



# Applications of Bayesian Classification to Data Management

*Christopher Lynnes*

*NASA/GSFC*

*Co-Authors:*

*S. Berrick, A. Gopalan, X. Hua, S. Shen, P. Smith, K. Yang*

*NASA/GSFC*

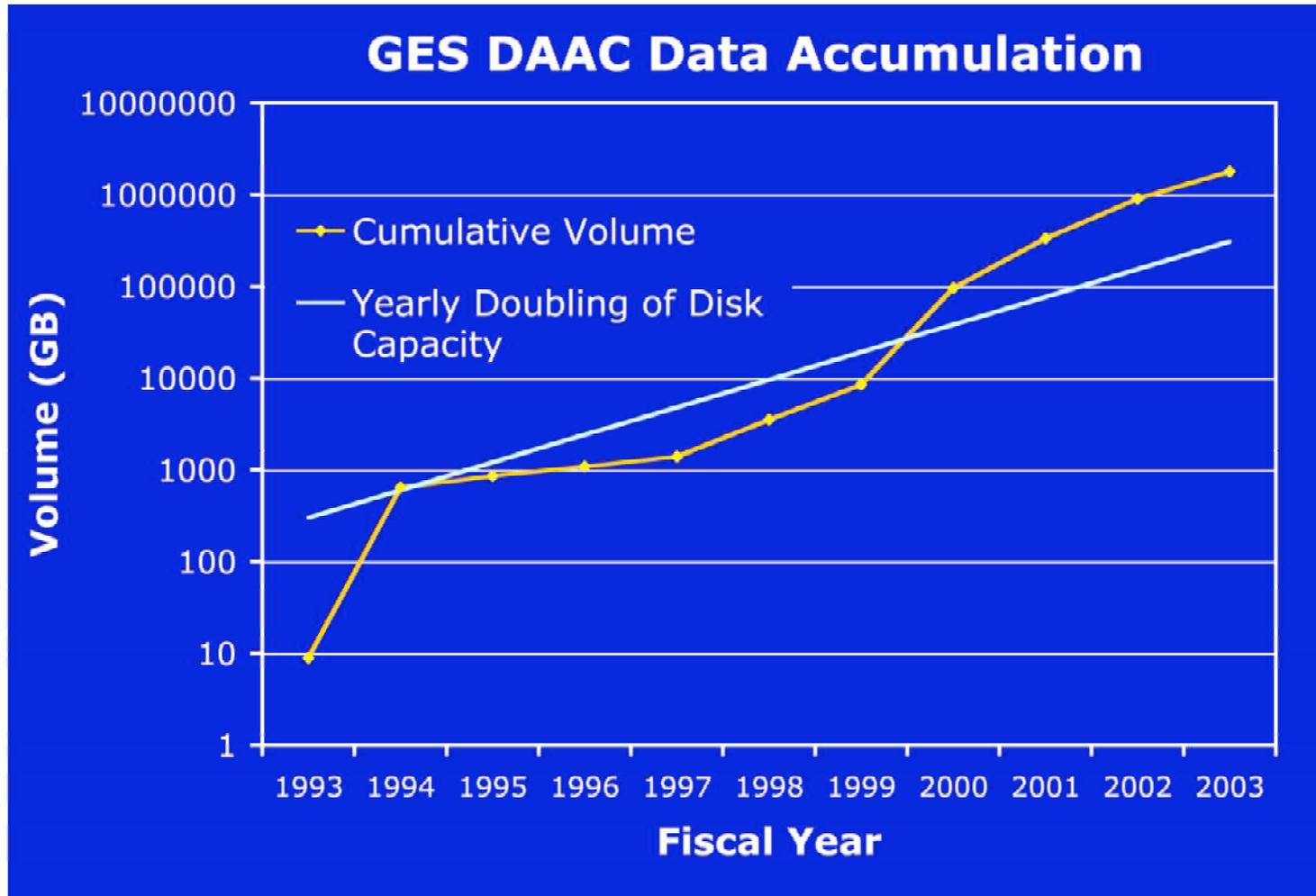
*K. Wheeler, C. Curry*

*NASA/ARC*



# Problem Statement

Science data volume keeps pace with technology.





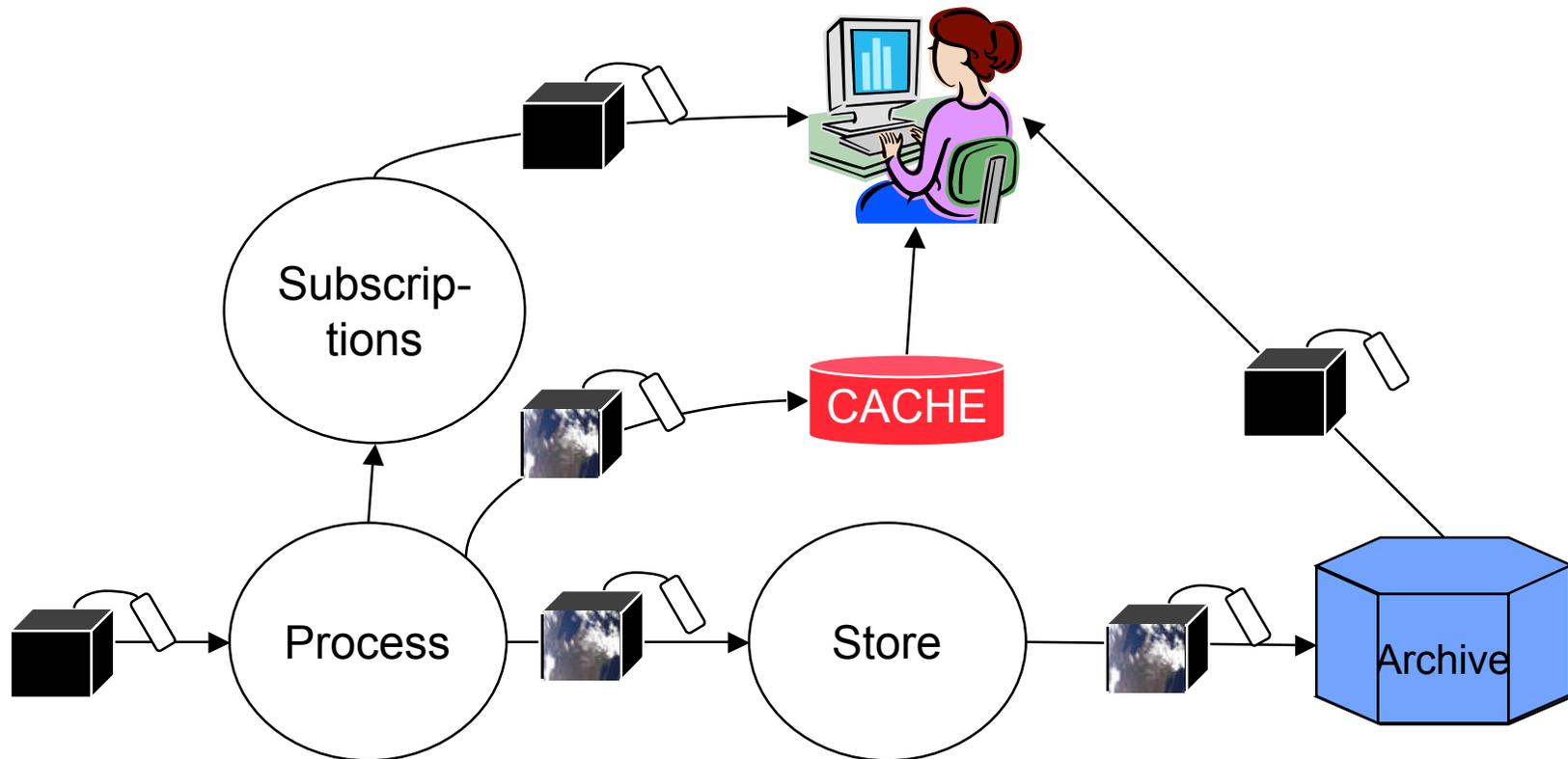
## Data demands are also increasing

- Lower Latency: driven by applications
- Online access: driven by machine-to-machine interfaces (e.g., models)
- Volume: driven by advances in computing and data mining
- A solution is to manage data according to their “usefulness”.



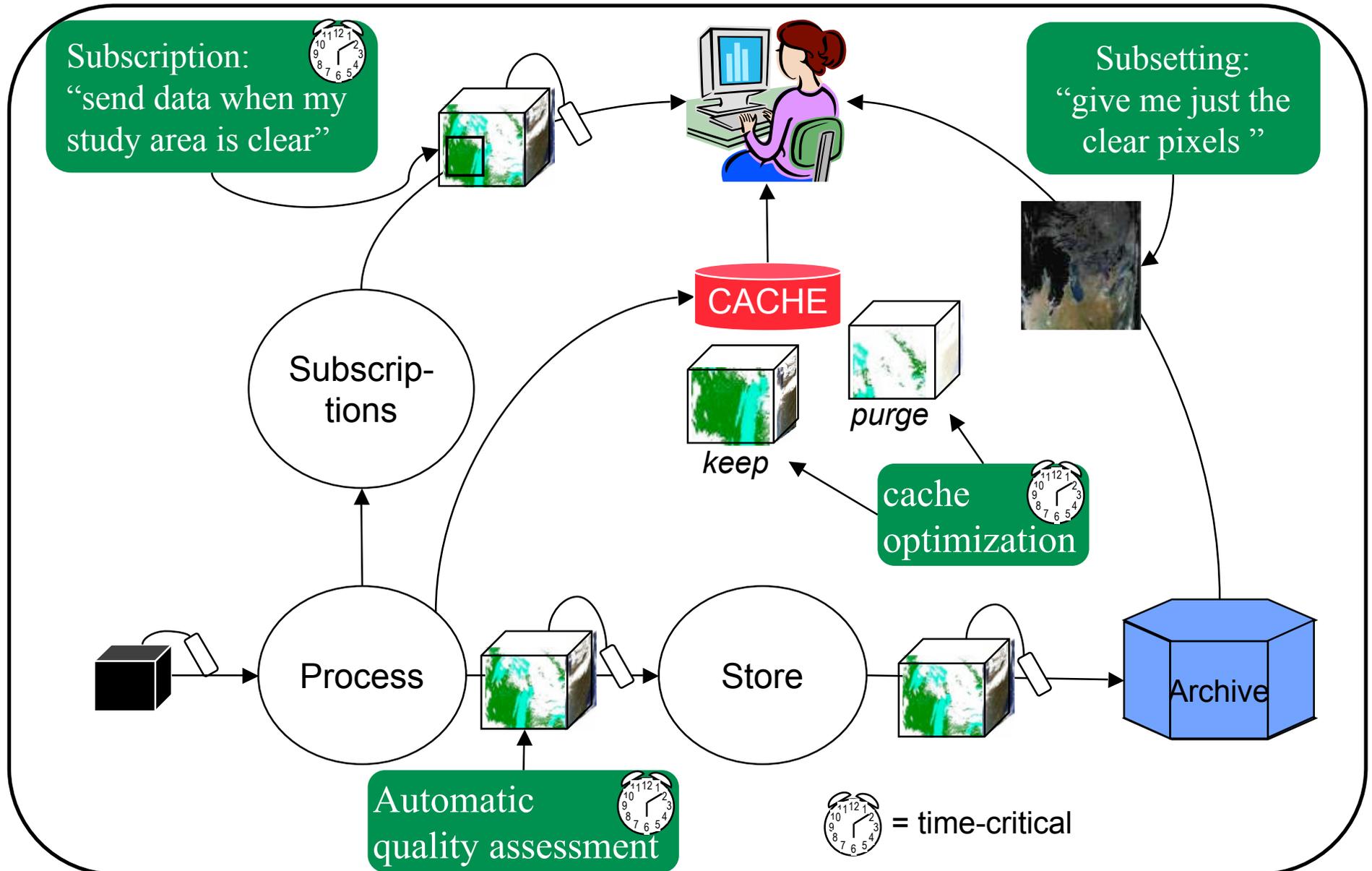
## Data Management Today: black-box paradigm

- Data are managed as largely opaque objects
  - albeit with labels (metadata) and “cover art” (browse)





# Content-based Data Management





## Usefulness is in the eye of the beholder

Study Type	Pixel Characteristics					
	Cloudy	Clear-Sky				
		Ocean	Sunglint	Land	Snow/Ice	Fire
Cloud Properties	X					
Aerosols		X	(X)	X		X
Ocean Color		X				
Land Vegetation				X		
Snow Cover/Sea Ice					X	
Wildfires						X



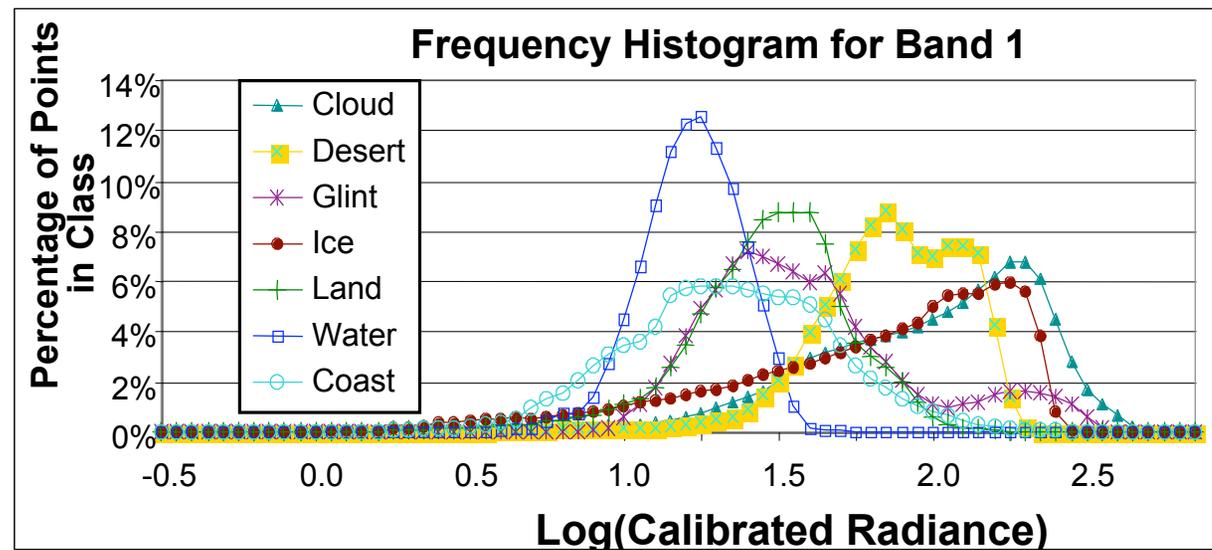
## Characterization of MODIS Calibrated Radiance

- Most popular product at Goddard DAAC
- Train algorithm to classify pixels
  - Cloud, glint, land, water, etc.
- Speed of the *forward* algorithm is critical.
  - However, we can afford time and CPU for training.
- Products from science algorithms train machine learning algorithms
  - Products as proxy for domain experts
  - Nearly unlimited supply of training and test data
  - Circular logic if we were making science products...
  - ...but in the decision support domain, it serves as a high-speed approximator to the science algorithm.



# Bayesian Classification Applied to MODIS Calibrated Radiance

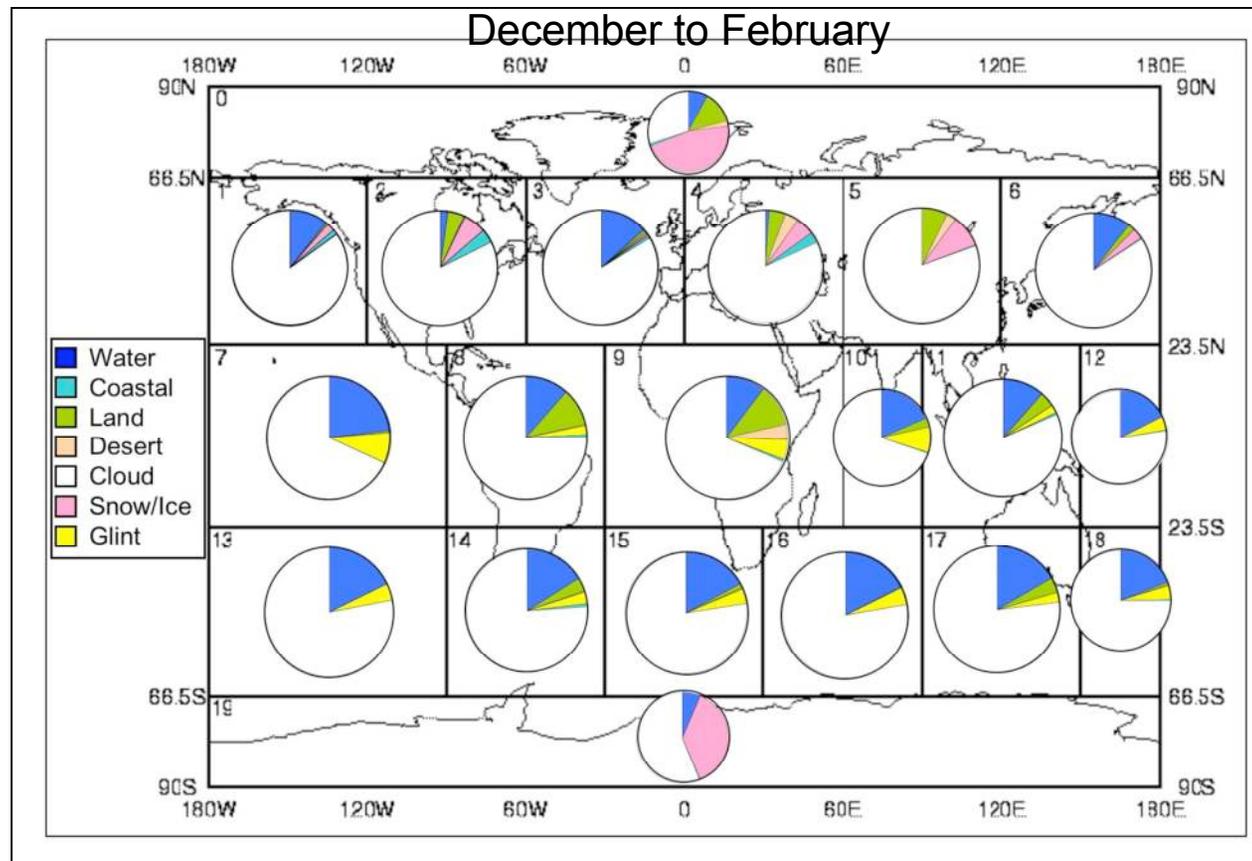
- Bayesian classification:
  - $\Pr(C|E) = \prod \Pr(E_i | C) \times \Pr(C) / \prod \Pr(E_i)$
  - Where C is a class
  - And  $E_i$  are measurements of independent variables (evidence).
  - $\Pr(C)$  is the prior probability
- Training: Compute frequency histograms for  $E_i|C$ 
  - MODIS cloudmask and ocean color products “train” the classifier.





## Prior Probabilities

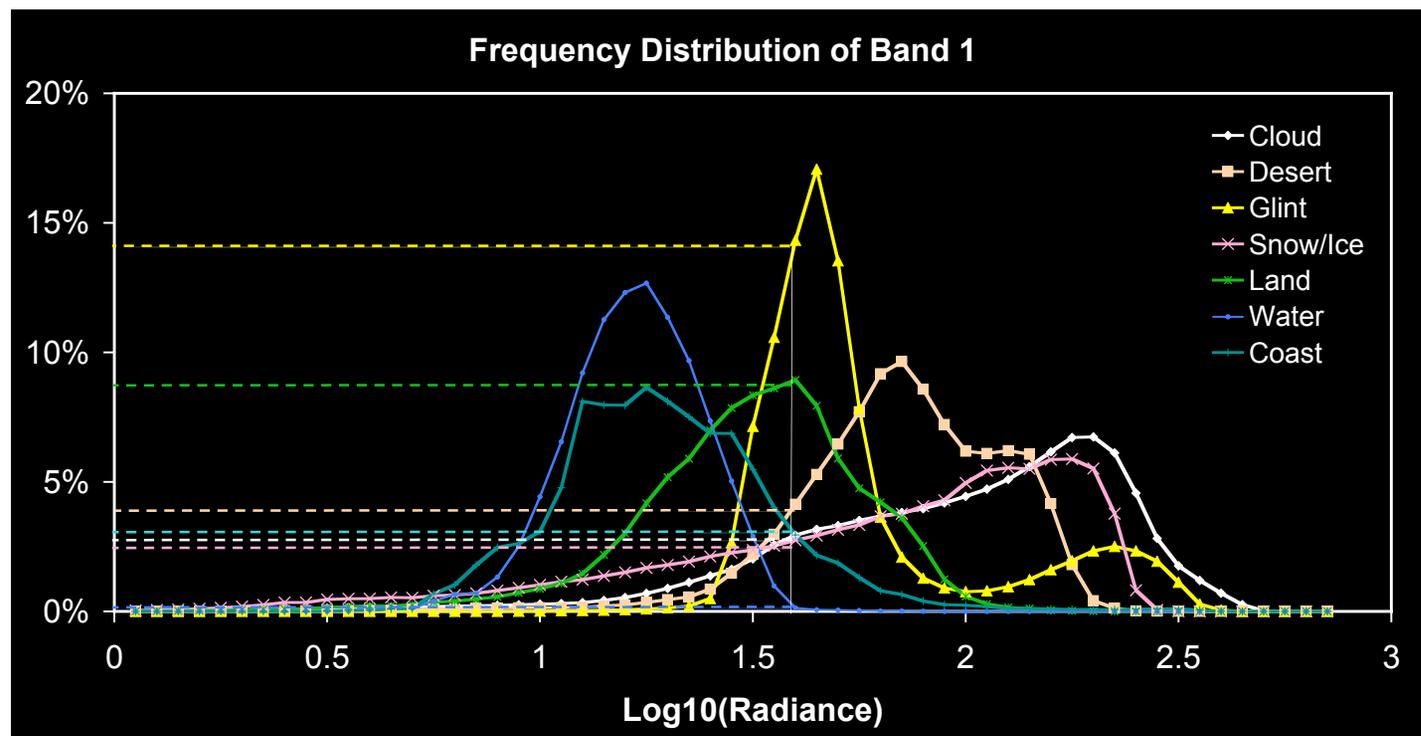
- Prior probabilities are “known” statistics for the earth
  - Regional and Seasonal variations
  - Derived from MODIS Level 3 gridded products





## Practical Classification - Application

- For each class:
  - Look up the probability for each band measurement in frequency histograms
  - Compute product to get the overall probability for membership in that class
  - Choose the class with the highest overall probability



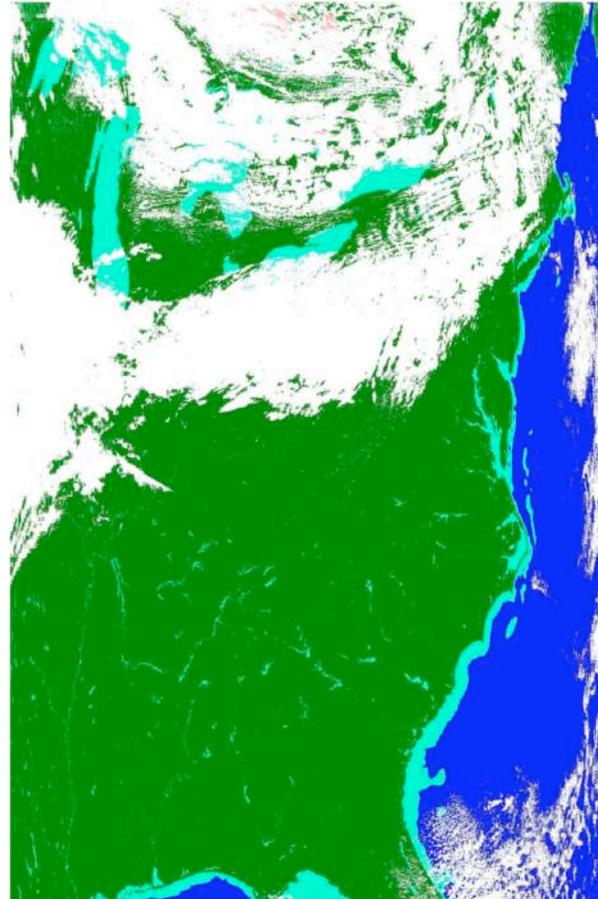


# Bayesian Classification Example

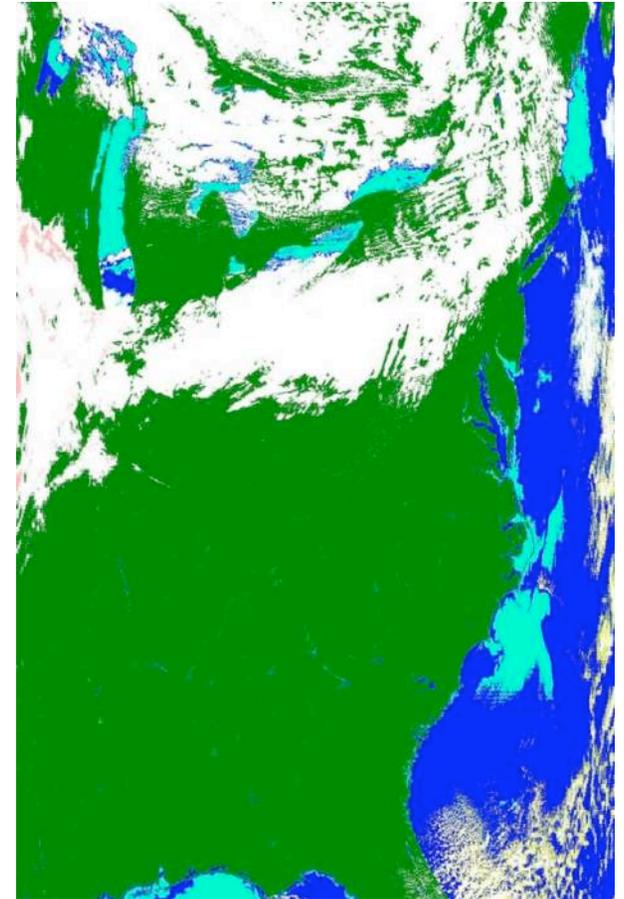
Terra/MODIS scene for  
16:20-16:25Z, 2003-10-16



MODIS Cloudmask Product



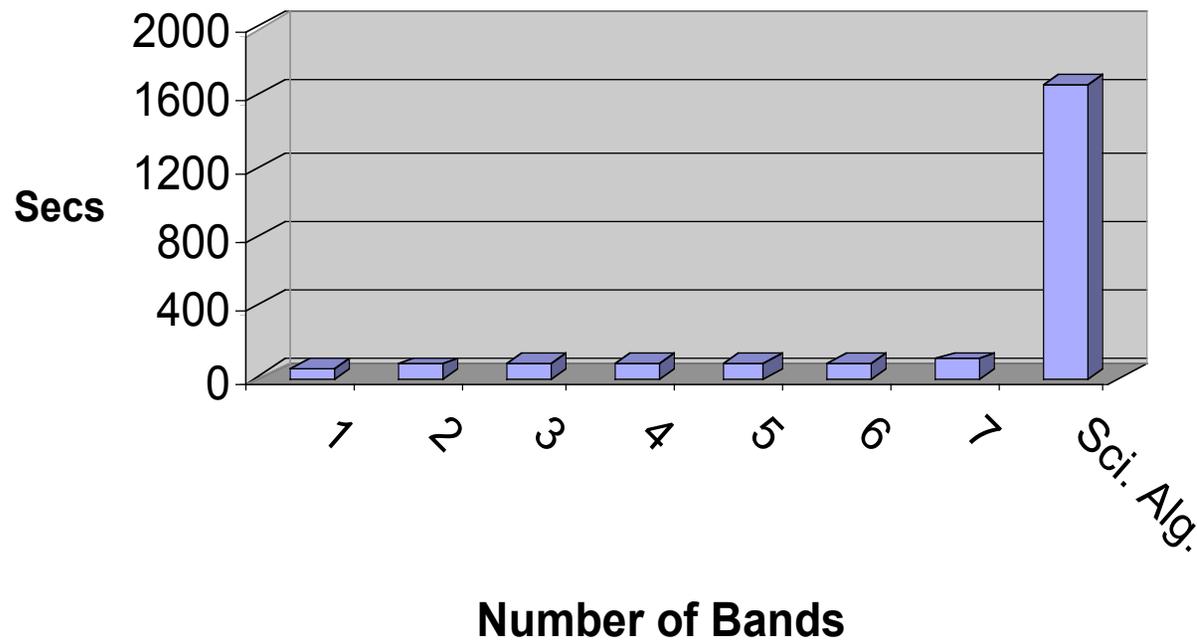
Bayesian classification using  
bands 1, 2, 2/1, 31, 32





# Timing Results

## Algorithm Timing for 300 s of Data

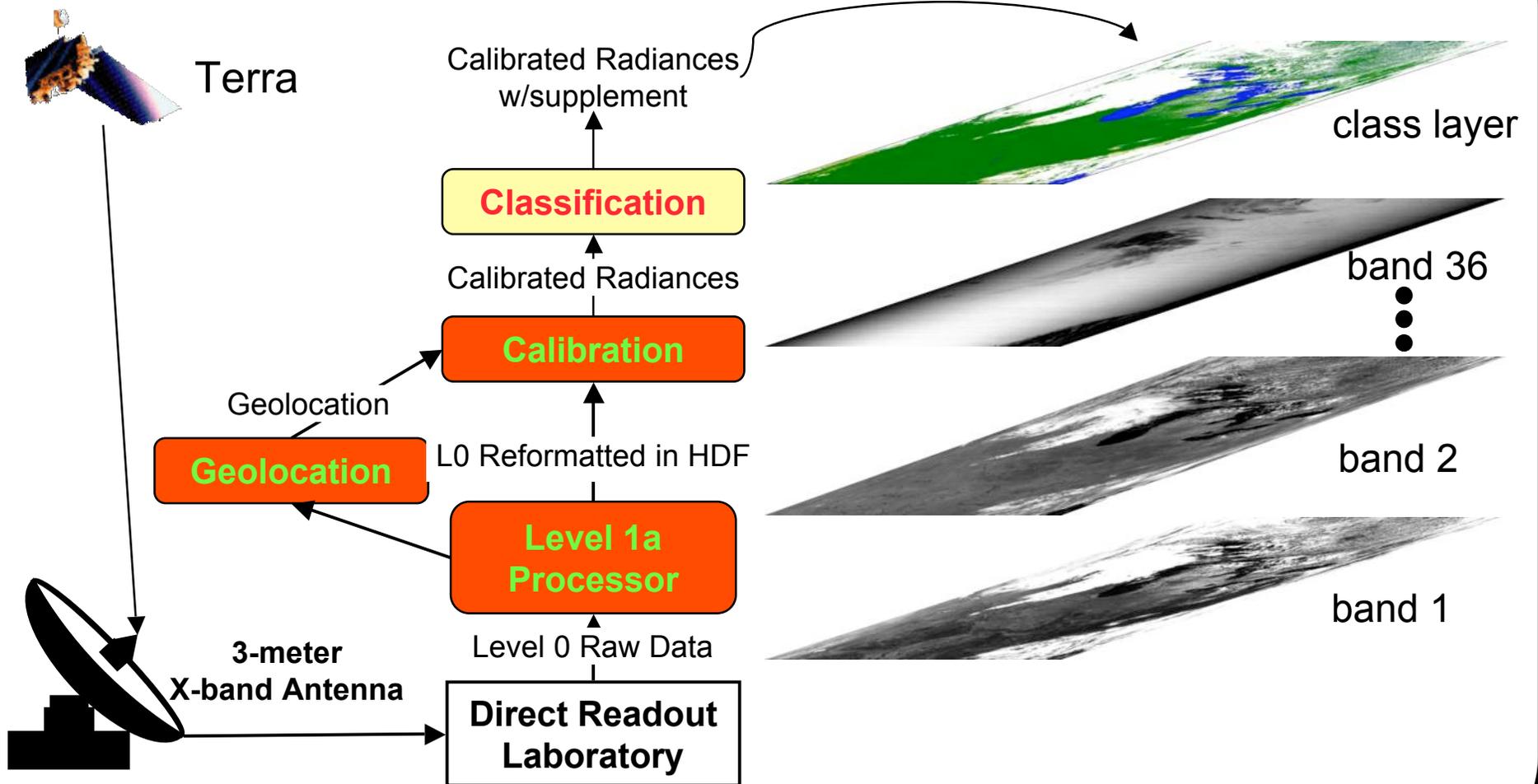


\*Bayesian classification on 250 MHz SGI, as a function of number of bands used



# Exploitation of Classification Results

- Add algorithm to Direct Broadcast processing stream





## Content-Based Subsetting

Deliver just the pixels likely to be useful

e.g., cloud-free

1. Classify using Bayesian classifier
2. Zero out pixels classified as cloud
3. Apply lossless compression

Currently implemented as an on-the-fly conversion in

WUSTL FTP, e.g.:

```
ftp g0dug03u.ecs.nasa.gov
```

```
>cd /datapool/OPS/user/MODB/RMT021KM.001
```

```
>ls
```

```
>cd 2004.06.13
```

```
>ls *.hdf
```

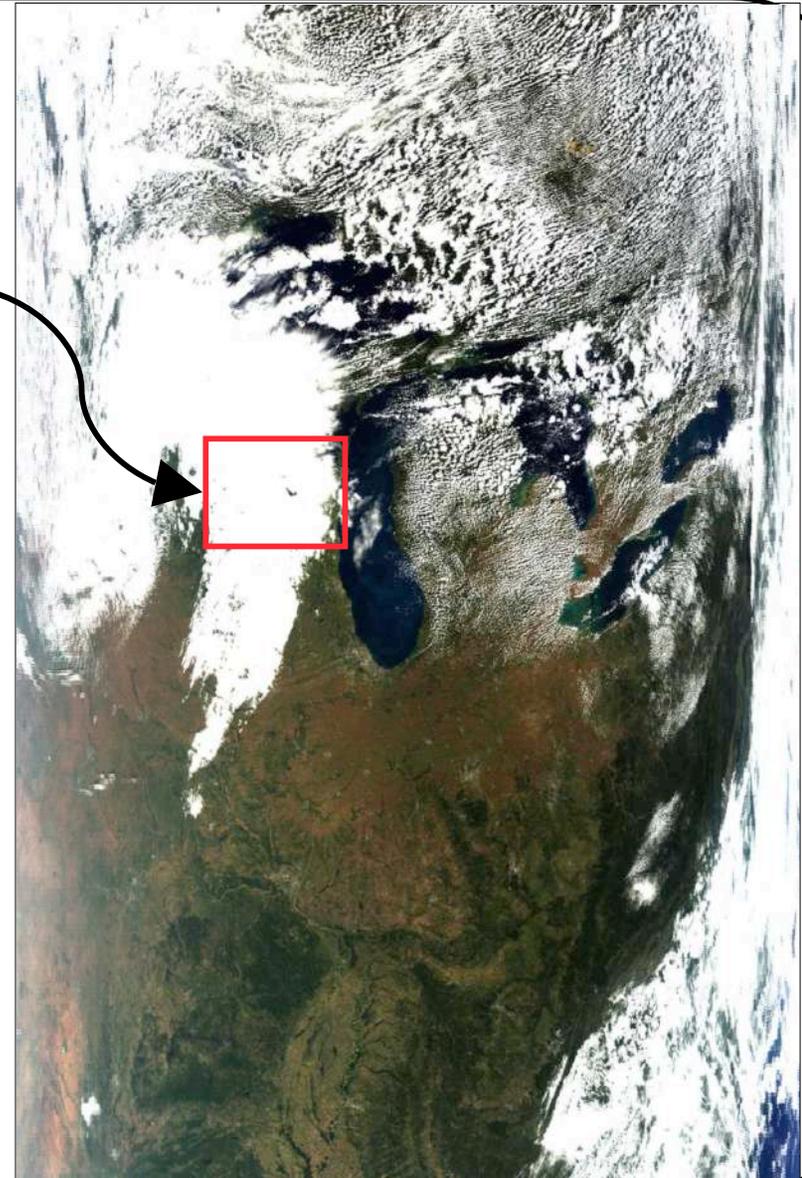
```
>get RMT021KM.A2004165.1843.001.2004166072602.hdf.clr
```





## Content-Based Data Selection

- Today: “select scenes where cloud cover  $< 50\%$ ”
  - Less than foolproof **study area**
- Tomorrow: “select scenes where Lake Winnebago is visible”
- Ad hoc indexing / queries are difficult, but...
- ...subscription queries should be tractable
  - “Is anyone looking for data that are clear for a particular area in this scene?”





## Automated Quality Assessment of Geolocation

- Compare observed land-water pattern with land-sea mask based on geolocation
  - Systematic geolocation error  $\Rightarrow$  systematic shift in pattern
- Technique:
  - Classify land/water/cloud from geolocated radiance
  - Assign +1.0 to land, -1.0 to water
    - Assign “unknown” classes a random number in the interval (-1.0, +1.0)
      - Cloud, snow/ice in classification
      - Ephemeral water in land-sea mask
  - Compute cross-correlation using 2-D FFT



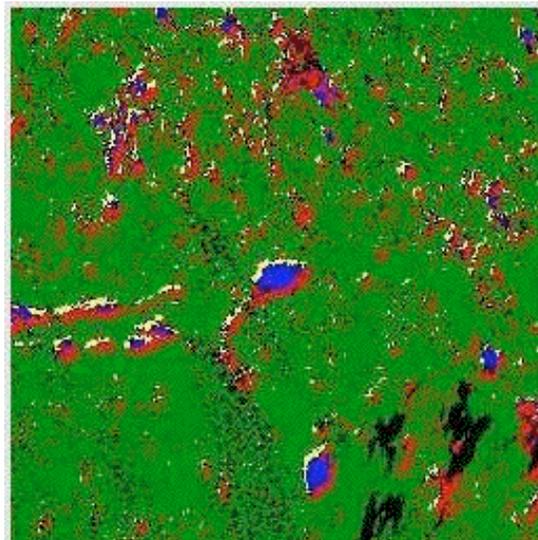
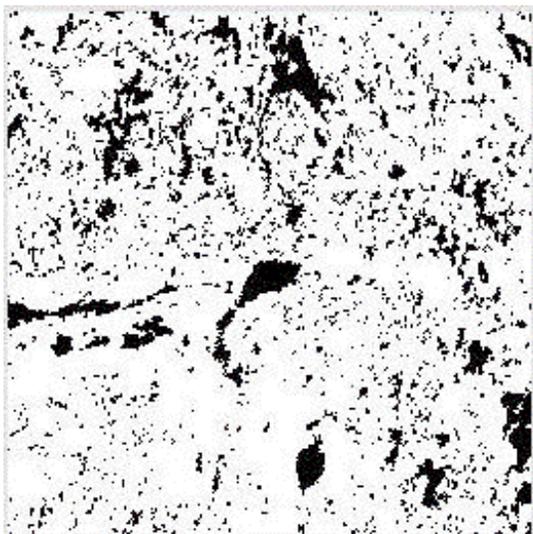
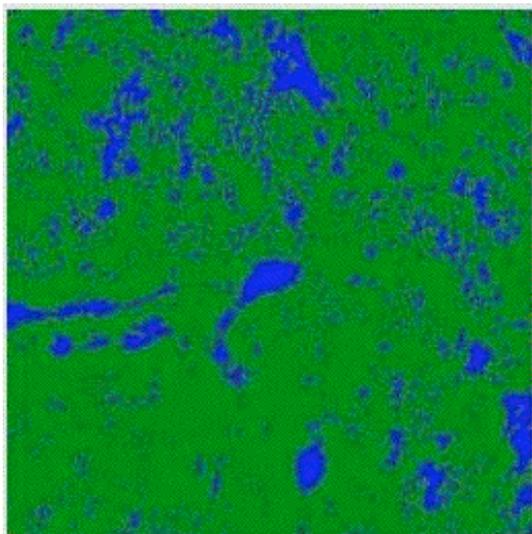
## Geolocation Case Study

- Terra/MODIS data for 19 June 2002 reprocessed with the usual onboard attitude and ephemeris
- But: a spacecraft maneuver made the onboard data inaccurate
  - Typically, definitive attitude/ephemeris are used in the vicinity of maneuvers
- Several months later...a group studying land cover change identified errors in geolocation



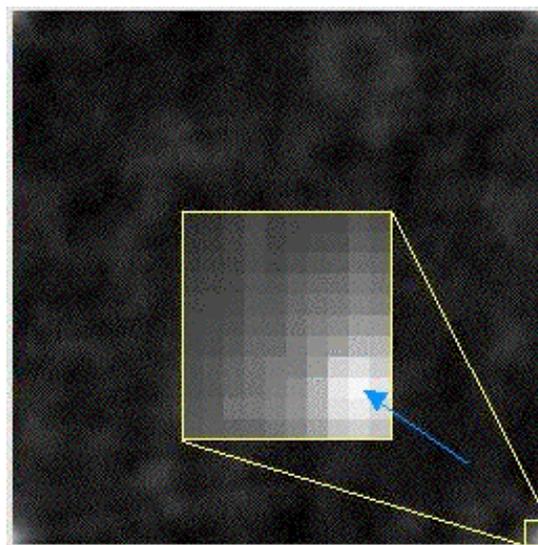
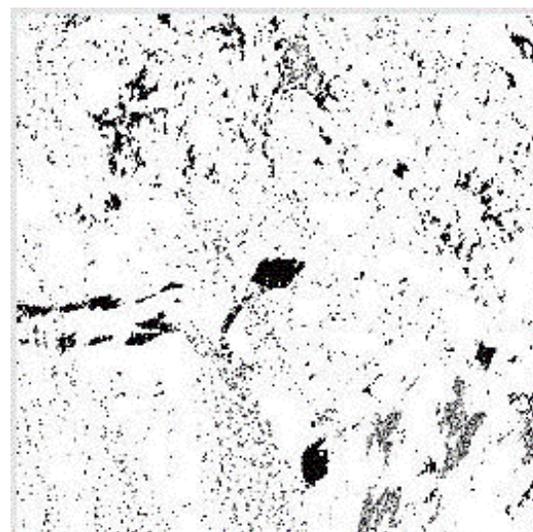
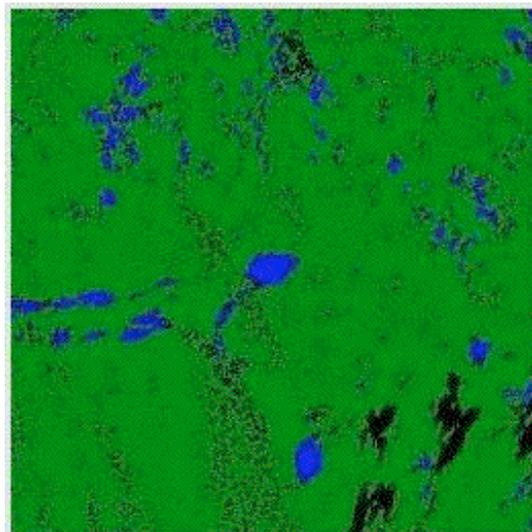
# Geolocation Shift Effect

Land-sea mask



Geolocation shift

Bayesian classification



Cross-Correlation