



*Land Processes  
Distributed Active Archive Center*



# Fundamentals of Data Analysis and Interpretation

**Brad Reed**

**SAIC\***

**USGS EROS Data Center**

**reed@usgs.gov**

# Outline

- **Visual Analysis and Interpretation Techniques**
- **Digital Image Processing**
  - Data Input
  - Pre-processing
  - Analysis

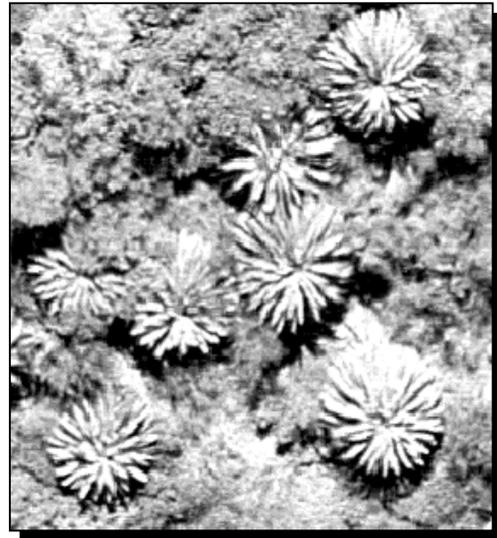
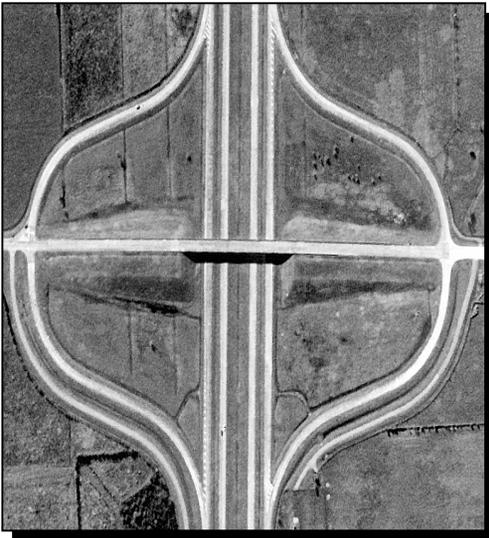
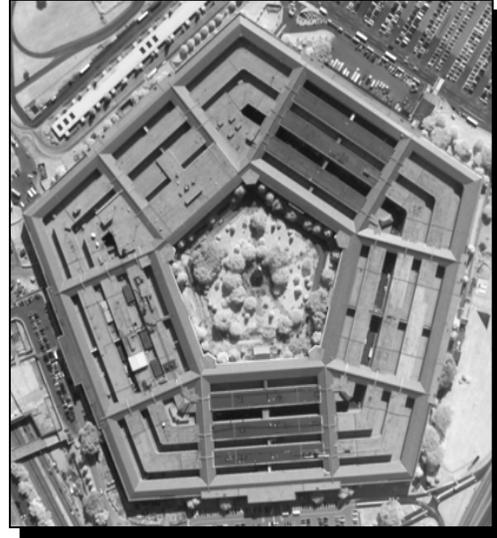
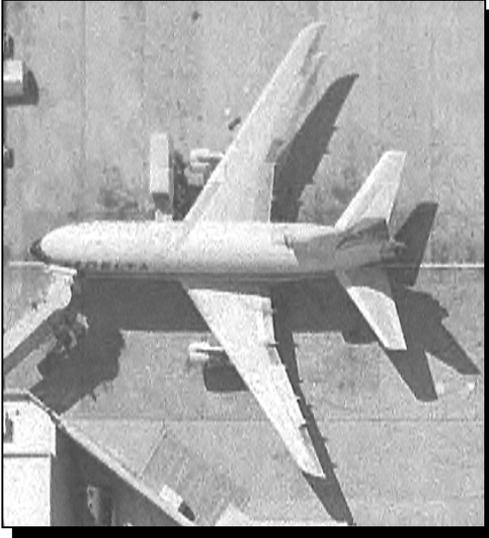
# Visual Analysis and Interpretation

- Traditionally applied to aerial photography interpretation
- May be extended to satellite imagery
  - as primary interpretation method
  - as “reality check” for computer analysis

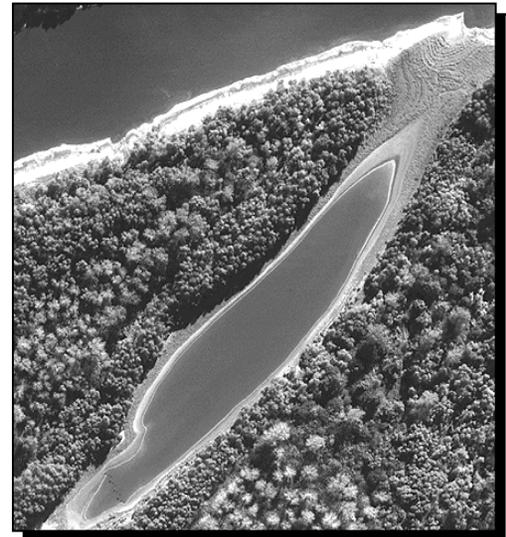
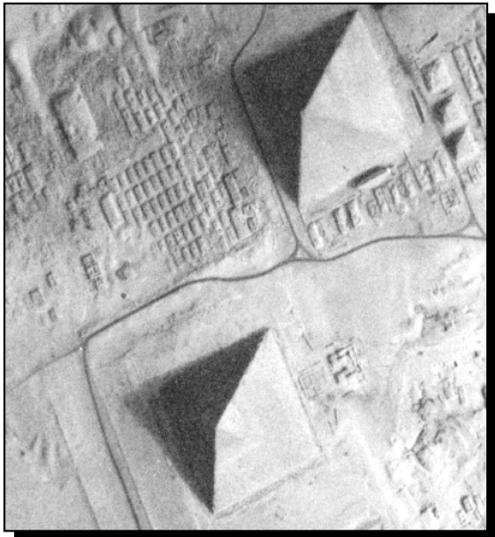
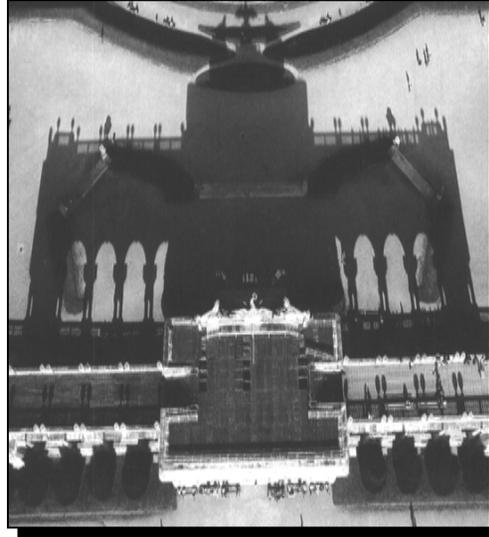
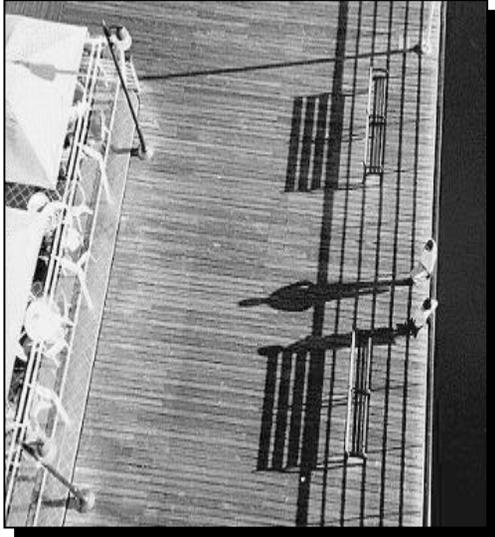
# Grand Forks, ND: April 1997



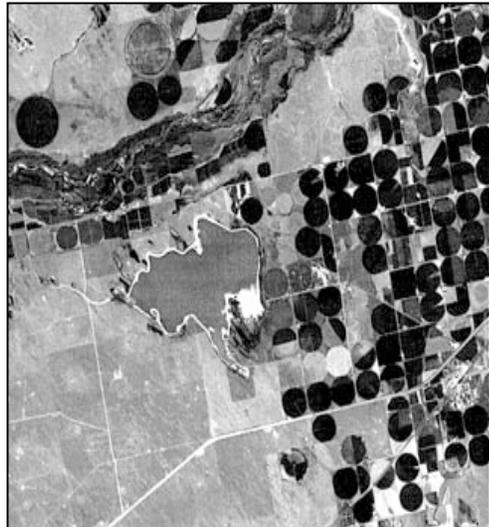
# Elements of Image Interpretation - Shape



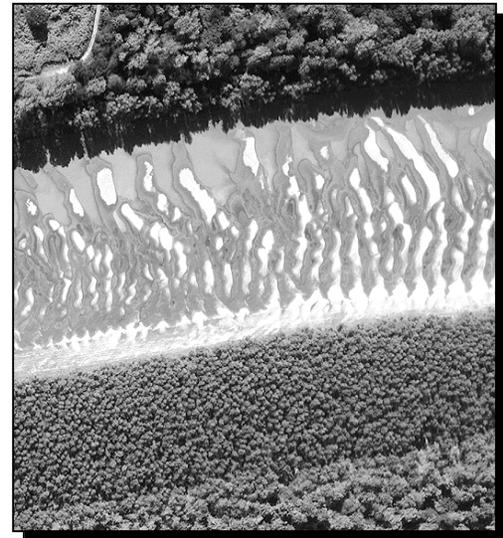
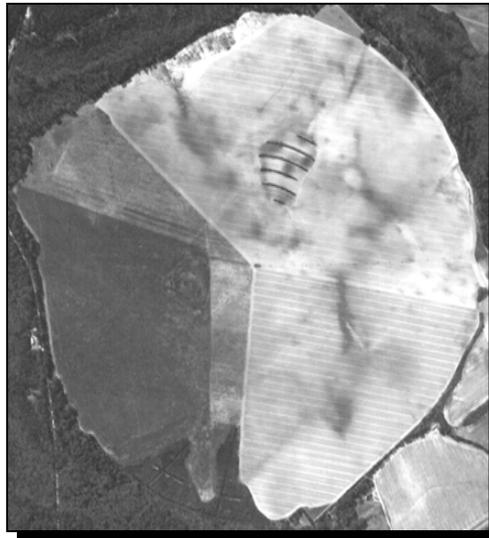
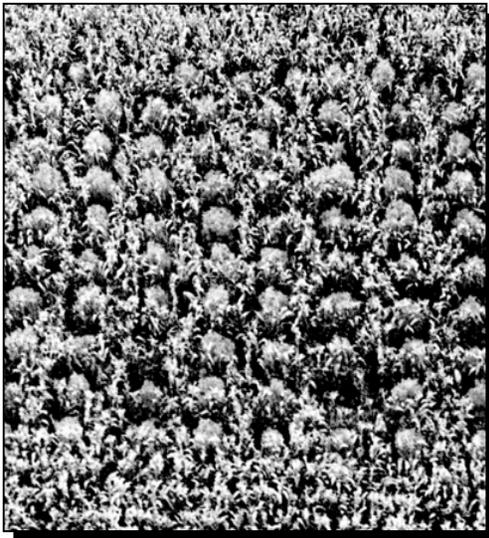
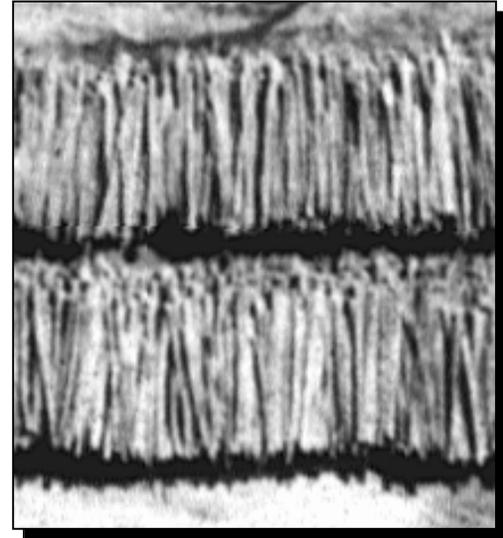
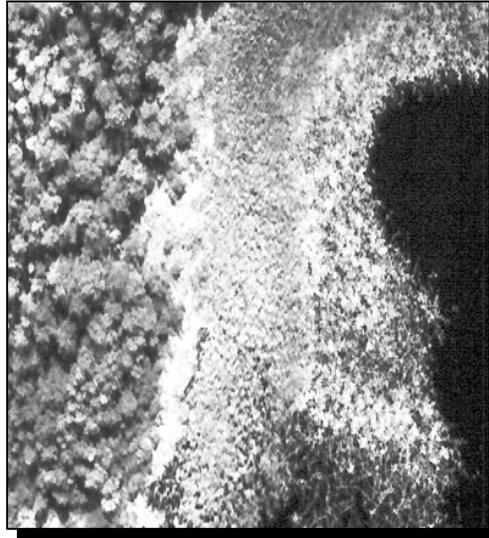
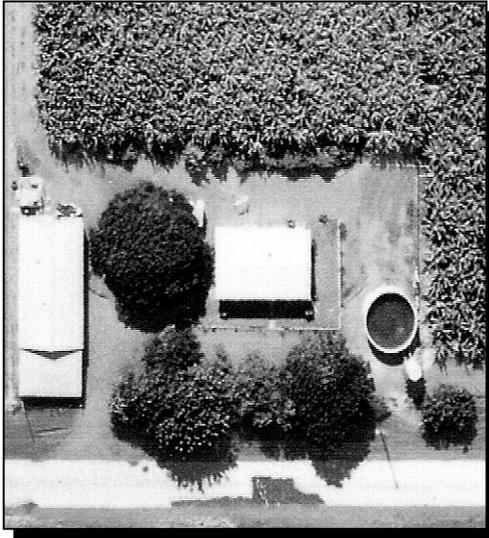
# Elements of Image Interpretation - Shadow



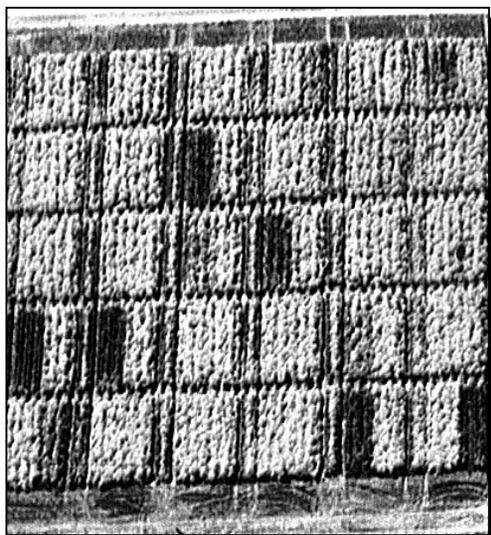
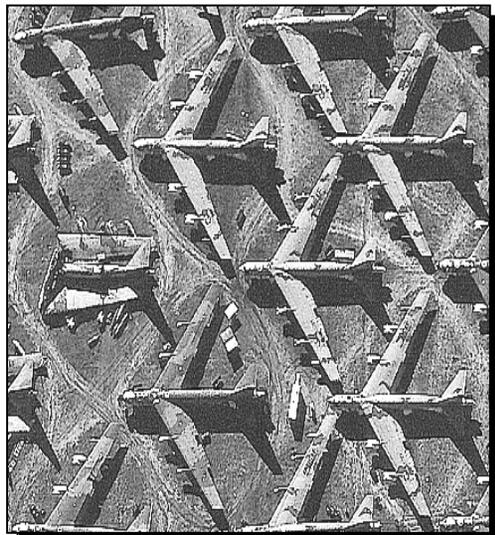
# Elements of Image Interpretation – Tone and Color



# Elements of Image Interpretation - Texture



# Elements of Image Interpretation - Pattern



# Elements of Image Interpretation - Site, Situation and Association



# Visual Interpretation

## Pop Quiz

1-meter resolution

▲ N 100 meters



Cemetery

1-meter resolution

▲  
N 100 meters

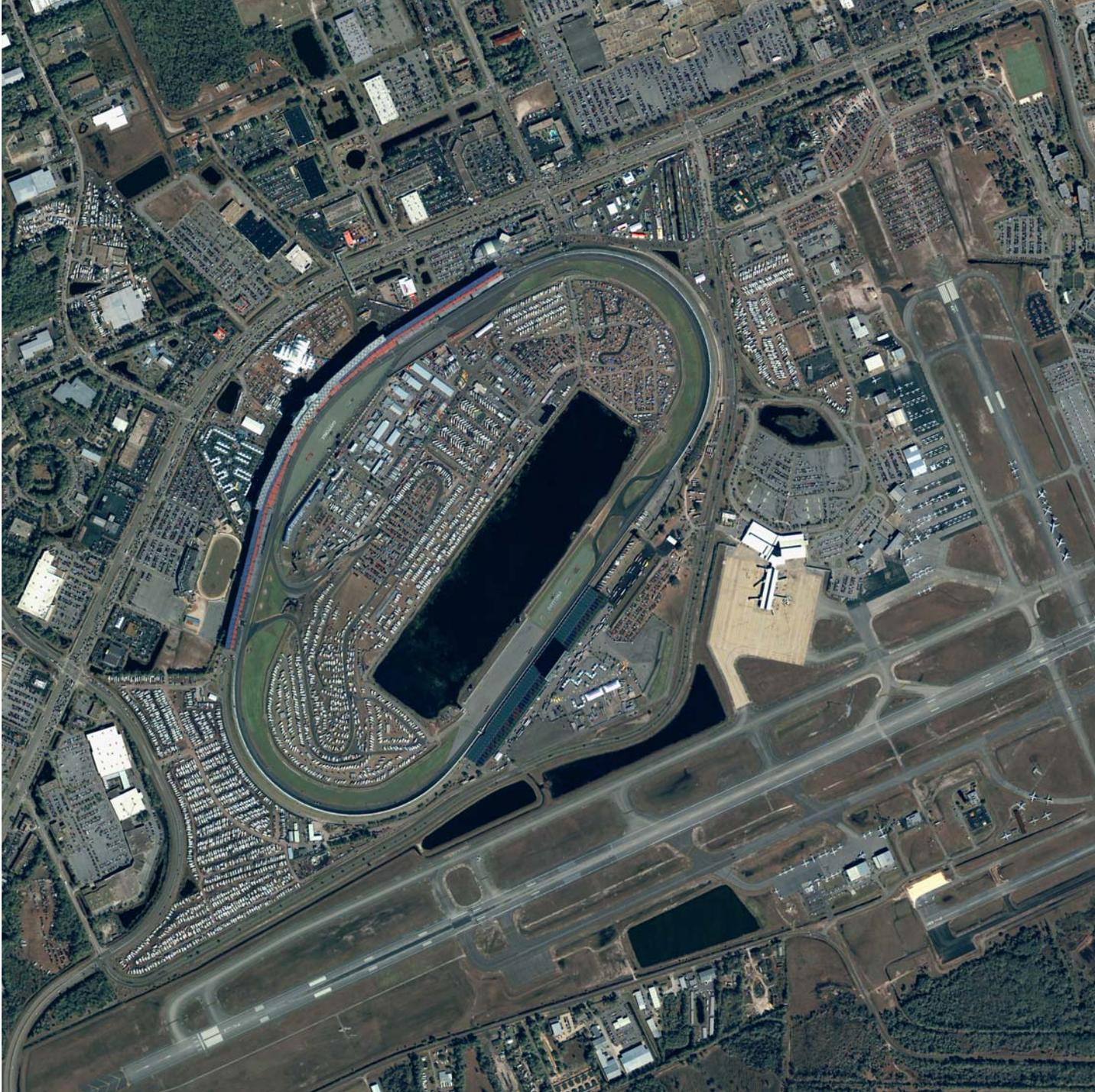


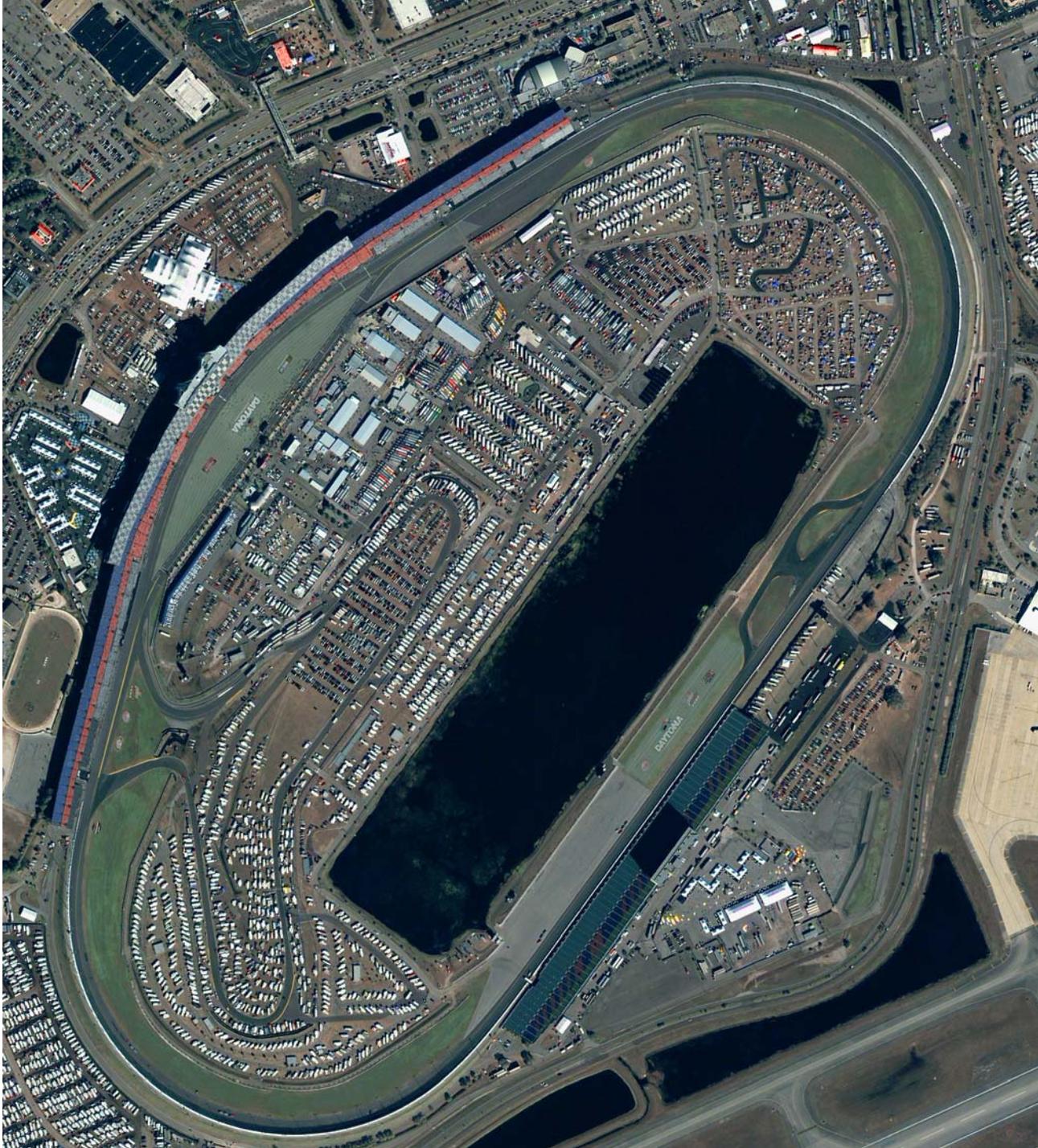
Dam

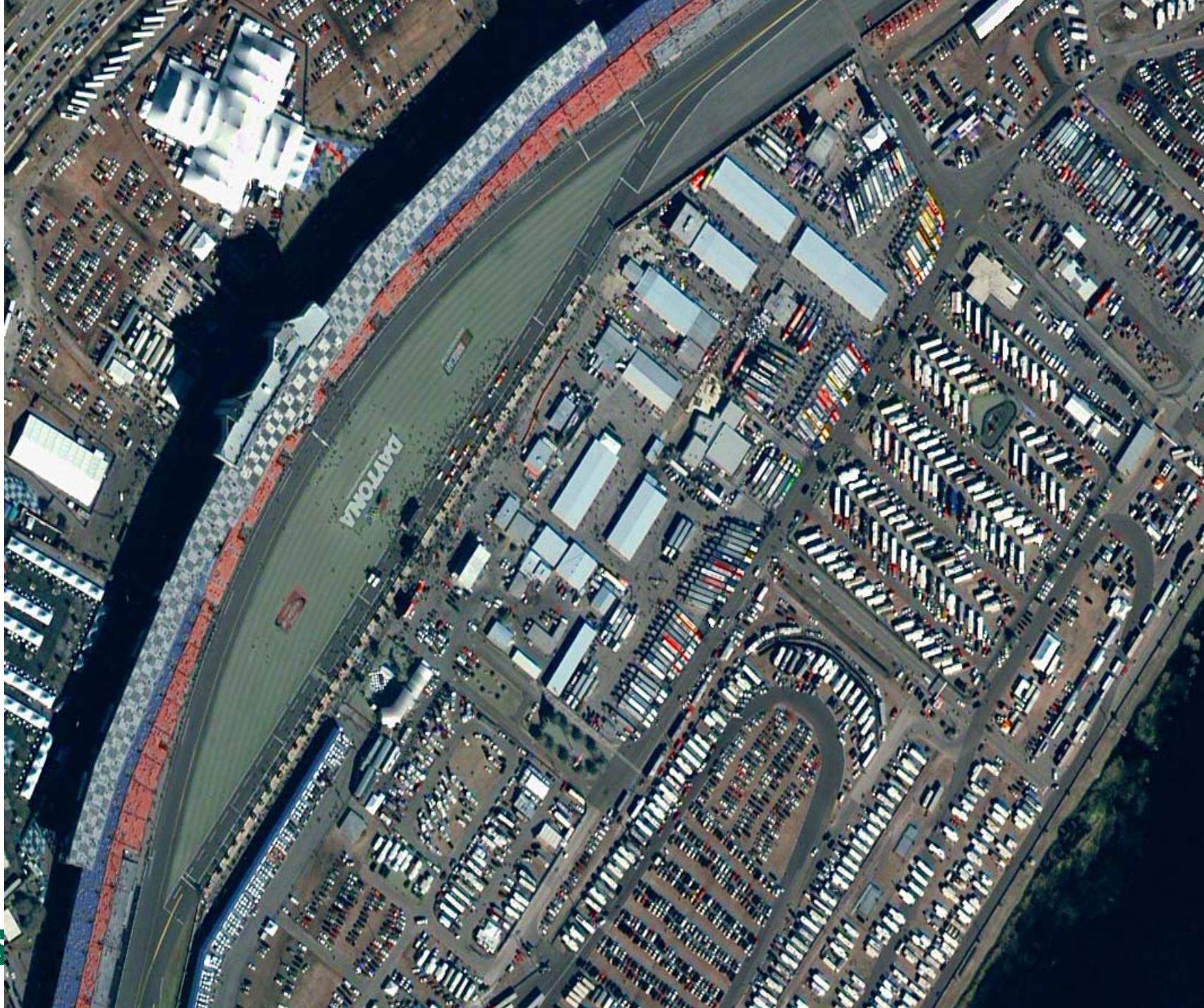
1-meter resolution



High School







# **Data Analysis and Interpretation – Typical steps**

- **Problem Definition**
- **Data Collection**
- **Data Preprocessing/Reduction**
- **Analysis/Data Integration**
- **Interpretation**
- **Discovery/Decision Support**

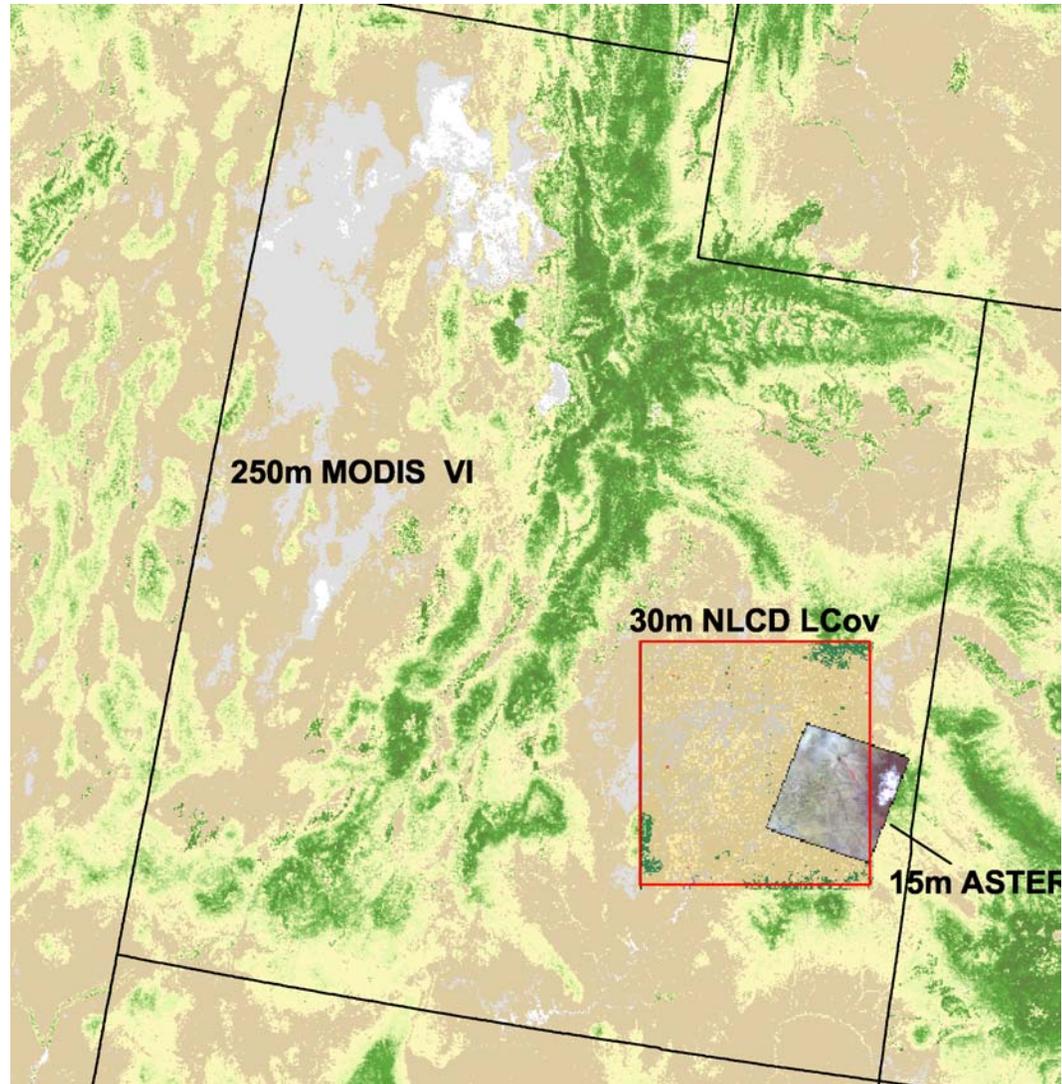
# Problem Definition

- **Clearly define what needs to be done**
- **Conceptualize your project**
- **Define tasks associated with project**
- **Set intermediate goals**

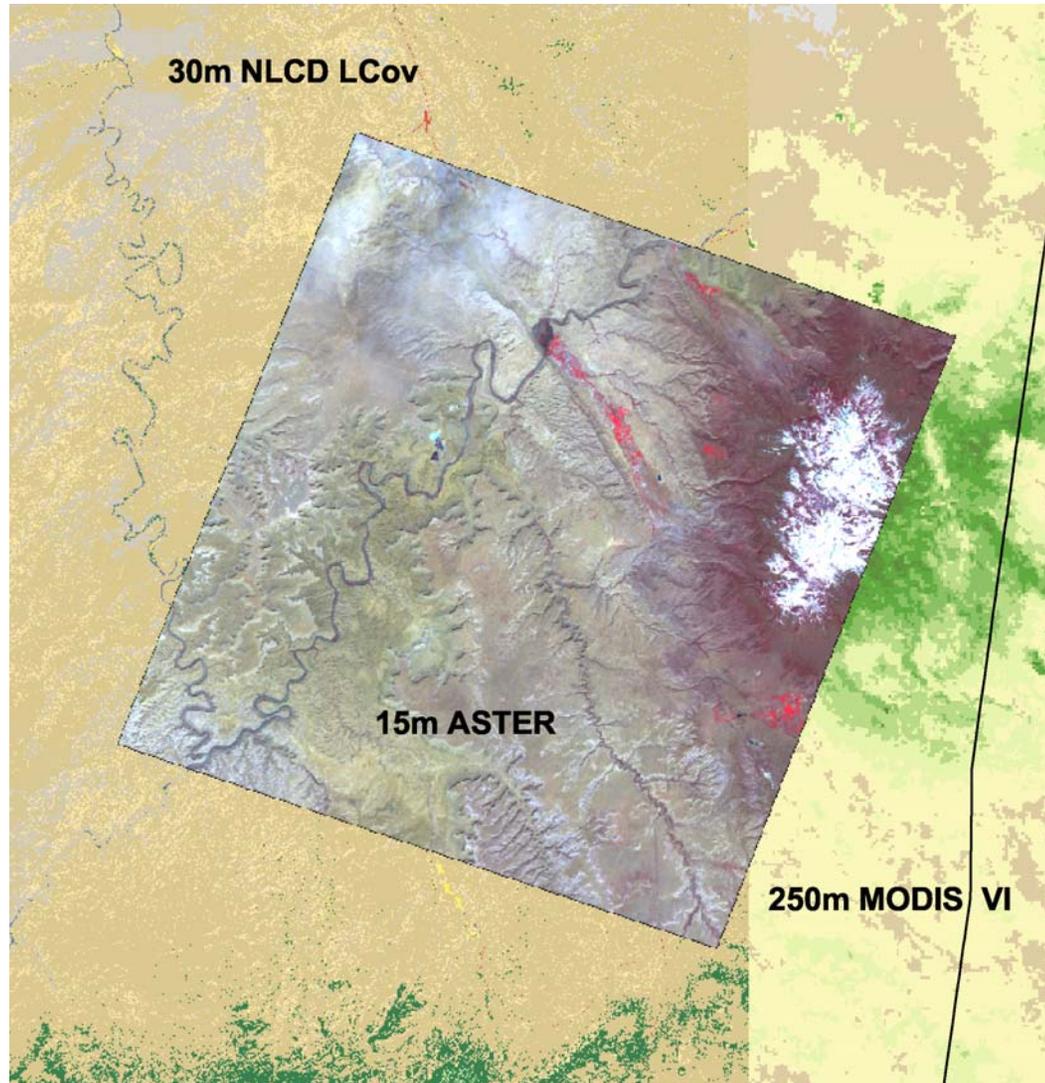
# Data Collection

- **Clearly define needs of project**
  - **Scale (IKONOS, ASTER, MODIS, etc.)**
  - **Characteristics of data (band definitions)**
  - **Availability (over study area, etc.)**
  - **Cost**
  - **Time**

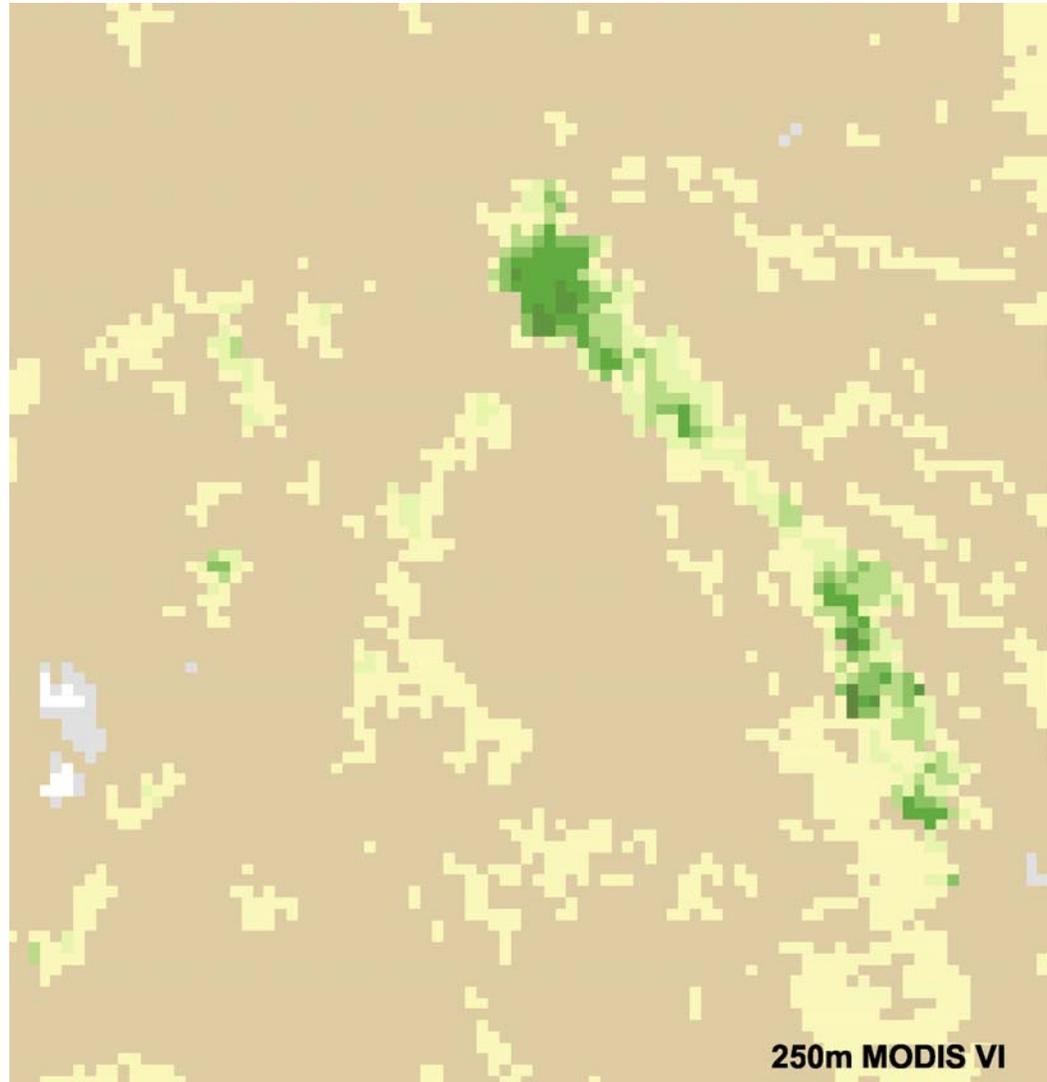
# Scale



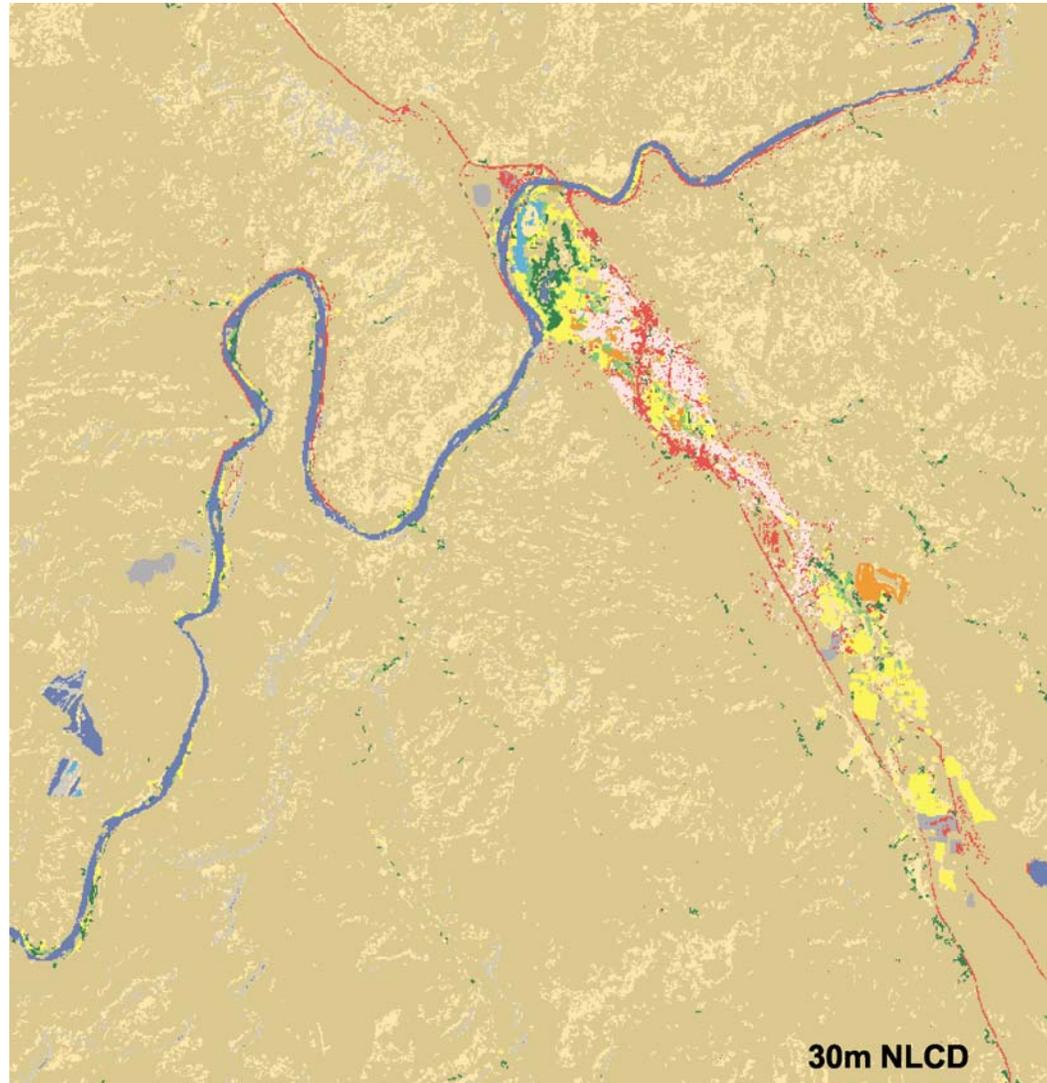
# Scale



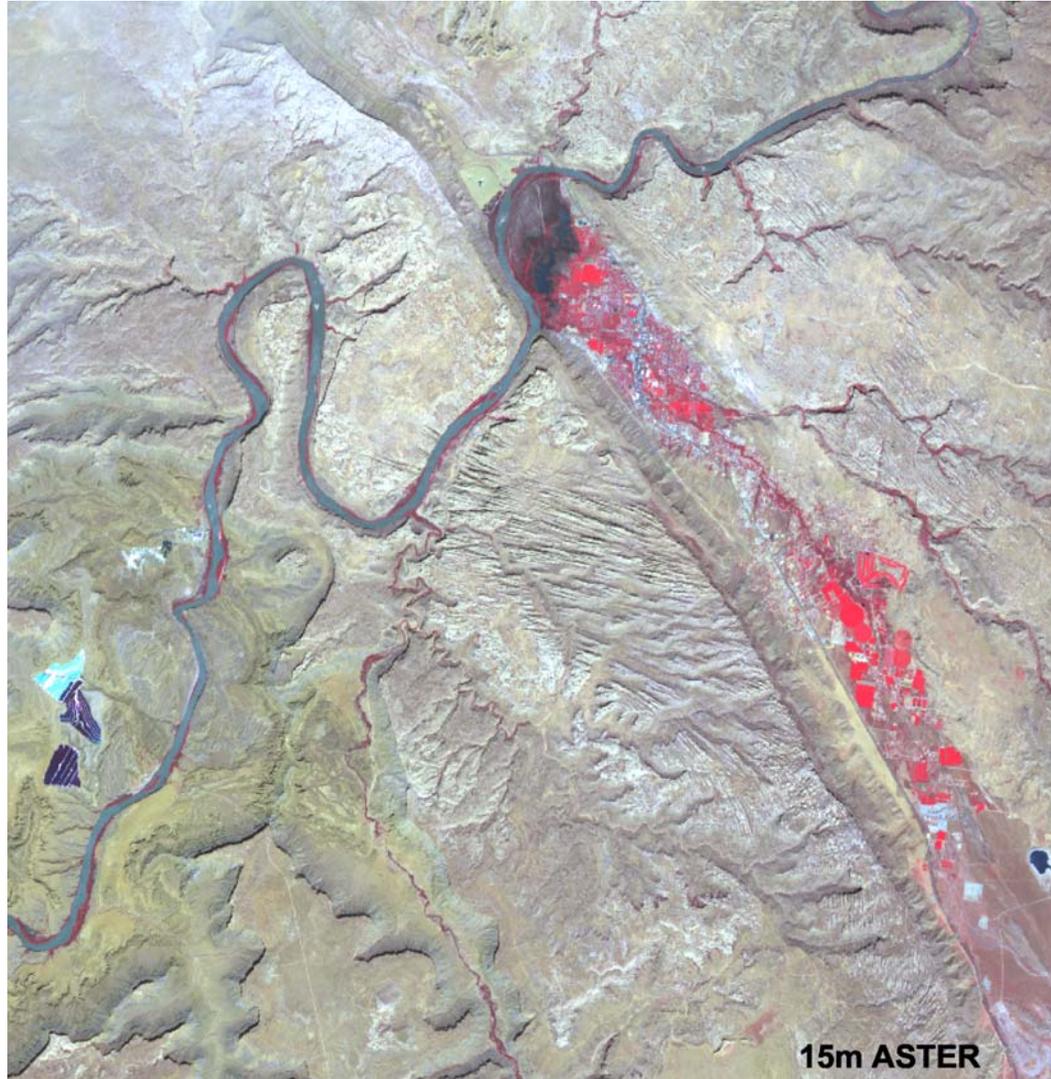
# Scale



# Scale



# Scale



# Data pre-processing

- **Data input**
  - **Media type (8mm tape, CD, DVD, DLT, ftp)**
  - **Data type: 8-bit, Integer (signed or unsigned), 32-bit**
  - **File format (generic binary, GeoTiff, HDF, HDF-EOS)**

# Media type

- **8mm tape: 5.0 Gb**
- **CD-ROM: 650-700 Mb**
- **DVD-ROM: 4.0 Gb**
- **DLT tape: 40 Gb**
- **ftp: limited by connection speed**

# Data type

- **8-bit: 0-255**
- **Integer: 65536 values**
  - **Signed: +/- 32768**
  - **Unsigned: 0-65535**
- **32-bit: 4294967296 values**
  - **Primarily used for bit-parsing in QA data**
  - **e.g. 10011000000011011001100000001101**

# Bit parsing

Bit	Description
0	adjacency correction performed;
	1 – yes
	0 – no
1	atmospheric correction performed;
	1 – yes
	0 – no
2-5	band 7 data quality four bit range;
	0000 – highest quality
	1000 – dead detector; data copied from adjacent detector
	1001 – solar zenith $\geq 86$ degrees
	1010 – solar zenith $\geq 85$ and $< 86$ degrees
	1011 – missing input
	1100 – internal constant used in place of climatological data
30-31	MODLAND QA bits;
	corrected product produced at
	00 – ideal quality all bands
	01 – less than ideal quality some or all bands
	corrected product not produced due to
	10 – cloud effects all bands
	11 – other reasons some or all bands may be fill value

# File Format

- **Generic binary: simple array with accompanying header file**
  - Almost universally accepted by software packages
- **GeoTiff: Projection information contained in file**
  - Widely accepted
  - Subtle differences in software packages

# File format - HDF

- **What is HDF?**
  - Hierarchical Data Format
  - “multi-object” file format
  - Created at the National Center for Supercomputing Applications (NCSA)
- **Why HDF?**
  - Supports common types of data
  - Efficient storage of large data sets
  - Many types of data can be included within a single HDF file (data, metadata, palettes, etc.)

# File format **HDF-EOS**

- **HDF-EOS**
  - **Specific to NASA Earth Observing System**
  - **Specific flags/data fields**

# Data Pre-processing

- **Atmospheric correction (usually already performed)**
- **Convert data to reflectance values**
- **Georeference data (remote sensing and ancillary data sets) to common reference system**
- **Data reduction**

# Atmospheric Correction

- **Performed to convert from radiance to reflectance**
- **Removes the effects of the atmosphere**
  - **Water vapor**
  - **Ozone**
  - **Atmospheric gases**
  - **Aerosols**

# Georeferencing

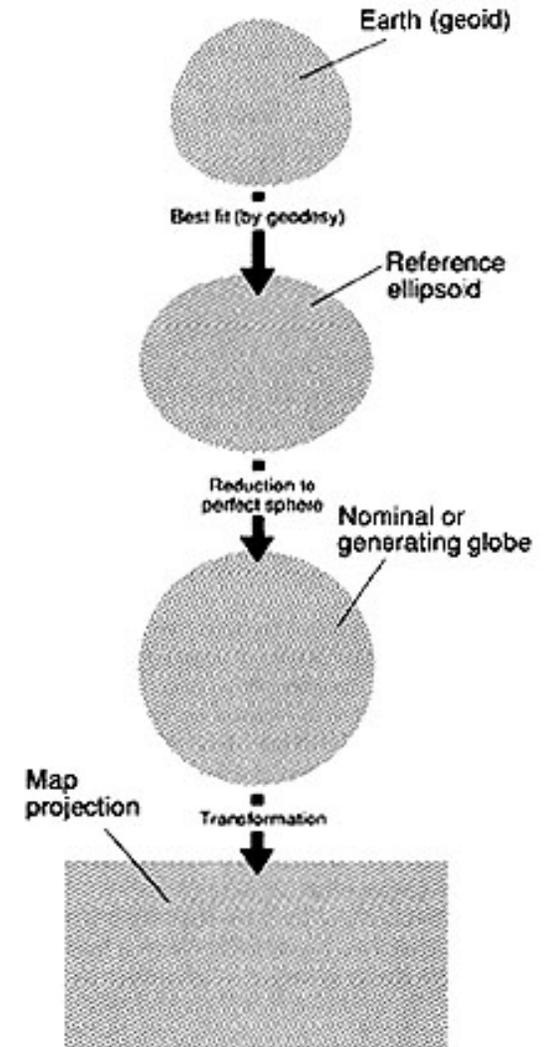
- **Reprojecting**
- **Resampling**
  - **Converts data into known reference units**
  - **Ability to work with ancillary data sets**
  - **Ability to make measurements**

# Map Projections

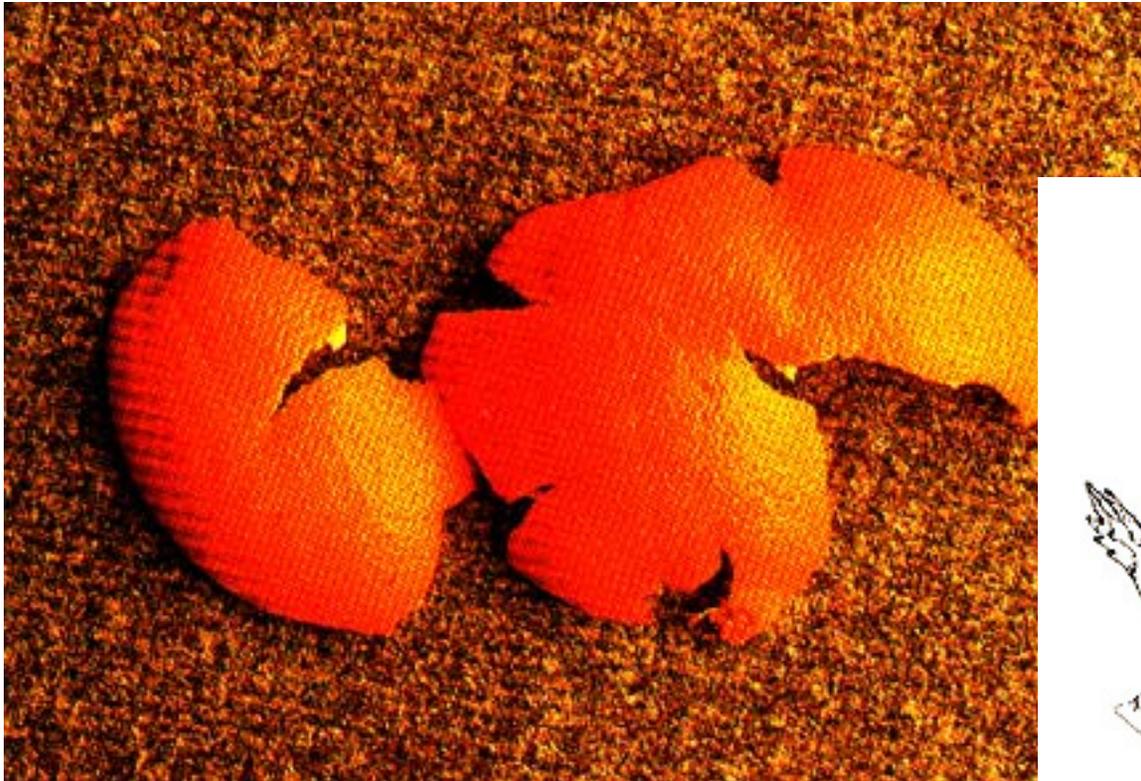
- **Why are they necessary?**
  - **Globes give the most realistic depiction of Earth, but...**
    - **Cannot see the entire Earth at once**
    - **Inconvenient**
    - **For practical use, size is a problem**
    - **Computations on a sphere are much more complex than those on a planar surface**
    - **Construction of globes is difficult**

# Map Projections

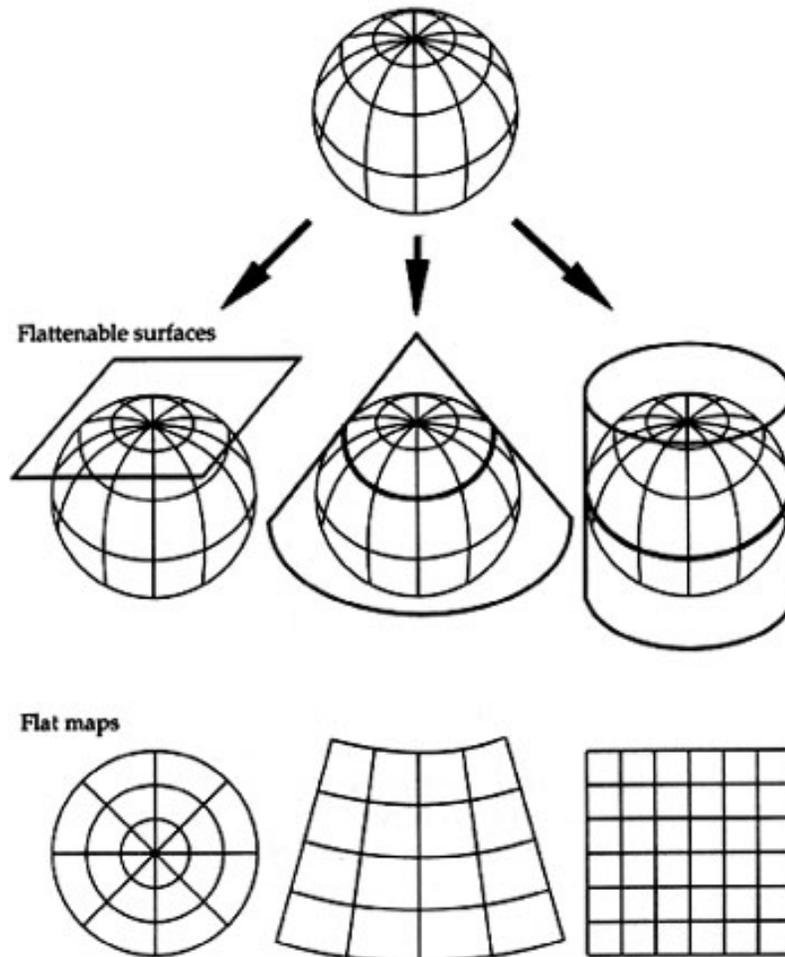
- **Globe “flattening”**
- **A series of distorting transformations is required, as are interruptions or breaks in true-Earth continuity**



# Flattening



# Types of “developable” surfaces



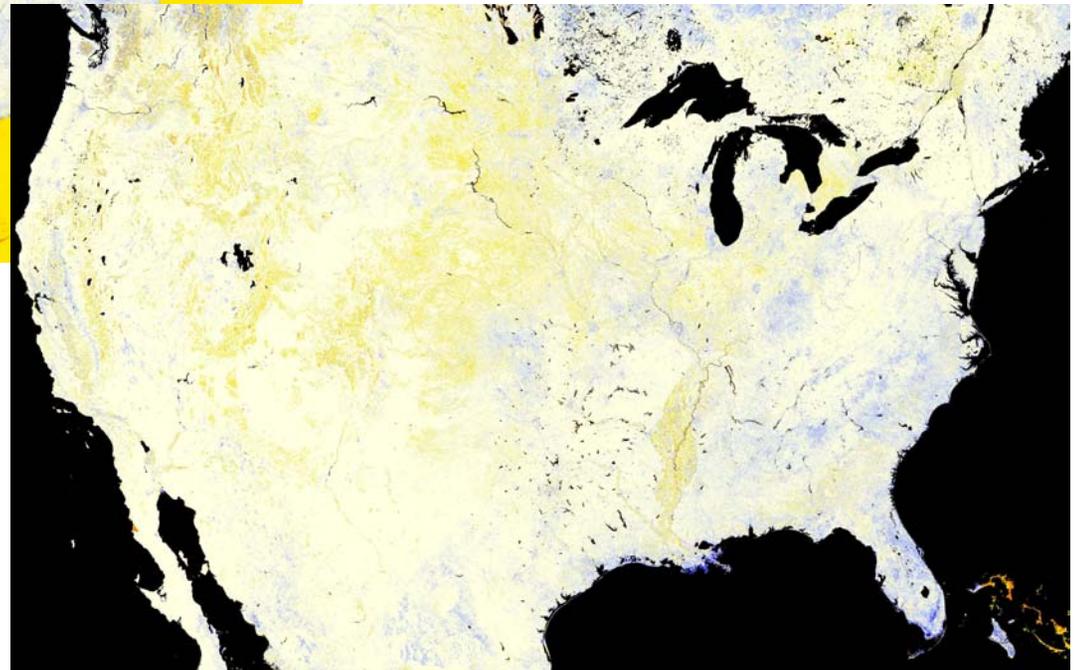
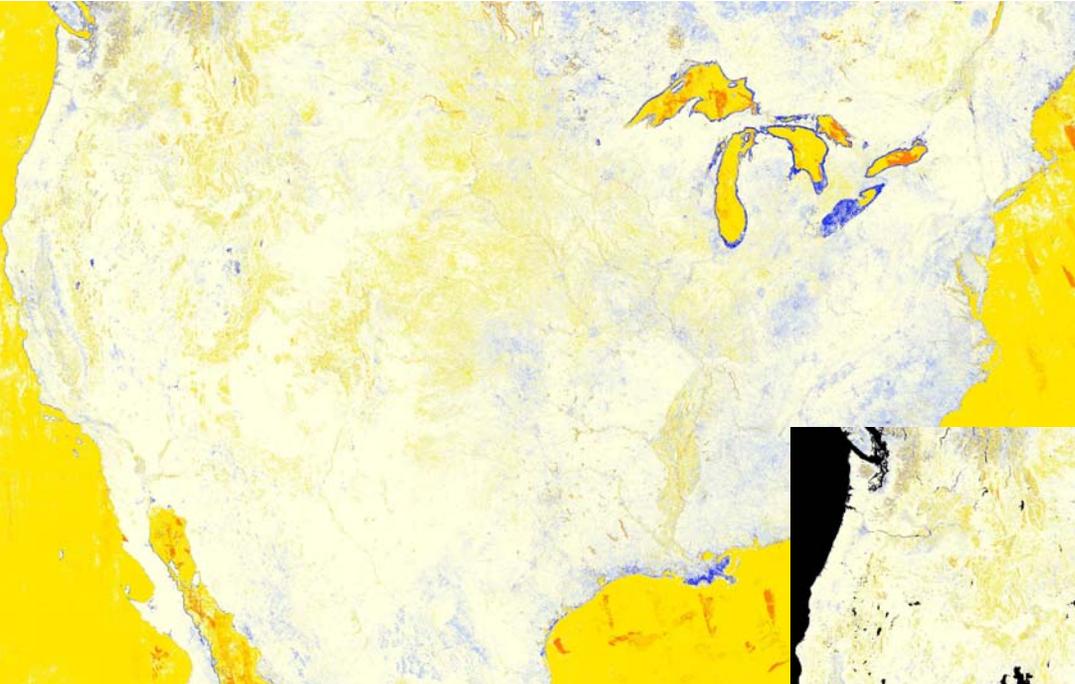
# Data Reduction

- **Mosaicking**
- **Resizing (spectrally, spatially)**
- **Masking**

# Mosaicking



# Masking



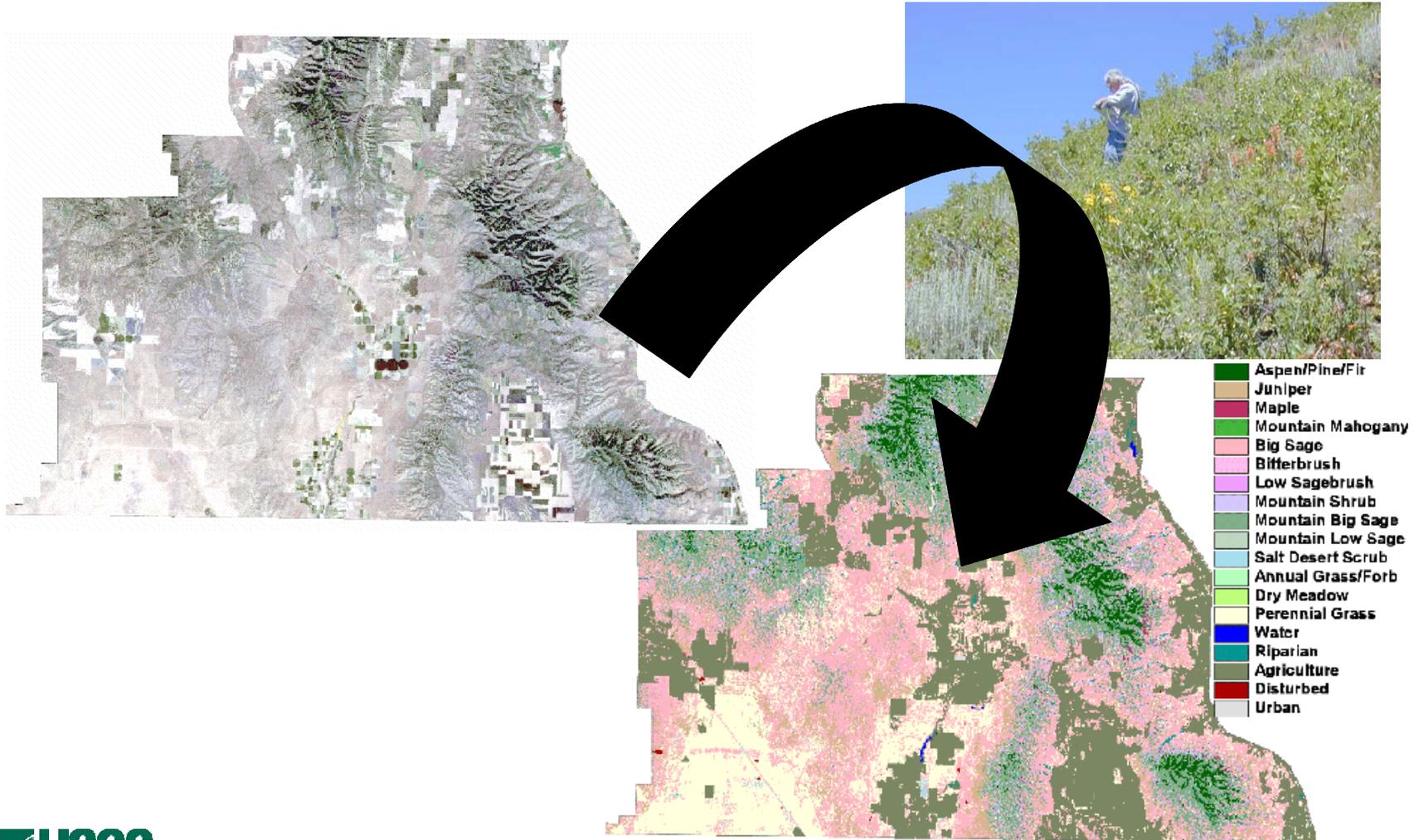
# Data Analysis and Interpretation

- **Analysis** – “the separating or breaking up of any whole into its parts”
- **Interpretation** – “the explanation of the meaning or significance of any part with respect to the whole”

# Data Analysis

- **Driven by needs of study and limited only by the skills and creativity of researcher**
- **Projects will give us experience, but as an introduction...**
  - **Classification**
  - **Band ratios**

# Thematic Information Extraction: Image Classification



# Classification

- **Supervised vs. Unsupervised**
- **Clustering Algorithms**
- **Classification Algorithms**

# Supervised Classification

- **The process is guided by the user**
- **Involves collecting training site data**
- **Applies algorithm to training sites and finds other similar regions**
- **Knowledge of field and spectral characteristics are required prior to classification**

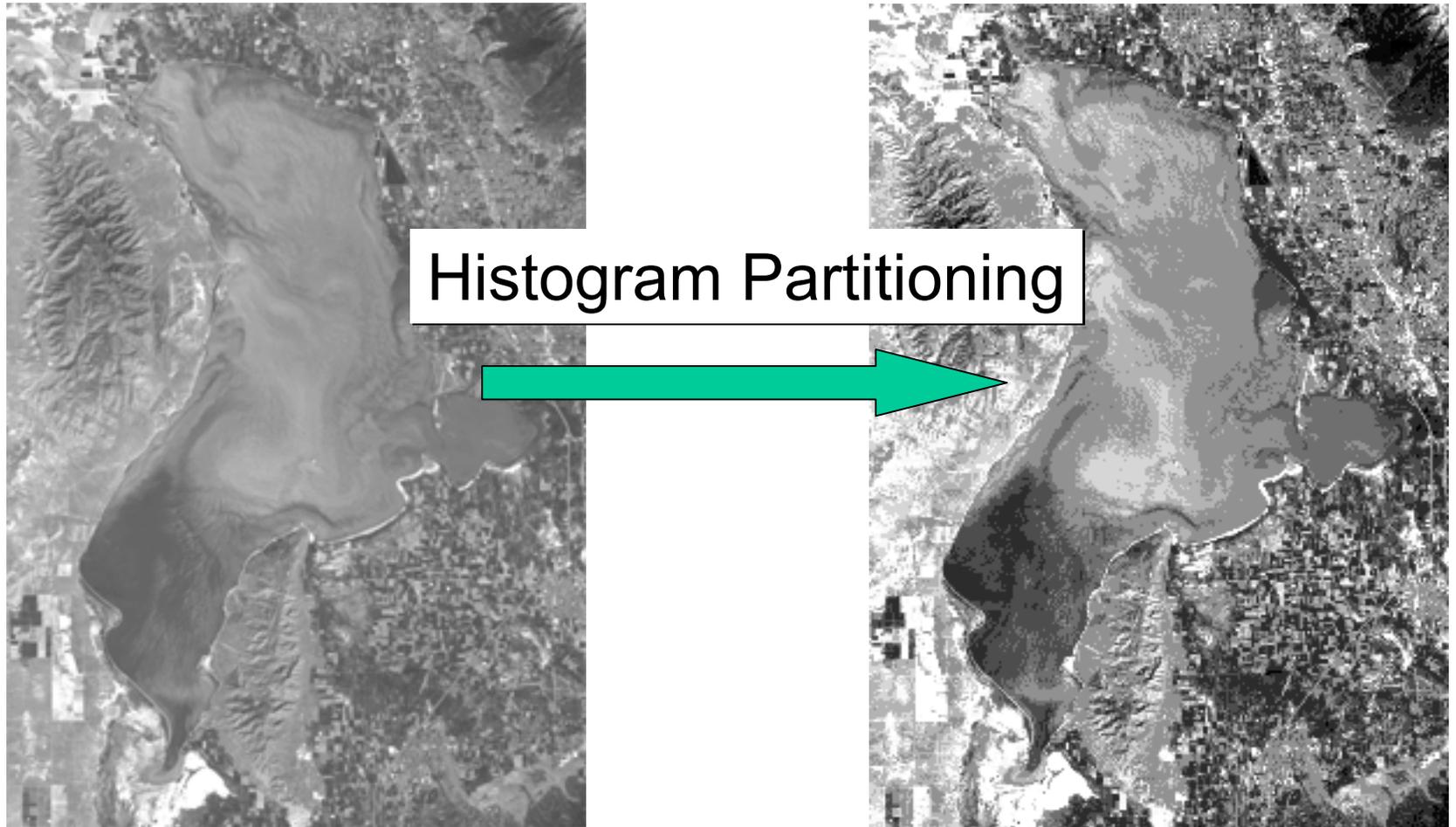
# Training Site

- **Physical location on the surface of the earth representing a specific class of interest (land cover, soil, etc.)**
- **Multiple training sites for each class are necessary to capture class variability**
- **Goal is to tie spectral response as recorded by the satellite to ground target characteristics**

# Classification Algorithms

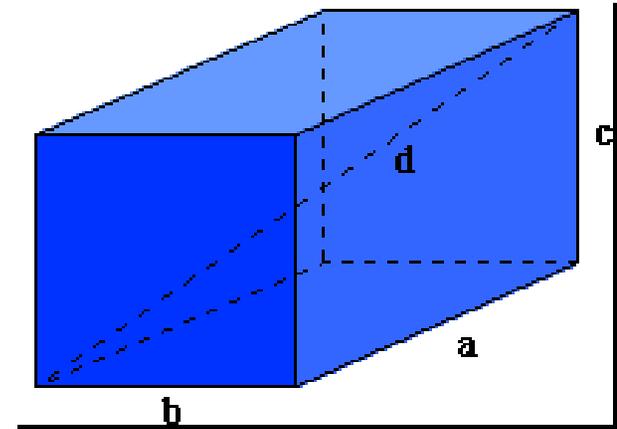
- **Density Slice (single band)**
- **Parallelepiped**
- **Minimum distance to mean**
- **Maximum likelihood**
  
- **Others...**

# Density Slice

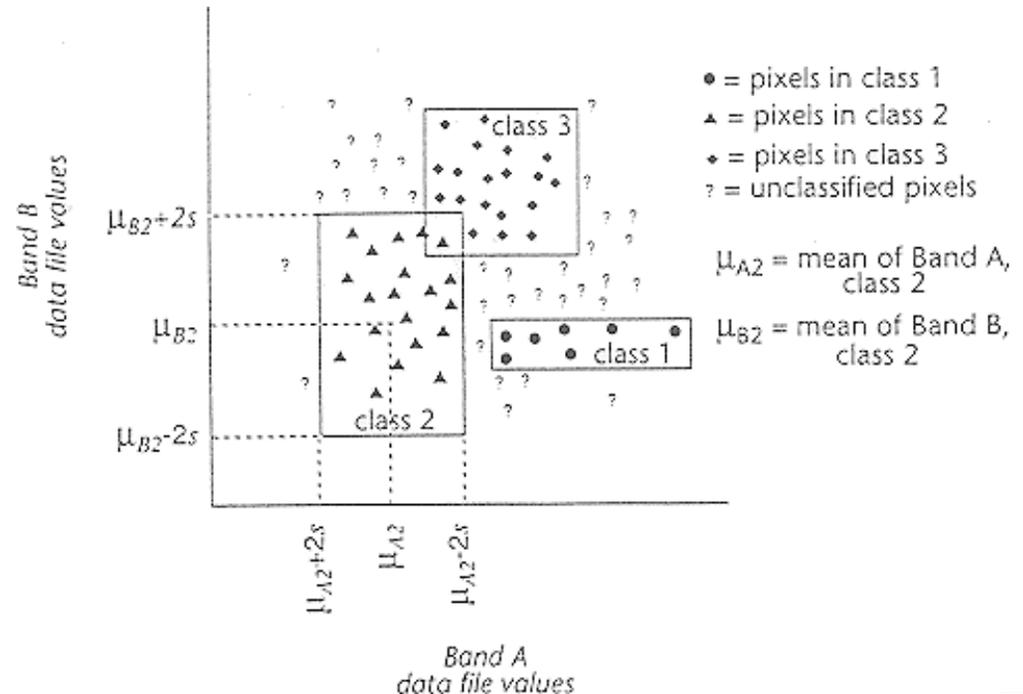


# Parallelepiped Classification

A three-dimensional figure all of whose face angles are right angles, so all its faces are rectangles and all its angles are right angles.



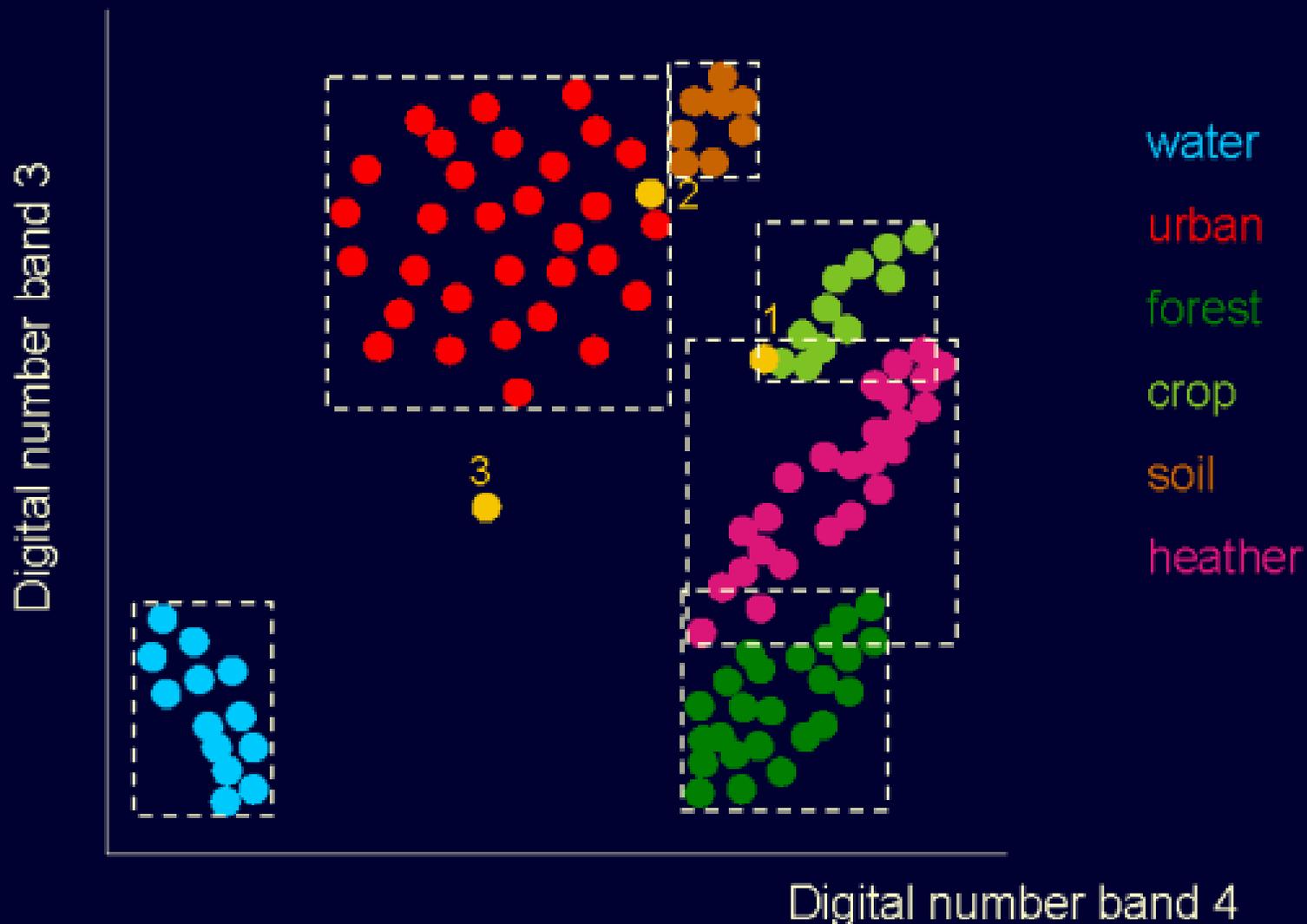
A parallelepiped classification is formed by defining the range of spectral values from each available band of the EM spectrum that define a surface feature. This forms an n-dimensional “cube” which defines the spectral class.



# Parallelepiped Classification (cont.)

- **Bounds (range of acceptable values in each band) are determined from training sets identified on the image with supporting field data.**
- **Training sets that define the spectral properties of a surface feature can be used to determine the standard deviation of the spectral information for that feature. This standard deviation can be used to determine boundaries.**
- **The algorithm tests a pixel to see if its spectral values fall within the n-dimensional bounds for each class. Each pixel is assigned to the first class that it fits.**
- **This procedure requires little computational rigor and is therefore very fast.**

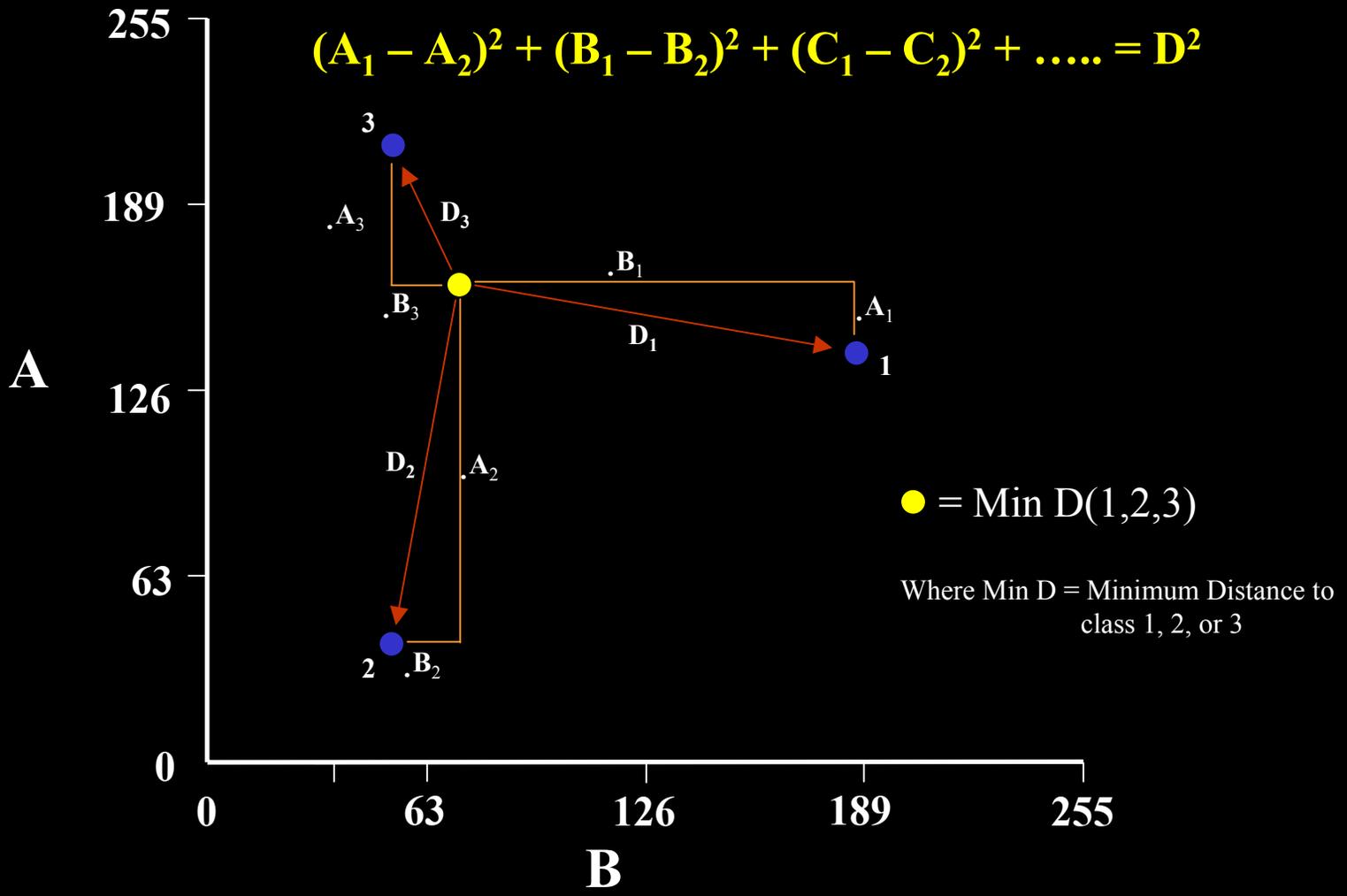
# Parallelepiped (box) classification



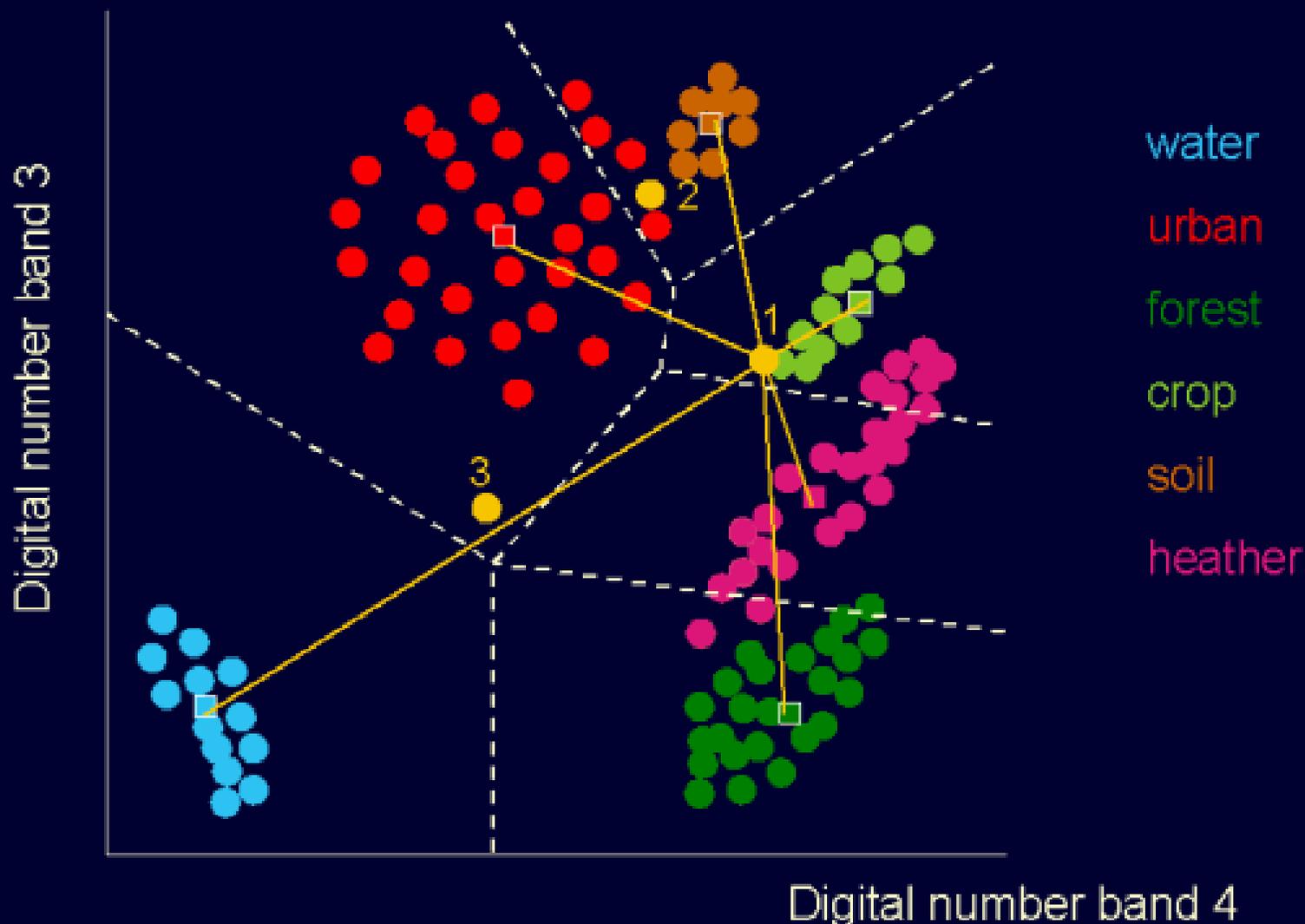
# Minimum Distance to Means

- **Similar to Parallelepiped classifier, but instead of bounding areas, the user supplies spectral class means in n-dimensional space and the algorithm calculates the distance between a candidate pixel and each class.**
- **The candidate pixel is assigned to the class with the smallest spectral Euclidian distance (minimum distance) to the candidate pixel.**
- **The distance is calculated using either an n-dimensional Pythagorean theorem, or a “Round-the-Block” measure**

$$(A_1 - A_2)^2 + (B_1 - B_2)^2 + (C_1 - C_2)^2 + \dots = D^2$$



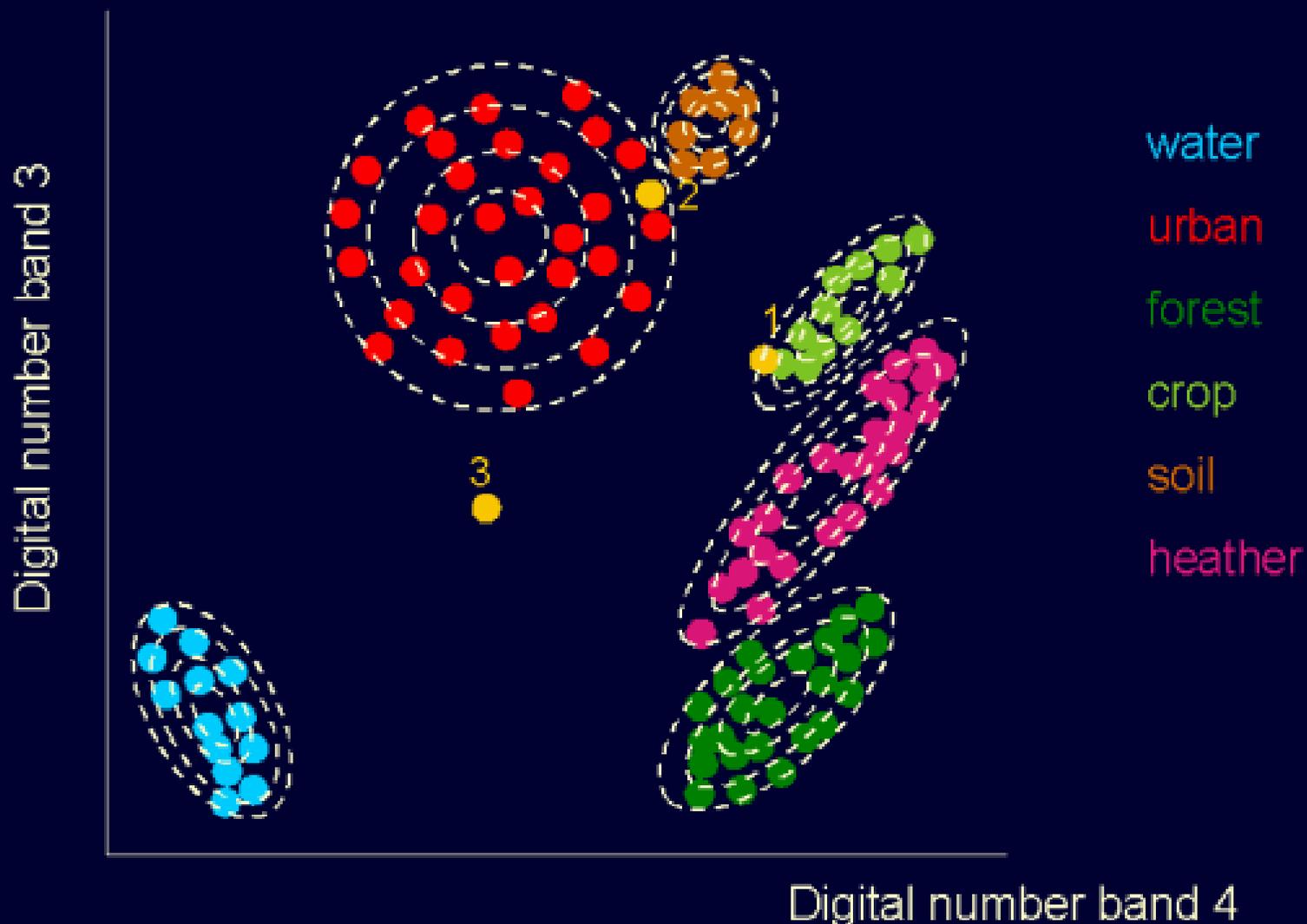
# Minimum distance to means classification



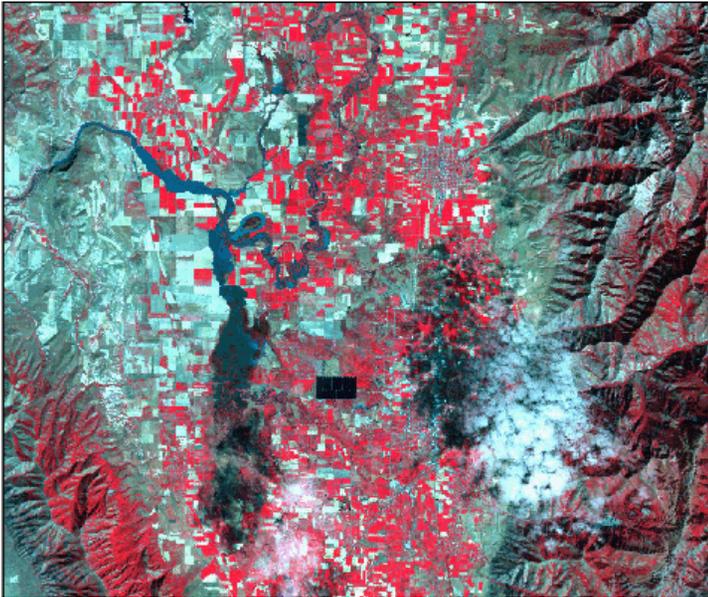
# Maximum Likelihood

- **Uses spectral class probabilities to determine class ownership of a particular pixel. Uses mean and variance and co-variance estimates for each spectral class.**
  - The probability is calculated for each class.
  - The pixel is assigned to the class with the largest probability
- **Therefore, if MDM uses  $D_{ab}$  as the measure for association, Maximum Likelihood uses  $P_{ab}$  which is the probability of pixel 'b' belonging to class 'a'**
- **Assumes that the statistics for each spectral class have a Gaussian (normal) distribution.**
- **Spectral classes with bi- or tri-modal distributions in any of the n bands imply that more than one ground class is represented in the training data.**

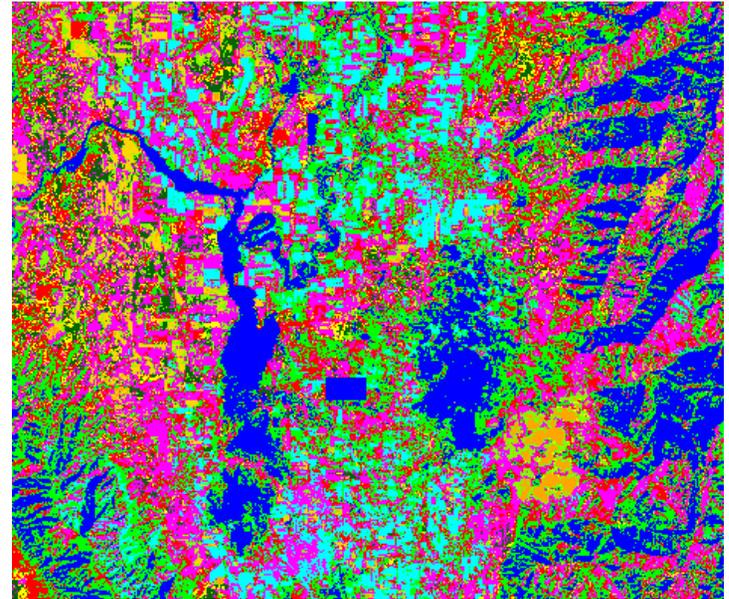
# Maximum likelihood classification



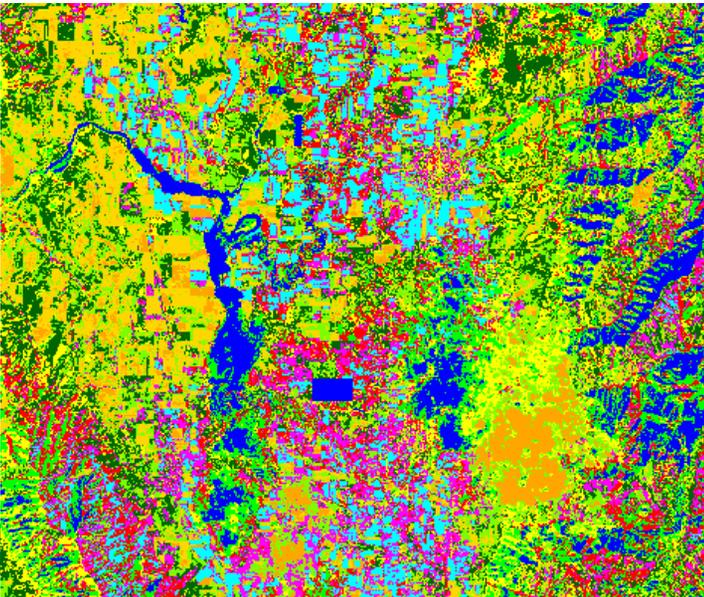
Original Landsat TM



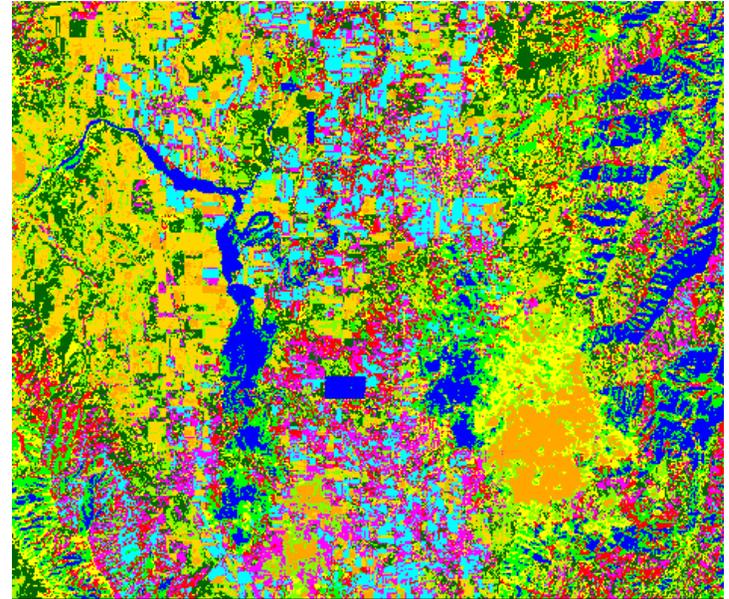
Parallelepiped Classification



Minimum Distance Classification



Maximum Likelihood Classification



# Unsupervised classification

- The unsupervised classification process generates “natural” clusters of spectral data in multi-dimensional space.
- The user has little input and “allows” the computer to generate spectral means, variances, and co-variances from the available imagery.
- Spectral clusters are assigned to ground based informational classes using *a posteriori* information.
- Spectral clusters will tend to better represent the range of variation present in the image as contrasted with the supervised approach.
- Spectral clusters may not be an exact match to informational classes that the user desires.

# Unsupervised classification (cont.)

- Unsupervised classification is usually a “two-pass” process of:
  1. Generating Spectral Clusters
  2. Minimum Distance to Means classifier (or other as desired)
- The first pass (cluster generation), can be a single stage operation of identifying spectral clusters from a sample of data, or it can be an iterative operation that generates clusters, and re-evaluates cluster assignments during repetitive steps.

# Isodata clustering

- **ISODATA – Iterative self-organizing data analysis technique**
- The ISODATA clustering method uses Euclidian distance in n-dimensional feature space to iteratively generate spectral clusters.
- This is an unsupervised classification approach that attempts to minimize spatial bias by utilizing feature space and not “image space” to generate spectral clusters.
- Initial spectral clusters are allocated (seeded) within n-dimensional feature space according to the standard deviation distance from the central mean.

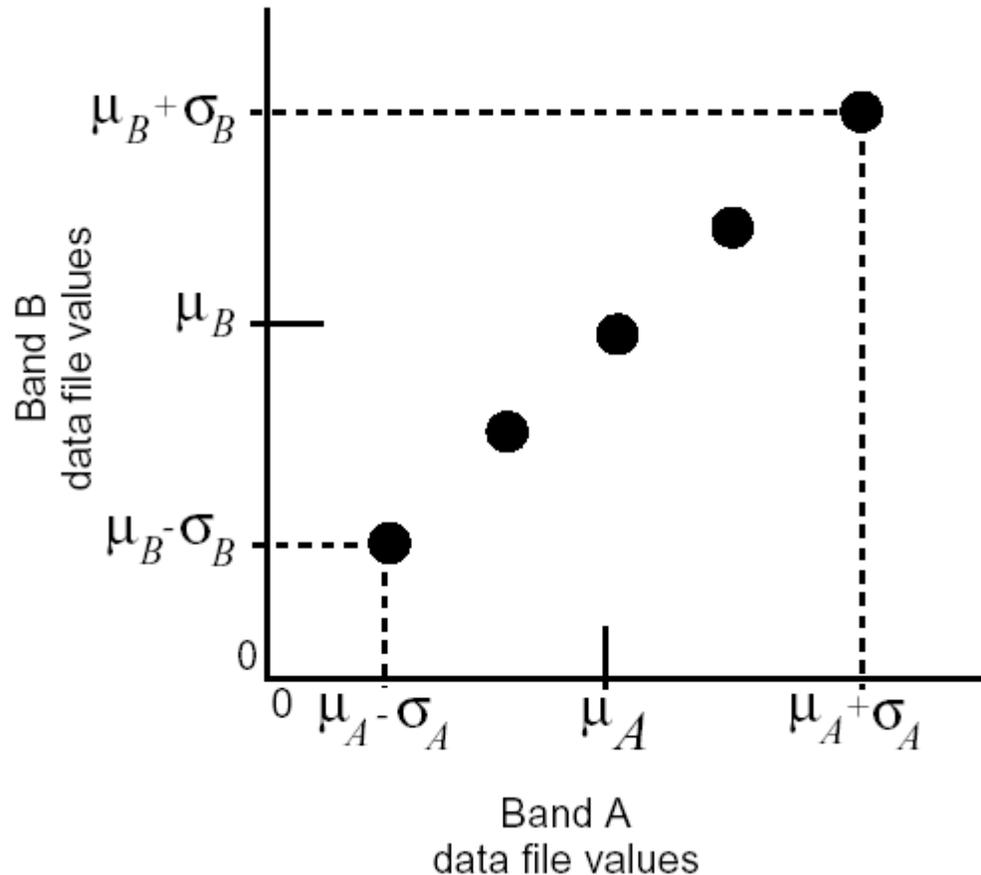
# ISODATA clustering

To perform ISODATA clustering, the user specifies:

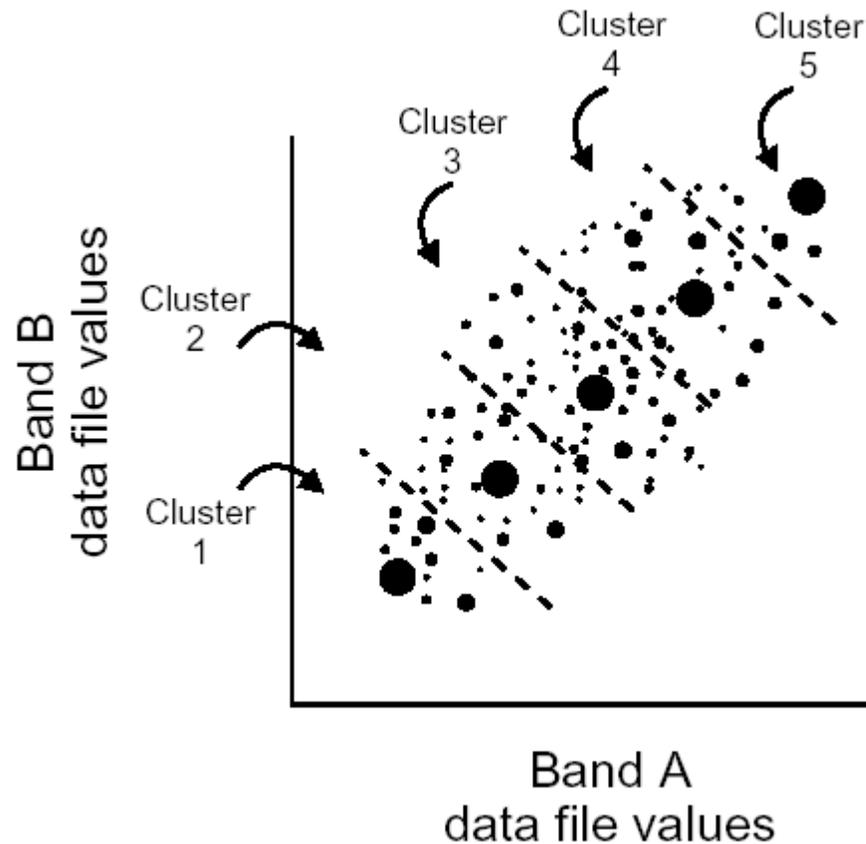
- **$N$**  - the maximum number of clusters to be considered. Since each cluster is the basis for a class, this number becomes the maximum number of classes to be formed. The ISODATA process begins by determining  $N$  arbitrary cluster means. Some clusters with too few pixels can be eliminated, leaving less than  $N$  clusters.
- **$T$**  - a convergence threshold, which is the maximum percentage of pixels whose class values are allowed to be unchanged between iterations.
- **$M$**  - the maximum number of iterations to be performed.

# Initial ISODATA Clusters

The initial cluster means are evenly distributed between  $(\mu_A - \sigma_A, \mu_B - \sigma_B)$  and  $(\mu_A + \sigma_A, \mu_B + \sigma_B)$ .

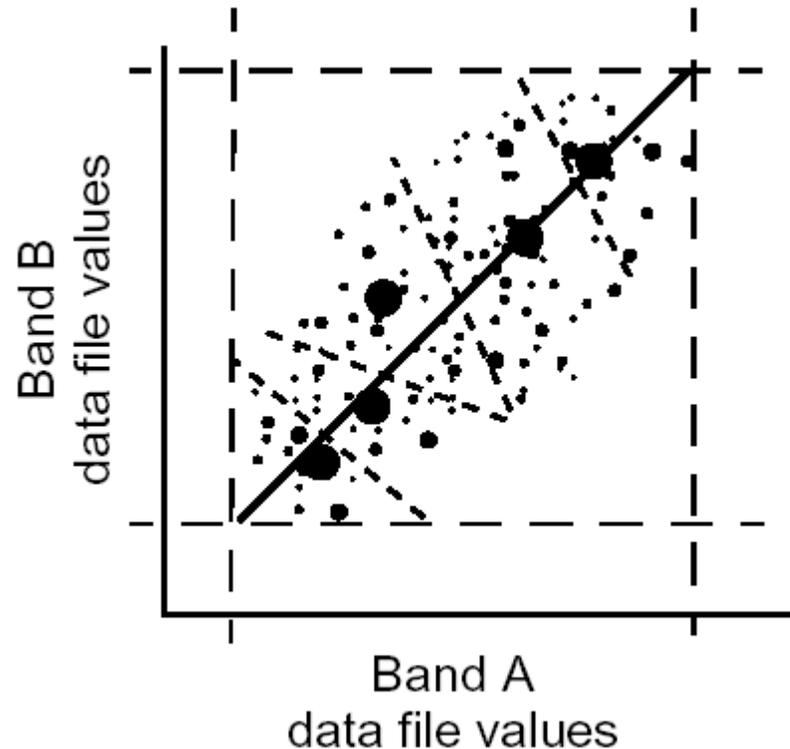


# ISODATA First Pass



Pixels are assigned to each initial spectral clusters based on the minimum distance. As pixels are assigned to clusters, the cluster mean is recalculated to include the influence of that pixel. As means are recalculated, the cluster position moves in feature space.

# ISODATA Second Pass



For the second iteration, the means of all clusters are recalculated, causing them to shift in feature space. The entire process is repeated—each candidate pixel is compared to the new cluster means and assigned to the closest cluster mean.

# ISODATA

## Post-clustering Assignment

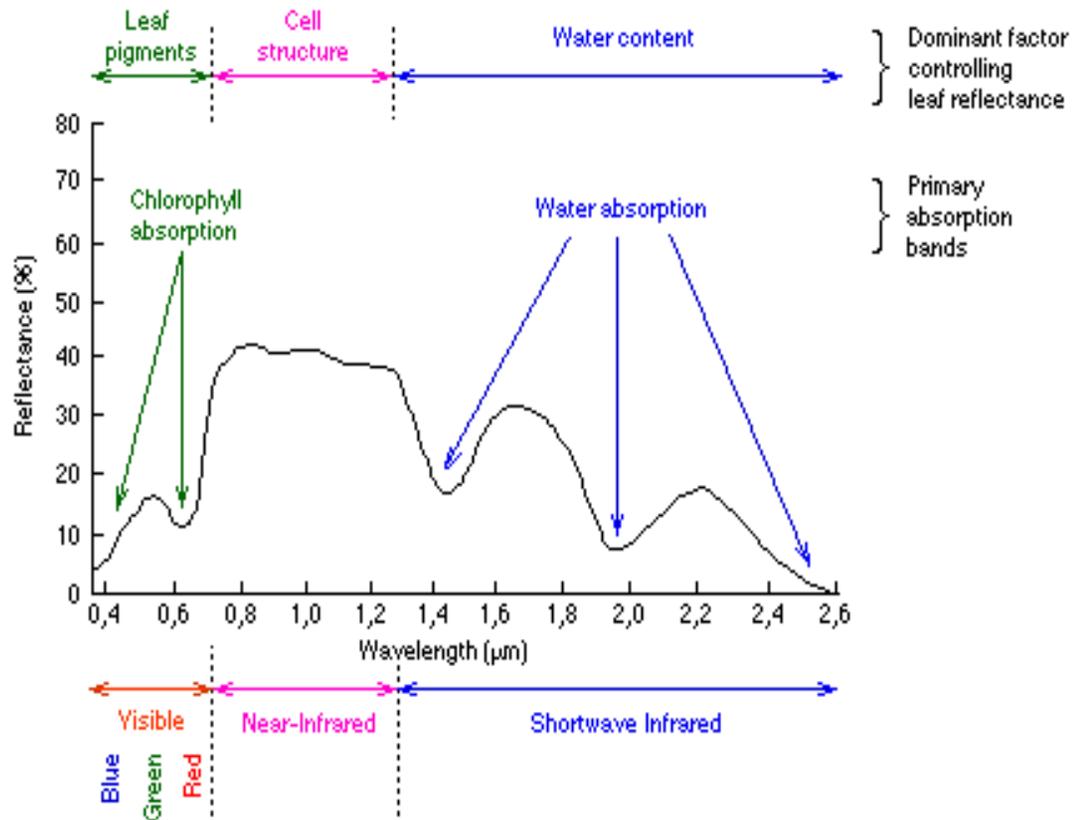
- Clusters are assigned to an information class based on available ancillary information (i.e., field reference data, maps, aerial photos, analyst experience)
- If clusters are confused between two or more information classes it is assigned to a “mixed” class
- Clusters are often split or merged

# ISODATA pros and cons

- **Pros**
  - Not geographically biased to any particular part of the image
  - Successful at finding inherent spectral clusters
- **Cons**
  - Need to know *a priori* number of classes
  - May be time consuming

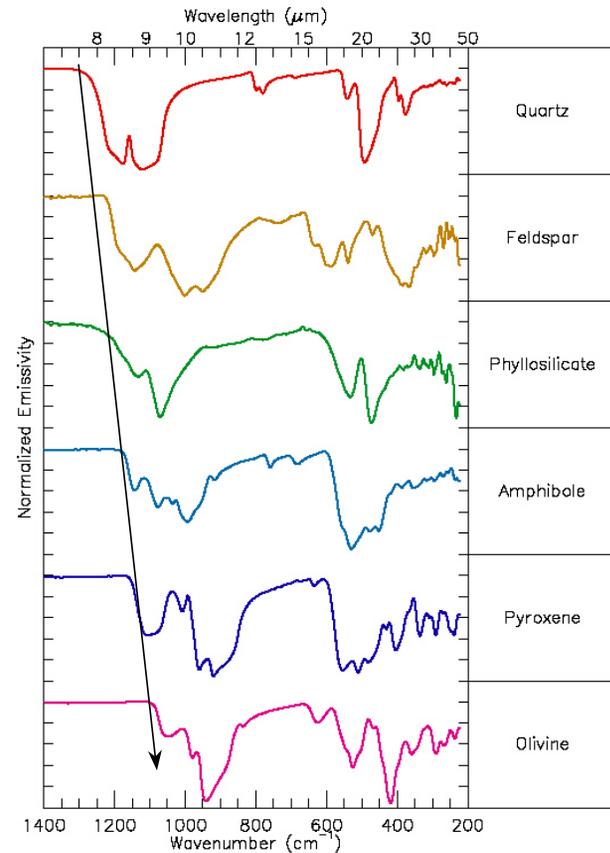
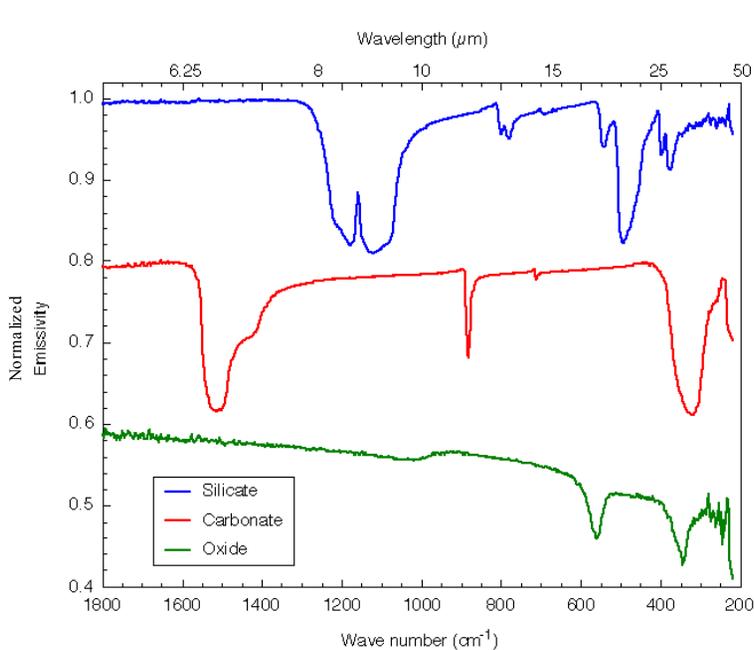
# Band Ratios

- **Enhances spectral differences between bands**
- **Usually, simply dividing one spectral band by another produces relative intensities**
- **Can create complex ratios (e.g., vegetation indices) by dividing mathematical equations or by multiple ratios**

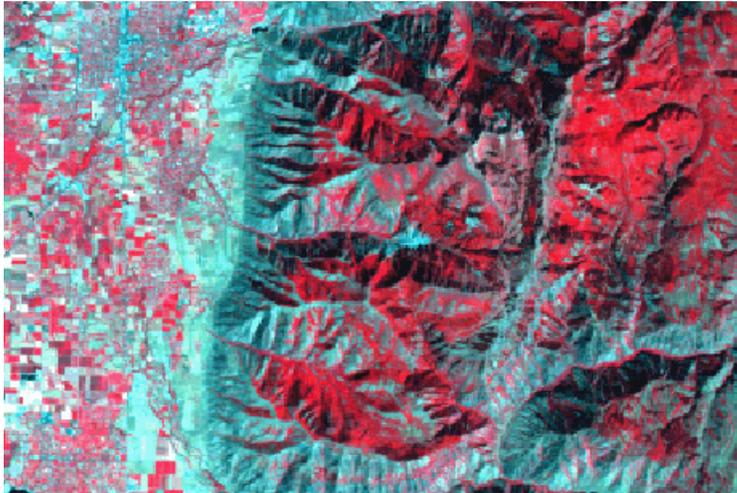


**Typical spectral response characteristics of green vegetation showing the spectral effects of leaf pigments, cell structure, and water content (Hoffer, 1978)**

# Emissivity Spectra of Some Rock Forming Minerals



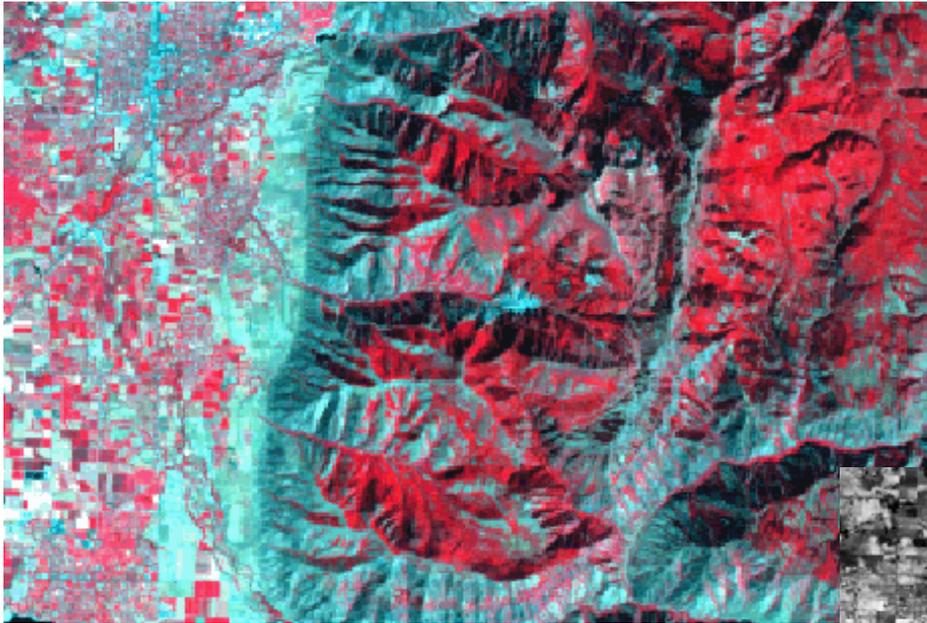
# Image Processing: Spectral Indices



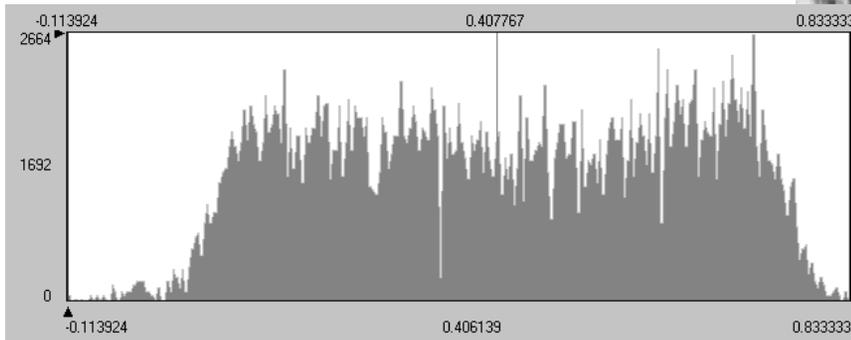
**Spectral indices are designed to convert spectral reflectance into biophysical information that can be interpreted directly by a user.**



# Normalized Difference Vegetation Index (NDVI)

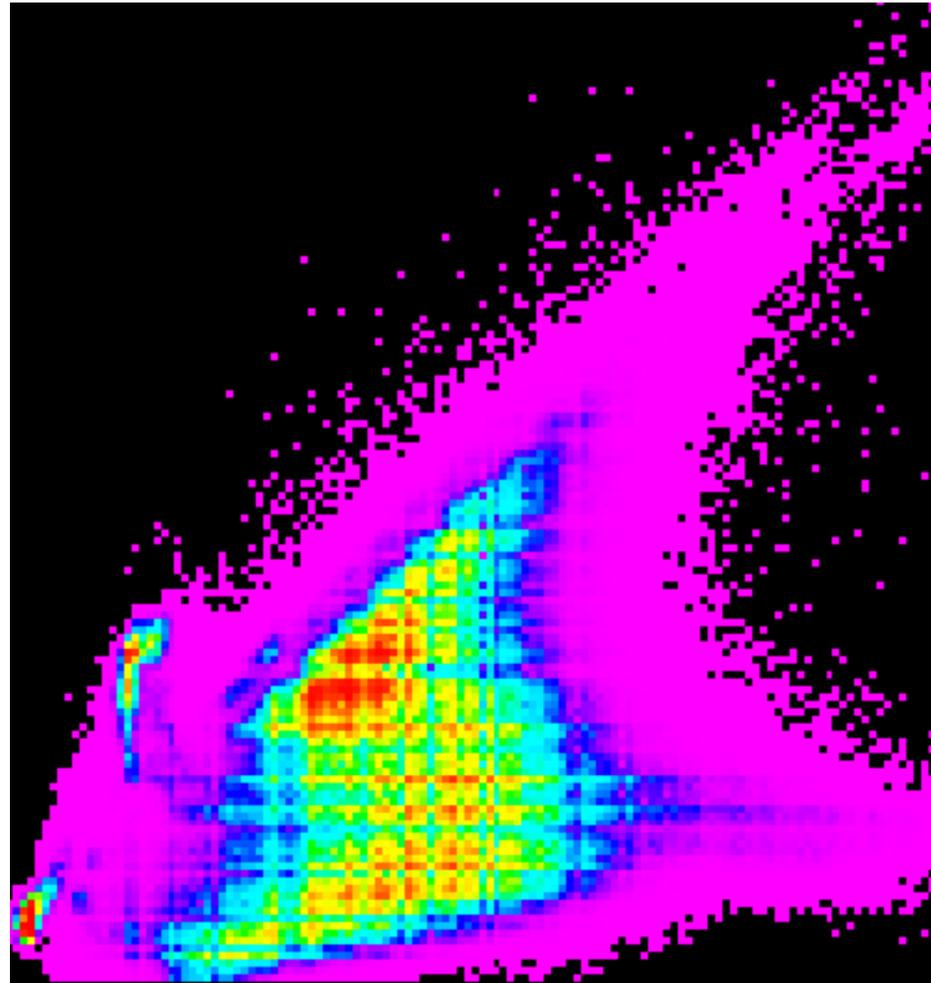


$$\frac{(NIR - Red)}{(NIR + Red)} = -1 \text{ to } 1$$

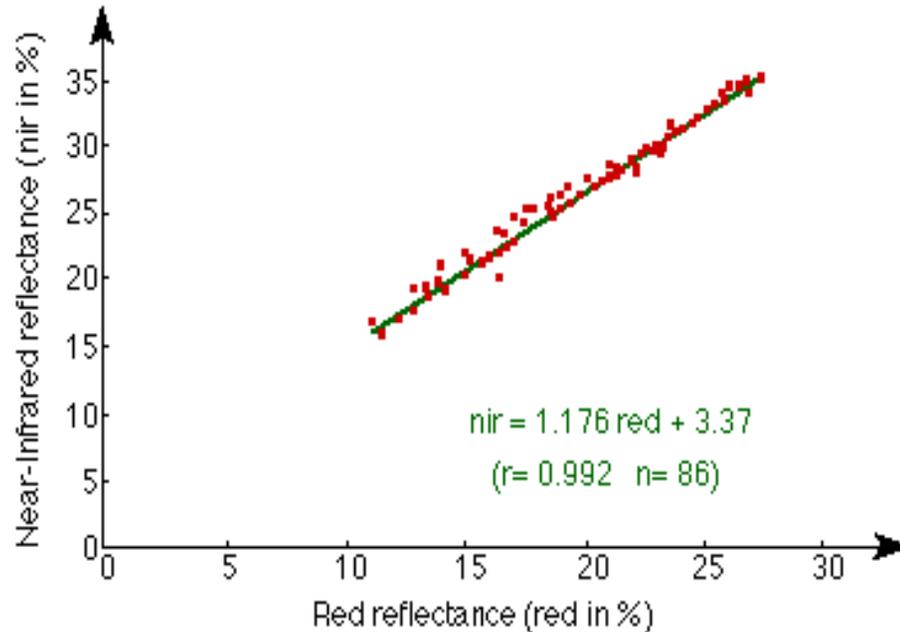


The feature space between Red and NIR reflectance show not only the effect of vegetation on spectral response but also the effect of soil characteristics.

NIR Reflectance

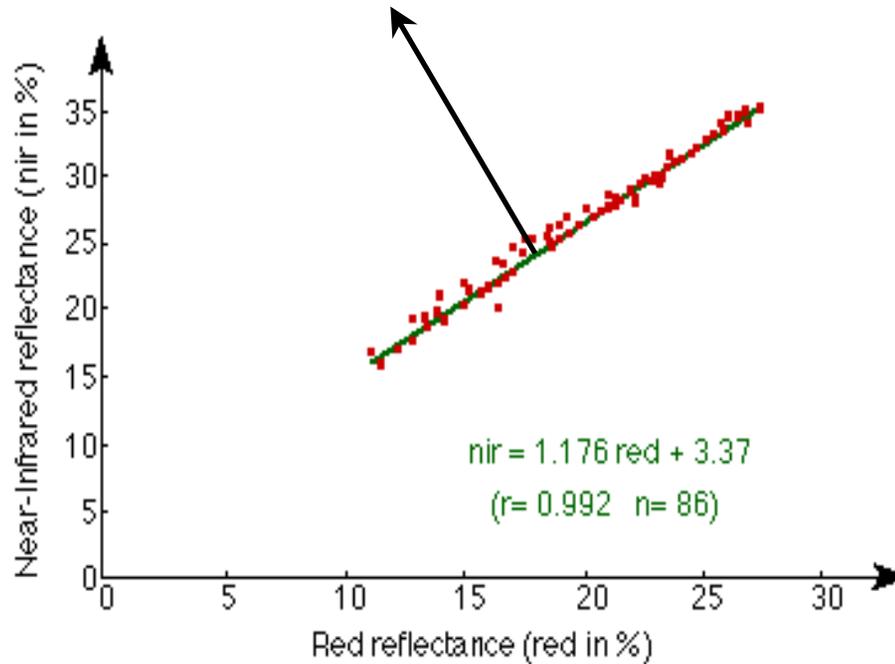


Red Reflectance



**Soil color, moisture, texture, and organic matter affect the relationship between NIR / Red reflectance in a linear fashion which can be described by a least-squares regression.**

**Within any given landscape, low and high points of this relationship may be composed of areas of deep shadow (low) and areas of snow or rock outcrop (high) as well as areas of different soil color and moisture.**



**Movement away from this line toward higher NIR reflectance and lower Red reflectance identify varying amounts of vegetation cover which absorbs Red light and reflect NIR.**

# Soil Adjusted Vegetation Index

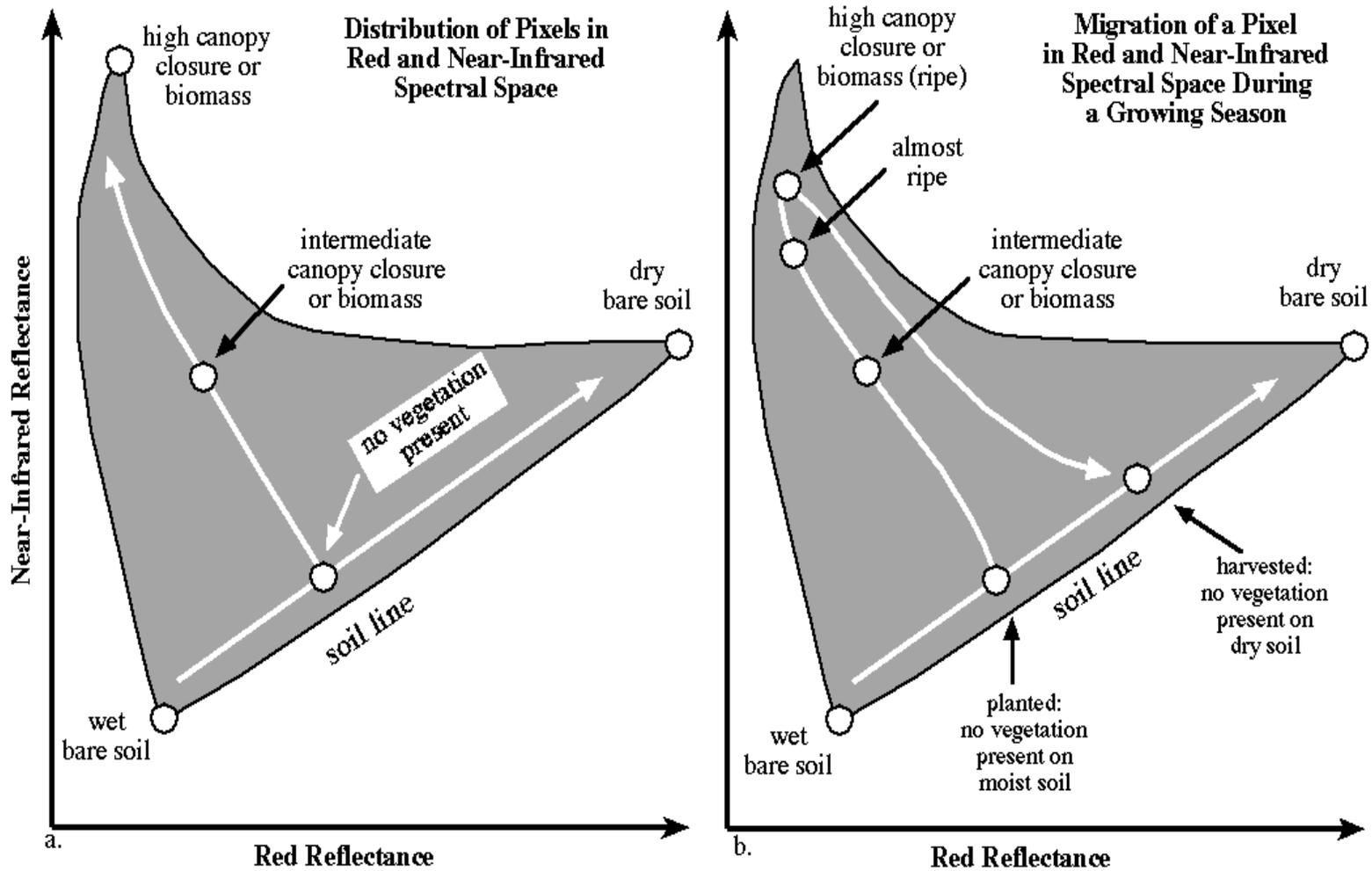
Huete (1988) suggested a vegetation index designed to minimize the effect of the soil background. He called this index the soil-adjusted vegetation index (SAVI).

$$\text{SAVI} = \frac{(\text{NIR}-\text{red})}{(\text{NIR}+\text{red}+L)} * (1+L)$$

Huete showed that depending on vegetation cover, the NDVI for different cover conditions did not converge at the same location. Huete established an L-factor which adjusted the NDVI so that different vegetation densities would intersect the soil line at the same location.

For high vegetation cover, the value of L is 0.0, and L is 1.0 for low vegetation cover. For intermediate vegetation cover L=0.5.

# Distribution of Pixels in a Scene in Red and Near-infrared Multispectral Feature Space



# Enhanced Vegetation Index (EVI)

Adapted from SAVI – more linear relationship to vegetation variables (e.g. biomass, LAI, etc.)

Addition of blue band information for scattering correction

$$\text{EVI} = G * (\text{NIR} - \text{Red}) / (\text{NIR} + \text{C1} * \text{Red} - \text{C2} * \text{Blue} + \text{L})$$

where G (gain factor) = 2.5

$$\text{C1} = 6$$

$$\text{C2} = 7.5$$

$$\text{L} = 1$$

# Spectral Unmixing

- The reflectance of each pixel is assumed to be a linear combination of each material (endmember) within the pixel
- For example, if 25% of a pixel is grass, 25% bare soil, and 50% trees, the reflectance (spectrum) for that pixel is a weighted average of 25% the spectrum of grass, 25% the spectrum of bare soil, and 50% the spectrum of trees.
- Linear unmixing is solving for the abundance values of each endmember for every pixel.
- The number of endmembers must be less than the number of spectral bands and all endmember in the image must be used.
- Results are highly dependent on quality of endmember input data. These can be created from image or taken from spectral libraries.

# Part 1. Preparing MODIS data for Analysis

- **Import data into Imagine**
- **View image**
- **Perform reprojection of image**
- **View reprojected image**
- **Perform band ratio'ing**
  - **Make NDVI**
  - **Make SAVI**

# Part 2. Preparing ASTER data for Analysis

- **Import ASTER data into ENVI**
- **Rotate image**
- **Perform Principal Components Analysis (PCA)**
- **Sharpen image**