Data Mining for CEOP Data

October 9, 2002

Shin-ichi Sobue/NASDA

Purpose of Data Mining

“The extraction of hidden predictive information from large databases.”

Hidden!? Data

Example - Medical Research

July 18, 2002

Healthy Lifestyle May Help Cut

Alzheimer’s Risk, Scientists Find

Study of Alzheimer’s - 1,449 people for 21 years.

High cholesterol and high blood pressure raise the risk of developing Alzheimer’s more than carrying ApoE-e4, a gene variation considered the most important genetic risk factor for the disease.

High cholesterol and high blood pressure!

Business Example - Beer and Diapers

On Thursdays

Diapers --> Beer

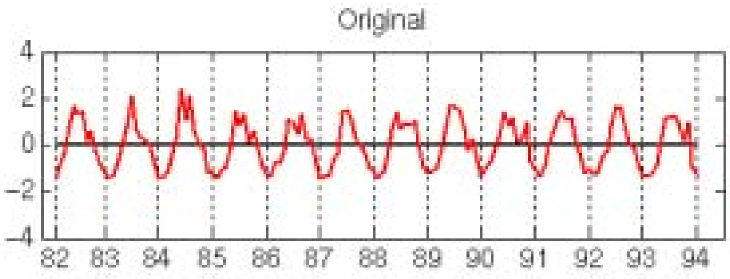
1. Displays of Beer and Diapers close together (near the registers).

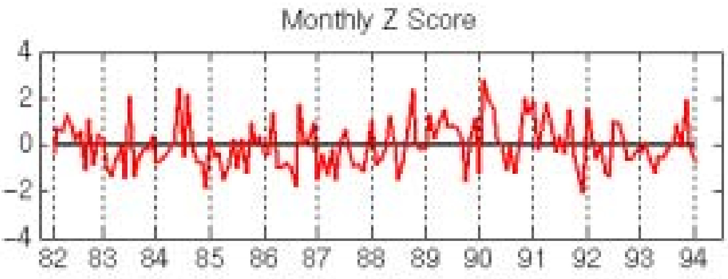
2. ?

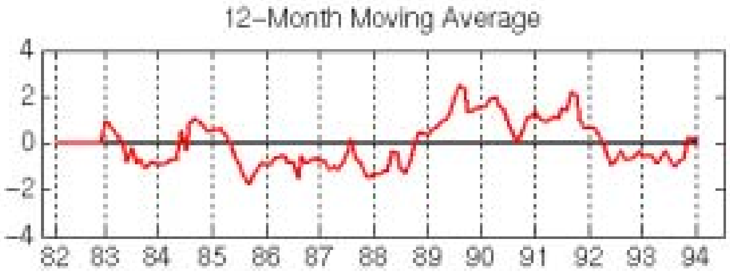
Preprocessing: Removing Seasonality

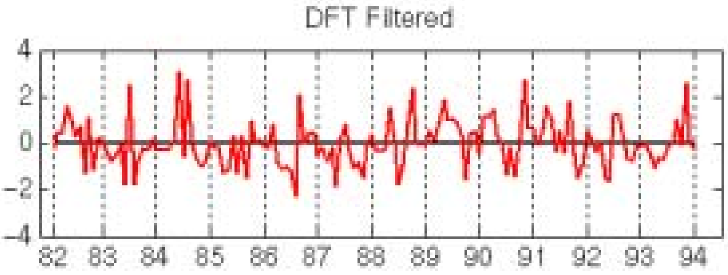
- Remove seasonality to see events (anomalies) of interest.
  - Monthly Z Score
    - Subtract monthly mean and divide by monthly standard deviation
  - 12 month moving average
  - Discrete Fourier Transform
  - Singular Value Decomposition

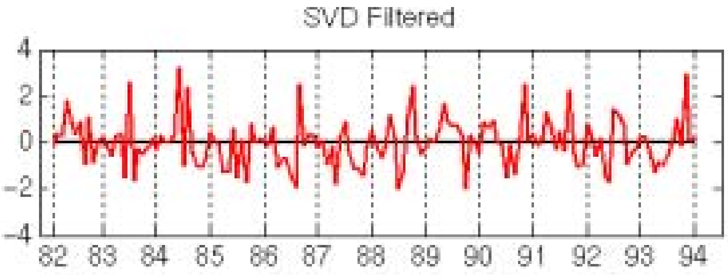
Preprocessing: Removing Seasonality


Original


Monthly Z Score

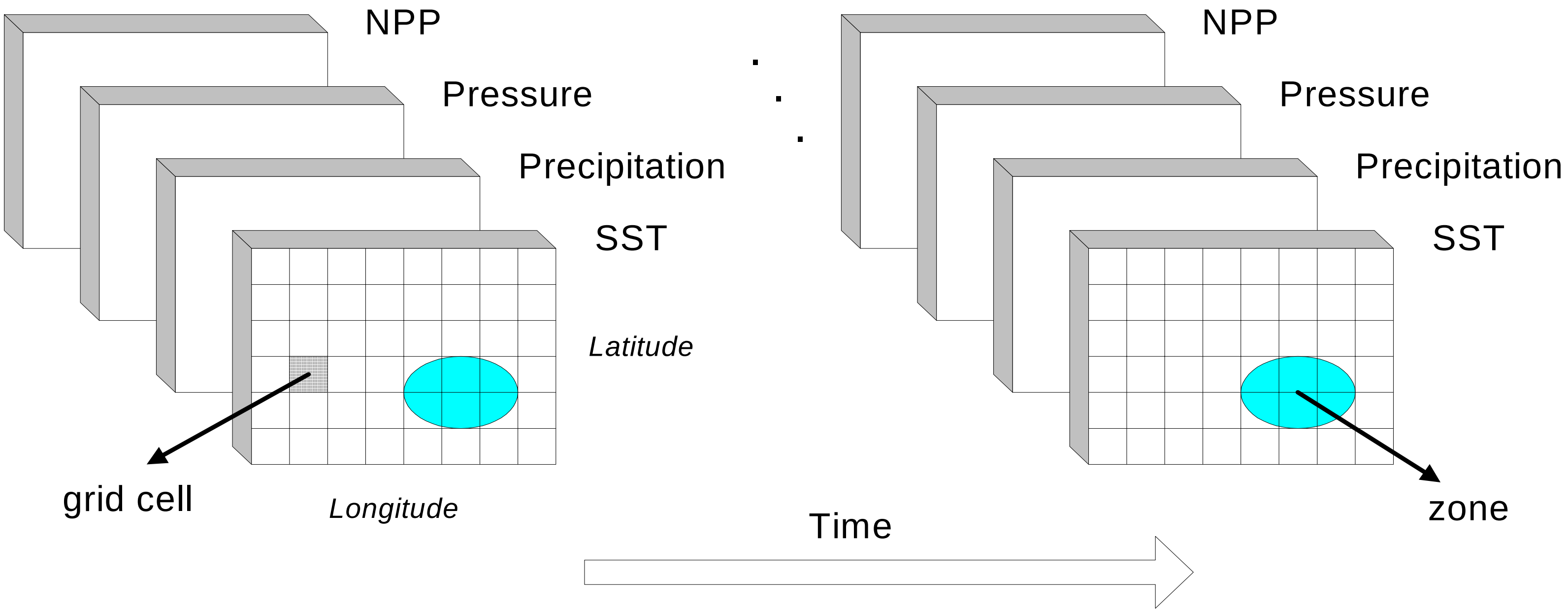

12−Month Moving Average


DFT Filtered


SVD Filtered

Time Series Data

- Global snapshots of values for a number of variables on land surfaces or water.
- Monthly values over a range of 10 to 50 years.
- Gridded values (NPP 0.5° x 0.5°, SST 1° x 1°)



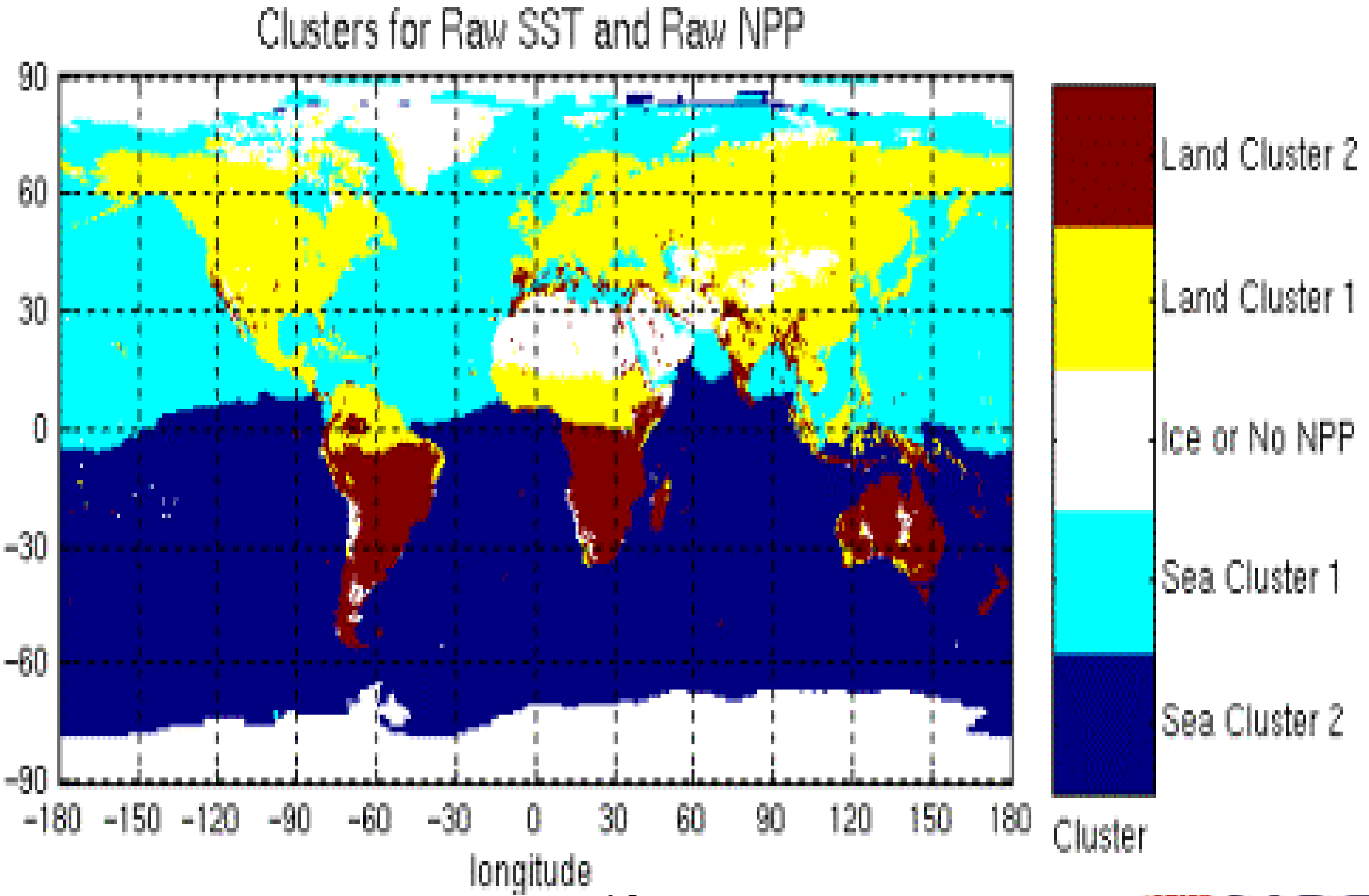
Data Clustering

- Cluster Formation
  - Find regions of the land or ocean which have similar time series behavior.
- Teleconnections
  - Teleconnections - time series behavior is (very) similar over widely separated points on the Earth.
- Interested in relationships between regions, not “points.”
- For ocean, clustering based on SST (Sea Surface Temperature), SLP (Sea Level Pressure), etc.
- For land, clustering based on NPP or other variables, e.g., precipitation, temperature.

K-means Clustering

- The K-means algorithm
  - “K” is number of (nonoverlapping) clusters
  - “means” - center of the cluster is the mean or median of the “nearness”of the points in its cluster, where “nearness” is defined by a similarity function (Pearson’s correlation coefficient).

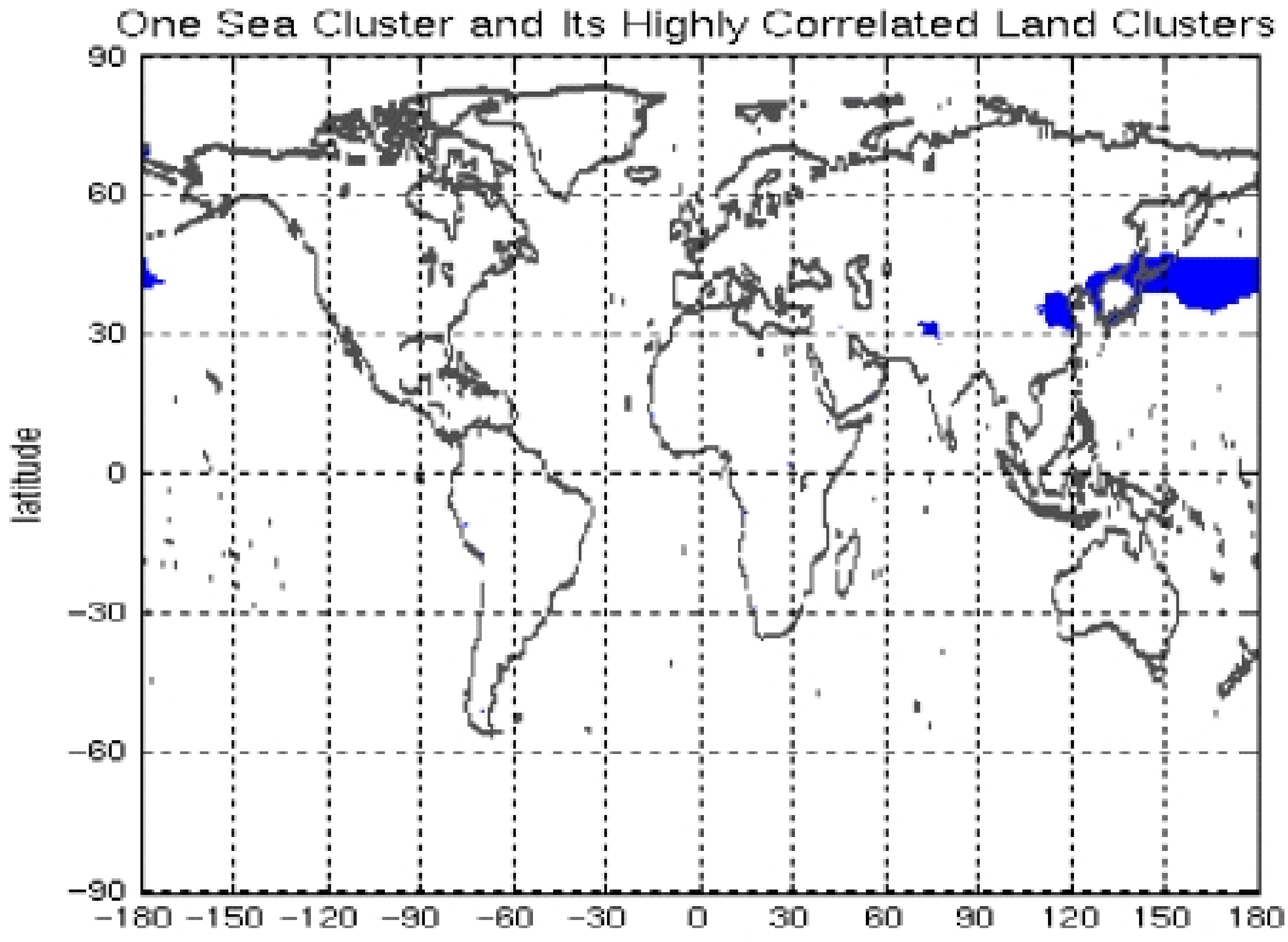
K-Means Clustering of Raw NPP and Raw SST



Discovering new Teleconnections

- Find land and sea clusters that are highly correlated, to identify new potential teleconnection patterns.
  - Produced 100 clusters for the land (NPP) and 100 clusters for the sea (SST).
  - Calculate correlations between sea and land clusters.
  - Next page shows diagram of sea cluster 19 and land clusters 56 and 58 (correlations of 0.56 and 0.50).
- Sea cluster 19 is highly correlated (-0.77) with the Pacific Decadal Oscillation

One Sea Cluster and Associated Land Clusters



Discovery of Ocean Climate Indices

- 1. SOI (Southern Oscillation Index)
- 2. NAO (North Atlantic Oscillation)
- 3. AO (Artic Oscillation)
- 4. PDO (Pacific Decadel Oscillation)
- 5. QBO (Quasi-Biennial Oscillation Index)
- 6 .CTI (Cold Tongue Index)
- 7. WP (Western Pacific)
- 8. ANOM12 (Normalized version of NINO12)
- 9. ANOM3 (Normalized version of ANOM3)
- 10. ANOM4 (Normalized version of NINO4)
- 11. ANOM34 (Normalized version NINO34)
- (Note: 1, 6, and 8-11 are El Nino related indices)

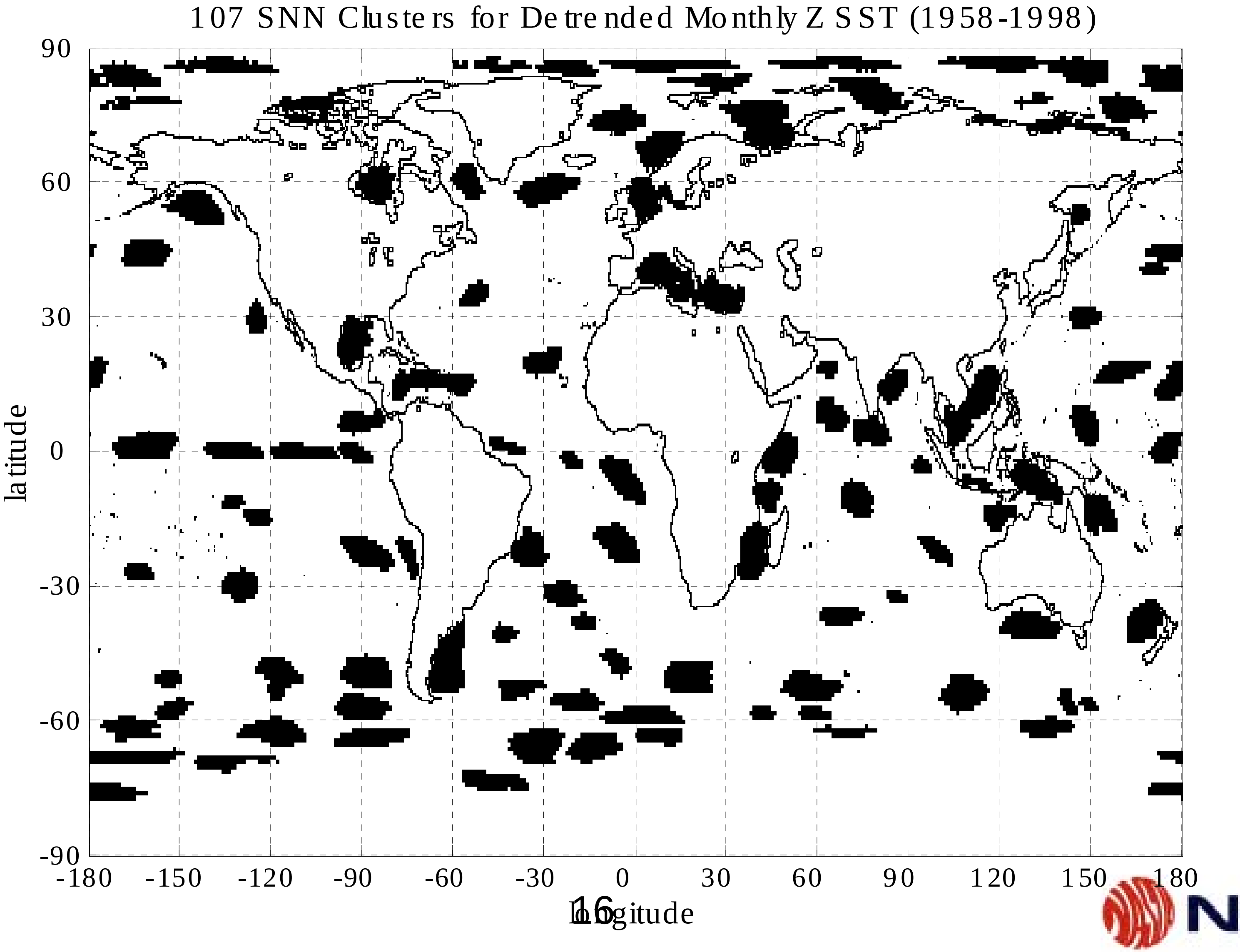
Discovery of Ocean Climate Indices

- **SST clusters are potential Ocean Climate Indices (OCIs).**
- **Determine if the clusters match known OCIs.**
- **Evaluate the influence of clusters (potential OCIs) on land points.**
- **For clusters that don’t match know OCIs, consider them to be potential OCIs and conduct further analysis to see if they are useful and interesting.**

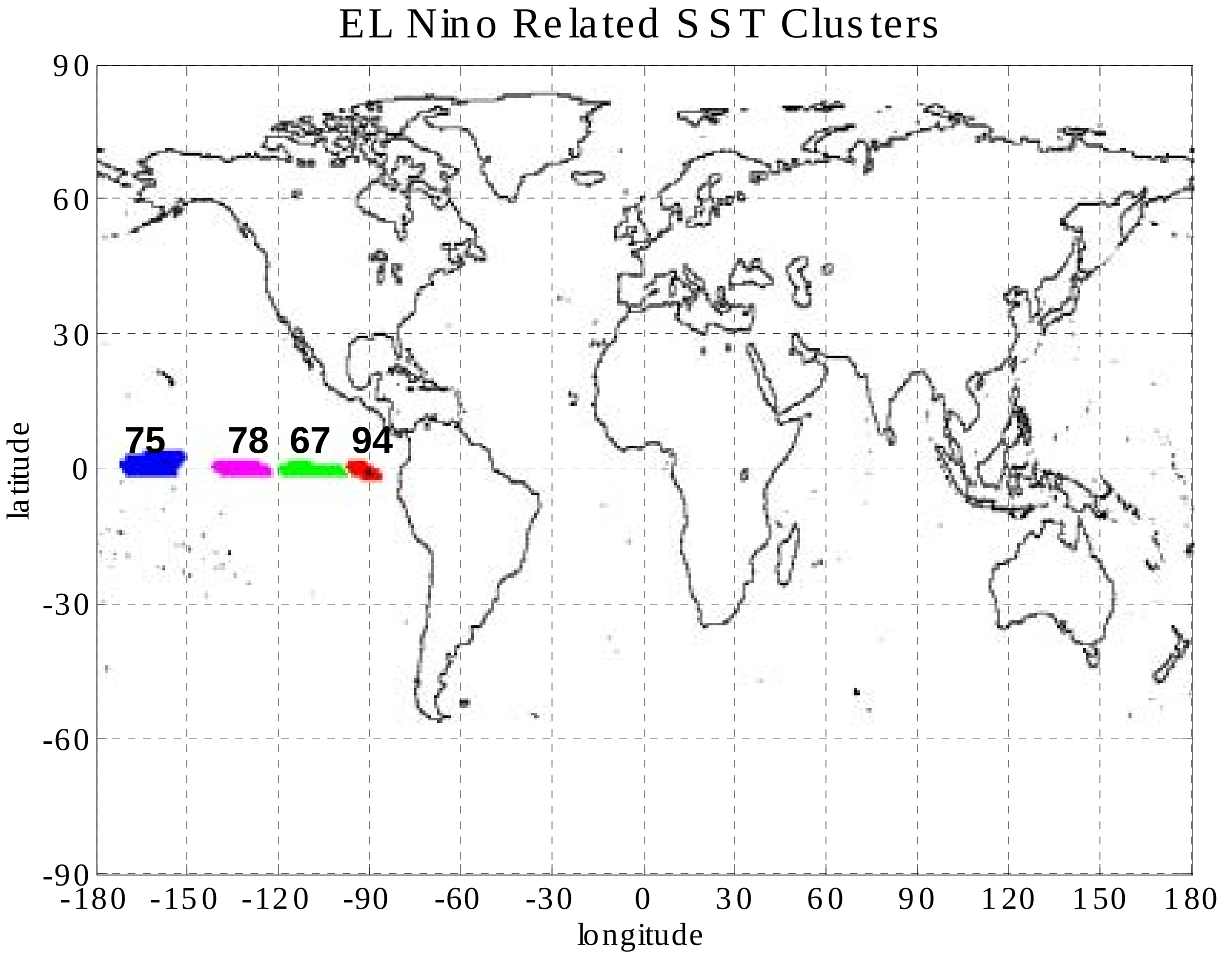
Shared Nearest Neighbor (SNN) Clustering

- **Find the nearest neighbors of each data point.**
  - **In this case data points are time series.**
  - **Examine the similarity between pairs of points in terms of how many nearest neighbors two points share.**
- **Eliminate noise, which are points with low density.**
- **Build clusters around the core points, which are points with high density.**

SST Clusters



SST Clusters that Correspond to El Nino Climate Indices



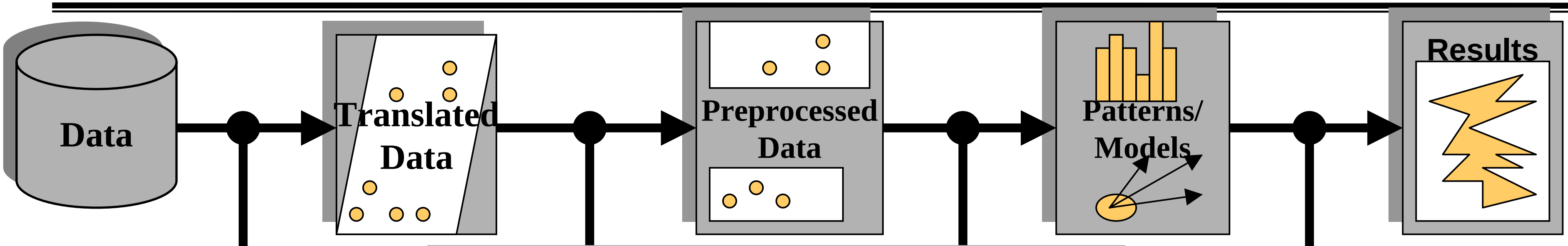
Niño Region	Range Longitude	Range Latitude
1+2 (94)	90°W-80°W	10°S-0°
3 (67)	150°W-90°W	5°S-5°N
3.4 (78)	170°W-120°W	5°S-5°N
4 (75)	160°E-150°W	5°S-5°N

El Nino Regions Defined by Earth Scientists

SNN clusters of SST that are highly correlated with El Nino indices, ~ 0.93 correlation.

ADaM

Algorithm Development and Mining System



Processing

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...

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...

ADaM - Input

Data Readers and Writers

- Binary
- GIF
- HDF
- HDF-EOS
- TIFF
- ASCII

ADaM - Preprocessing

- Grid Operations
- Subsetting
- Image Processing

Collage
Crop
Dilate
Equalize
Erode
Gabor
Filter

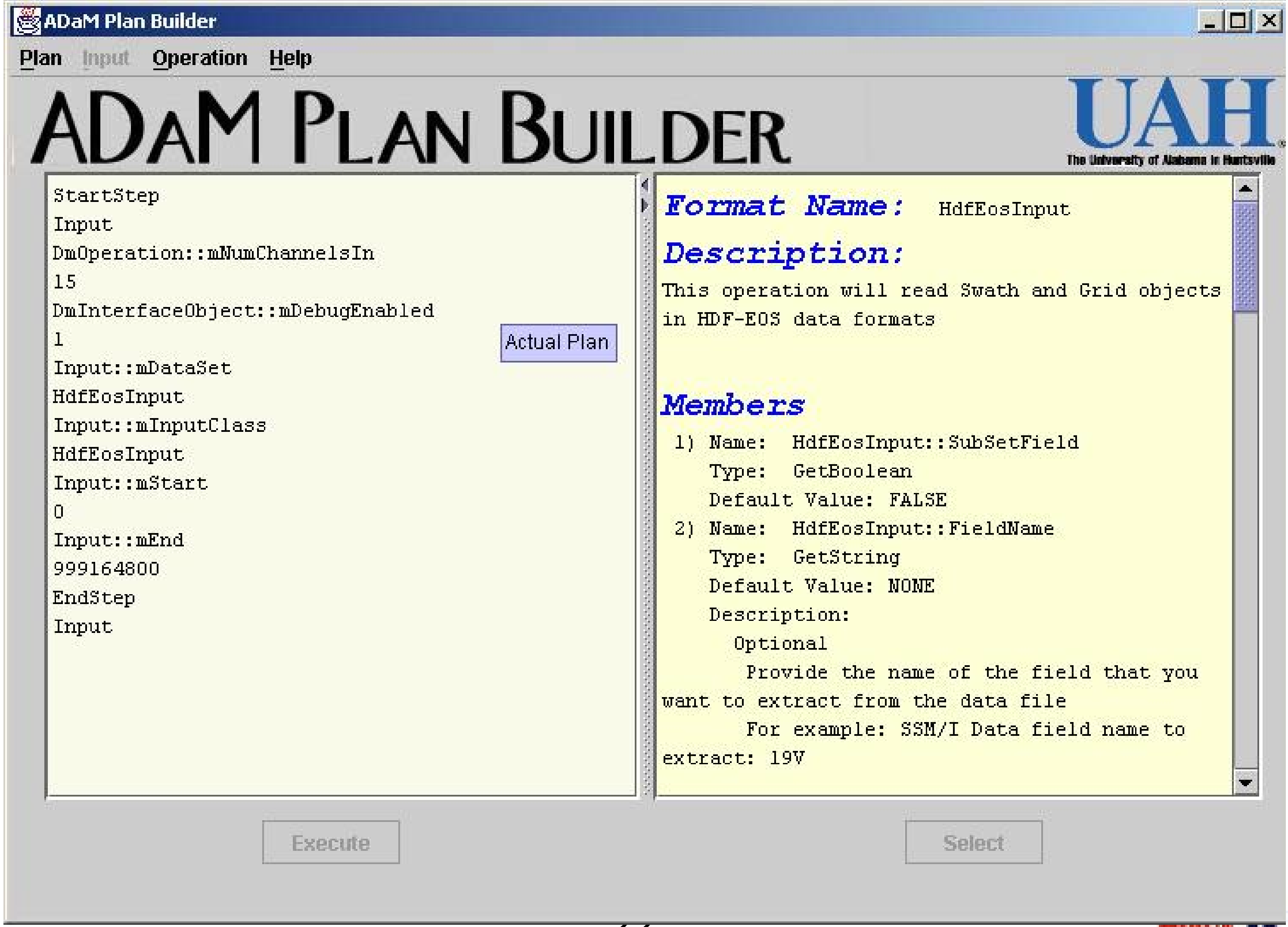
Invert
Overlay
Pulse Coupled Neural Network
Quantize
Rotate
Spatial Filters
Statistics

ADaM - Analysis

1. Genetic Algorithms
2. Pattern Recognition

- Bayes Classifier
- Isodata clustering operation
- K Means
- Max/Min Operation
- Multiple Prototype Minimum Distance Classifier
- Decision Tree Classifier
- Oblique Decision Tree
- Recursively Splitting Neural Network
- Minimum Distance Classifier

ADaM - Plan Builder Client



```
StartStep
Input
DmOperation::mNumChannelsIn
15
DmInterfaceObject::mDebugEnabled
1
Input::mDataSet
HdfEosInput
Input::mInputClass
HdfEosInput
Input::mStart
0
Input::mEnd
999164800
EndStep
Input
```

Format Name: HdfEosInput

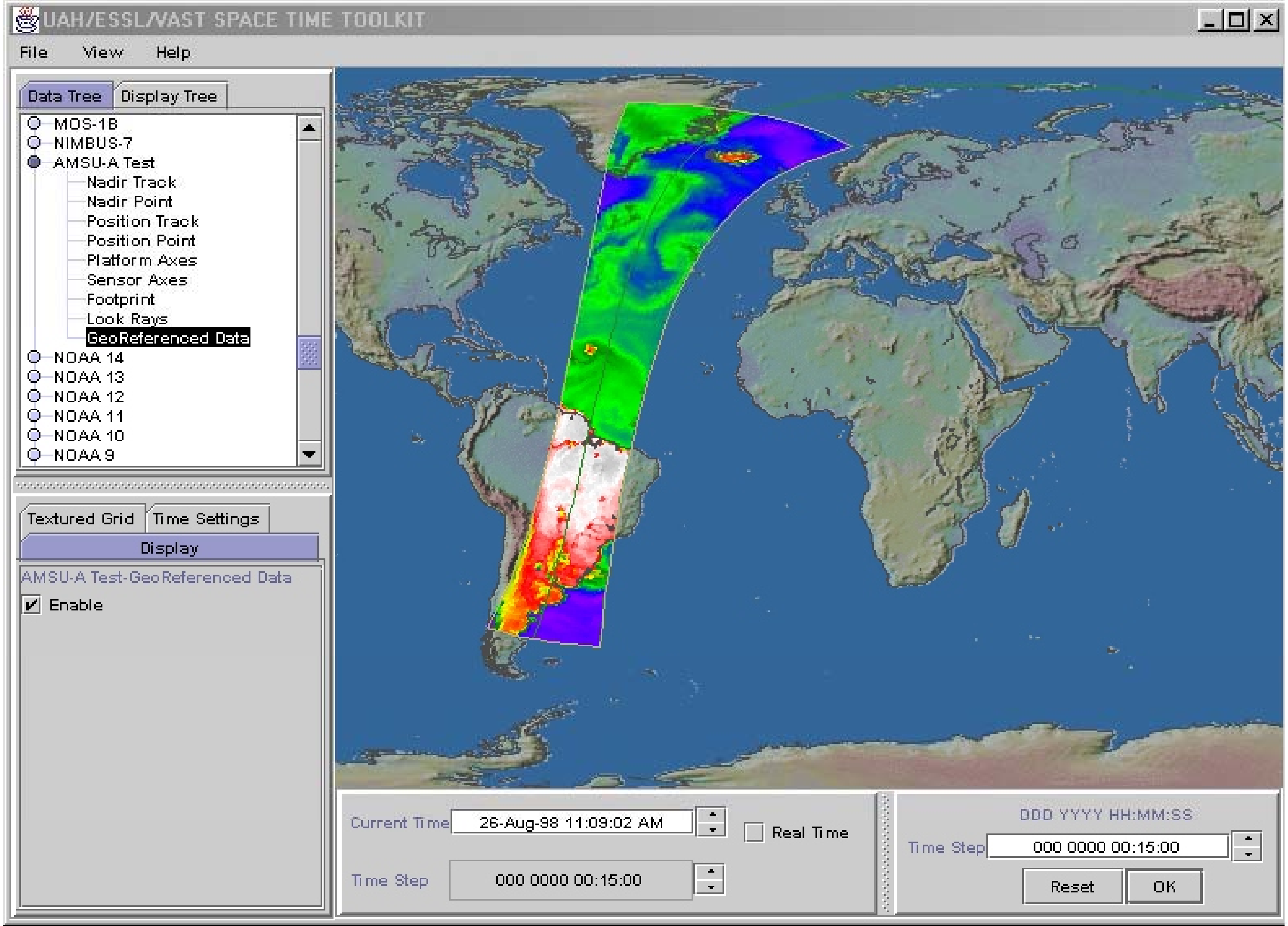
Description:

This operation will read Swath and Grid objects in HDF-EOS data formats

Members

```
 1) Name:  HdfEosInput::SubSetField
    Type:  GetBoolean
    Default Value: FALSE
 2) Name:  HdfEosInput::FieldName
    Type:  GetString
    Default Value: NONE
    Description:
      Optional
       Provide the name of the field that you
want to extract from the data file
       For example: SSM/I Data field name to
extract: 19V
```

Space Time Toolkit



UAH/ESSL/VAST SPACE TIME TOOLKIT

- ○ MOS-1B
- ○ NIMBUS-7
- ● AMSU-A Test
  - Nadir Track
  - Nadir Point
  - Position Track
  - Position Point
  - Platform Axes
  - Sensor Axes
  - Footprint
  - Look Rays
  - GeoReferenced Data
- ○ NOAA 14
- ○ NOAA 13
- ○ NOAA 12
- ○ NOAA 11
- ○ NOAA 10
- ○ NOAA 9

AMSU-A Test-GeoReferenced Data

☑ Enable

Current Time: 26-Aug-98 11:09:02 AM

□ Real Time

Time Step: 000 0000 00:15:00

DDD YYYY HH:MM:SS

Time Step: 000 0000 00:15:00

Space Time Toolkit

Can visually integrate virtually any geographic data set, regardless of differences in their spatial and temporal representation.

Low-level sensor data

Gridded products

Vector based GIS data

3D model output, etc.

“Types” of Data Mining

- **Relationship Mining** - mining for various types of relationships includes most types of data mining.
  - **Cause or Indicator Relationships** - Find the actual causes of an “event” (e.g. determine if land clearing has disturbed hydrologic runoff and/or flood frequency).
  - **Effect Relationships** - The analysis of events having regional climate impact (e.g., volcanic eruptions, desert flash floods).
  - **Correlation Relationships** - Study of earthquakes (e.g. the space-time correlations involving triggered slips, foreshocks and aftershocks).
  - **Linkage Relationships** - An event at one time and place is related to an event at another time and place (e.g., El Nino and drought in Indonesia). Feature Transformation
  - **Prediction Relationships** - Predict storm tracts and changes in intensity.

“Types” of Data Mining (continued)

- **Exploratory Pattern Mining** - Look for unexpected spatial/temporal patterns long-term data that cover a longer period (e.g. three to five years),
- **Complex Process Characterization** - Develop predictive statistical models that can be applied to areas such as seismic activity or the spreading of fire. Physical/mechanistic models that can be applied to areas such as earthquake fault modeling, interactions between crust/mantle and stress transfer.

WTF-CEOP Activities - next 6 months (1/2)

1. WTF-CEOP website: list information on research in data mining of Earth Science data.
2. WTF-CEOP website: list examples of data mining.

Example of data mining query:

- Investigation: It has been shown that warming in the equatorial Pacific Ocean in the El Nino area affects the strength of winds at the 200 mb level in the tropical Atlantic. During drought years in West Africa, winds increase from west to east at the 200 mb level, in the Central Atlantic.
- Query: The user wants to find the time periods where the SST anomaly in the El Nino region, is less than -1°C and winds, at 200 mb, in the Central Atlantic, are greater than   -5.00 m/s. For those time periods the user then wants to study the precipitation and vegetation data, for the western Sahel.

WTF-CEOP Activities - next 6 months (2/2)

3. Survey CEOP scientists for what type of data mining would be useful to them. Add these queries to item 2 above.
4. WTF-CEOP website: develop an on-line library of examples of visualization techniques for 3D/4D data and data mining.
5. Prepare a CEOP database testbed (has the same software environment as CEOP). This database testbed will be used to adapt and test CEOS data mining algorithms before they are installed on CEOP.
6. Identify “candidate” CEOP algorithms - algorithms that have been developed by (NASA, ESA, etc.) scientists that can be used by CEOP. NASDA will install and test the algorithms on the CEOP database testbed.
7. Survey COTS Data Mining software.