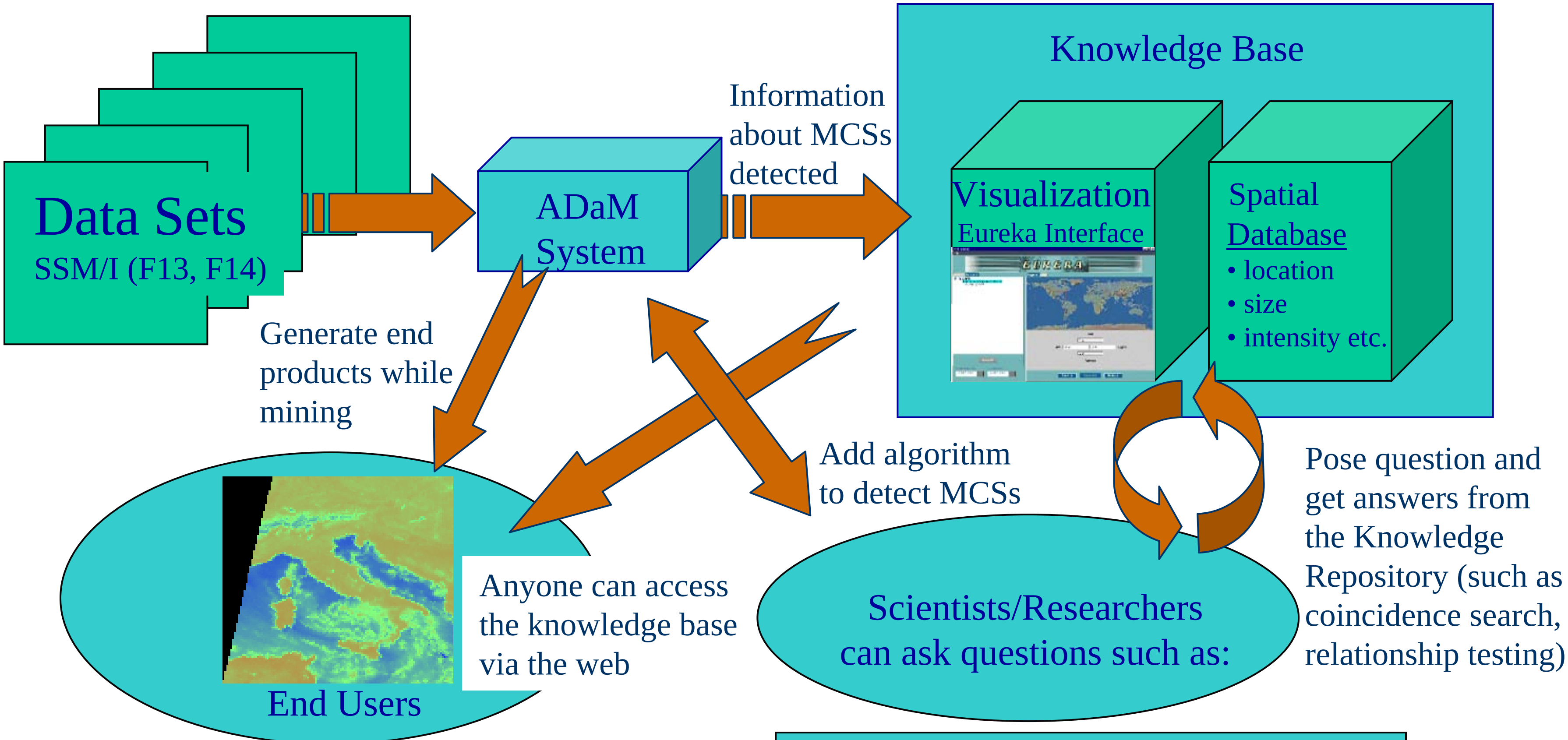
Access knowledge base for

- **Data**
- **Knowledge**
- **Information**



Earth Science Example of Developing a Knowledge Network:

Collaborative Research in Mesoscale Convective Systems



- Generate information useful to the general public (students, researchers, policy makers etc)
  - Images
  - Forecast aids
  - General Science information
- Answer the practical side of the problem

- What is the latitudinal distribution of MCSs?
- Which continent has more MCSs?
- What is the seasonal distribution of MCSs?
- What is the relationship between the number of MCSs and their intensity?

UAH Information Technology and Systems Center

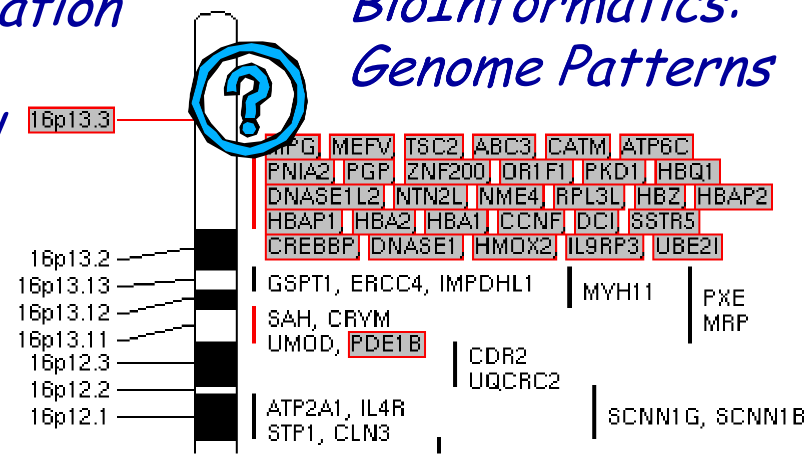
Data Discovery ?

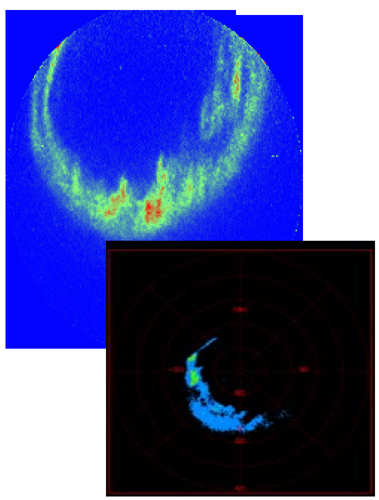
Information Technology and Systems Center

Data Mining in Action

Grid Mining:
- *NASA Information Power Grid*
- *NSF TeraGrid*

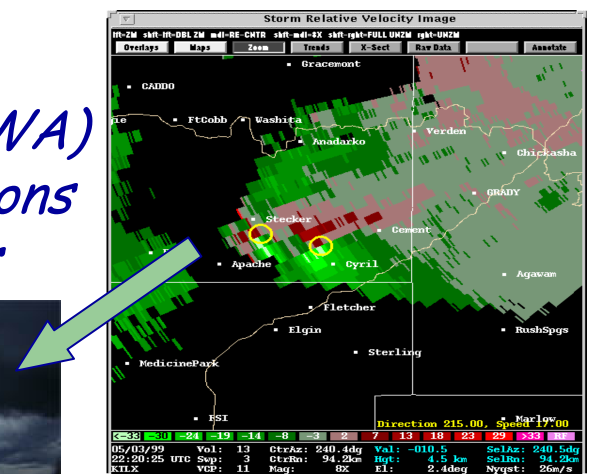
BioInformatics: Genome Patterns





Space Science: Polar Cap Boundary in Auroras

Earth Science:
- *Mining Model Data (Ames, Goddard, SWA)*
- *Satellite Observations*
- *Radar Observations*



Information Technology and Systems Center

UAH

Mining on Data Ingest: Tropical Cyclone Detection

Advanced Microwave Sounding Unit (AMSU-A) Data

Calibration/ Limb Correction/ Converted to Tb

Data Archive

ADaM Mining Environment

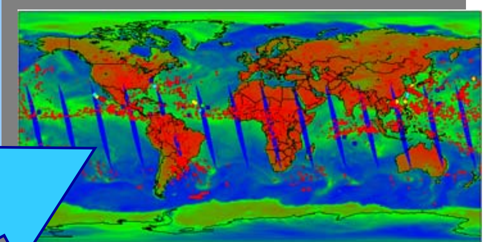
Knowledge Base

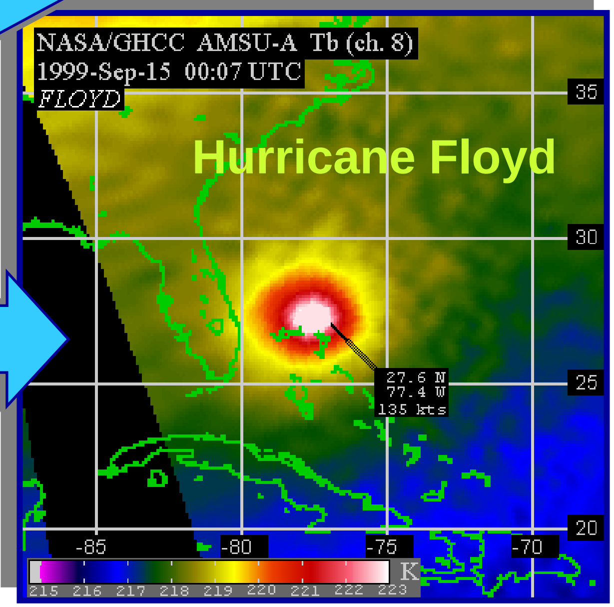
Result

Mining Plan:

- Water cover mask to eliminate land
- Laplacian filter to compute temperature gradients
- Science Algorithm to estimate wind speed
- Contiguous regions with wind speeds above a desired threshold identified
- Additional test to eliminate false positives
- Maximum wind speed and location produced

Further Analysis

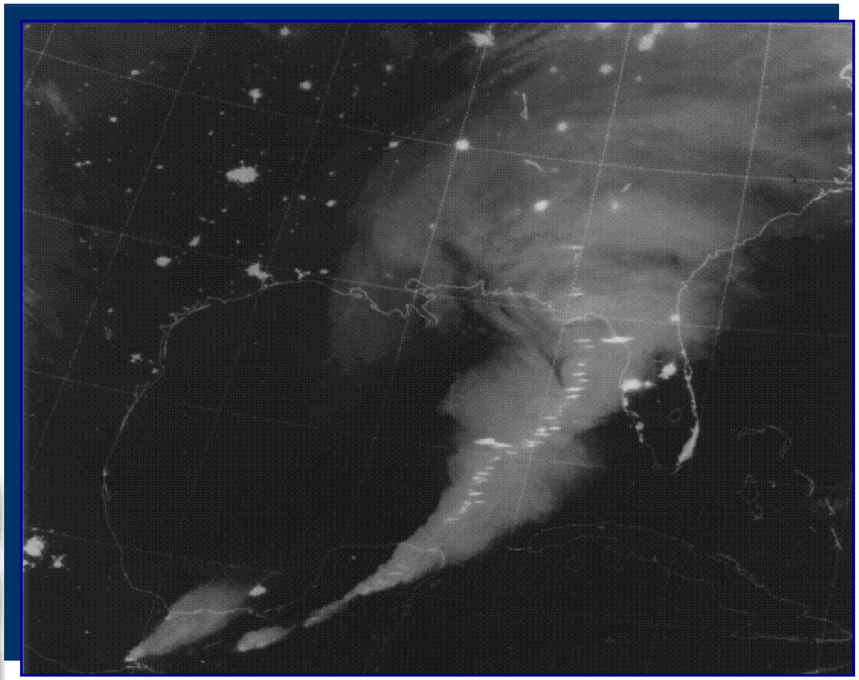




Results are placed on the web, made available to National Hurricane Center & Joint Typhoon Warning Center, and stored for further analysis

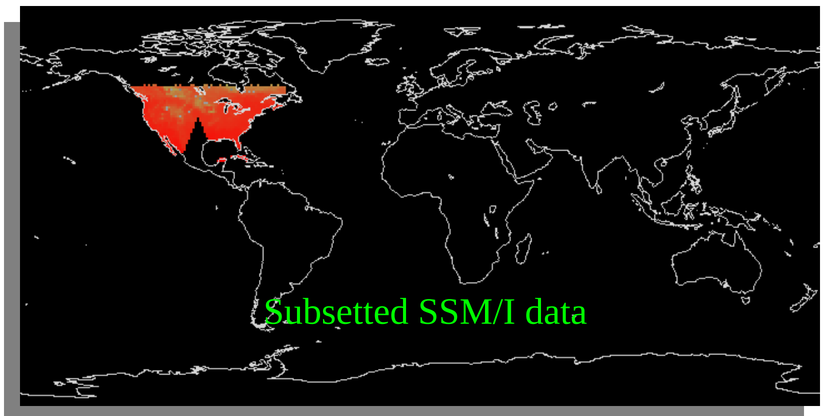
http://pm-esip.msfc.nasa.gov/cyclone

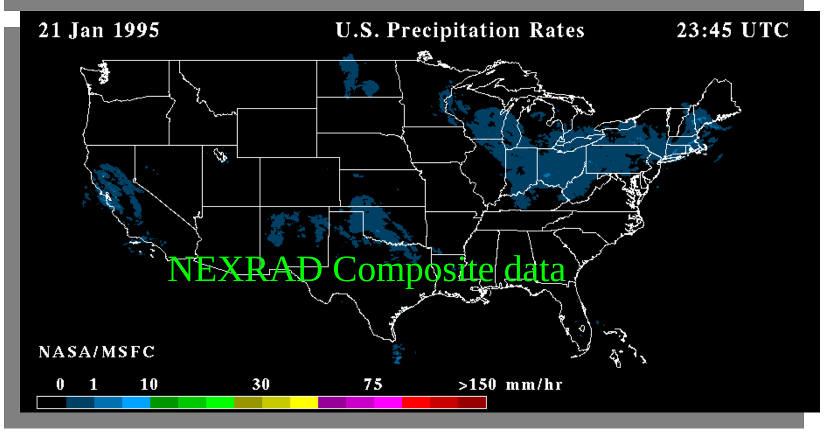
Using Morphological Filtering to Detect Lightning in Operational Linescan System (OLS) Images



- Scientist: Dr. Steve Goodman (GHCC/MSFC NASA)
- To identify lightning streaks in night time portions of OLS images
- OLS is carried by DMSP satellites and produces a visible and thermal image
- Lightning shows up as bright horizontal streaks as do city lights and moonlight reflected off the clouds
- Approach based on morphological filtering and gradient detection was selected
- Both visible and thermal band used

Using Trainable Classifiers for Rainfall Estimation and Identification in SSM/I





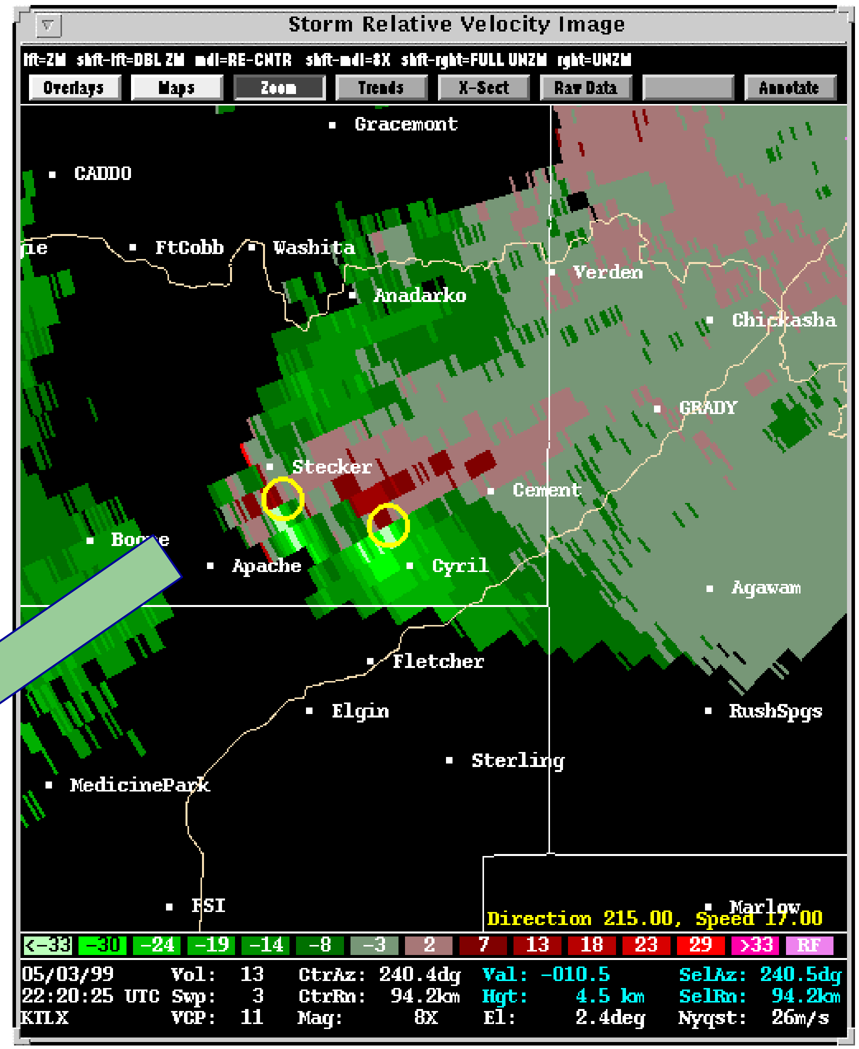
- Scientist: Dr. Steve Goodman (GHCC/MSFC NASA)
- To determine whether generic pattern recognition techniques could be applied to SSM/I data to detect rain
- Minimum Distance Classifier, Back-propagation Neural Network and Discrete Bayes Classifier were compared against a Science Algorithm (WetNet PIP Algorithm)
- US Composite rainfall product was used as ground truth

Cumulus Cloud Classification

- Science Rationale: Man-made changes to land use cause changes in weather patterns, especially cumulus clouds
- ADaM allows comparison of many different classification techniques based on accuracy of detection and amount of time required to classify
- Best algorithm can be used to create cloud mask product

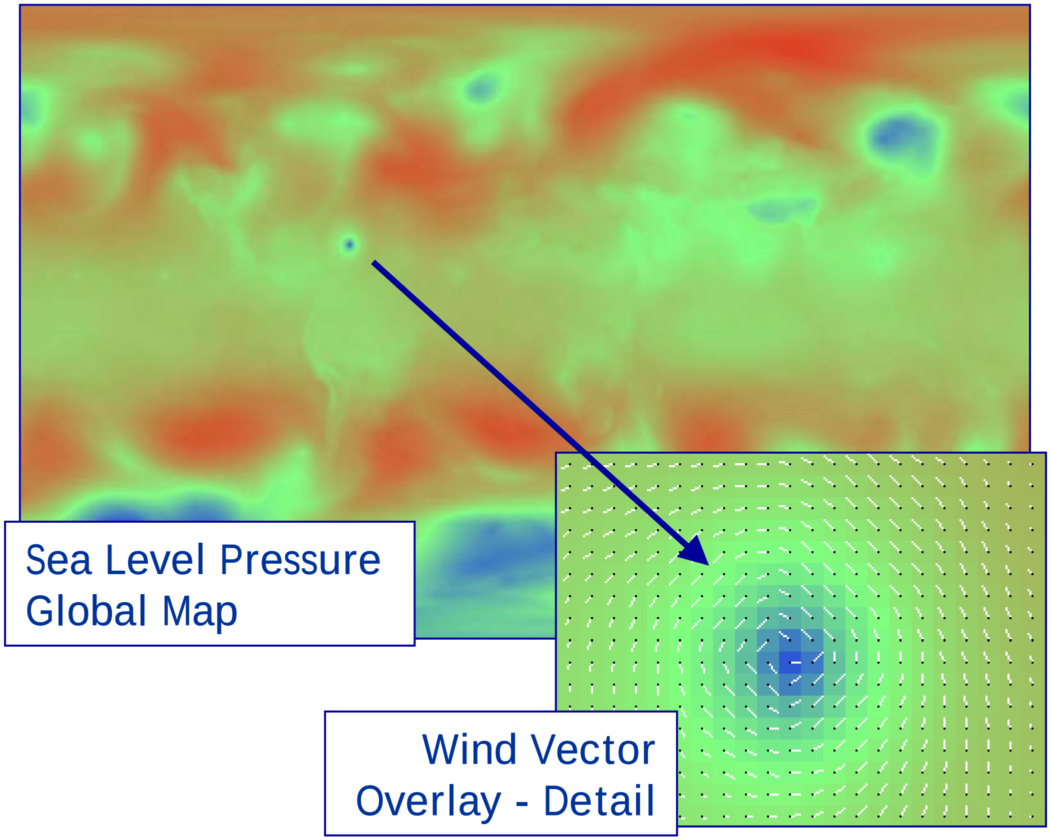
Mesocyclone Signatures

- *Problem*: Detecting mesocyclone signatures in Radar data
- *Science Rationale*: Improved accuracy and reduced false alarm rate for indicators of tornadic activity
- *Technique*: Developing an algorithm based on wind velocity shear signatures



Mining Model Data

To advance the capacity for information extraction from models, NASA/ARC, the Global Modeling and Assimilation Office at NASA/GSFC, ITSC and Simpson Weather Associates are applying data mining frameworks for the analysis and extraction of information from numerical model output data generated or archived at the GMAO. The team is conducting experiments focusing on the automated detection and mining of atmospheric phenomena relationships within the model data.



Tropical Cyclone Identification

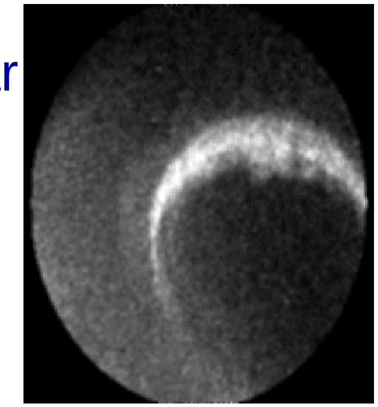
- The heuristic procedure considered all tropical ocean pixels and accepted those that:
  - Had surface pressure below a certain threshold (990)
  - Had vorticity above a certain threshold (15)
- As an alternative to the heuristic procedure, a clustering algorithm was used to derive the signature of the cyclones
  - Using pressure, vorticity
  - Using pressure, vorticity, temperature, cloud total
  - Using pressure, vorticity, cloud low

Information Technology and Systems Center

UAH

Automated Data Analysis for Boundary Detection and Quantification

- *Problem*: Analysis of polar cap auroras in large volumes of spacecraft ultraviolet images
- *Scientific Rationale*: Auroras can be used to predict geomagnetic storms which may damage satellites and disrupt radio connections
- *Technique*: Developing different mining algorithms to detect and quantify polar cap boundary



ORIGINAL IMAGE

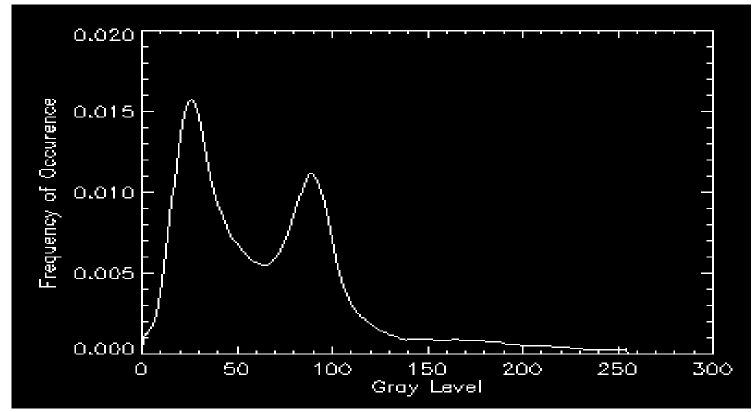
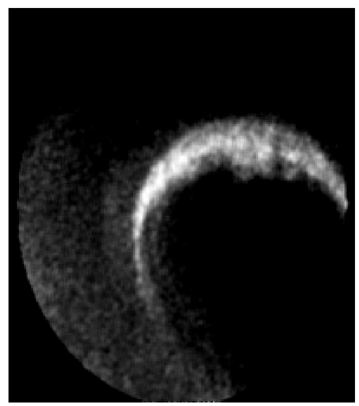
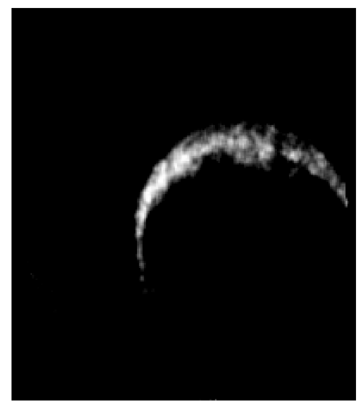


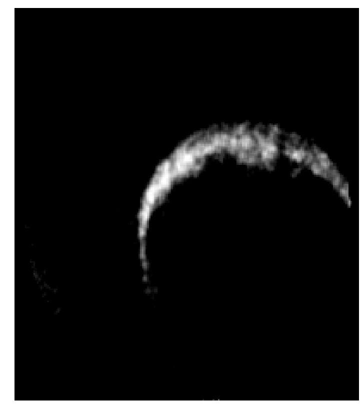
IMAGE HISTOGRAM



MIXTURE MODELING (64)

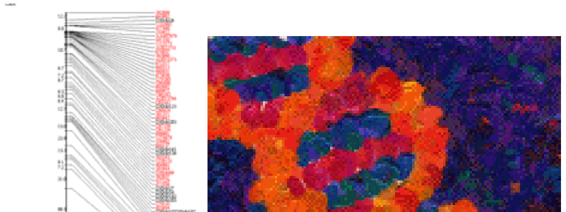


FUZZY SETS (132)

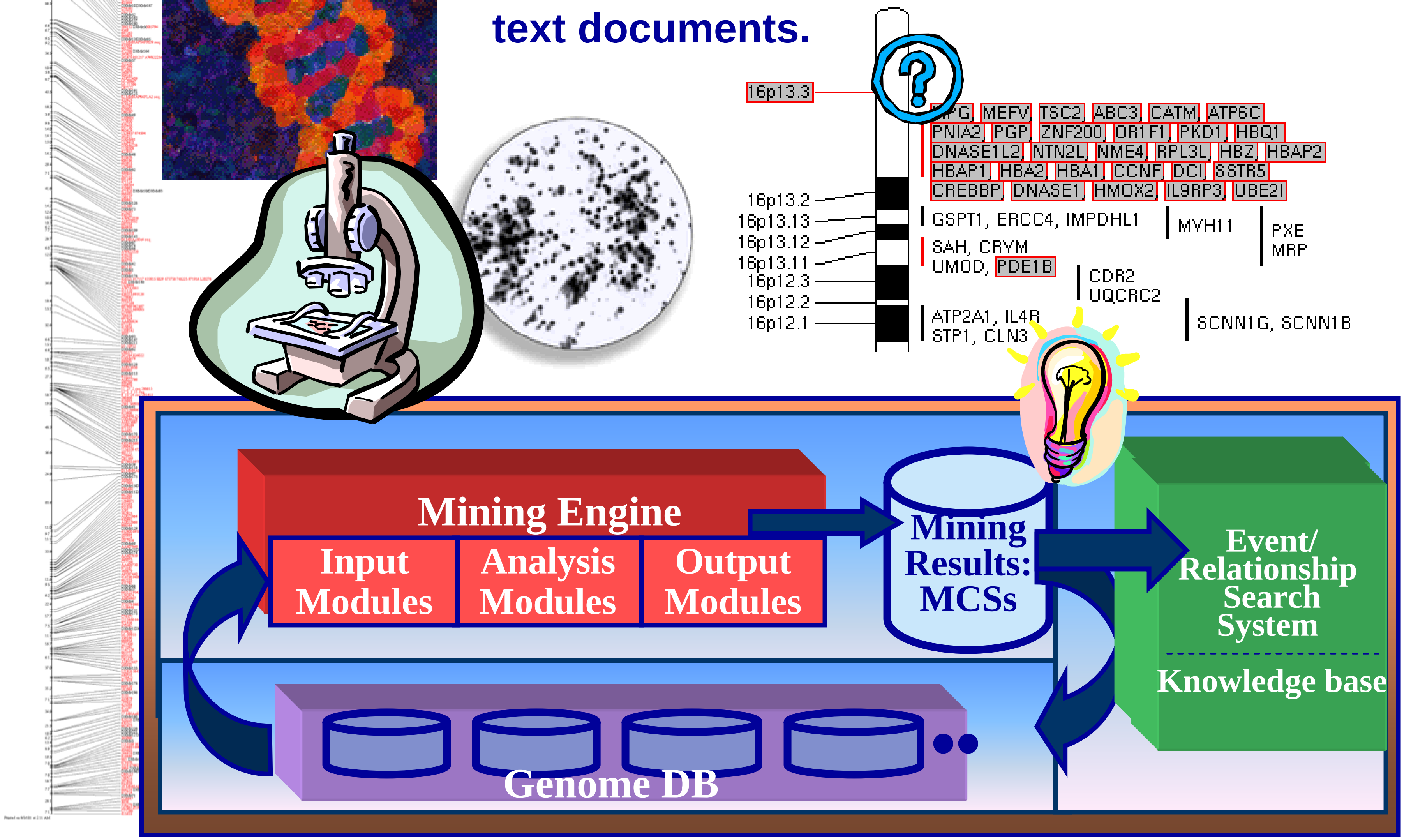


ENTROPY (122)

BioInformatics: Genome Patterns



<u>Text Pattern Recognition</u>:
Used to search for text patterns in bioscience data as well as other text documents.



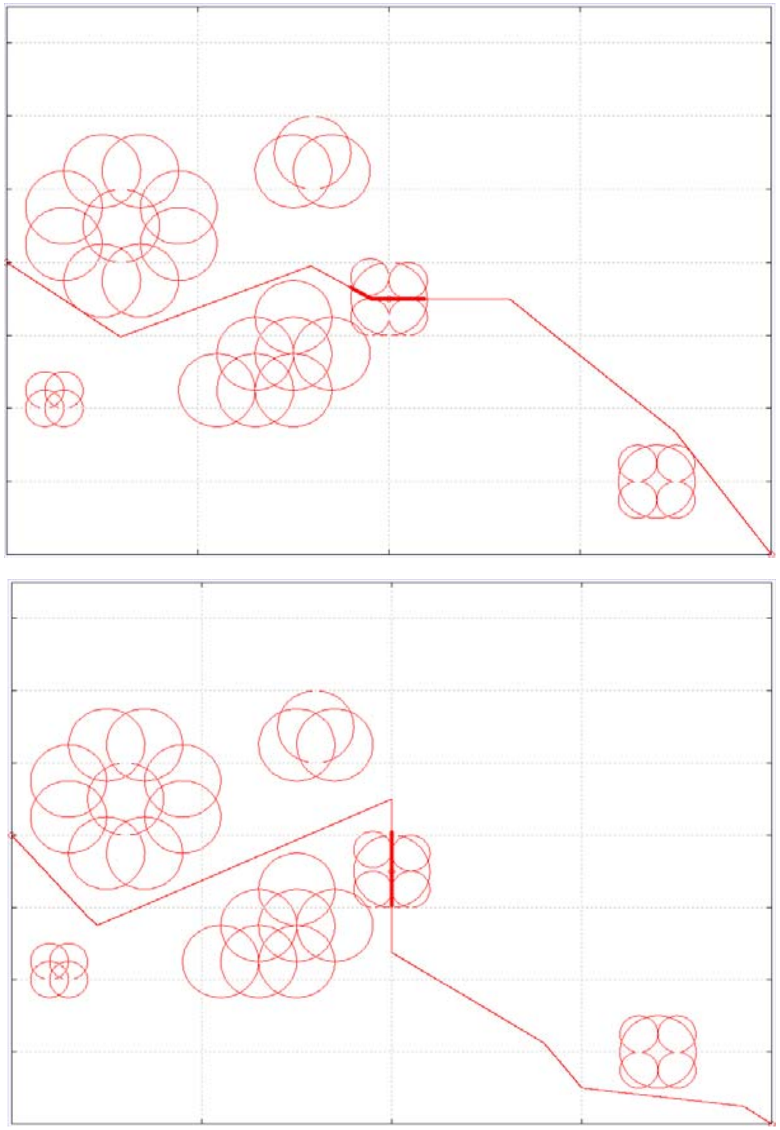
Data Mining and Optimization Methods for Wargames

- Playing a wargame involves making decisions at many levels, from overall strategies to pursue down to actions for individual units.
- Creating a realistic and challenging Artificial Intelligence (AI) for a wargame requires the solution of many different problems.
- Problems of interest include both assessment of current dispositions and planning of actions
- Data mining and optimization methods can be used to solve some of these problems.

Optimization Methods for Flight Path Planning

Problem: Plan a flight path from a source to a target and then an destination. Minimize risks posed by enemy air defenses while not exceeding fuel allowance.

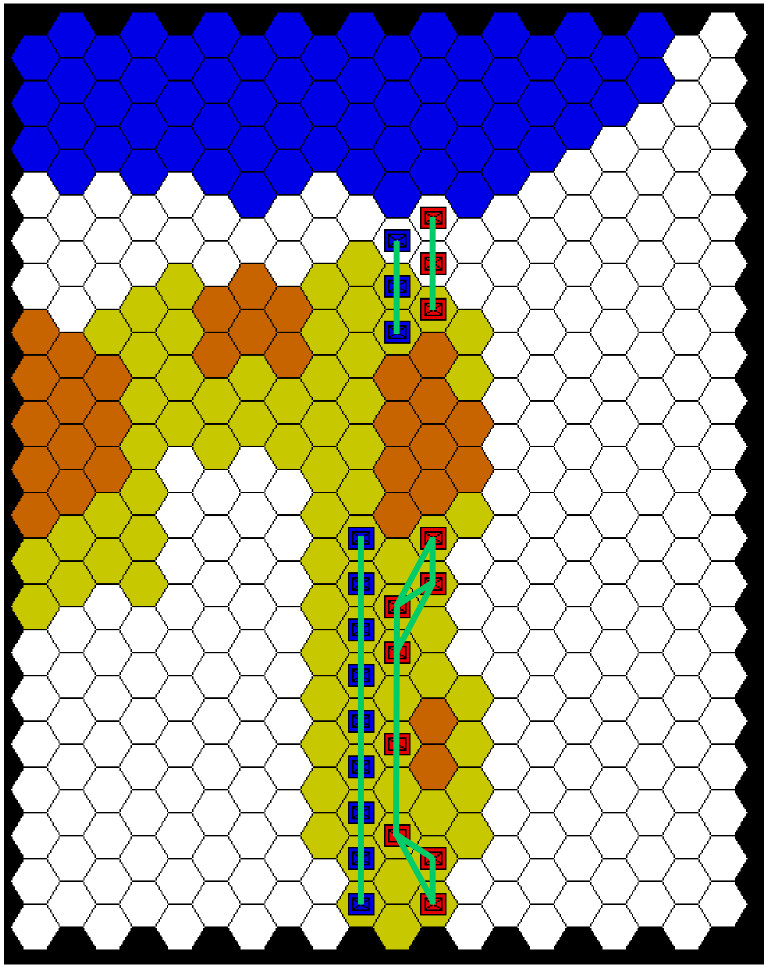
- ***Techniques:*** Genetic Algorithm and Greedy search methods for minimizing risk
- ***Enhancements:*** Encoding of flight paths using bit strings, computation of risk as intersection of path segments with air defenses
- ***Result:*** Flight paths plotted as lines, solid where afterburner applied. Enemy air defenses as circles.



Density Based Clustering For Determination of Front Lines

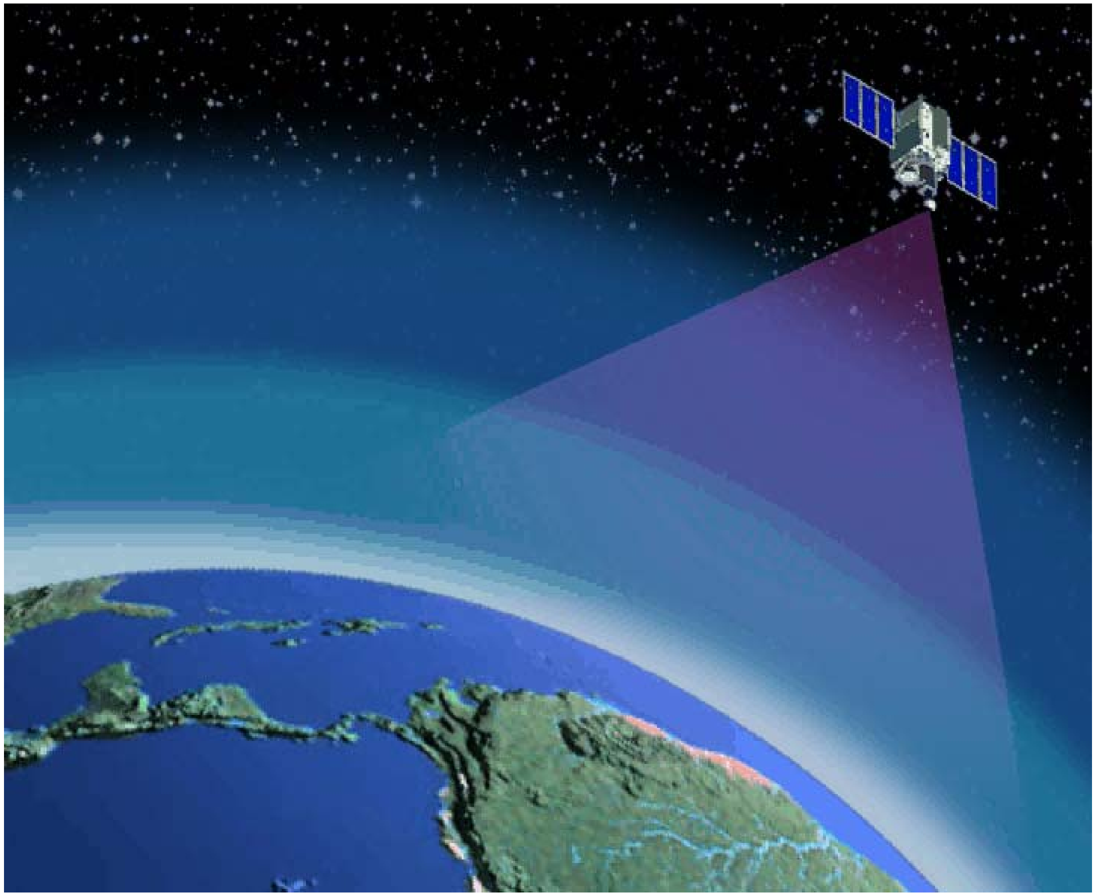
Problem: Define front lines, contiguous groups of units that can provide mutual support and prevent enemy movement. Groups of units placed with sufficient density constitute a front line.

- ***Mining technique:*** Density based clustering algorithm that identifies contiguous dense regions of points
- ***Enhancements:*** Compute distances on hexagonal grids, factoring in variable terrain and obstructions
- ***Result:*** Links units that are mutually supporting (green lines in figure at right)



On-Board Real-Time Processing Sensor Control/Targeting

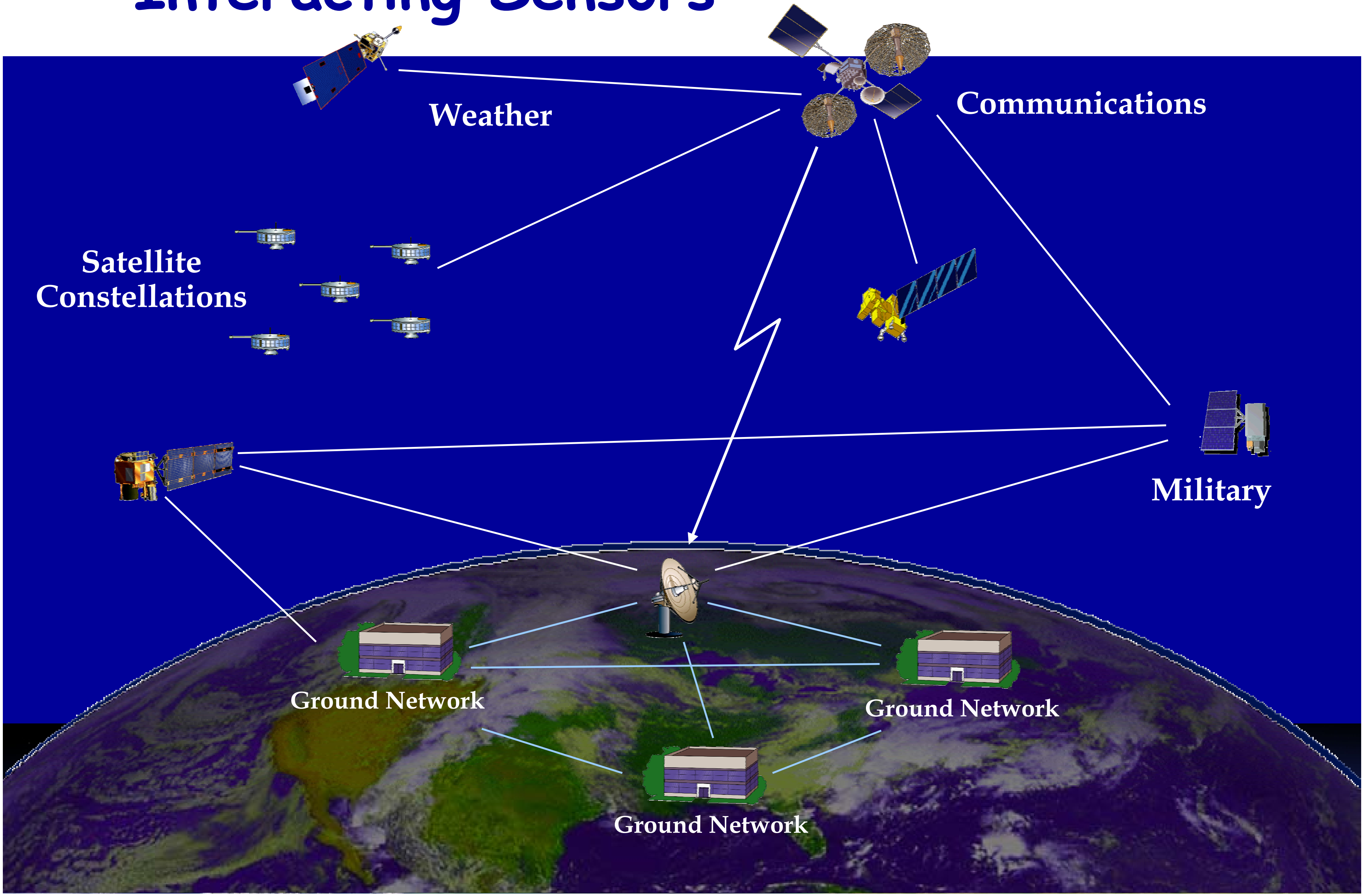
EVE – Environment for On-board Processing



www.itsc.uah.edu/eve

- **Anomaly detection**
- **Data Mining**
- **Autonomous Decision Making**
- **Immediate response**
- **Direct satellite to Earth delivery of results**

A Reconfigurable Web of Interacting Sensors

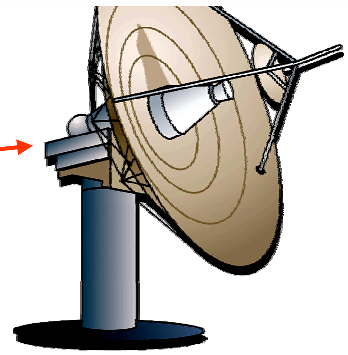


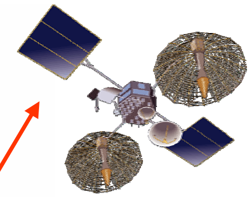
UAH Information Technology and Systems Center

Example Application of EVE Technology: Lightning Detection During Tornadic Activity

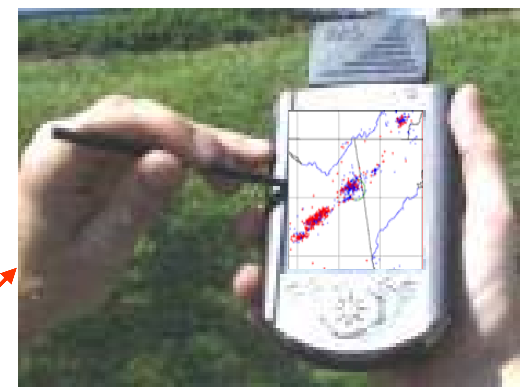
1) The user creates a mining plan using the EVE editor

2) The Ground Station uploads the plan to multiple on-board platforms





3) On-board Platform 1 uses its sensor to watch for lightning events



4) Platform 1 notifies Platform 2 of the event

5) Platform 2 requests subsetting web services from an NSSTC server

NSSTC Core Facility

6) The results are sent back to Platform 1 for display and further processing

Emerging Cyberinfrastructures and Research Communities

NOMADS
The NOAA Operational Model Archive and Distribution System

GriPhyN
Data Intensive Science

Homeland Security

GEON
CYBERINFRASTRUCTURE FOR THE GEOSCIENCES

NVODS
The National Virtual Ocean Data System

TERAGRID

MADIS

NSDI

DLESE
A community-centered resource for anyone interested in learning more about the Earth.
Supported by the National Science Foundation

earthscope

NVO
NATIONAL VIRTUAL OBSERVATORY

Earth System Grid

Building the National Virtual Collaborat
for Earthquake Engineering Research
NEESgrid

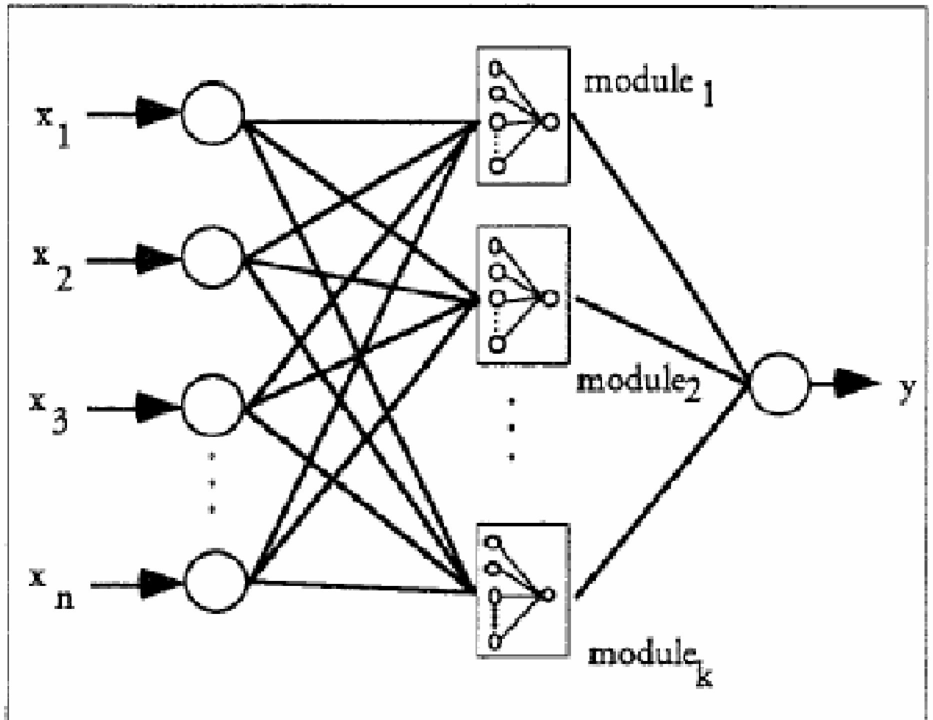
ESIP FEDERATION

nbii
National Biological Information Infrastructure

VGEE
Visual Geophysical Exploration Environ.
Home

Self Partitioning Neural Networks (SPNN) for the TeraGrid Expedition

- Opposing forces within the training set are responsible for most of the training problems in a Back-Propagation Neural Network (Ranganath and Kerstetter, 1995)
- Dissimilar targets organize into groups and oppose each other leading to little or no learning
- SPNN partitions the target patterns into co-operative groups and trains each of the target groups on a sub-network

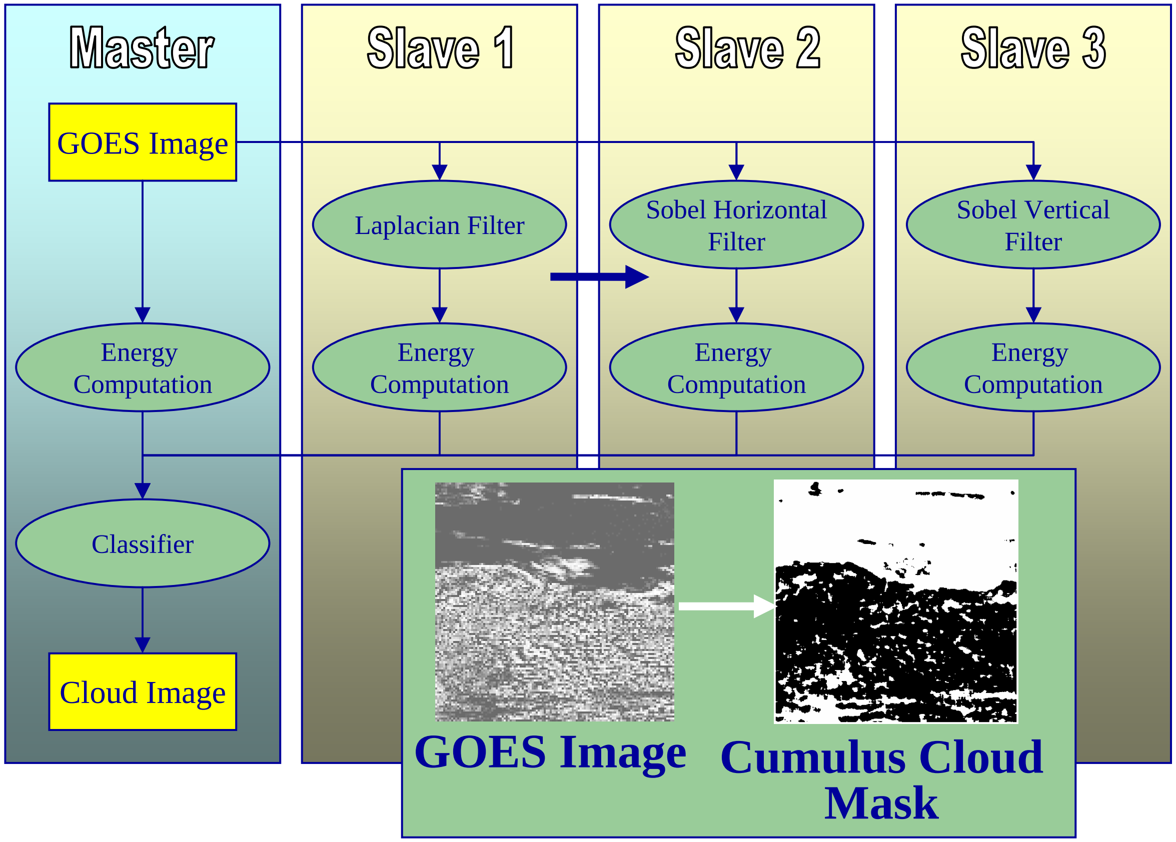


The SPNN architecture.

- SPNN architecture lends itself for parallelization
  - Fine grained for training
  - Fine or coarse grained for classification

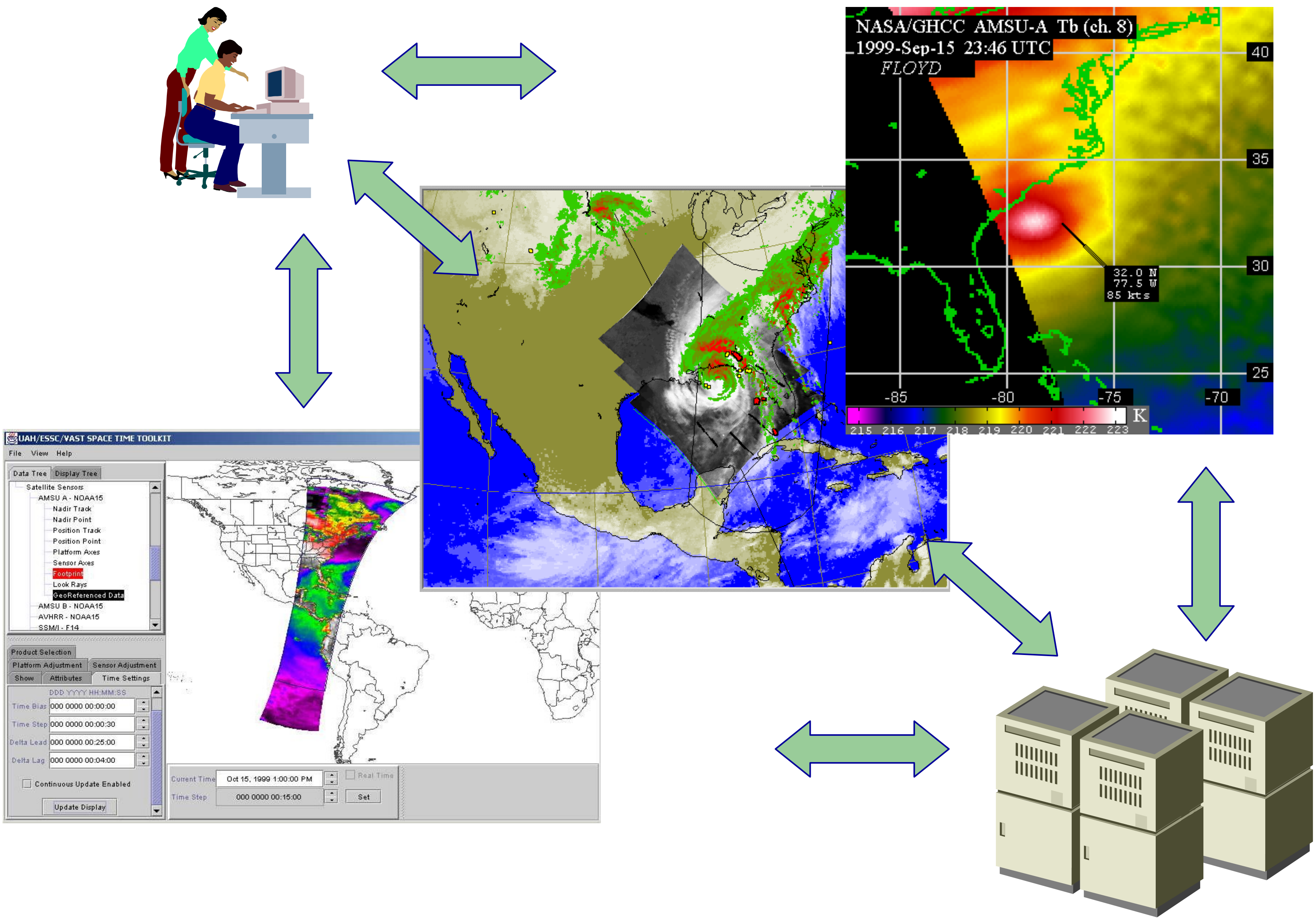
Parallel Version of Cloud Extraction

- GOES images can be used to recognize cumulus cloud fields
- Cumulus clouds are small and do not show up well in 4km resolution IR channels
- Detection of cumulus cloud fields in GOES can be accomplished by using texture features or edge detectors



- Three edge detection filters are used together to detect cumulus clouds which lends itself to implementation on a parallel cluster

Online Access through Tools and Services



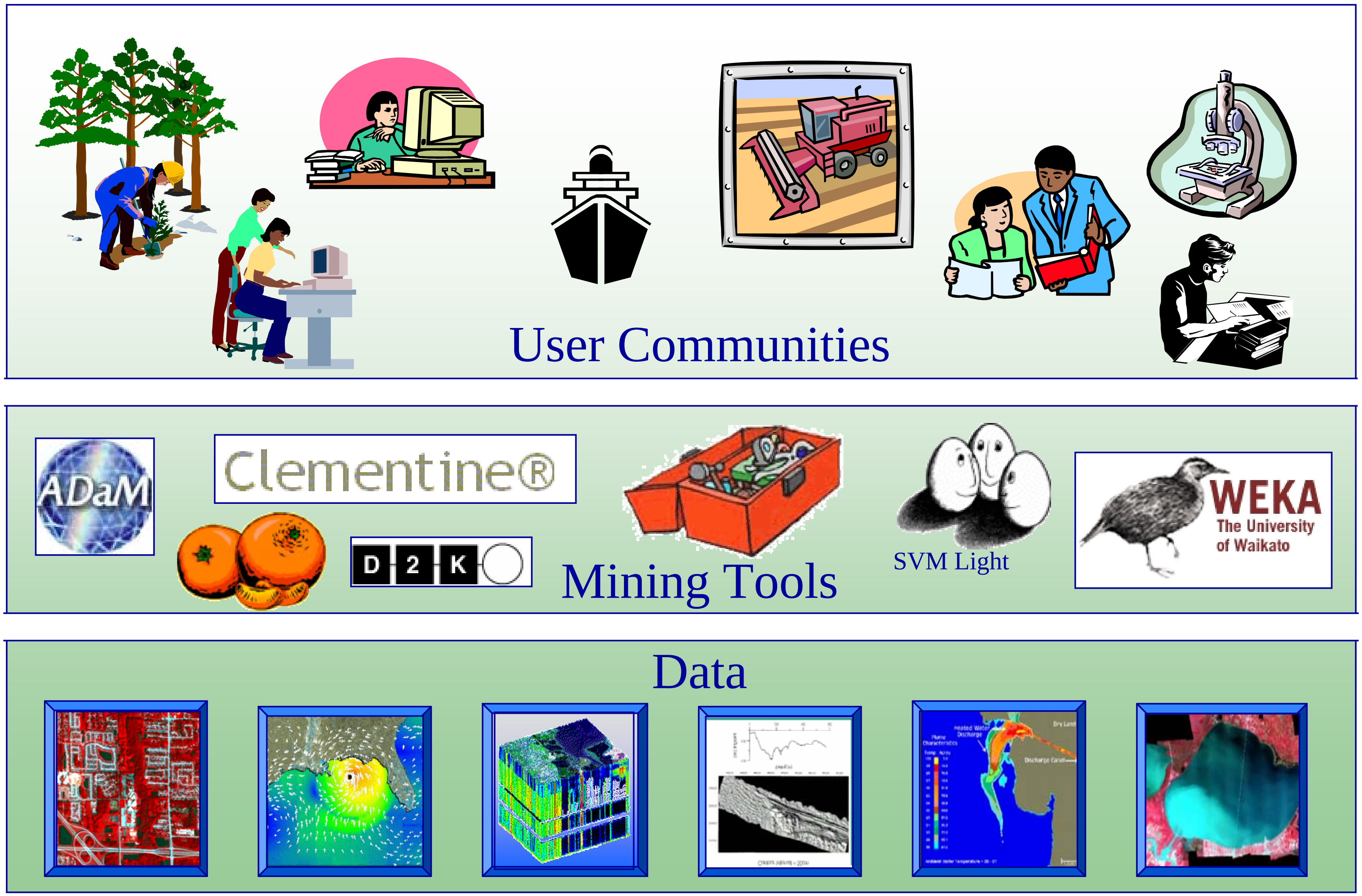
UAH Information Technology and Systems Center

Improving Data Usability

- **Advanced Applications Development**
  - *Data organization and management* for archival and analysis
  - *Data Mining* in real-time and for post run analysis
  - *Interchange Technologies* for improved data exploitation
  - *Semantics* to transform data exploitation via intelligent automated processing
- **Infrastructure Development**
  - *Grid technologies* for seamless access to multiple computational and data resources into a virtual computing environment
  - *Cluster technologies* for high speed parallel computation, for multiple agent computations, and other applications
  - *High-performance networking* for advanced applications development and high performance connectivity
  - Next generation technologies in *videoconferencing and electronic collaboration*

Achieving Usability ?

Meeting the Data Usability Challenge



UAH Information Technology and Systems Center

NSF Cyberinfrastructure Report

2003

Advanced Cyber Infrastructure

Testimony for the

NSF Advisory Committee on Cyber Infrastructure

January 22, 2002

Sara J. Graves
Director, Information Technology and Systems Center
Professor, Computer Science Department
University of Alabama in Huntsville
Director, Information Technology and Research Center
National Space Science and Technology Center
256-824-6064
sgraves@itsc.uah.edu

Revolutionizing Science and Engineering Through Cyberinfrastructure:

Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure

January 2003

Daniel E. Atkins, Chair
University of Michigan

Kelvin K. Droegemeier
University of Oklahoma

Stuart I. Feldman
IBM

Hector Garcia-Molina
Stanford University

Michael L. Klein
University of Pennsylvania

David G. Messerschmitt
University of California at Berkeley

Paul Messina
California Institute of Technology

Jeremiah P. Ostriker
Princeton University

Margaret H. Wright
New York University

Key Questions:

- What is the most effective approach to developing an integrated framework and plan for an interdisciplinary environmental cyberinfrastructure?
- What organizational structure is needed to provide long-term support for data storage, access, model development, and services for a global clientele of researchers, educators, policy makers, and citizens?
- How will effective interagency and public-private partnerships be formed to provide financial support for such an extensive and costly system?
- How can communication and coordination among computer scientists and environmental researchers and educators be enhanced to develop this innovative, powerful, and accessible infrastructure?

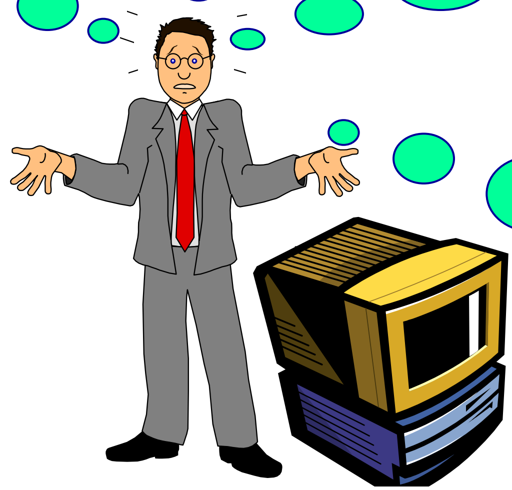
Mining Environment: When, Where, Who and Why?

WHEN
- Real Time
- On-Ingest
- On-Demand
- Repeatedly

WHERE
- User Workstation
- Data Mining Center
- GRID

WHO
- End Users
- Domain Experts
- Mining Experts

WHY
- Event
- Relationship
- Association
- Corroboration
- Collaboration



Data Mining

Challenges

- Develop and document common/standard interfaces for interoperability of data and services
- Design new data models for handling
  - real-time/streaming input
  - data fusion/integration
- Design and develop distributed standardized catalog capabilities
- Develop advanced resource allocation and load balancing techniques
- Exploit the Grid for enhanced data mining functionality
- Develop more intelligent and intuitive user interfaces
- Develop ontologies of scientific data, processes and data mining techniques for multiple domains
- Support language and system independent components
- Incorporate data mining into scientific curricula

Data Integration and Mining: From Global Information to Local Knowledge