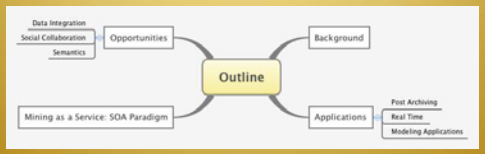
Mining 3.0: Injecting Semantics into the Mining Process

and MORE!

Rahul Ramachandran
Information Technology and Systems Center
University of Alabama Huntsville
rramachandran@itsc.uah.edu
http://geoinformatics.itsc.uah.edu/

UAHuntsville

Presentation Outline



Background

"Dealing with data deluge"

" By the year 2000, satellites deployed by the NASA will be transmitting 1 terabytes of data to earth every day"

"If stored on magnetic tape, this daily deluge would require a stack of reels as high as the 169 meter Washington monument in the U.S. Capital"

"But when the data are so complex and voluminous and require complicated algorithms for processing and visualization, important phenomena described by them may be **overlooked**"

N. Gershon, G. Miller, 1993: Dealing with data deluge, IEEE Spectrum

Problem still remains but now we are talking in petabytes, exabytes!

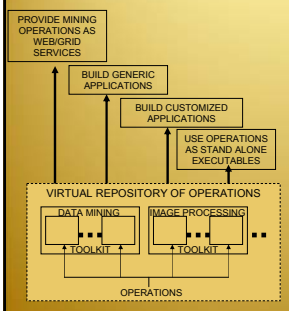
NASA Data Mining Workshop

- "Issues in the Application of Data Mining to Scientific Data"
- Hosted by UAHuntsville in 1999
- "Data mining techniques are key elements in making use of the prodigious volume of data that NASA and other earth remote sensing organizations will be capturing in the coming millennium."
- "..investments be made in developing advanced data mining tools and technologies."

J. Behnke, E. Dobson: NASA Workshop on Issues in the Application of Data Mining to Scientific Data, SIGKDD Explorations

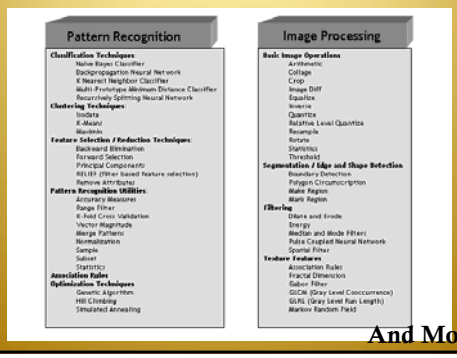
- Synergy between Scientists & Data Miners
- Data mining tools for Scientists

ADaM Toolkit



- Component based where each component is provided with a C++ application programming interface (API), an executable in support of scripting tools (e.g. Perl, Python, Tcl, Shell)
- ADaM components are lightweight and autonomous, and have been used successfully in a grid environment
- ADaM has several translation components that provide data level interoperability with other mining systems (such as WEKA and Orange), and point tools (such as libSVM and svmLight)
- ADaM toolkit is available via the web

ADaM Components

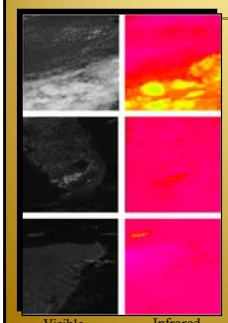


And More !

Applications: Post Archive

1. Cumulus Cloud Classification in GOES Imagery
2. Evidence of Satellite Fragmentation by Orbital debris
3. Dust detection in MODIS Imagery

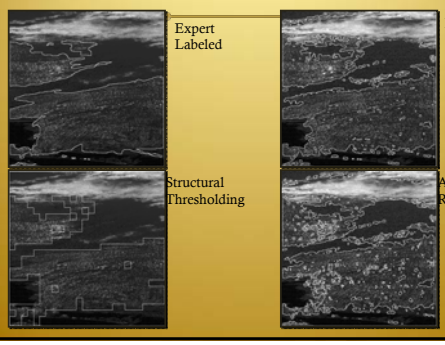
(1) GOES Cumulus Cloud Classification



Visible Infrared

- Motivation
  - Understanding of cumulus cloud climatology, important for climate change studies
  - Developing realistic models for cloud field structure for use in radiative transfer models
  - Understand the influence of land use on regional climate
  - Existing methods had problems in detecting boundary layer cumulus
- Compare the accuracy and performance of different methods:
  - Structural Thresholding
  - Classifiers based on
    - Texture features (GLCM, GLRL, Association Rules)
    - Spectral features
    - Edge detection techniques
- Utilize the best method and mine GOES data for certain time period

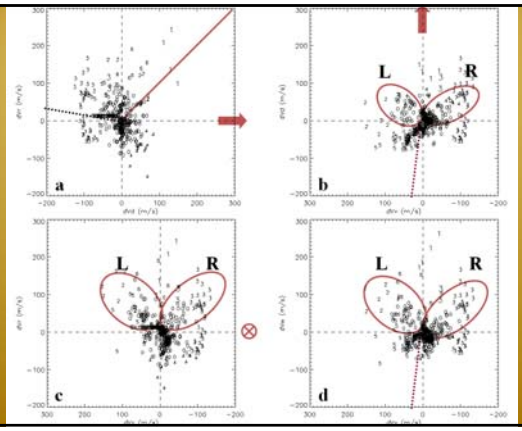
GOES Cumulus Cloud Classification



(2) Evidence of Satellite Fragmentation by Orbital Debris

- Since 1961, the number of satellite fragmentations in space had escalated to a cumulative total of 170 by 2001.
- These fragmentations have created hazardous orbital debris and pushed the number by trackable objects in orbit to over 8,900 by 2001.
- Most of the fragmentations were explosions of rocket bodies due to ignition of residual fuel; many were due to deliberate actions taken by the former Soviet Union; at least one was the result of a U. S. Anti-satellite (ASAT) experiment; and few were suspected to be associated with the Soviet ASAT program of the past.
- Our analysis finds compelling circumstantial evidence that satellite fragmentation by orbital debris may have already taken place.
- Clustering algorithm was used as part of ***Exploratory Data Analysis*** and was critical in identifying the high velocity particles (outliers)!!

Tan, A. and R. Ramachandran, 2005: Velocity Pertubations Analysis of the Spot 1 Ariane Rocket Fragmentation. *Journal of the Astronautical Science,* ***53, 39-50.***



(3) Detect Dust Storms in MODIS Imagery

- Motivation
  - Operational implications: Civilian and Military
  - Science
    - Dust storms from arid and semi arid regions transport dust containing iron to remote ocean locations, which spur the growth of plankton
    - Phytoplankton act as an important sink of atmospheric carbon dioxide
    - Phytoplankton also emit of Dimethly Sulfide (DMS) which increases cloud condensation nuclei (CCN); increase in CCN results in more but smaller size cloud droplets in marine stratus thereby increasing cloud albedo and creating a cooling effect
    - *Dust storms indirectly impacts global climate*

Ramachandran, R., X. Li, S. Movva, S. Graves, U. S. Nair, and C. Lynnes, 2007: Investigating Data Mining Techniques to Detect Dust Storms in MODIS Imagery. *32nd International Symposium on Remote Sensing of Environment, San Jose, Costa Rica.*

CNN.com

WORLD

Sandstorm derails train in China; 4 die

POSTED: 8:51 a.m. EST, February 26, 2007

STORY HIGHLIGHTS

- 11-car train had just left a station in Turpan, in the Xinjiang region's east
- Sandstorm cracked windows and knocked train cars off track, passenger says
- At least 100 rescue workers, doctors and police were at the scene

BEIJING, China (AP) -- A sandstorm with hurricane-strength wind gusts has derailed a train in China's far west, killing at least four people and injuring another 30, state media says.

The 11-car train had just left a station in Turpan, in the Xinjiang region's east, when it was hit by wind at 2 a.m. Wednesday, the Xinhua News Agency said.

"A strong sandstorm cracked window panes soon after the train left Turpan, and blew some of the cars off the rail when we were trying to plug up the windows," Xinhua quoted passenger Su Chuanyi as saying.

Sandstorms fed by the deserts of Xinjiang and inner Mongolia blow toward Beijing and the eastern seaboard each spring powered by vigorous winds. Sometimes, the dust blows out across the Pacific, clouding the skies of South Korea and occasionally drifting as far as the western coast of the United States.

At least 100 rescue workers, doctors and police were at the scene of the derailment, which disrupted operations of the Southern Xinjiang Railway, it said.

The train had been running between the capital city of Urumqi and Aksu, in the west.

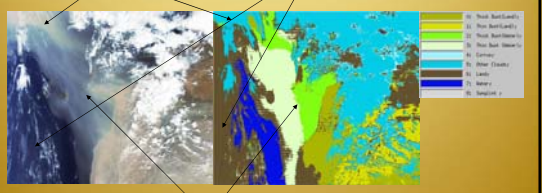
Al Asad, Iraq 2005 – Wall of Sand traveling at 60 mph

Problem Specification

- Very few studies have utilized remote sensed satellite imagery to study fertilizations of oceans by dust storms and the resulting plankton growth
- To study these connections, a robust algorithm is required to detect and identify Saharan dust passage over remote locations in the Atlantic
- Existing methods in the literature use radiative properties of dust to constraint the feature space (use only specific spectral bands)
- Thresholds are derived based on the analysis of limited cases over specific region
- Effectiveness of thresholds can vary from region to region and from different studies for a specific region.
- Thresholds use a simple linear decision boundary

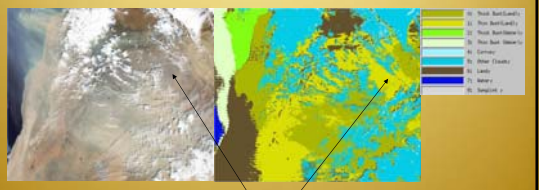
- **Objective:**
- Utilize Data Mining/ Machine learning techniques to create a classifier that would work over land, water and different regions
- Couple feature selection algorithms with a classifier to decide the best features

Forward Elimination/Bayes Classifier Results (1)



Dust over water detections looks reasonable

Forward Elimination Results (2)



Dust over land results are difficult to judge

Applications: Real Time

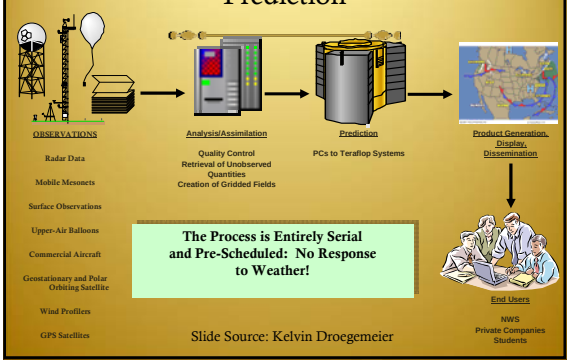
1. Storm Detection and Clustering for Dynamic Model Forecast Initialization
2. Sensor Management for Applied Research Technologies for On-Demand Modeling

(1) Linked Environments for Atmospheric Discovery (LEAD)

- A ***cyberinfrastructure*** for mesoscale meteorology where scientists, students, and operational practitioners interact with the weather.
  - Real-time, on-demand, and ***dynamically adaptive*** needs for mesoscale weather research
  - High volume data sets and streams
  - Computationally demanding numerical models and data assimilation systems

Droegemeier et al., 2005: Service-Oriented Environments in Research and Education for Dynamically Interacting with Mesoscale Weather. *IEEE Computing in Science & Engineering*, ***7, 24-32.***

Conventional Numerical Weather Prediction

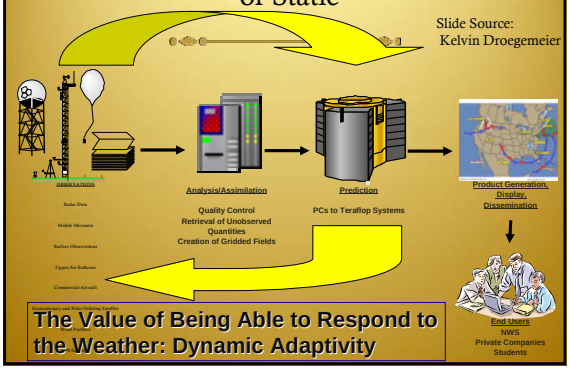


The Process is Entirely Serial and Pre-Scheduled: No Response to Weather!

Slide Source: Kelvin Droegemeier

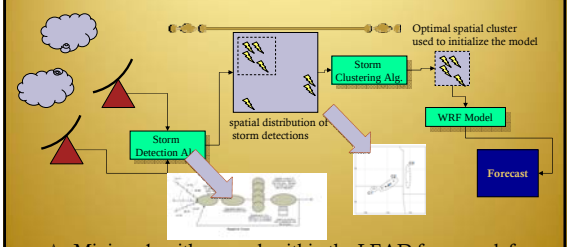
The LEAD Vision: No Longer Serial or Static

Slide Source: Kelvin Droegemeier



The Value of Being Able to Respond to the Weather: Dynamic Adaptivity

LEAD Proof of Concept: Storm Detection and Clustering for Dynamic Model Forecast Initialization



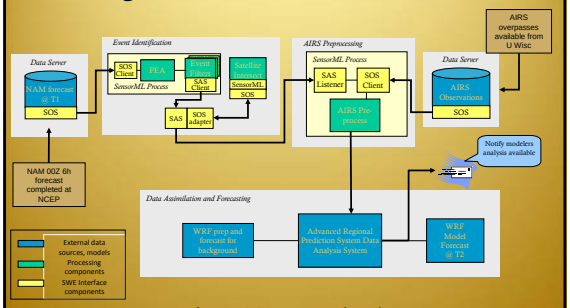
- Mining algorithms work within the LEAD framework for a dynamic and adaptive response by running a workflow of services (model forecasts) triggered by weather events

Li, X., B. Plale, N. Vijayakumar, R. Ramachandran, S. Graves, and H. Conover, 2008: Real-time Storm Detection and Weather Forecast Activation through Data Mining and Events Processing. *Earth Science Informatics.*

(2) Sensor Management for Applied Research Technologies for On-Demand Modeling (SMART-ODM)

- Explore the use of OGC SWE capabilities to integrate both satellite observations and forecast model outputs into new data acquisition and assimilation strategies to improve regional weather forecast models
- Intelligently assimilate Atmospheric Infrared Sounder (AIRS) satellite temperature and moisture retrievals into a regional Weather Research and Forecasting (WRF) model over the southeastern United States.
- The addition of current weather can improve the accuracy of a WRF forecast
  - Assimilating voluminous satellite data is computationally expensive.
  - Only worth assimilating during the significant weather
- SMART system ***detects weather phenomena***, namely precipitation and convection, and to determine the coincidence of observations from additional satellite instruments

Near Real Time AIRS Assimilation Integrated with SPoRT Processes



http://smart.uah.edu

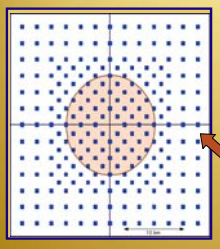
Applications: Numerical Modeling

1. Mining Ensemble Model Output
2. Solving Assimilation Problem Using Intelligent Data Thinning

(1) Mining Ensemble Model Output

- Science Objective:
  - To investigate different thunderstorm cell interactions favorable for subsequent tornado (mesocyclone) formation
- Approach:
  - Use idealized WRF model simulations with different initial conditions (ensemble model runs)
  - Create a large parameter space of thunderstorm cell interaction and storm behavior
  - Mine this parameter space for interesting patterns and trends
    - Mining Algorithm to detection Mesocyclones
    - Tracking algorithm to characterize Mesocyclones
    - Cluster different runs based on a derived set of features

WRF Initializations

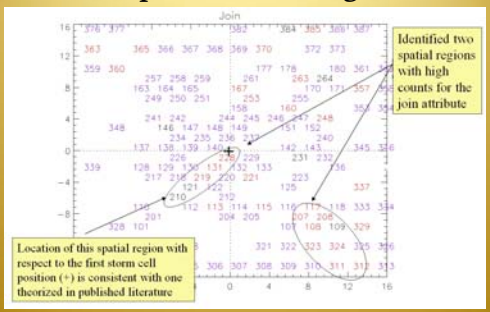


Matrix of WRF simulations

- 230 WRF runs were made, + two control (single-cell)
- Each corresponded to a particular arrangement of a pair of initial storm cells

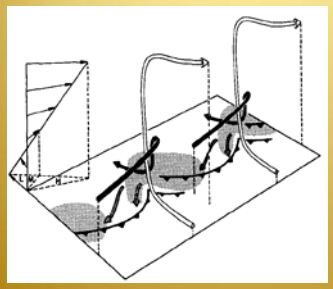
Figure Legend:
- Each square represents a simulation run
- 1st storm cell is always initialized in the middle; 2nd cell can be at any of the blue squares
- The center cell stronger

Spatial Clustering



Ramachandran, R., X. Li, S. Movva, J. Rushing, S. Tanner, S. Graves, B. Jewett, and R. Wilhelmson, 2004: Mining Ensemble Model Data. *Seventh Workshop on Mining Scientific and Engineering Datasets, SIAM International Conference on Data Mining, Lake Buena Vista, Florida.*

Line orientation Theory



Rotunno, Klemp, and Weisman (RKW, 1988)

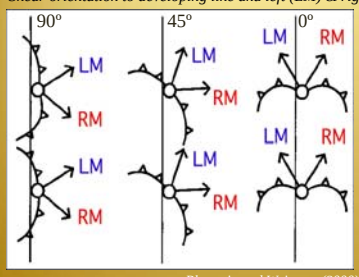
RKW (1988) studied the conditions needed for long-lived squall lines, including the now-famous "optimum shear" between the cold pool & environment.

They also found that shear at a 45º angle led to supercells; split cells didn't interact.

Slide Source: Brian Jewett

Early work: line orientation

Shear orientation to developing line and left (LM) & right (RM) cell motion.



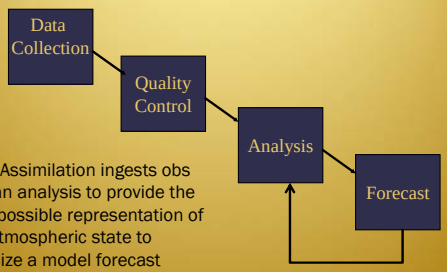
Bluestein & Weisman found that when shear was perpendicular to the line, destructive cell interaction took place along the line.

As in RKW, shear at a 45° angle led to supercells - the LM moved into cold air.

Bluestein and Weisman (2000)

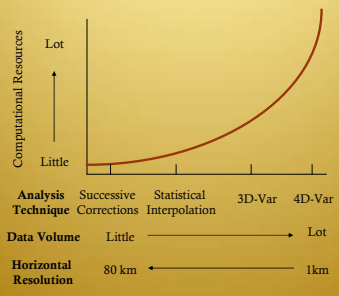
Slide Source: Brian Jewett

(2) Intelligent Data Thinning for Satellite Imagery



Data Assimilation ingests obs into an analysis to provide the best possible representation of the atmospheric state to initialize a model forecast

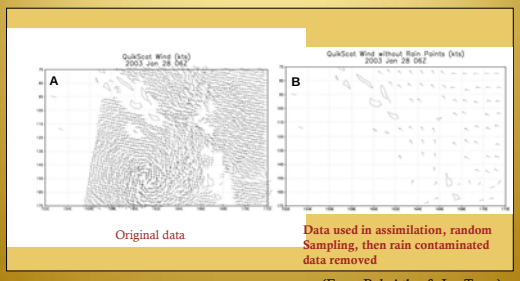
Computational Resources Required for Data Assimilation



Need for 'new' Data Thinning Techniques

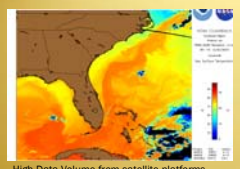
- Current data thinning approaches
  - Sub-sampling
  - Random Sampling
  - Super-Obing (subsampling with averaging)
- Limitations
  - All data points are treated equally
  - Miss important features by sub-sampling features of interest such as cyclones
- ***Intelligent Data Thinning Algorithms***
  - Reduces number of data points required for an analysis
  - Maintains fidelity of the analysis (keeps the most important data points)

Example

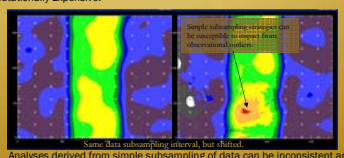


(From Bob Atlas & Joe Terry)

Example



High Data Volume from satellite platforms
(e.g. infrared based SST, scatterometer winds) carry redundant data. Computationally Expensive!



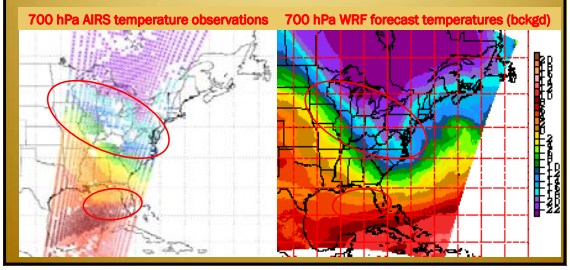
Analyses derived from simple subsampling of data can be inconsistent and are not optimal in efficiency.

Compared Thinning Strategies (10% of full)

- Subsample:
  - Takes profile with most retrieved levels within a 3x3 box
- Random:
  - Searches observations and ensures that retained observations are thinned to a user-defined distance
  - 10 permutations performed to create an ensemble
- mDADT:
  - modified Density Adjusted Data Thinning (DADT)
  - DADT builds a thinned set of observations from an initially empty set using a priority queue consisting of variance intensity between observations and their neighbors
  - thins on 2-D pressure levels using equivalent potential temperature; then levels are recombined to form 3-D structure
  - modified to use thermal front parameter to determine gradients and anchor points; retains those first
  - once these have been selected it takes observations in the homogeneous regions in reverse order of the priority queue (least variant)

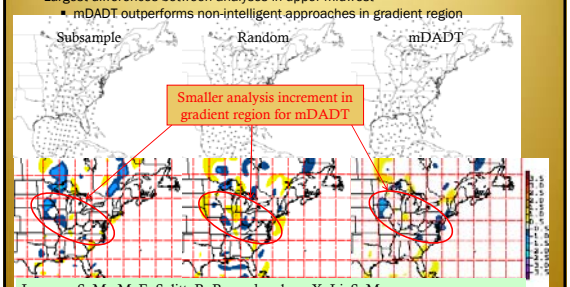
Case Study Day: 12 March 2005

- 700 hPa temperature gradient in observations and background over midwest and northern Gulf of Mexico
- Observations and background show similar patterns



700 hPa Temperature Analysis Comparison

- Overall analysis increments are ±1.5°C over AIRS swath
- Largest differences between analyses in upper midwest
  - mDADT outperforms non-intelligent approaches in gradient region



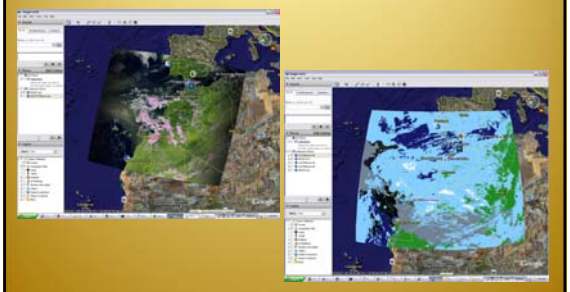
Lazarus, S. M., M. E. Splitt, R. Ramachandran, X. Li, S. Movva, and S. Graves, Submitted April, 2009: Evaluation of Data Reduction Algorithms for Real-time assimilation and analysis. *Weather Analysis and Forecasting.*

Mining as a Service:

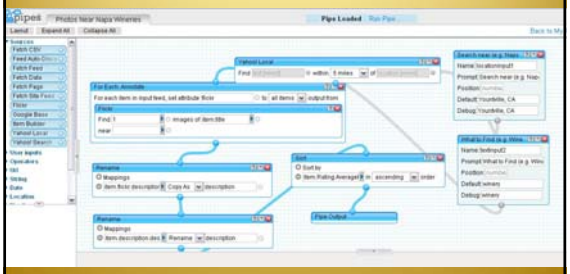
Data mining in the new paradigm of distributed sources and data

CyberInfrastructure, SOA, eScience

Conducting Science using Internet as the Primary Computer

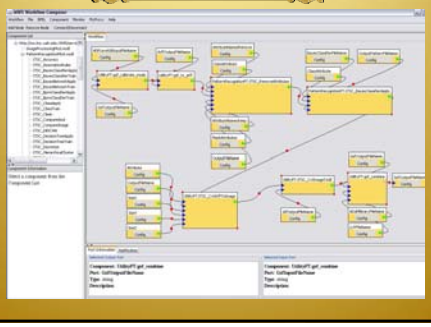


Mash-ups Example: Yahoo Pipes



The use of mash-ups to process and filter information is common in the private industry

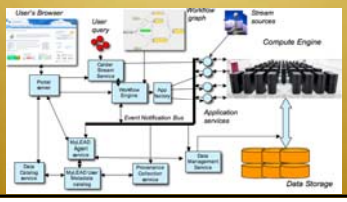
Data Mining in the 'new' Distributed Data/Services Paradigm



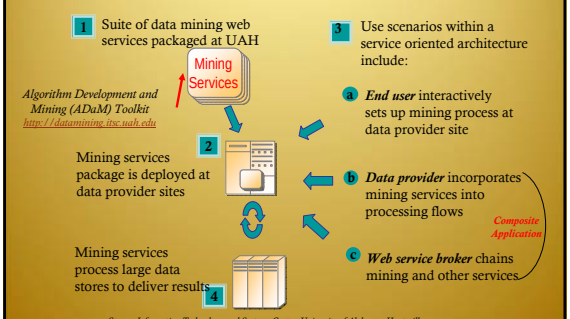
Linked Environments for Atmospheric Discovery (LEAD)

An *integrated framework* for identifying, accessing, preparing, assimilating, predicting, managing, analyzing, ***mining***, and visualizing meteorological data, *independent of format and physical location*

- Utilizes Teragrid supercomputing resources

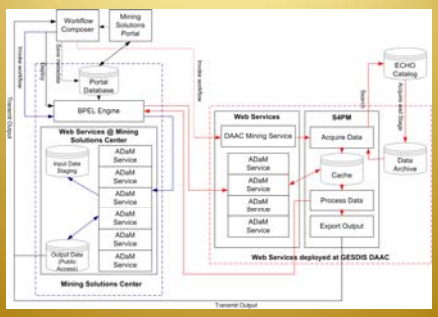


Deployable Suite of Mining Services for Online Science Data Repositories



Source: Information Technology and Systems Center, University of Alabama–Huntsville

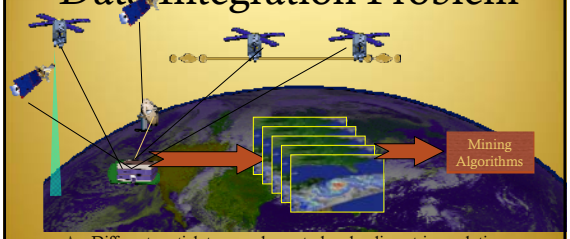
Prototype Implementation



Opportunities

Solution to Data Integration, Social Collaboration and Semantics!

Data Integration Problem

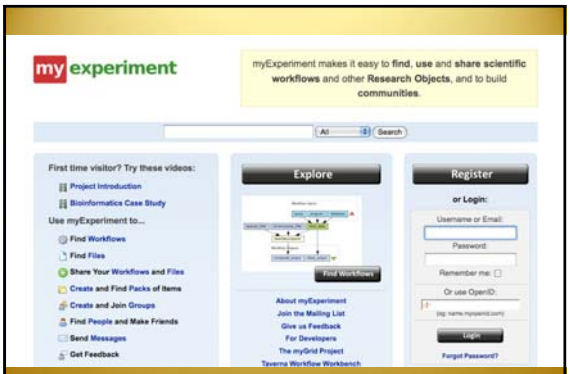


- Different spatial, temporal, spectral and radiometric resolutions
- Different formats, calibration needs
- Huge problem: Integrating data from multiple sensors at a large scale!

SOA can provide provide a solution?

Social collaboration and Mining

- SOA model of chaining services to create analysis workflows provides the research community unprecedented opportunity to ***collaborate***:
  - sharing their workflows with one another
  - ***reproducing*** and analyzing research results
  - leveraging colleagues' expertise to expedite the process of scientific knowledge discovery.
- In many cases, the output of one workflow can be an input to others, leading to chained workflows with components shared by two or more researchers.



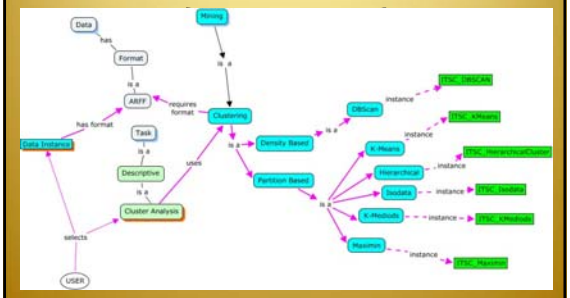
Social collaboration has already begun!

Mining 3.0: Using Semantics via Ontology

- Ontologies can be used for data, service and workflow search.
  - Discover related useful resources via linked data
- Ontologies can be used to automate data integration process
- Ontologies can perform semantic checks while creating mining workflows
- Ontologies can also be used to perform logic based classifications
- Ontologies can be used to assist users explore the mining solutions space.
  - "When confronted with a new problem, data miners, even data mining experts, often do not explore the design space of DM processes thoroughly" [1]
  - KDD Cup 1998 Analysis - to select a subset of customers to whom to mail solicitations, in order to maximize profit (revenues minus the cost of mailing). [1]
  - The winners did not use more complicated mining algorithms. Rather, they used a different DM process, one that is known by specialists to be particularly effective for target marketing. [1]

[1] Bernstein, A., F. Provost, and S. Hill, 2005: Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. *Knowledge and Data Engineering, IEEE Transactions on, 17, 503-518.*

Example!



Contact Information

Rahul Ramachandran
University of Alabama Huntsville
rramachandran@itsc.uah.edu
http://geoinformatics.itsc.uah.edu/

UAHuntsville