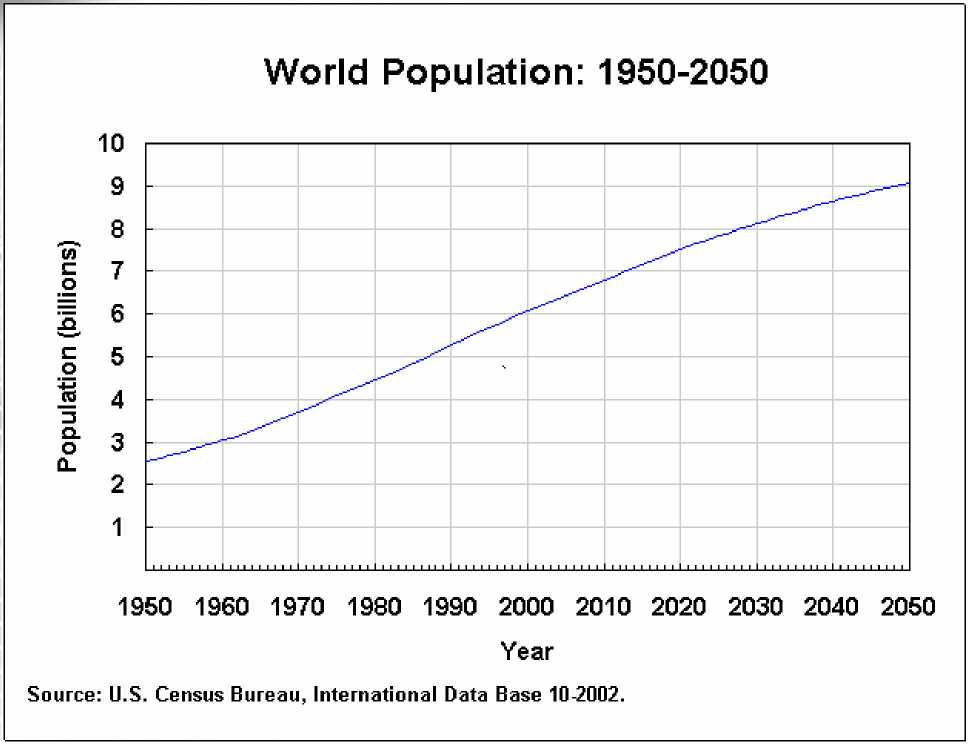
Image Information Mining and Semantic Webs for Knowledge Discovery

Roger L King

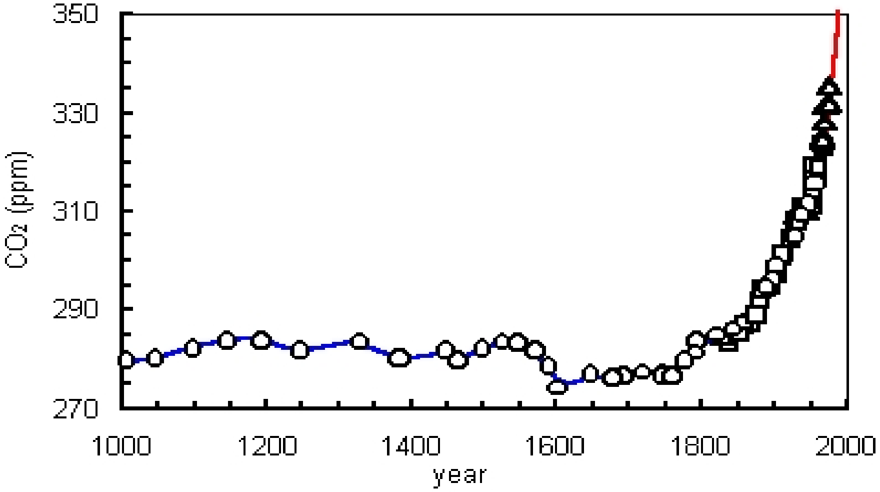
William L. Giles Distinguished Professor

Earth’s Population Continues to Grow


World Population: 1950-2050

Source: U.S. Census Bureau, International Data Base 10-2002.

- 6 billion – 1999
- 8 billion - 2025



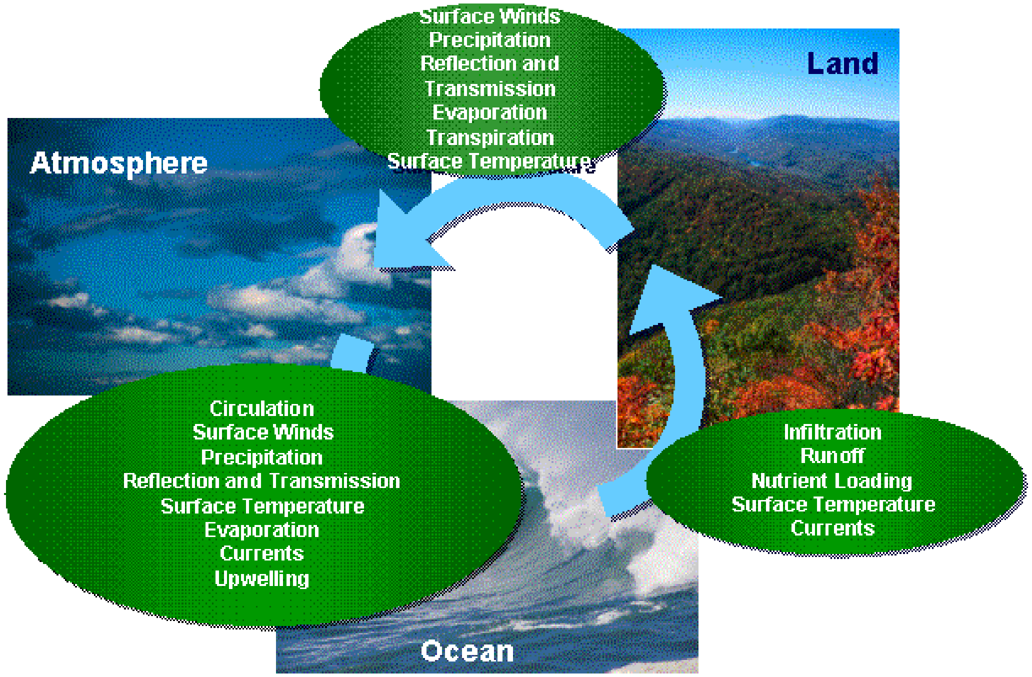
- However, there are consequences.

Fresh Water – Finite Resource

- View of Earth from space as a “Blue Marble” gives mankind a false sense of security. Only a small fraction of the planet’s water wealth can be tapped and that share must sustain life for mankind plus numerous other species.
- Nearly 1 out of every 3 people in the developing world - some 1.2 billion people in all – do not have access to a safe and reliable supply for their daily needs.

Freshwater and the Food Supply

- Evaporation - 500,000 cubic kilometers
- An equal amount falls back to earth as rain, sleet, or snow, but in a different distribution.
- However, all of this precipitation cannot be captured.
- The net captured for use by mankind is ~14,000 cubic kilometers. This serves as the planet's stable freshwater supply.
- Water available per person - 2,222 cubic meters – 2003; 1750 cubic meters - 2025. (low meat diet for one person for a year - 1100 cubic meters)



All of Earth’s Natural Resources are Finite

- The same can be said of Earth’s other resources (finite and not evenly distributed)
  - Energy
  - Land resources
    - Pasture
    - Cropland
    - Forest products
  - Atmosphere
  - …

- However, the world’s expanding population requires the use of these resources for a basic quality of life.
- Therefore, there must be a vision to discover the knowledge to ensure sound (fiscal & environmental) management of these resources.

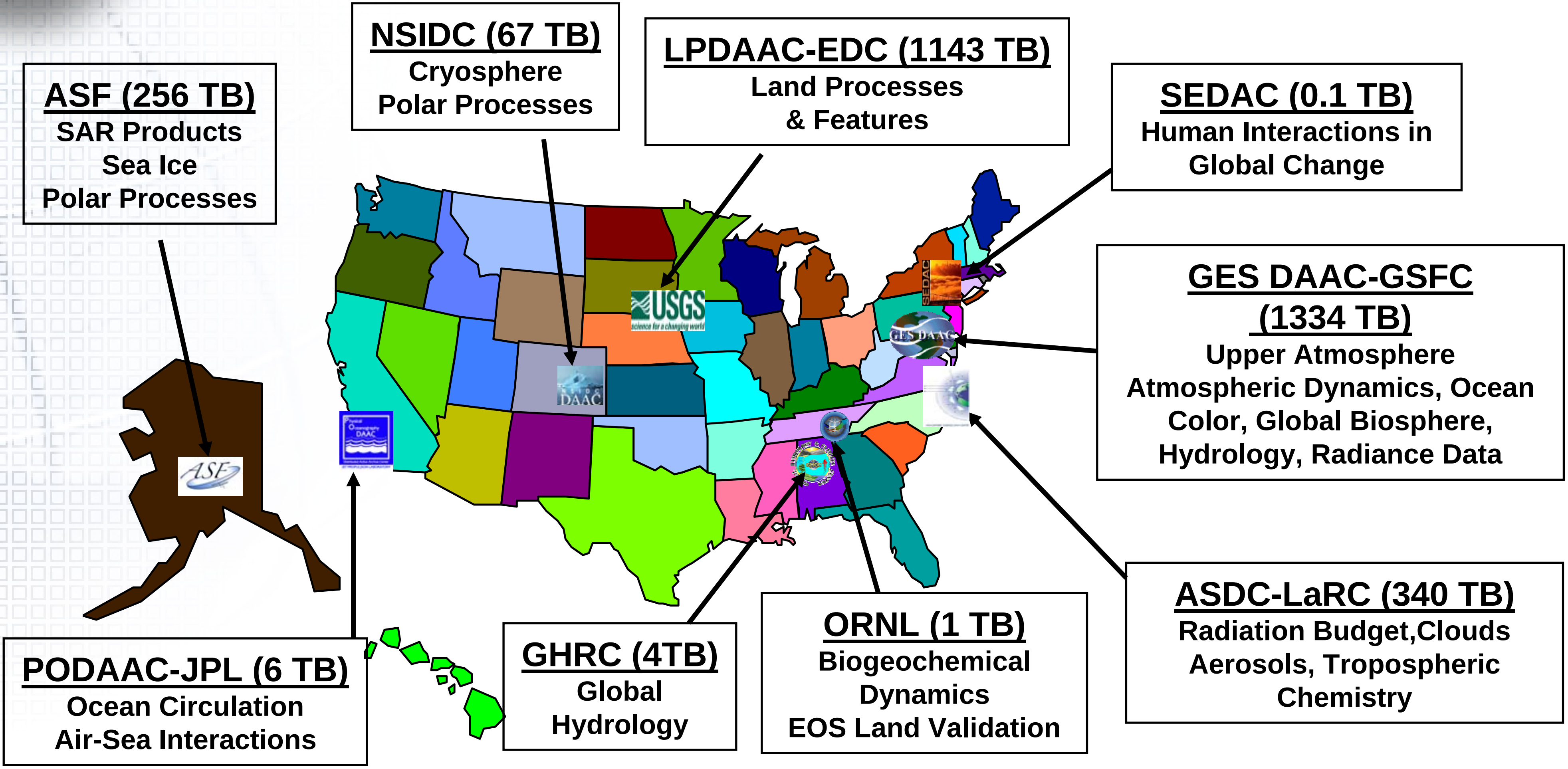
Earth Observation from Multiple Vantage Points

- ❑ A key to this vision is Earth observations.
- ❑ Multiple vantage points for Earth observation leads to multiple archives of imagery and other datasets.
- ❑ Earth observation community needs to begin to address the need for accessibility to archived data sets and to the development of tools to mine the world’s archives of measured data.
- ❑ Results from this activity can lead to discovery of new knowledge and understanding that will assist policy makers in meeting the needs of the Earth’s burgeoning population.

Earth Observation Archive Holders in the USA

- ❑ Civil global observations within the United States fall primarily under the auspices of three agencies
  - ❖ National Aeronautics and Space Administration (NASA),
  - ❖ National Oceanic and Atmospheric Administration (NOAA)
  - ❖ United States Geological Survey (USGS).
- ❑ Petabytes of Earth observation archives held by these three civilian agencies

NASA’s Earth Science Discipline Focused DAACs



ASF (256 TB)
SAR Products
Sea Ice
Polar Processes

NSIDC (67 TB)
Cryosphere
Polar Processes

LPDAAC-EDC (1143 TB)
Land Processes
& Features

SEDAC (0.1 TB)
Human Interactions in
Global Change

GES DAAC-GSFC (1334 TB)
Upper Atmosphere
Atmospheric Dynamics, Ocean Color, Global Biosphere, Hydrology, Radiance Data

ASDC-LaRC (340 TB)
Radiation Budget,Clouds
Aerosols, Tropospheric
Chemistry

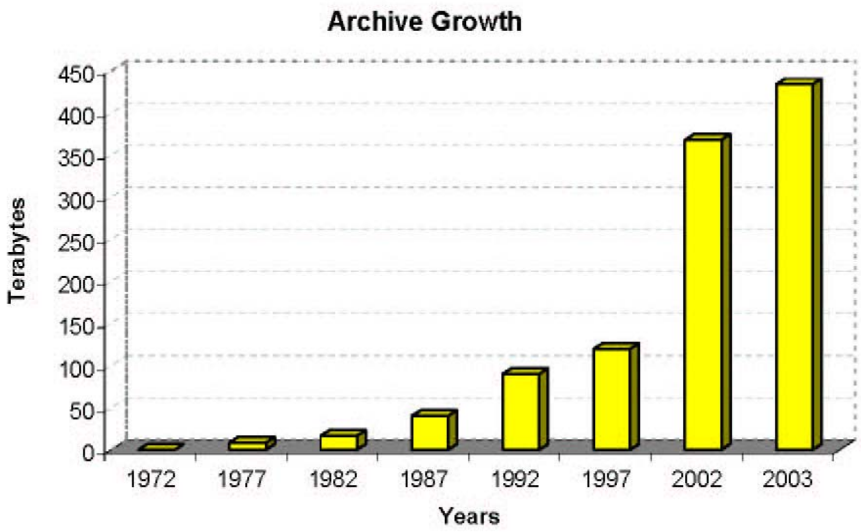
ORNL (1 TB)
Biogeochemical
Dynamics
EOS Land Validation

GHRC (4TB)
Global
Hydrology

PODAAC-JPL (6 TB)
Ocean Circulation
Air-Sea Interactions

NOAA and USGS Archives

- ❑ NOAA
  - ❖ National Climatic Data Center
  - ❖ National Geophysical Data Center
  - ❖ National Oceanographic Data Center
  - ❖ National Coastal Data Development Center
- ❑ USGS
  - ❖ Earth Resources Observation Systems (EROS) Data Center (EDC)



Archive Growth at EDC

31 years of Landsat 1-5 165 terabytes	4 years of Landsat 7 269 terabytes

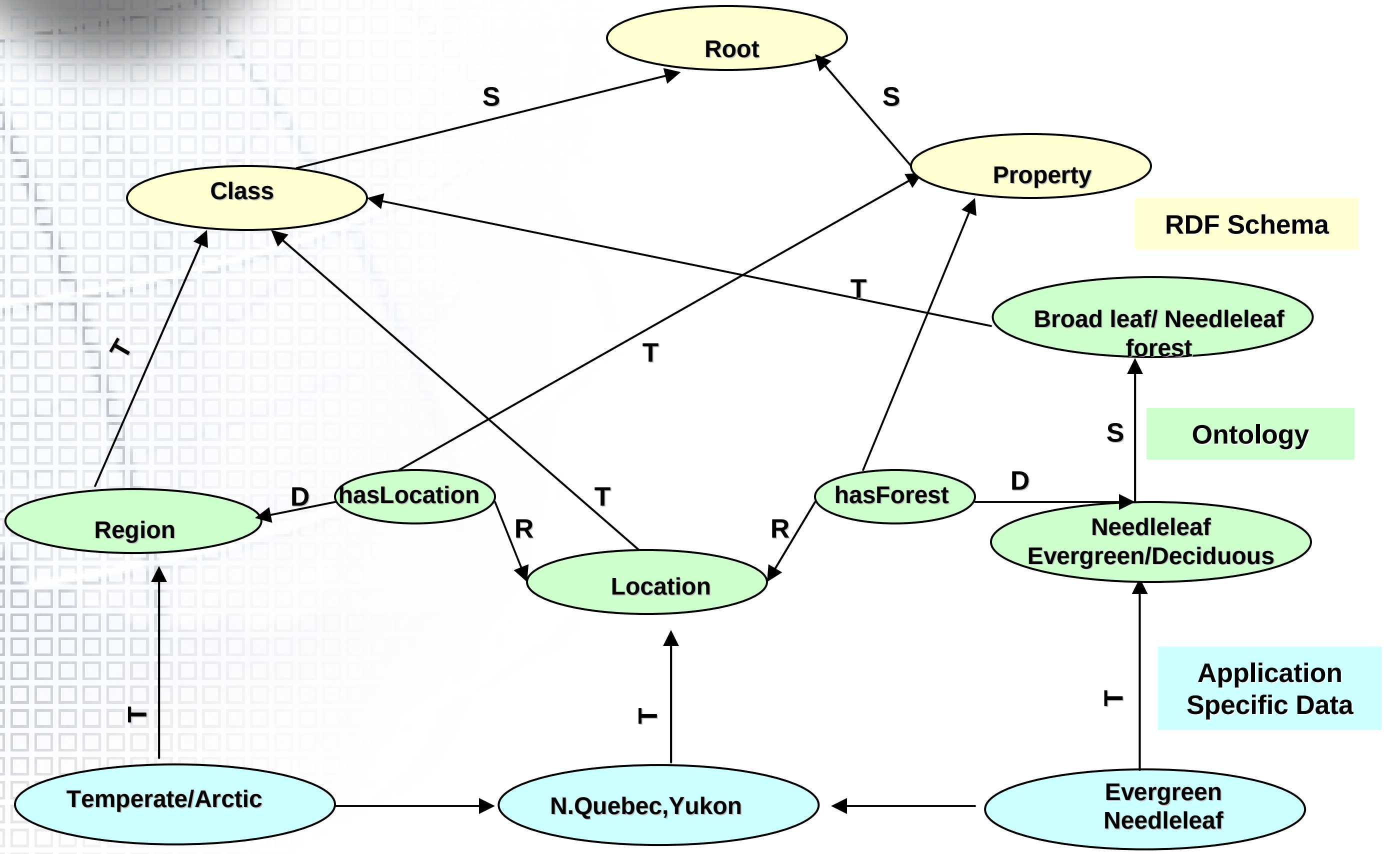
The Undiscovered Country

- ❑ Hamlet dreaded the *undiscovered country* that faces us all after death.
  - ❖ *Hamlet* Act III, Scene 1

- Chancellor Gorkon of the Klingon High Council saw the *undiscovered country* as a future where peace between the Klingon Empire and the United Federation of Planets reigned.
  – Star Trek VI: The Undiscovered Country

Image Archives: The Undiscovered Country

- For those studying Earth system science seeking solutions for the use of Earth’s natural resources, we too have an *undiscovered country*.
- Within our *undiscovered country,* hidden in the imagery and data archives of the global observation community, may lie important new knowledge about the Earth system.
- Hamlet and Gorkon voiced the fear we have in traveling into unknown lands, but from them we can learn that the reward that lies beyond may greatly surpass our present state.
- Therefore, a real need exists to develop the theory and applications of knowledge driven image information mining for exploring our *undiscovered country*.



Root
Class
Property
RDF Schema
Broad leaf/ Needleleaf forest
Ontology
hasLocation
hasForest
Region
Location
Needleleaf Evergreen/Deciduous
Application Specific Data
Temperate/Arctic
N.Quebec,Yukon
Evergreen Needleleaf
S
S
T
T
T
T
S
D
D
R
R
T
T
T

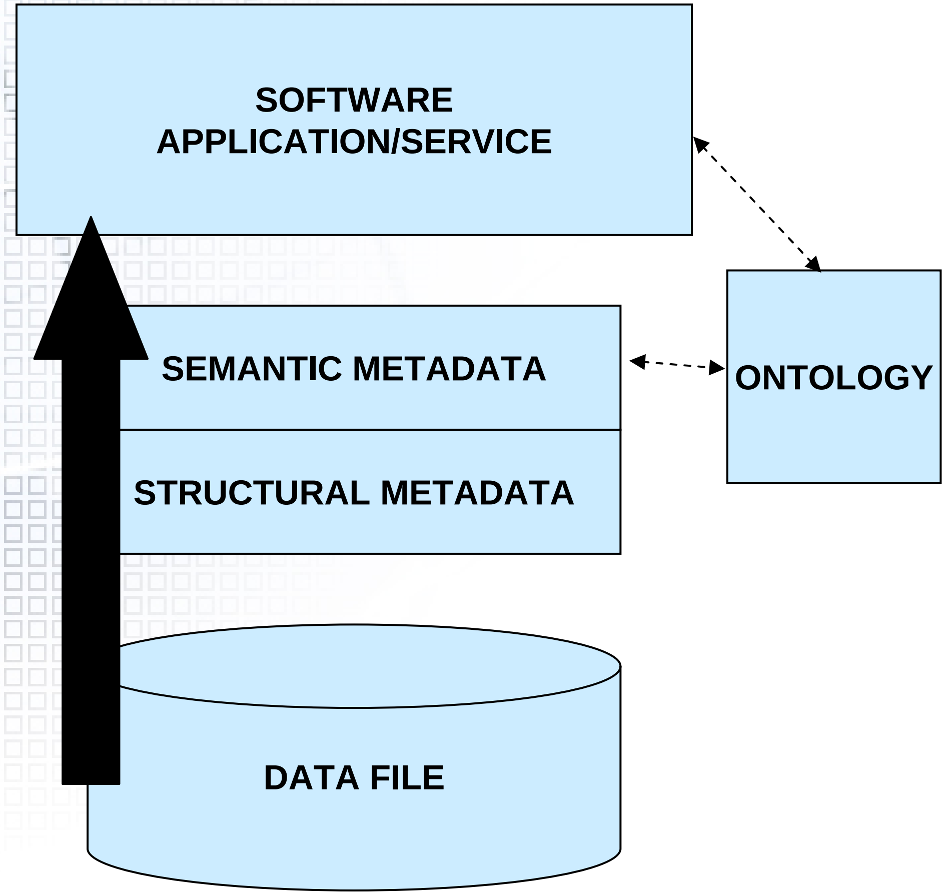
S:subclassOf
R:range
D:domain
T:instanceOf

Why ontologies ?

- •Share common understanding of domain
- •Reuse domain knowledge
- •Make domain assumptions explicit
- •Separate domain knowledge from the operational knowledge
- • Analyze domain knowledge

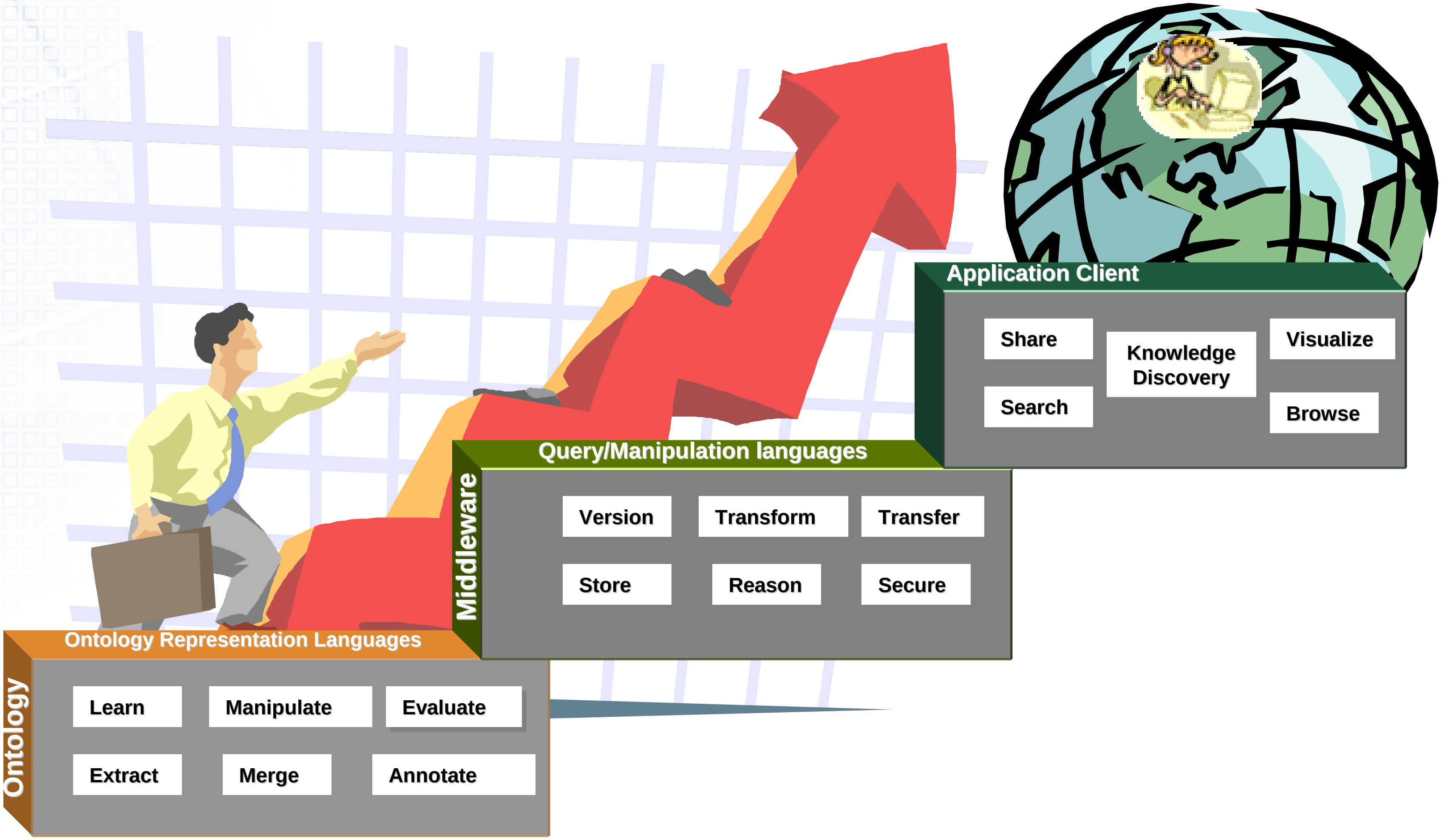
A Domain Ontology Perspective

Knowledge Mining in Earth Observation Data Archives



- ❑ Services require a rich set of metadata
  - ❖ Structural metadata: to provide full description of the physical parameters of the data file
  - ❖ Semantic metadata: to provide meaning of the data along with a context, to allow application to understand what it has read and how to use it
- ❑ Intelligence/Knowledge– to be able to make decisions via ontologies and a reasoning engine or via a machine learning algorithm or via heuristic algorithm

Ontology Driven Applications



Application Client

Share

Search

Knowledge Discovery

Visualize

Browse

Middleware

Query/Manipulation languages

Version

Transform

Transfer

Store

Reason

Secure

Ontology

Ontology Representation Languages

Learn

Manipulate

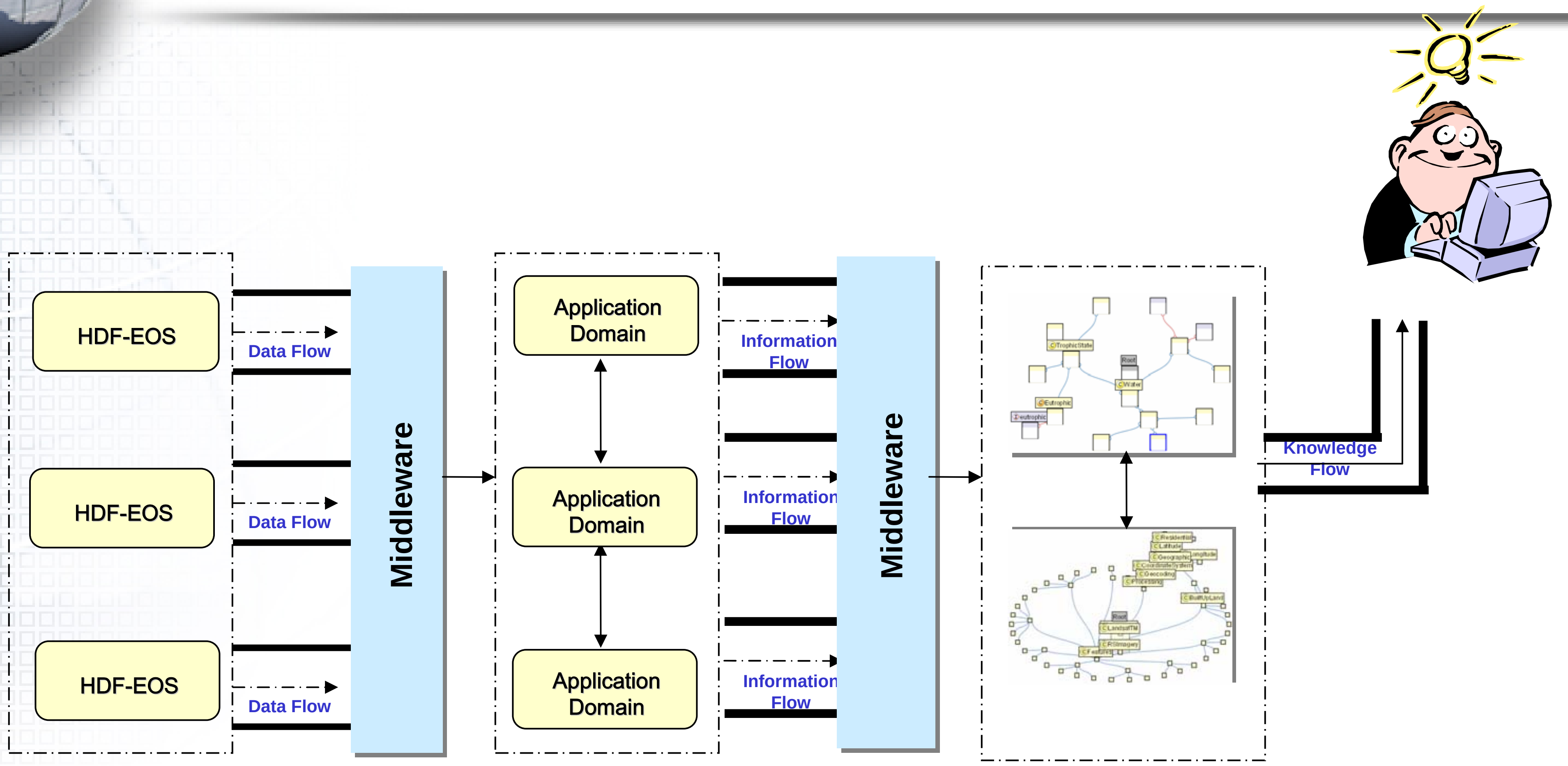
Evaluate

Extract

Merge

Annotate

Data Transformation



Distributed Active Archive Centers (DAAC's)

Resource discovery, metadata access, browse data pool

Distributed Data Analysis Centers (Research labs, Universities, etc)

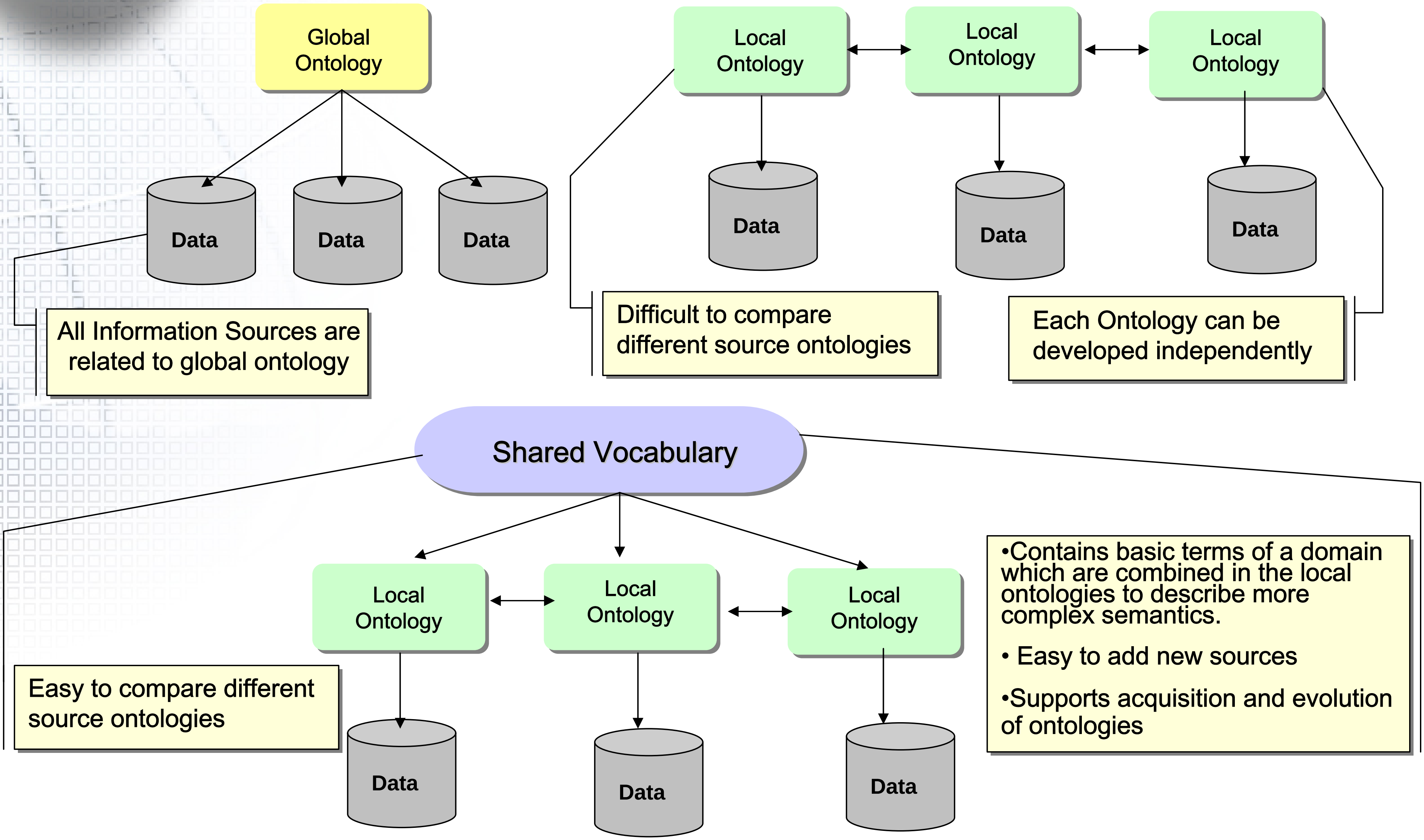
Resolve information heterogeneity (semantic,syntactic, format,etc)

Domain Specific knowledge building through ontological Modeling (OWL,DAML+OIL,etc)

Research Objectives

- Develop Middleware for Ontology Driven Brokering(MOB)
  - Translate metadata to semantic metadata
    - Enables identification of information and relevant knowledge (entities such as sensor type, geographic locations) and their relationships
    - Resource discovery, mediation and transformation
  - Ontology design, integration and deployment
    - Assert Inter-ontology relationships
    - Compute, integrate class hierarchy/consistency.
  - Provide tools for Image knowledge retrieval
    - Development of Application ontology (domain specific)
    - Image segmentation,primitive features, components extraction
    - Apply machine learning for feature classification
- Develop client side tools
  - Functionality to gather information at different levels of granularity, from the sub category to the specific data level

Ontology Integration-architectures



Global Ontology

Data

Data

Data

All Information Sources are related to global ontology

Local Ontology

Local Ontology

Local Ontology

Data

Data

Data

Difficult to compare different source ontologies

Each Ontology can be developed independently

Shared Vocabulary

Local Ontology

Local Ontology

Local Ontology

Data

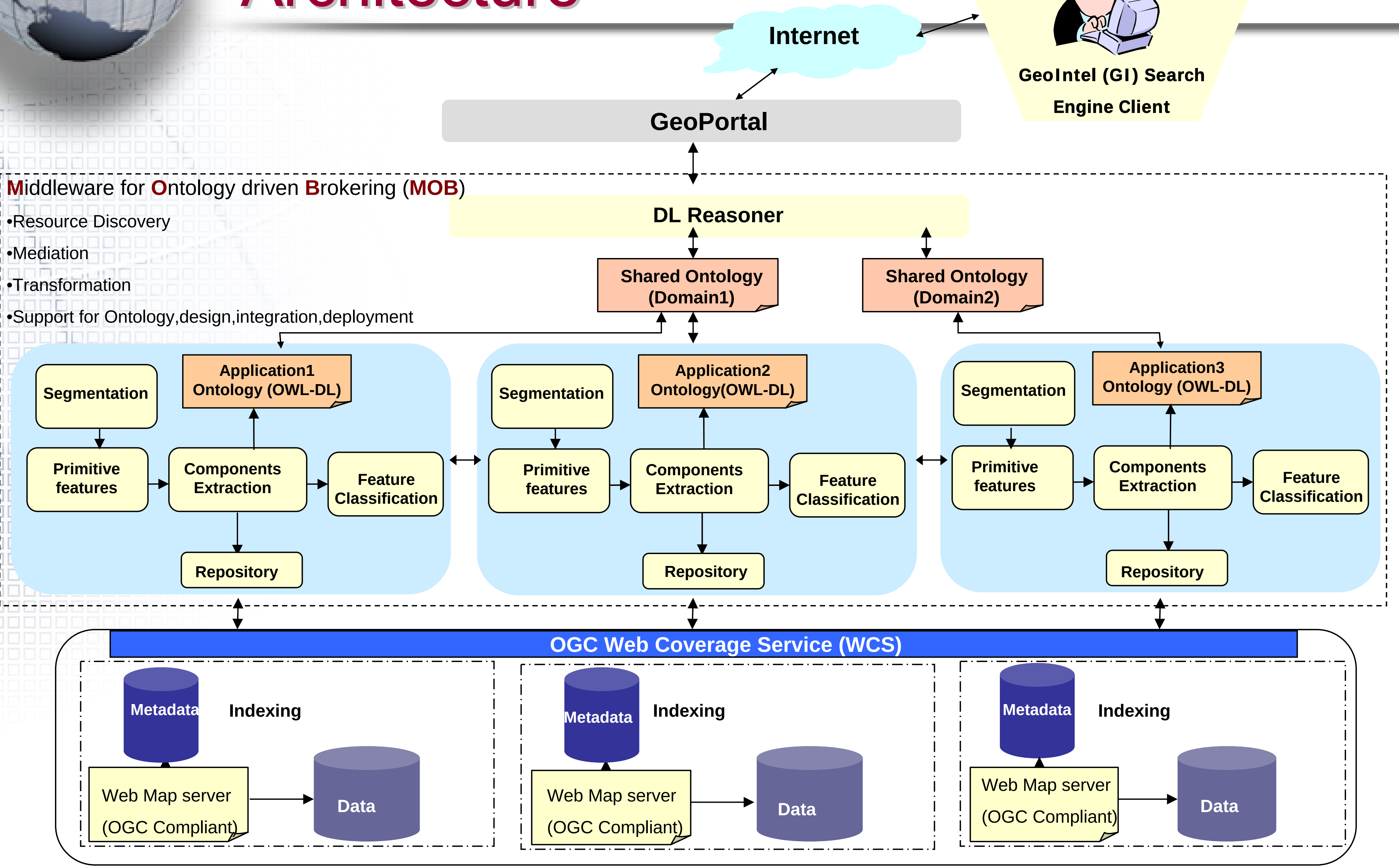
Data

Data

Easy to compare different source ontologies

- Contains basic terms of a domain which are combined in the local ontologies to describe more complex semantics.
- Easy to add new sources
- Supports acquisition and evolution of ontologies

Architecture



Internet

GeoIntel (GI) Search Engine Client

GeoPortal

Middleware for Ontology driven Brokering (MOB)

- Resource Discovery
- Mediation
- Transformation
- Support for Ontology,design,integration,deployment

DL Reasoner

Shared Ontology (Domain1)

Shared Ontology (Domain2)

Segmentation

Application1 Ontology (OWL-DL)

Primitive features

Components Extraction

Feature Classification

Repository

Segmentation

Application2 Ontology(OWL-DL)

Primitive features

Components Extraction

Feature Classification

Repository

Segmentation

Application3 Ontology (OWL-DL)

Primitive features

Components Extraction

Feature Classification

Repository

OGC Web Coverage Service (WCS)

Metadata

Indexing

Web Map server (OGC Compliant)

Data

Metadata

Indexing

Web Map server (OGC Compliant)

Data

Metadata

Indexing

Web Map server (OGC Compliant)

Data

Ontology Web language (OWL)

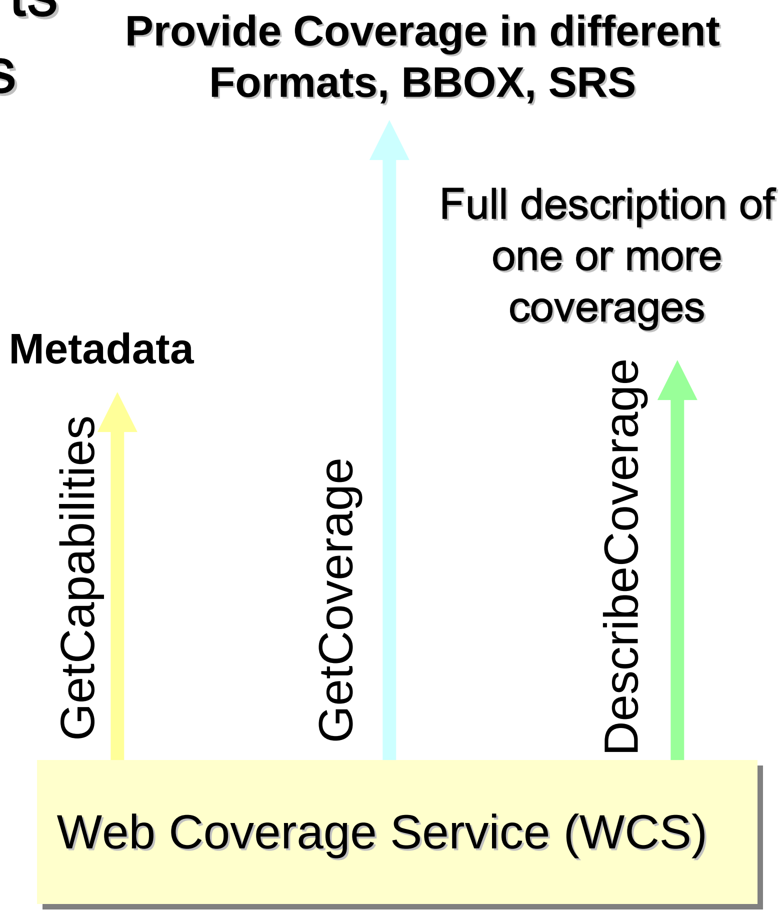
- ❑ "Language for defining structured, web-based ontology"- OGC definition.
  - ❖ Richer integration
  - ❖ Interoperability of data across application domains
- ❑ OWL applications
  - ❖ Web portals
  - ❖ Agents and services
  - ❖ Ubiquitous computing
  - ❖ Multimedia collections

Inside Ontology

- ❖ Classes+class hierarchy
- ❖ Instances
- ❖ Slots/values
- ❖ Inheritance
- ❖ Restrictions on slots (type, cardinality)
- ❖ Properties of slots
- ❖ Relations between classes

Web Coverage Service (WCS)

- The Web Coverage Service (WCS) supports electronic interchange of geospatial data as “coverages”- that is, digital geospatial information representing space varying phenomenon-OGC definition
- WCS provides
  - Spatial querying (grid spatial request)
  - Reprojection
  - Multiple output
  - Range subsetting



Provide Coverage in different Formats, BBOX, SRS

Full description of one or more coverages

Metadata

GetCapabilities

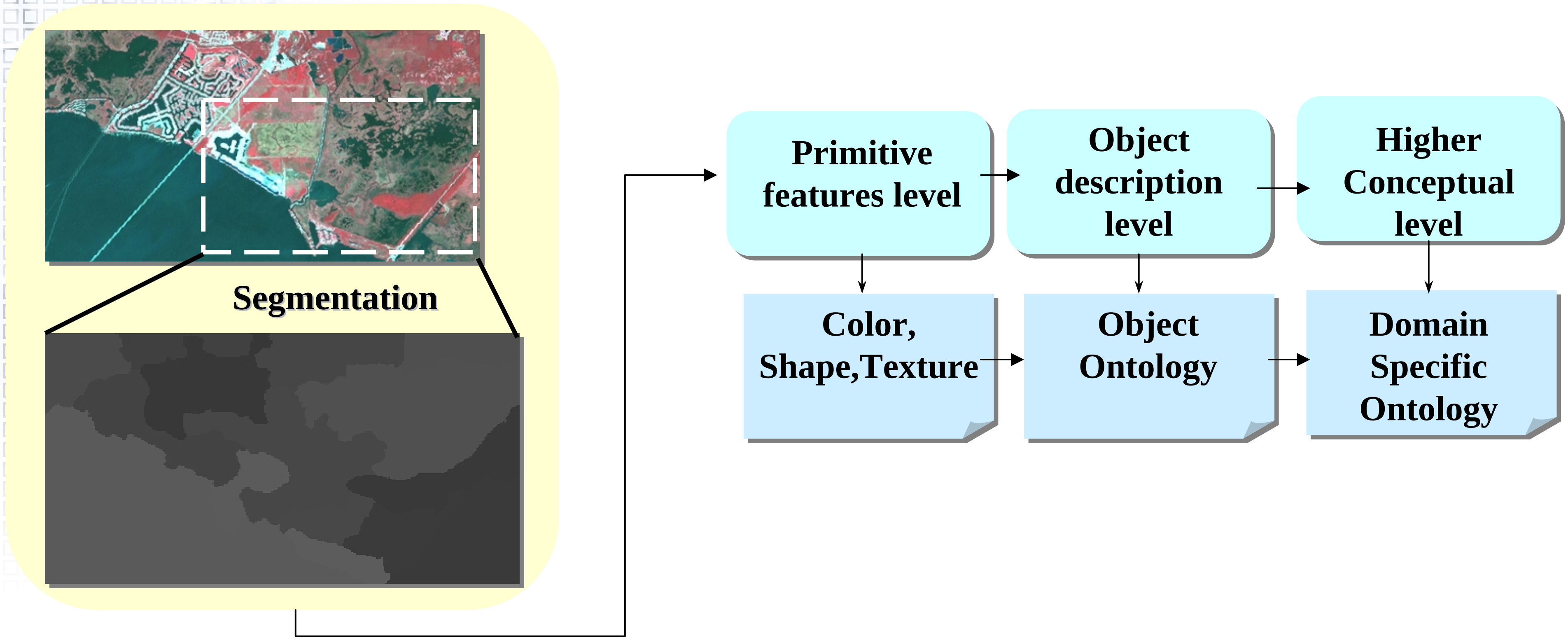
GetCoverage

DescribeCoverage

Web Coverage Service (WCS)

Feature Extraction

- ❑ Three level processing sequence consisting of Primitive Features Level (PFL), Intermediate Object Description Level (ODL) and a Higher Conceptual Level (HCL)



Segmentation

Primitive features level

Object description level

Higher Conceptual level

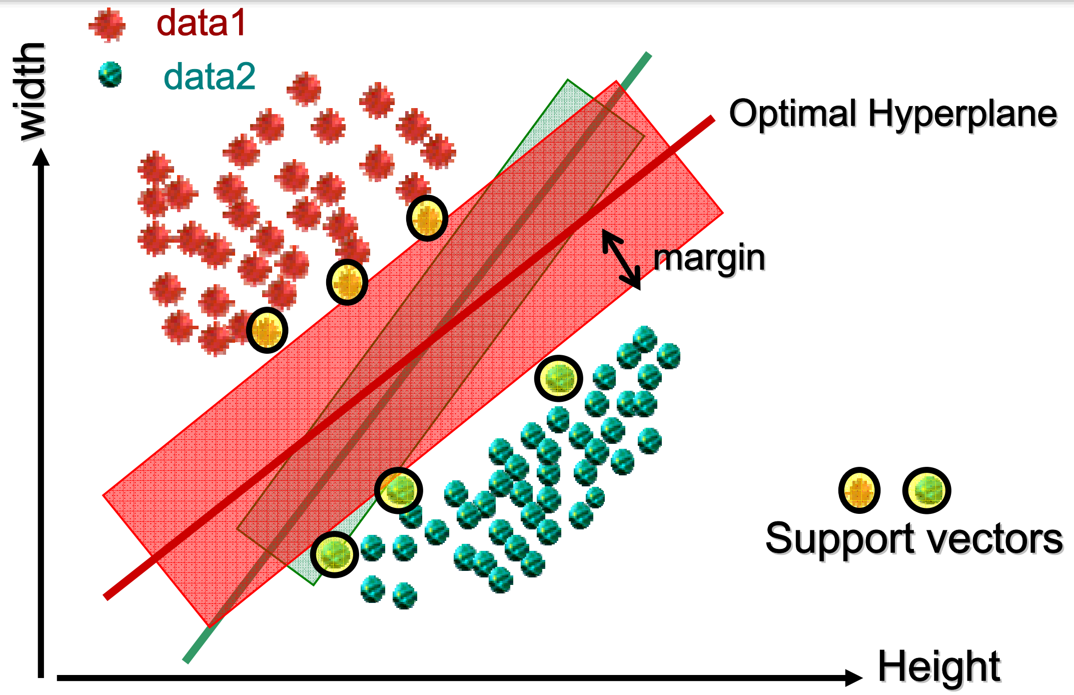
Color, Shape,Texture

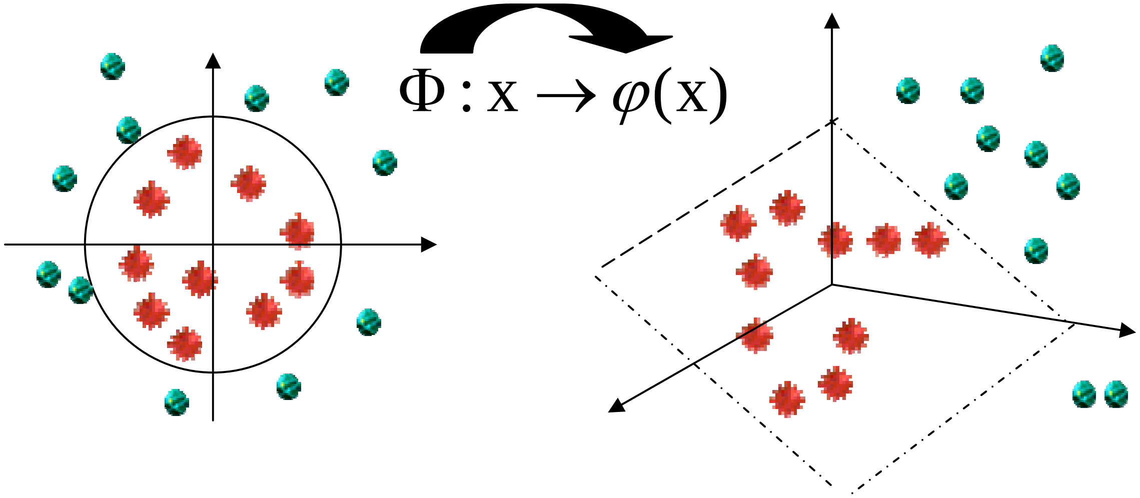
Object Ontology

Domain Specific Ontology

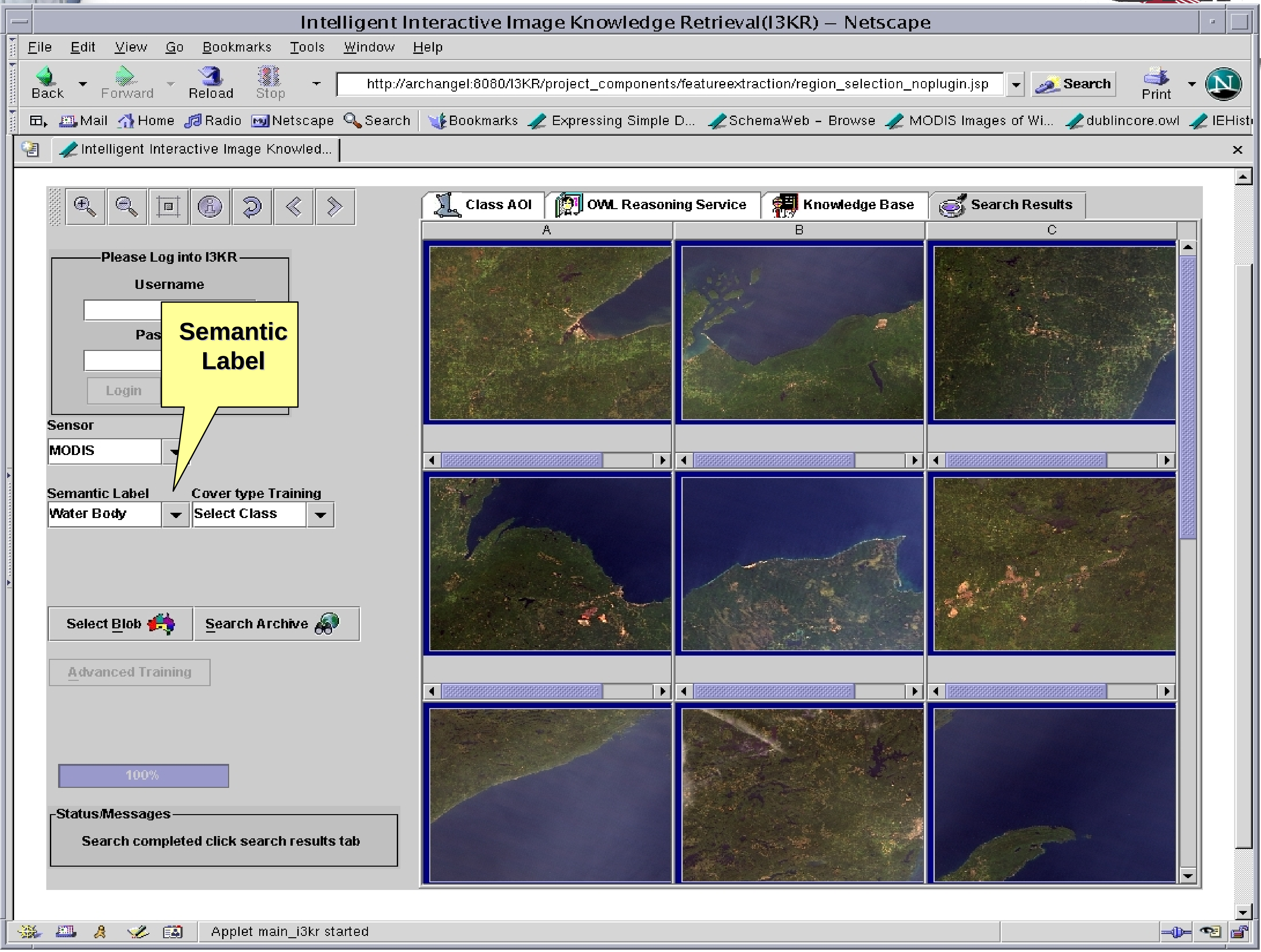
Support Vector Machines

- Support Vector Machine (SVM) is a powerful classification method which has shown outstanding classification performance in practice.
  - Simple, and always trained to find global optimum
- In its simplest form an SVM is a hyperplane that separates the positive and negative training samples with maximum margin.
- In the nonlinear case the original feature space is mapped to some higher dimensional feature space where the training set is separable




$\Phi : \mathrm{x} \rightarrow \varphi(\mathrm{x})$

Results



Semantic Label

Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape

http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp

Sensor: MODIS

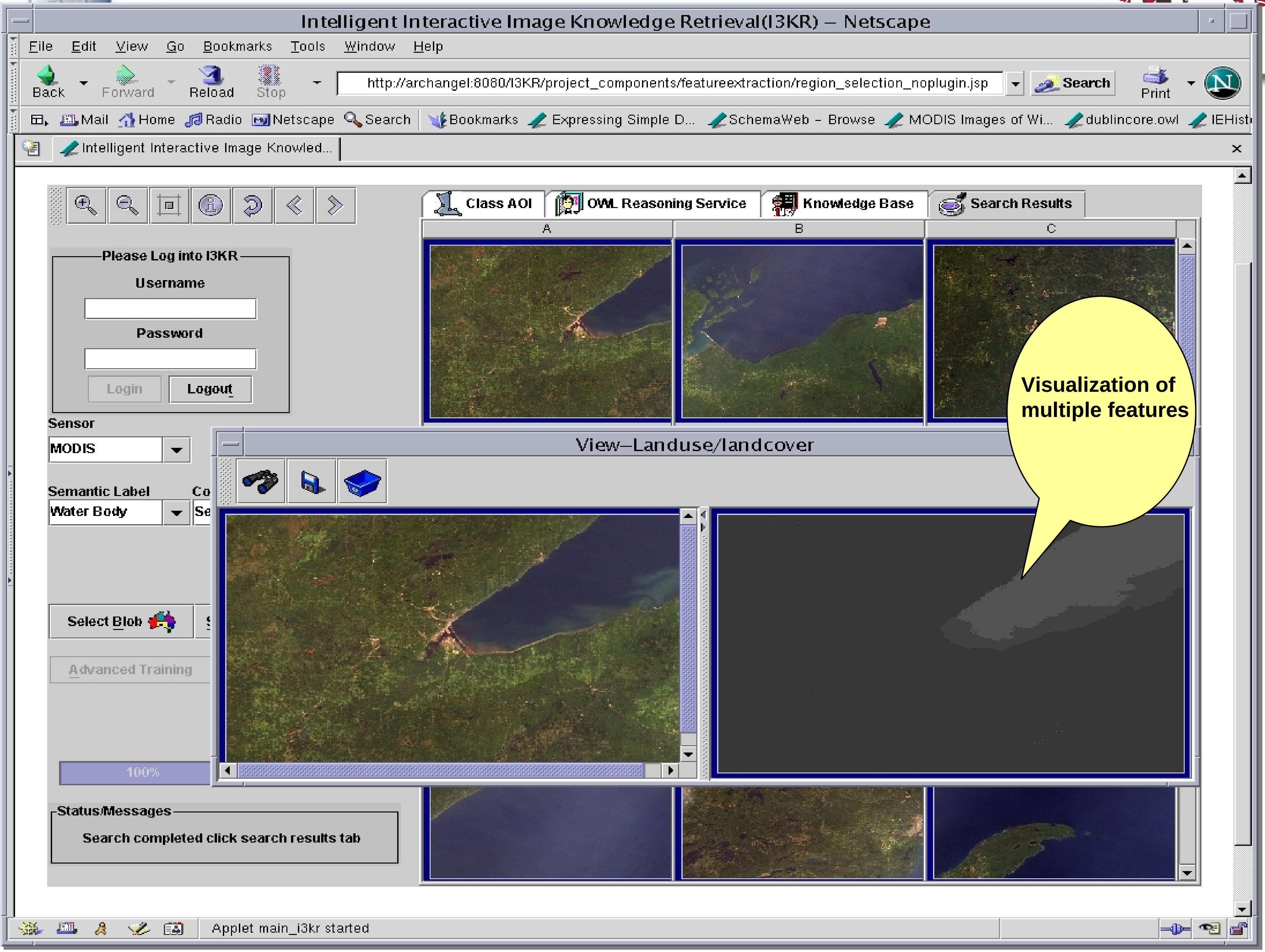
Semantic Label: Water Body

Cover type Training: Select Class

100%

Status/Messages: Search completed click search results tab

Applet main_i3kr started



Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape

http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp

Please Log into I3KR

Username

Password

Sensor: MODIS

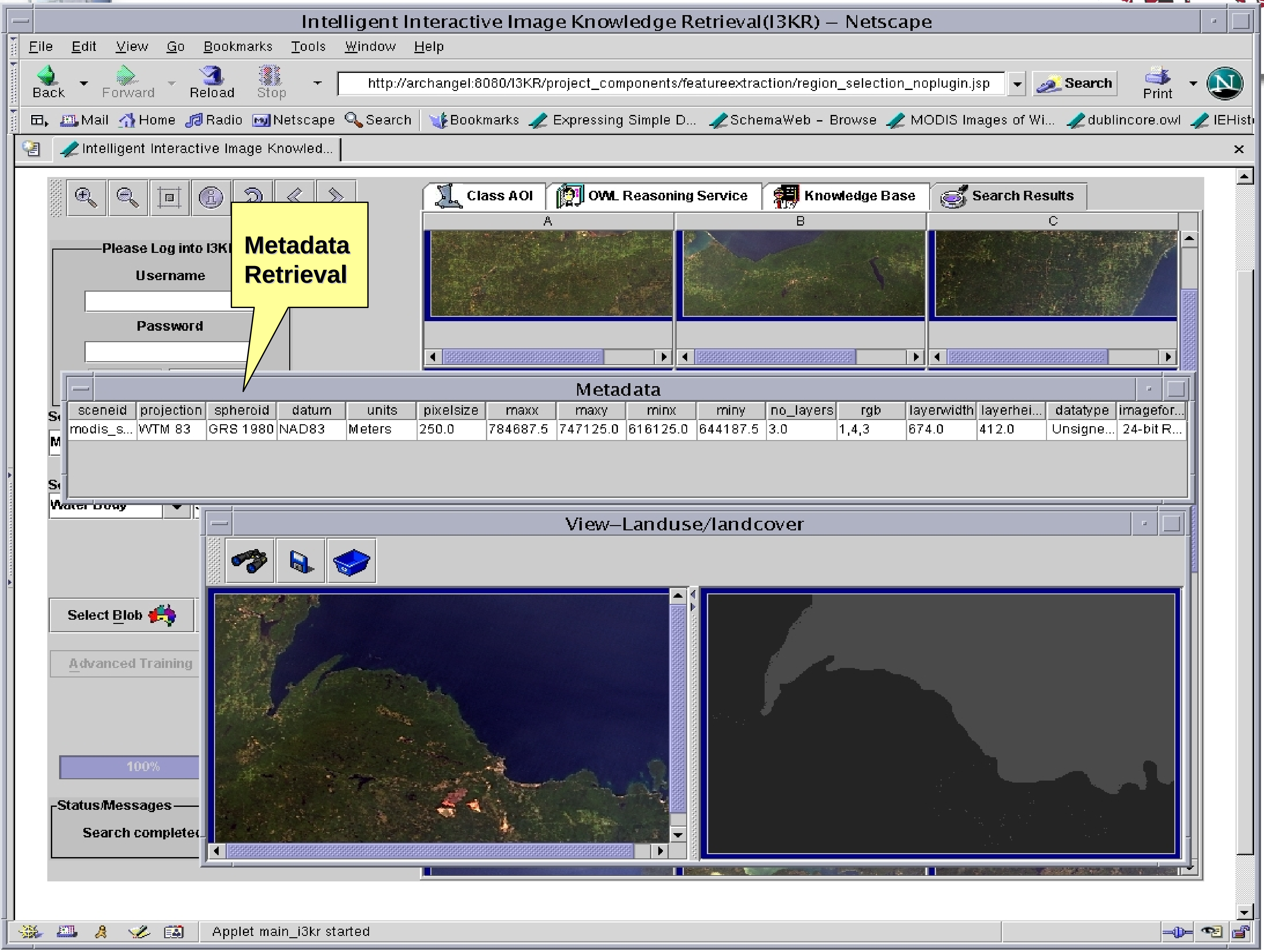
Semantic Label: Water Body

Status/Messages

Search completed click search results tab

Applet main_i3kr started

Visualization of multiple features



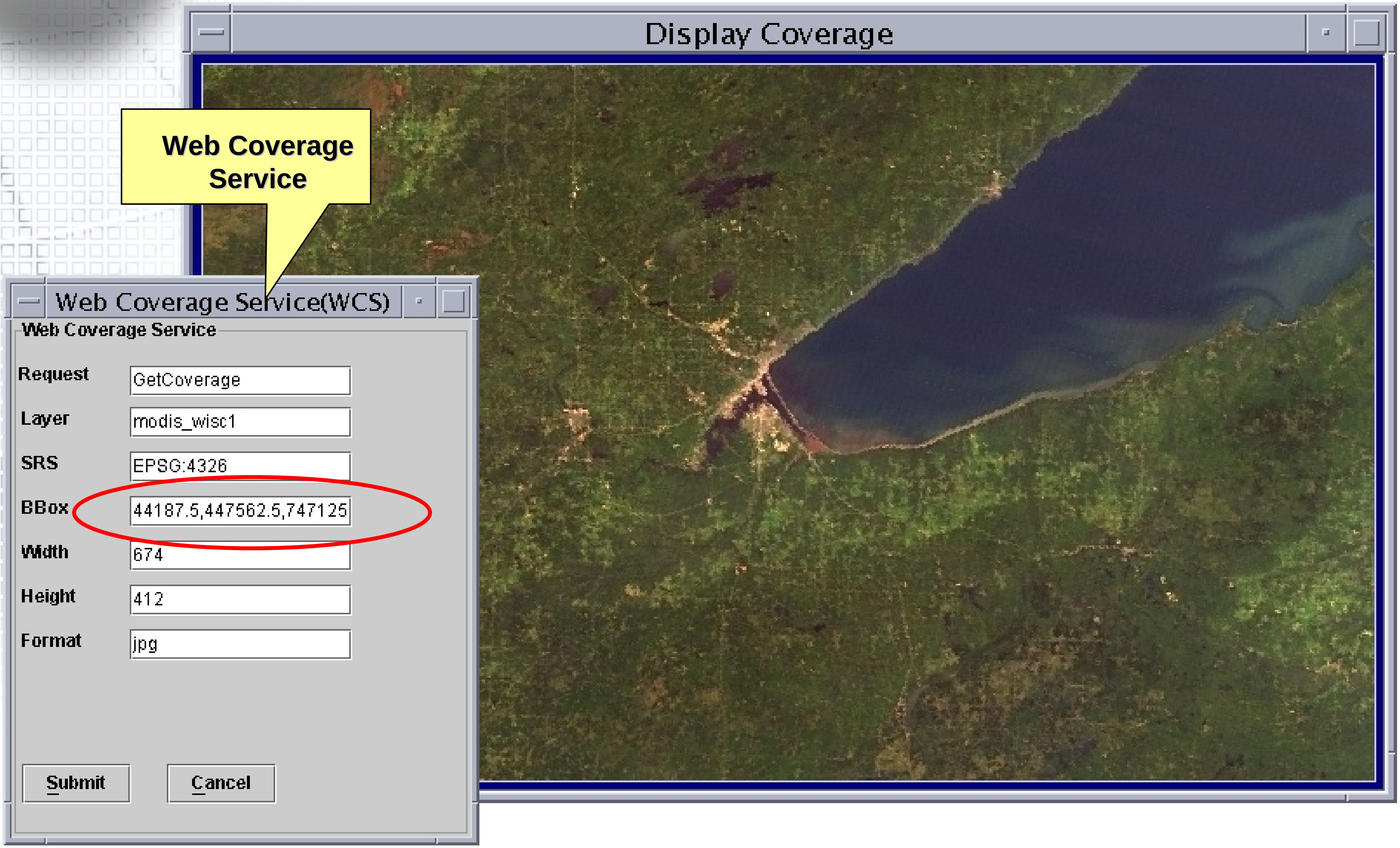
Metadata Retrieval

Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp
Search
Print
Mail Home Radio Netscape Search Bookmarks Expressing Simple D... SchemaWeb – Browse MODIS Images of Wi... dublincore.owl IEHist
Intelligent Interactive Image Knowled...
Class AOI
OWL Reasoning Service
Knowledge Base
Search Results
A
B
C
Please Log into I3KR
Username
Password
Water Body
Select Blob
Advanced Training
100%
Metadata
View–Landuse/landcover
Applet main_i3kr started

sceneid	projection	spheroid	datum	units	pixelsize	maxx	maxy	minx	miny	no_layers	rgb	layerwidth	layerhei...	datatype	imagefor...
modis_s...	WTM 83	GRS 1980	NAD83	Meters	250.0	784687.5	747125.0	616125.0	644187.5	3.0	1,4,3	674.0	412.0	Unsigne...	24-bit R...

Status/Messages

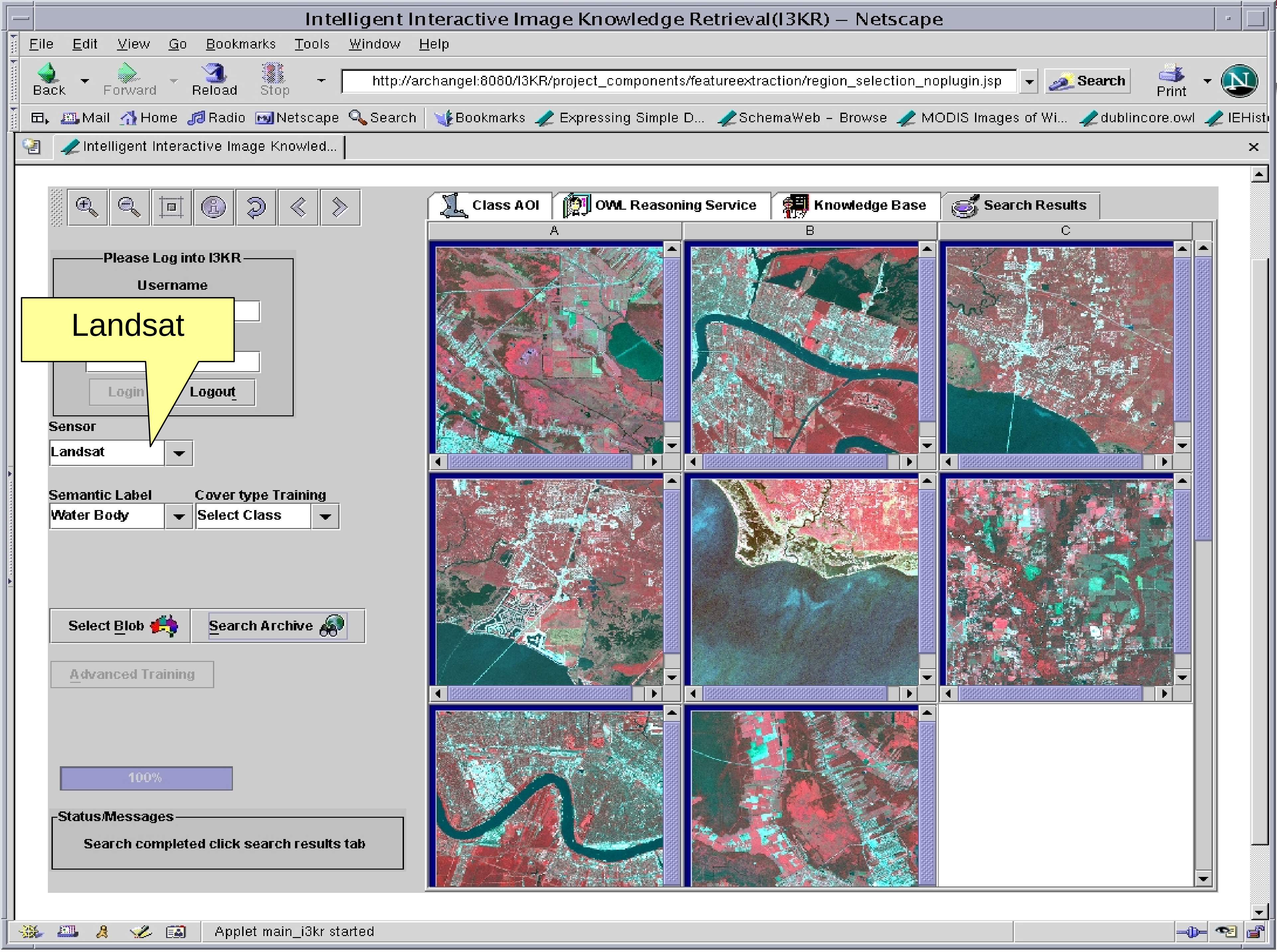
Search complete



Web Coverage Service

Web Coverage Service

Request	GetCoverage
Layer	modis_wisc1
SRS	EPSG:4326
BBox	44187.5,447562.5,747125
Width	674
Height	412
Format	jpg



Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape

http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp

Please Log into I3KR

Landsat

Sensor: Landsat

Semantic Label: Water Body

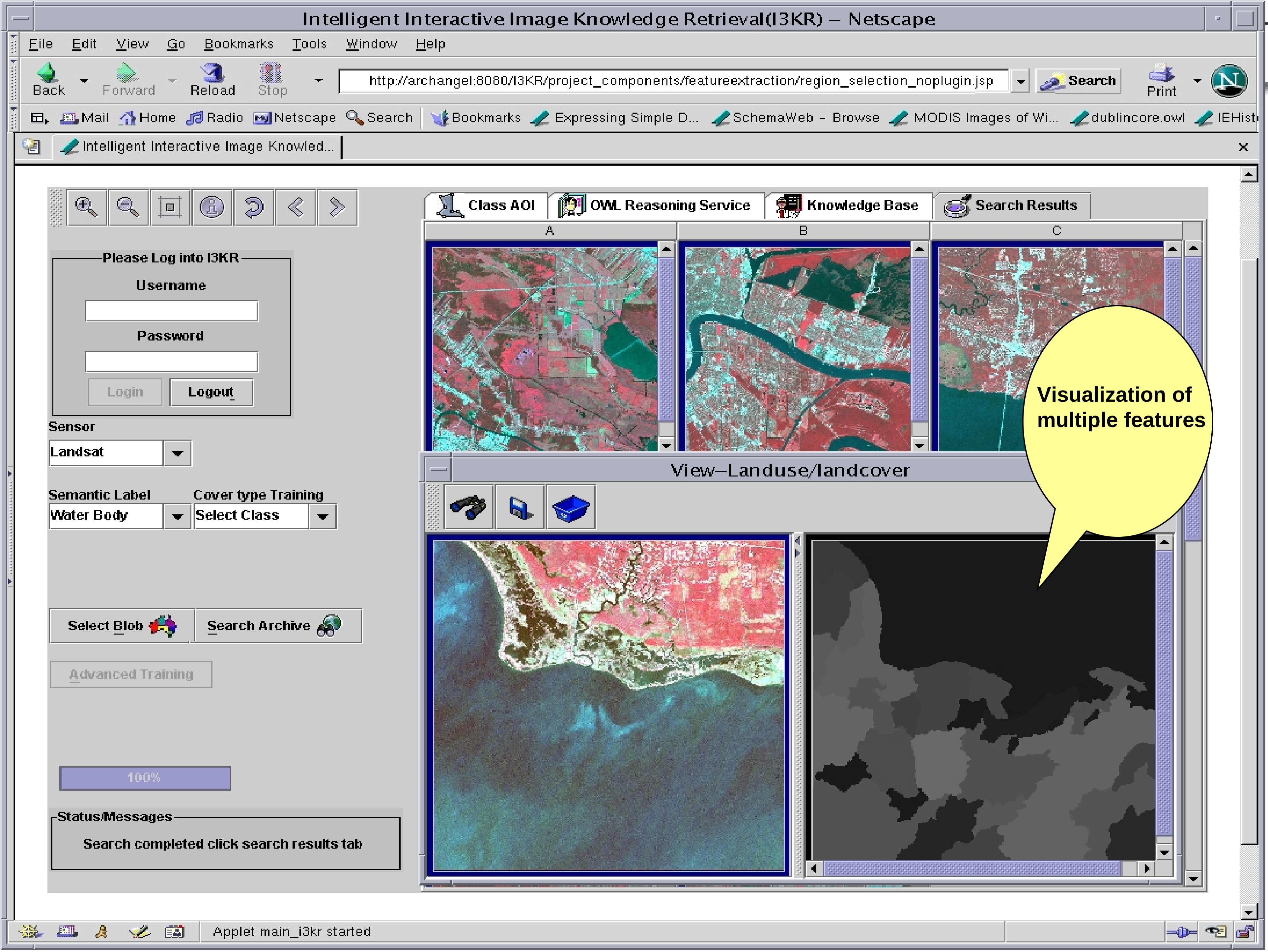
Cover type Training: Select Class

100%

Status/Messages

Search completed click search results tab

Applet main_i3kr started



Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape

http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp

Intelligent Interactive Image Knowled...

Please Log into I3KR

Username:

Password:

Sensor: Landsat

Semantic Label: Water Body

Cover type Training: Select Class

100%

Status/Messages

Search completed click search results tab

View–Landuse/landcover

Applet main_i3kr started

Visualization of multiple features

Problem

- ❑ Many realistic problems require expertise or data sources from globally distributed resources.
- ❑ These same problems may require processing of data into information from specific resources.
- ❑ How do I bring together geospatially diverse resources to facilitate sharing and knowledge discovery?
- ❑ How do I fuse data from a variety of sources (sensors, databases, images) into a useful information product?
- ❑ Constrained – compute horsepower, bandwidth, etc.

Cyberinfrastructure as a Solution

- Cyberinfrastructure is a NSF term used to refer to computational infrastructure that consists of:
  - computer hardware systems
  - application software and service
  - data and metadata management facilities and services
  - sociocultural elements of community building.
- The NSF recognizes that just as “infrastructure is required for an industrial economy, …, cyber infrastructure is required for a knowledge economy.”

Cyberinfrastructure Technologies

- ❑Grid computing and Peer to Peer computing (p2p)
- ❑Virtual Organization (VO) - An organization with its resources geographically distributed or when different organizations with different resources agree to share their resources in order to achieve a common goal.
- ❑Resources - Data, CPU power, domain specific software, or human experts in their fields.

Grid versus P2P Computing

- Pros:
  - Complex applications (grid computing evolved at universities, research labs).
  - Well accepted standards (OGSA) for middleware.
  - QoS – Nontrivial quality of service.
- Cons:
  - Scalability is not addressed well.

- Pros:
  - Dealing with a large number of peers and intermittent presence has led p2p to successfully address failure management.
- Cons:
  - No single middleware standard yet.
  - No QoS concept.
  - Applications are primarily file sharing and CPU sharing.

Cyberinfrastructure For Image Information Mining

❑ The cyberinfrastructure of choice often appears to be determined by the community of users.

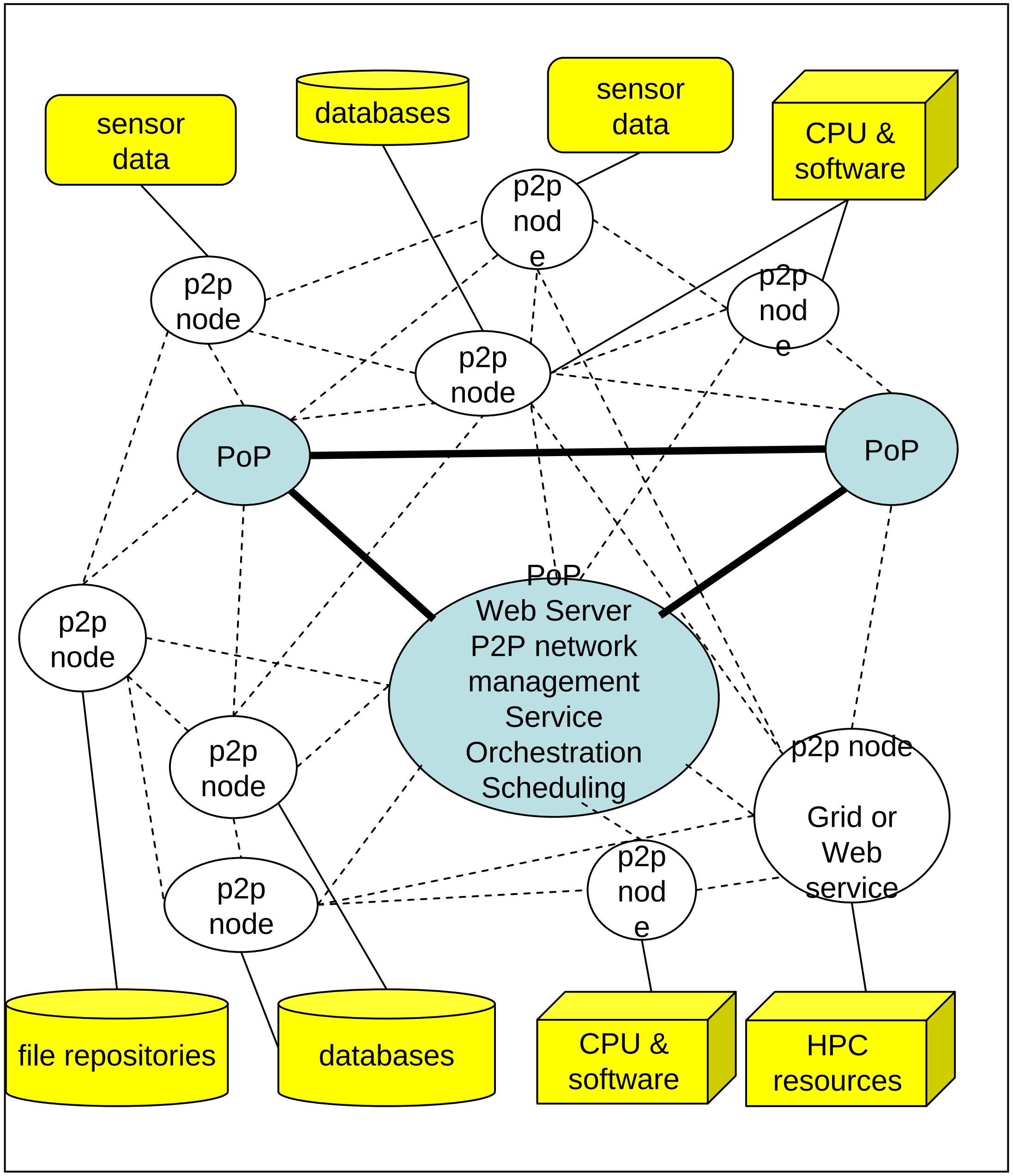
❑ If the users are primarily sharing files they will opt for the less formal p2p approach.

❑ For example, a p2p based Java application to share satellite imagery with various researchers SATELLA (http://satella.geog.umd.edu/)

❑ DIGITAL PUGLIA pools together resources from University of Lecce, Italy, San Diego Supercomputing Center (USA) and California Institute of Technology (USA) in an active digital library of remote sensing data.

Cyberinfrastructure For Image Information Mining

- ❑ An I2M cyberinfrastructure must allow for on-demand aggregation of computational resources (such as computational servers, instruments and sensors, databases, and data repositories) across administrative domains at any time.



Conclusions

- ✓ Image archives offer new sources of knowledge for today's discoverers
- ✓ Framework for content and semantic based information retrieval from remote sensing data archives
- ✓ Middleware for Ontology Driven Brokering (MOB)
- ✓ Web coverage service integration
- ✓ Machine learning methods for image information retrieval
- ✓ Early results from the prototype application using Landsat and MODIS data

Conclusions

- ✓ Results from this web service can lead to discovery of new knowledge and understanding.
- ✓ Image information mining services will become critical middleware components of the cyberinfrastructure of the future.
- ✓ The task of this middleware will eventually be to operate not only on archived datasets, but also on data streams in near real time.
- ✓ The two vying approaches for a cyberinfrastructure application for image information mining both offer strengths.
- ✓ Research should be focused on using the best aspects from grid technologies and p2p computing for building an image information mining cyberinfrastructure.

Thank You !

Backup

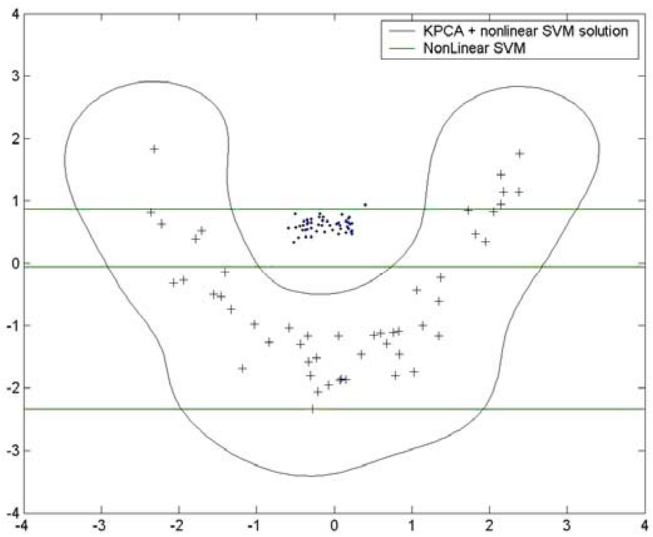
Kernel Principle Component Analysis

- As developed by Scholkopf et al., Kernel PCA (KPCA) is a technique for nonlinear dimension reduction of data with an underlying nonlinear spatial structure

- Kernel PCA is based on the formulation of PCA in terms of dot product matrix instead of covariance matrix

- It is possible to extract non-linear features using kernel functions by solving an eigenvalue problem like for PCA

- KPCA is used to extract structure from high dimensional data set



Support Vector Machines

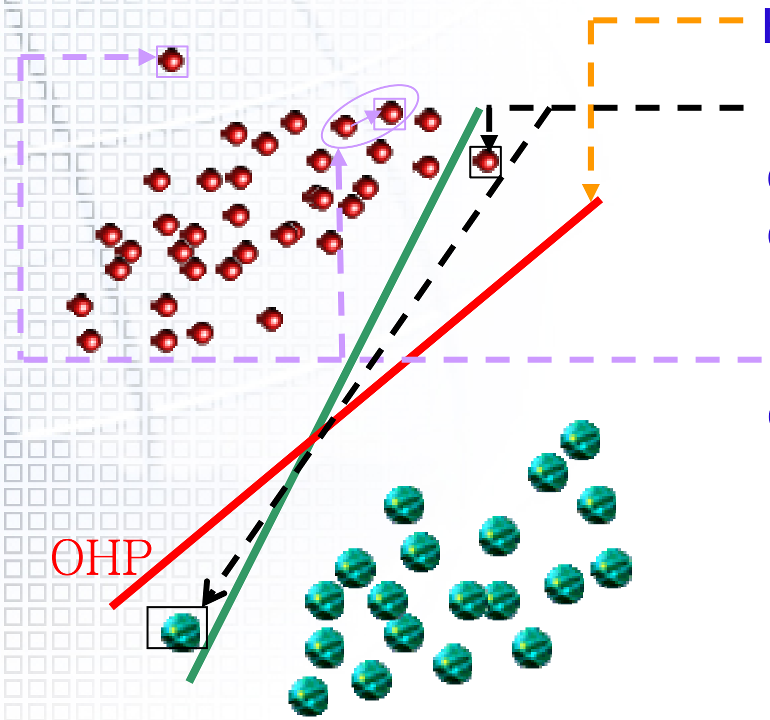
- ❑ Support Vector Machine (SVM) is a powerful classification method which has shown outstanding classification performance in practice.
  - ❖ **Simple, and always trained to find global optimum**

- ❑ It is based on a solid theoretical foundation-structural risk minimization.

- ❑ In its simplest form an SVM is a hyperplane that separates the positive and negative training samples with maximum margin.
  - ➢ The decision function of an SVM is $f(x) = \langle w \bullet x \rangle + b$, where $\langle w \bullet x \rangle$ is the dot product between w (the normal vector to the hyperplane) and x (the feature vector representing the example)
  - ➢ The margin for an input vector x_i is $y_i f(x_i)$ where $y_i \in \{-1,1\}$s the correct class label for x_i .
  - ➢ In the linear case , the margin is geometrically the distance from the hyperplane to the nearest positive and negative examples.
  - ➢ Seeking the maximum margin can be expressed as a quadratic optimization problem:

Minimizing $\langle w \bullet w \rangle$ subject to $y_i(\langle w \bullet x \rangle + b) \geq 1, \forall i$

Support Vector Machines



Intuitively feels safest

Hyperplane is really simple

If we've made a small variation in the location of the boundary this gives us least chance of causing a misclassification

Robust to outliers since the model is immune to change/removal of any non-support vector data points