Per-pixel classification confidence mapping using R and GRASS

Scott W Mitchell
Carleton University
Scott_Mitchell@carleton.ca

Carleton UNIVERSITY

Tarmo K Remmel
York University
remmelt@yorku.ca

YORK U UNIVERSITÉ UNIVERSITY

Mike Wulder
Pacific Forest Research Centre

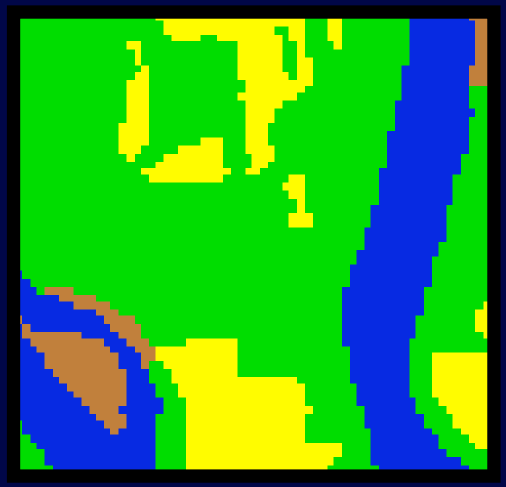
CFS CANADIAN FOREST SERVICE
cfs.nrcan.gc.ca

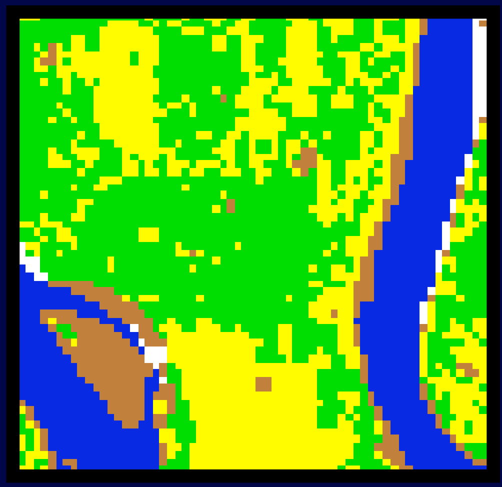
My plan this morning...

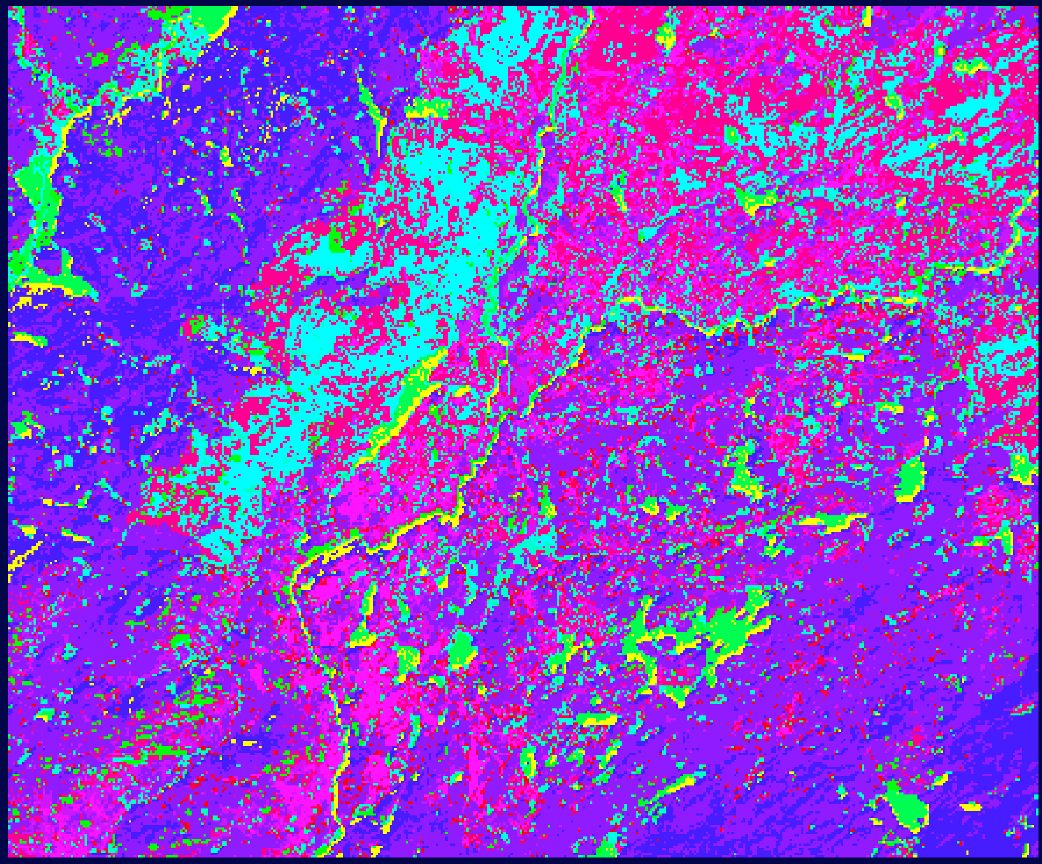
- Introduction / background
  - land use / land cover maps - where they come from
  - per-pixel classification: probability of class membership
  - CFS / Prince George demonstration project
    - work leading up to this part of the project
- Explain methods
- Proof of concept from R implementation
- Development of GRASS module

Land use / cover maps

- Important to a wide variety of applications
- often created using per-pixel image classification
  - convenient
  - standardized methods
  - data availability



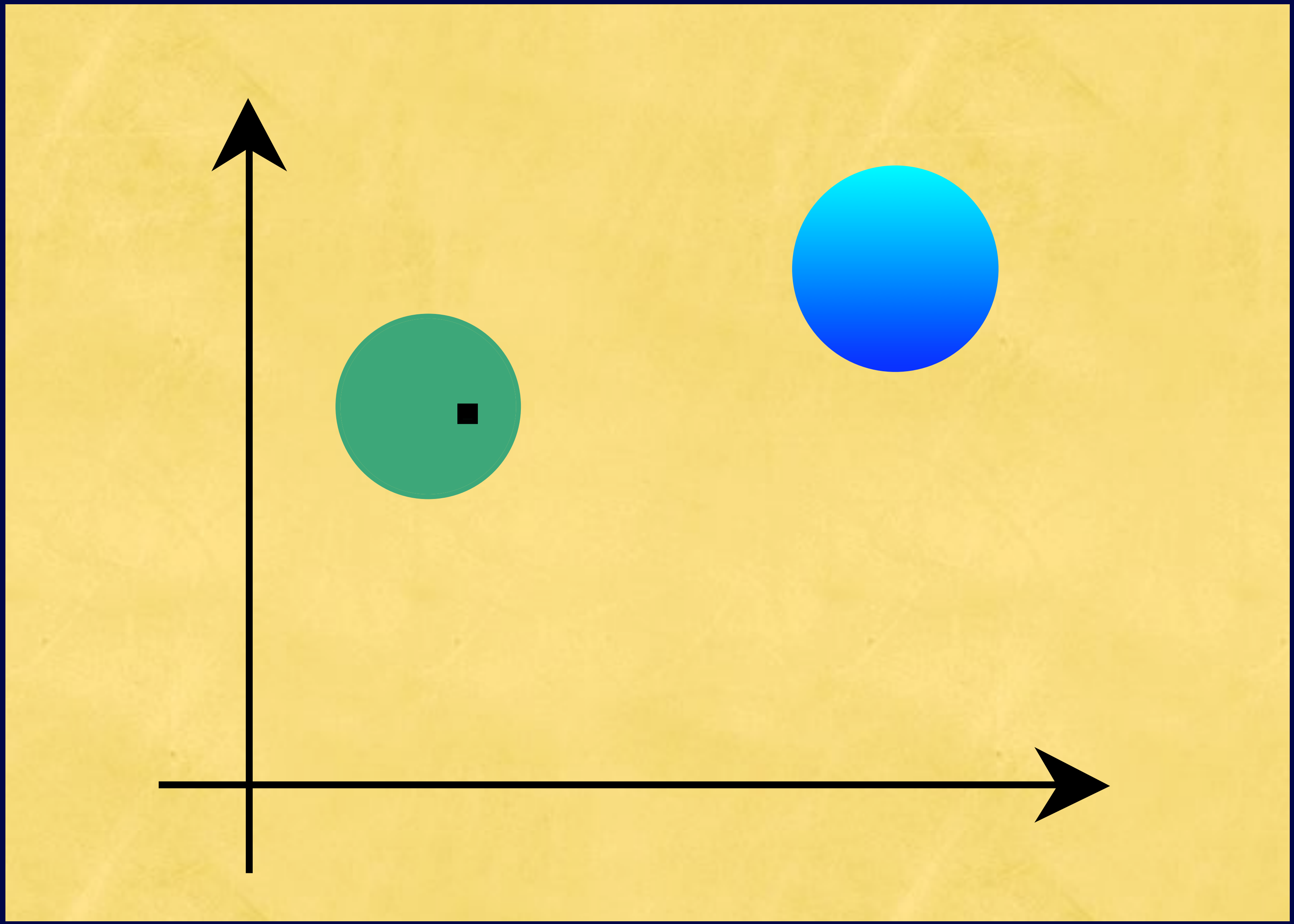




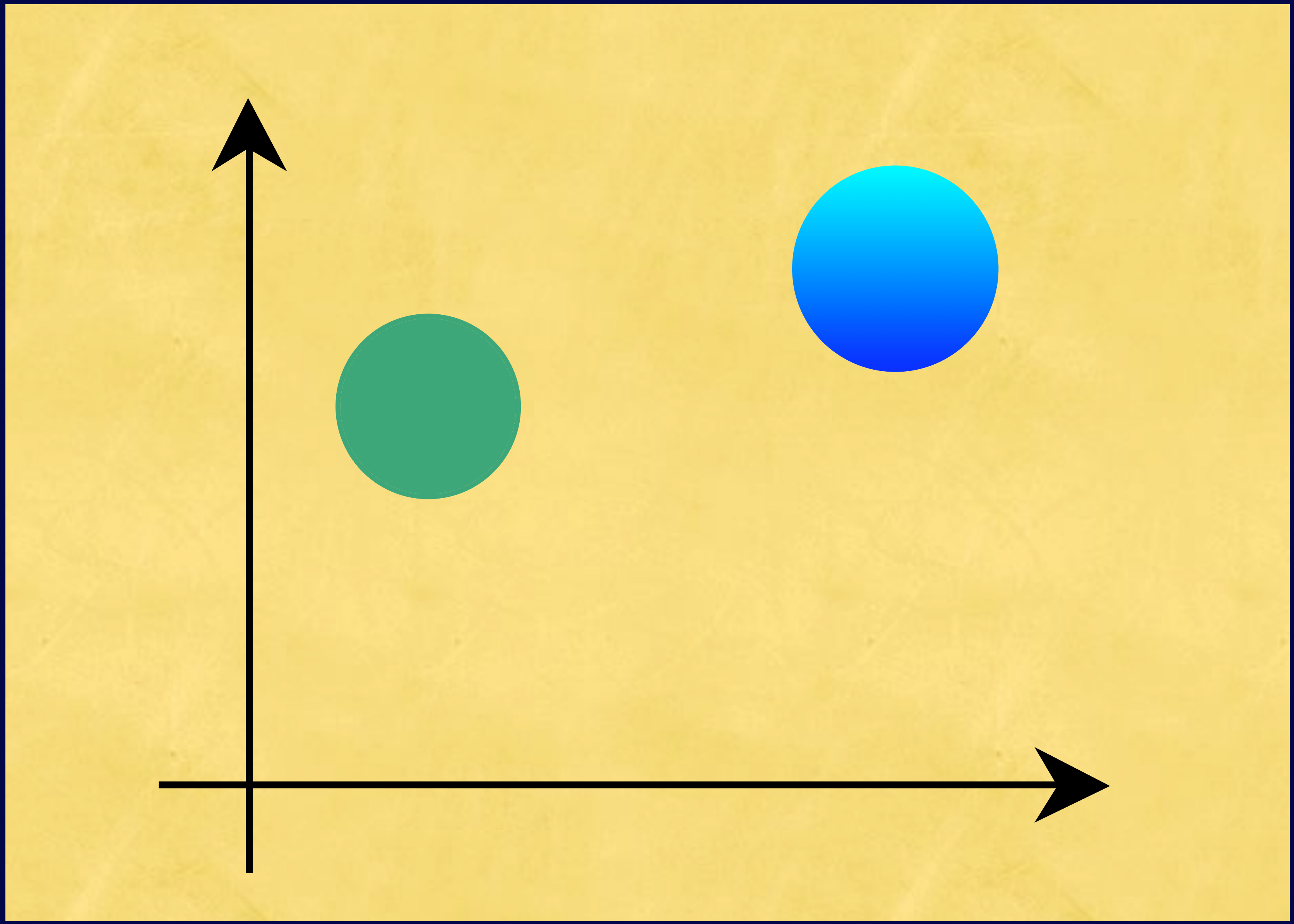
Where do the maps come from?

- some classification scheme which identifies clusters in n-D space
- normally only the final assignment to one class is used
- assessments typically global; useless for judging how reliable a prediction is at a particular point
- for each pixel, there are likelihoods of belonging to all possible classes based on distance away from the clusters

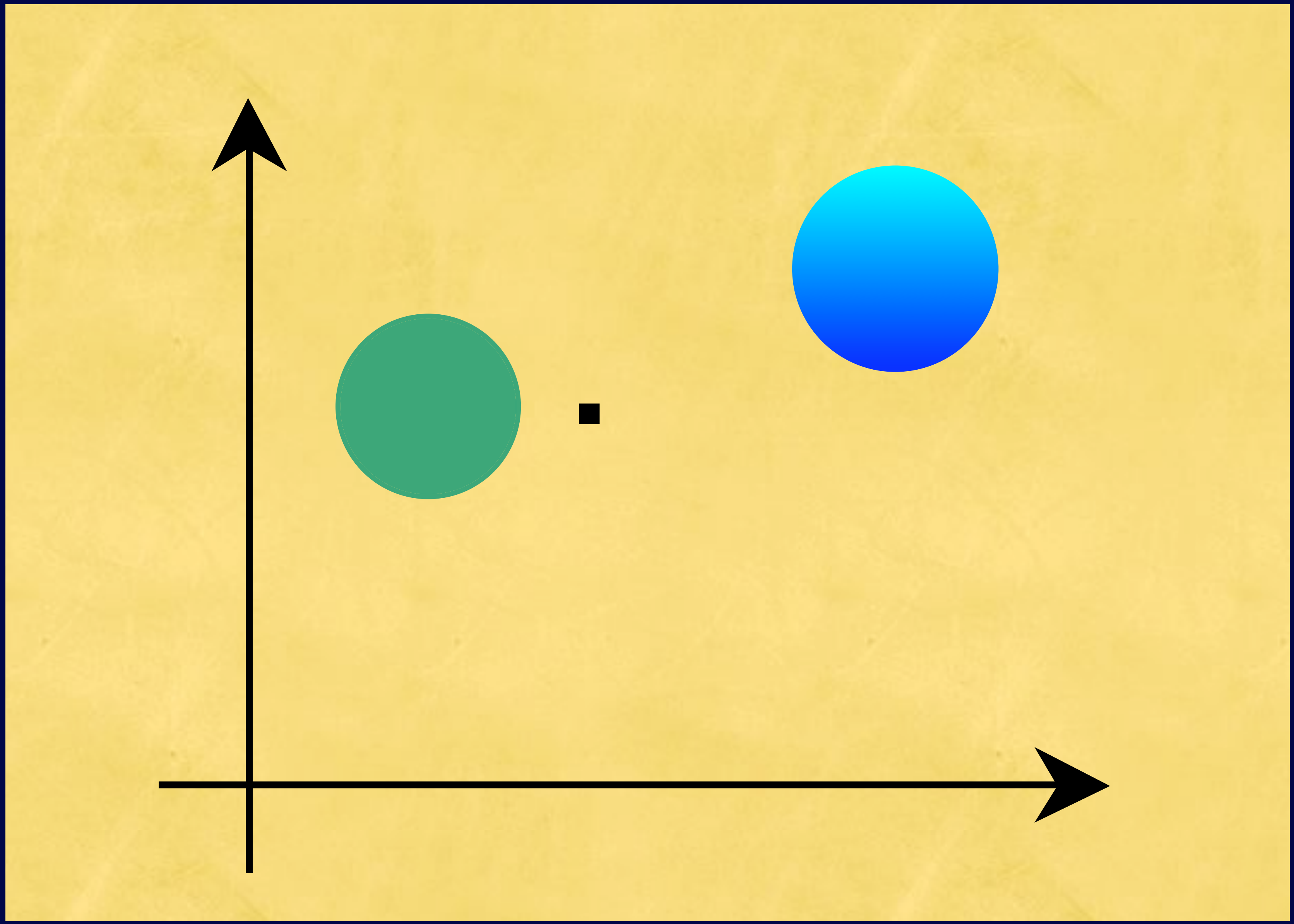
Can we get more information than that? Why would we want to???



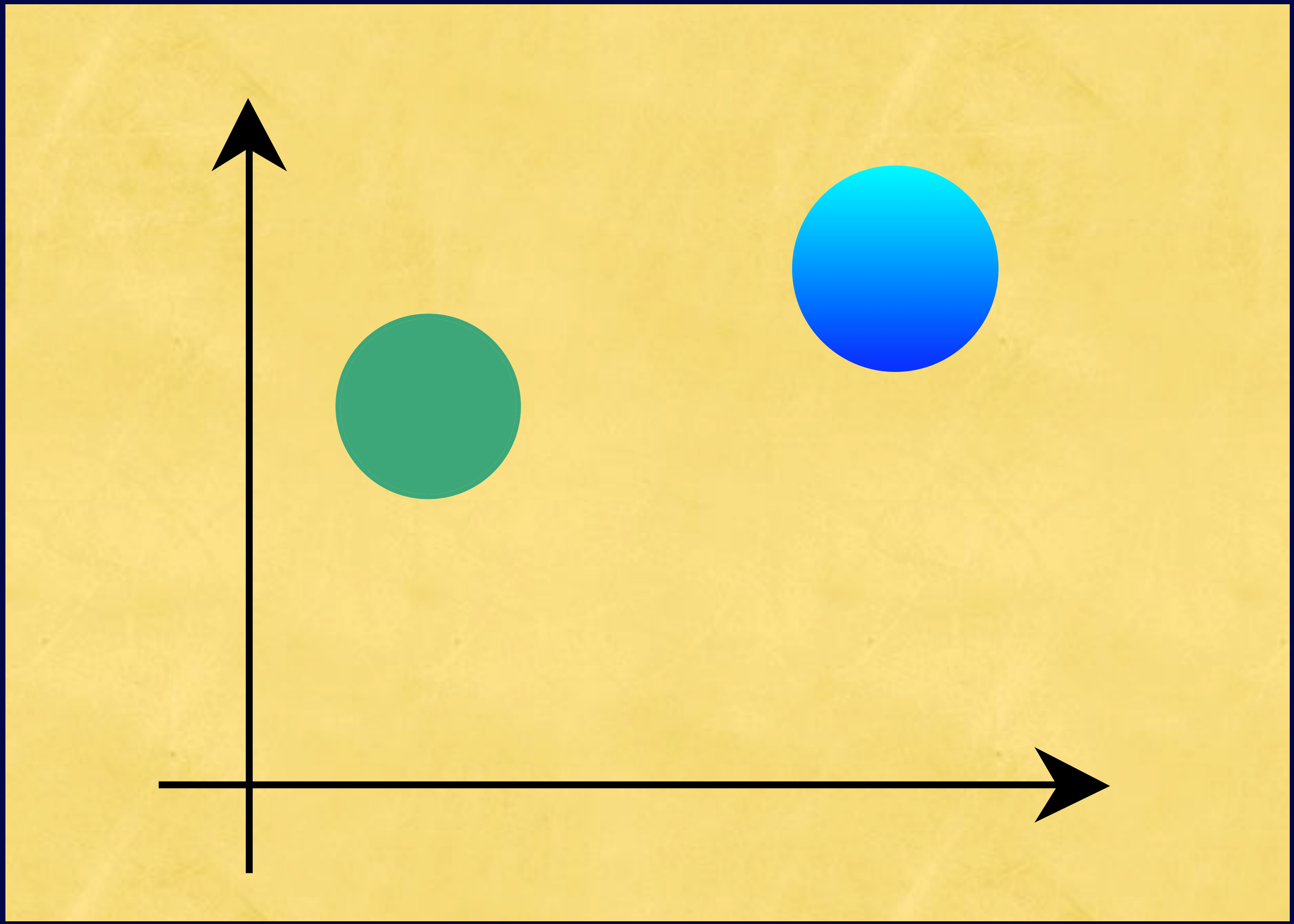
Can we get more information than that? Why would we want to???



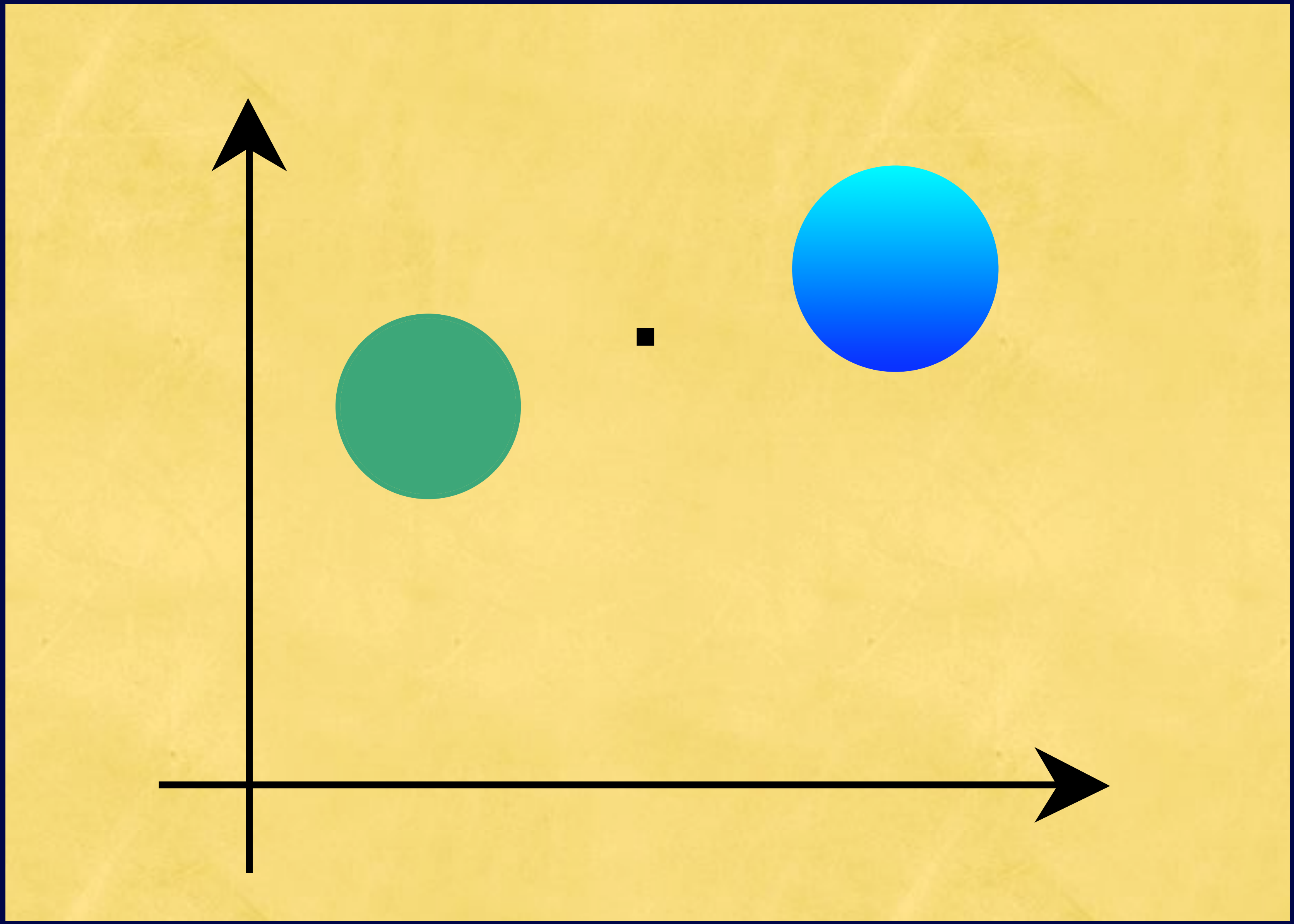
Can we get more information than that? Why would we want to???



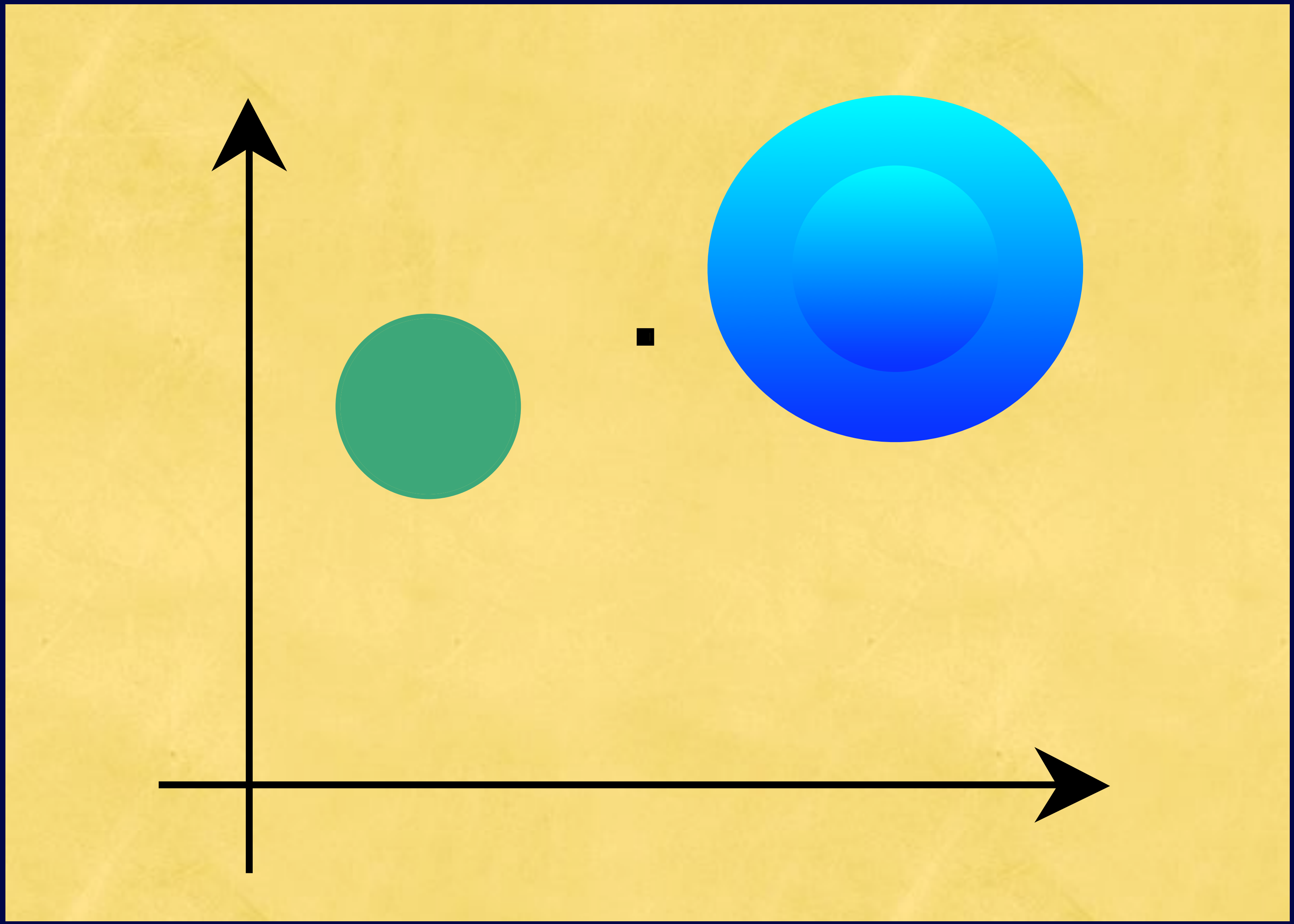
Can we get more information than that? Why would we want to???



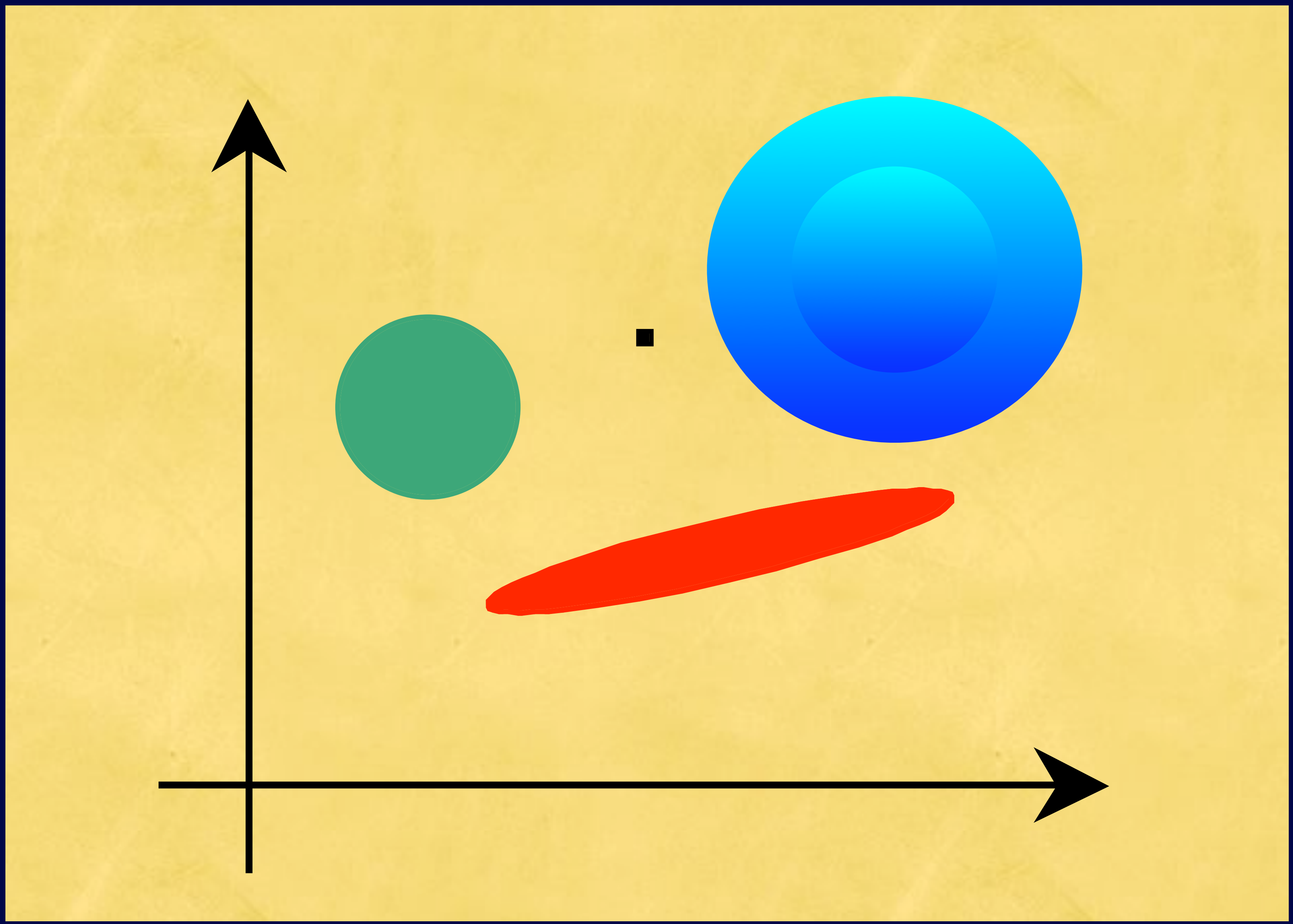
Can we get more information than that? Why would we want to???



Can we get more information than that? Why would we want to???



Can we get more information than that? Why would we want to???



Canada's Forests

- Canada: ≈1 billion ha (998 Mha)
- 402 Mha of forested and wooded land
- 183 Mha timber productive
- 148 Mha accessible

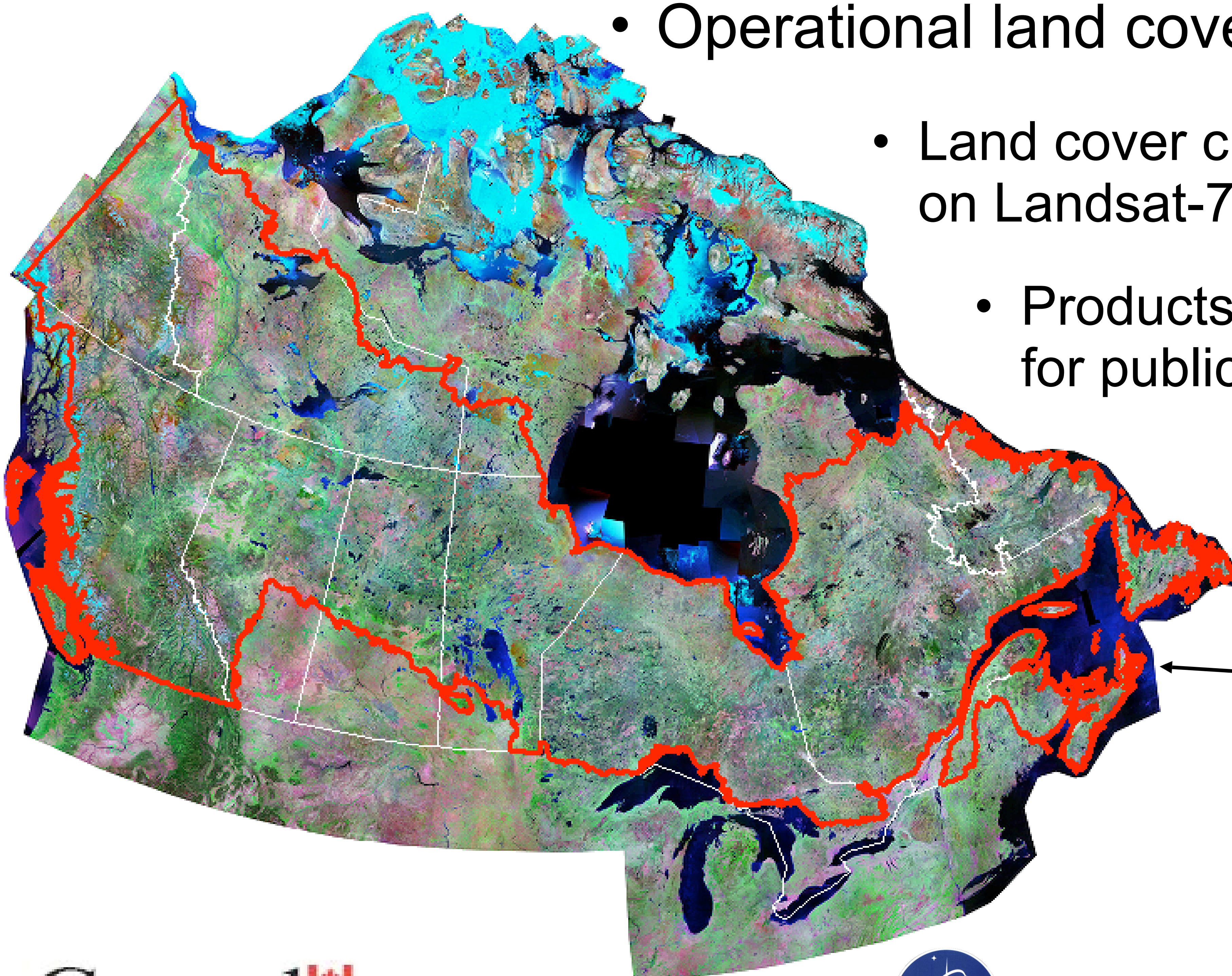
- Map source: Lowe et al. 1996

Canada

Natural Resources Canada / Ressources naturelles Canada
Canadian Forest Service / Service canadien des forêts

EOSD Land Cover Activity Area

- Operational land cover mapping program
  - Land cover classification is based on Landsat-7 ETM+ data.
    - Products are being developed for public access.



Forested area being mapped enclosed in red.

Canada

CSA ASC

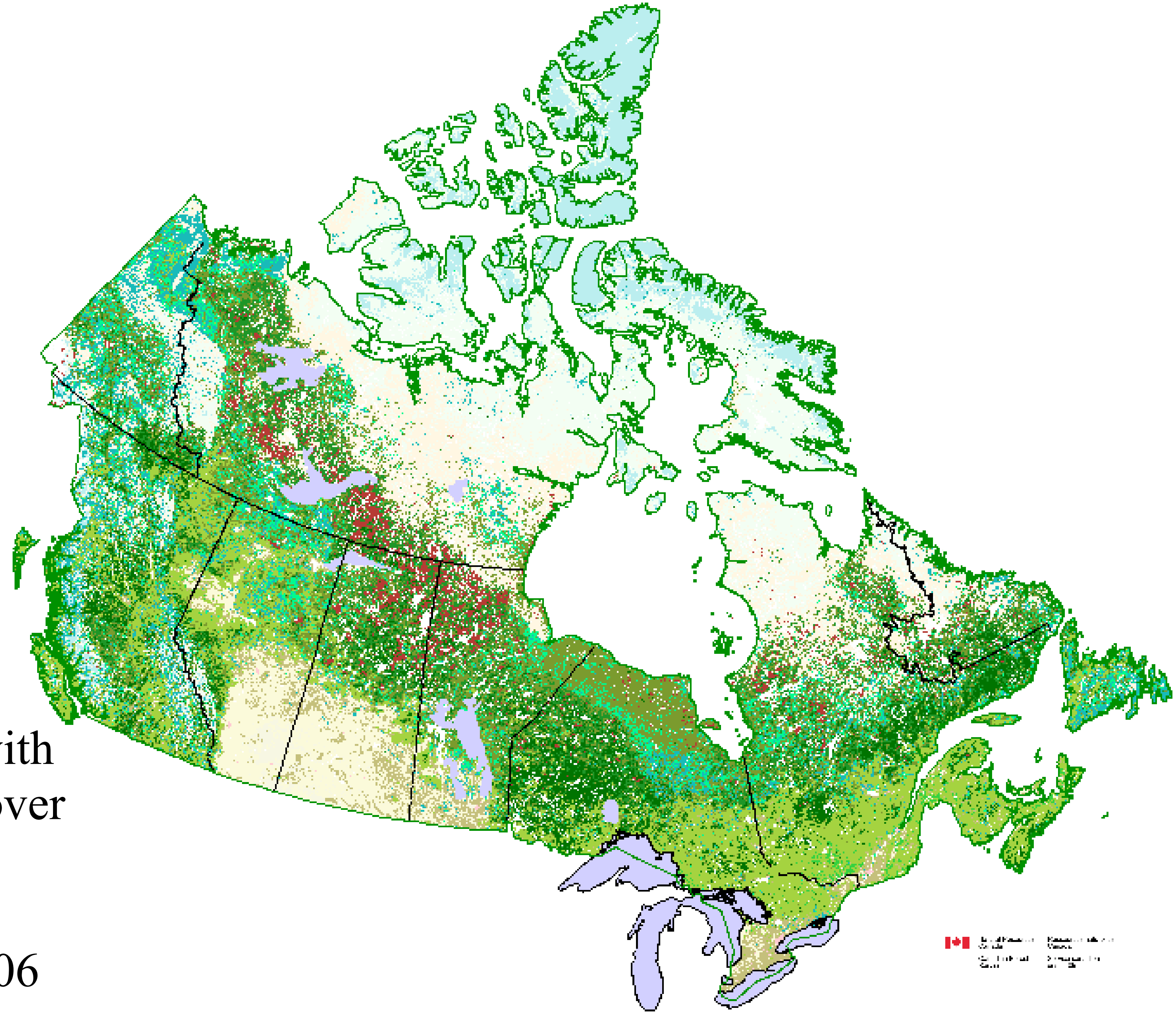
/ 14p

Natural Resources Canada
Canadian Forest Service

Ressources naturelles Canada
Service canadien des forêts

Earth Observation for Sustainable Development of Forests: Land Cover

NBIOME - AVHRR

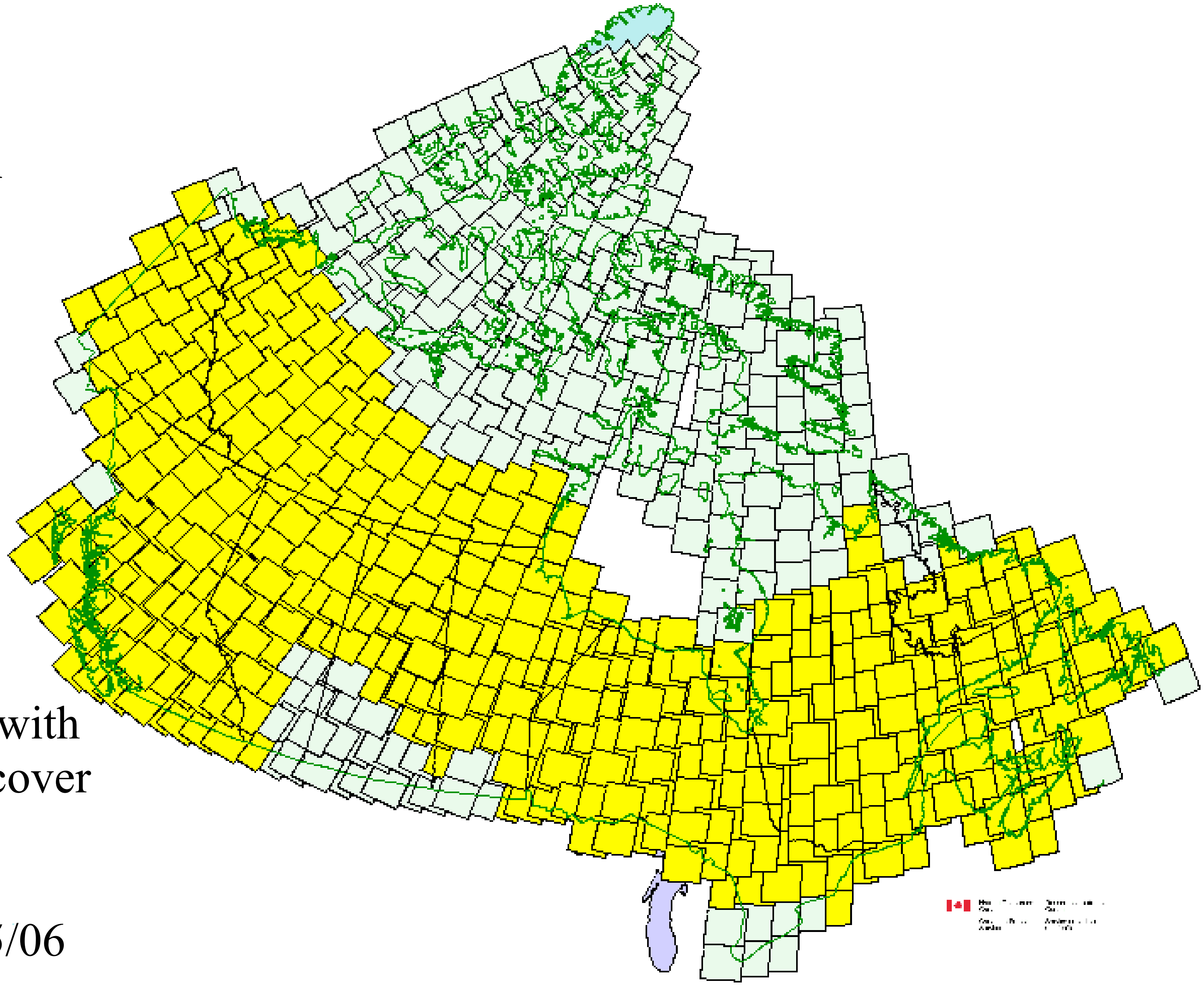


- > 1200 Landsat images
- More than 450 images with greater than 10% forest cover
- circa 2000 imagery
- For completion in 2005/06
- Hyperclustering and labeling; 6 optical channels, plus intra-pixel pan variance

Earth Observation for Sustainable Development of Forests: Land Cover

NBIOME - AVHRR

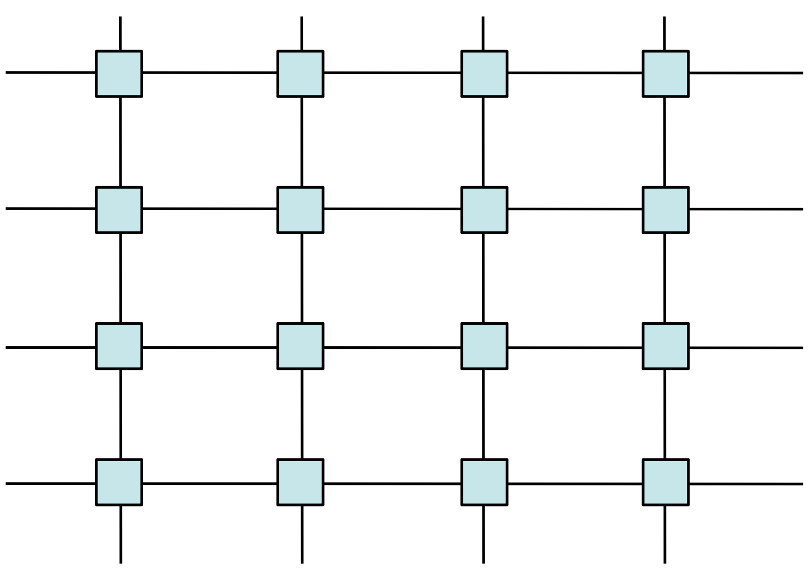
WRS with > 10% forest cover



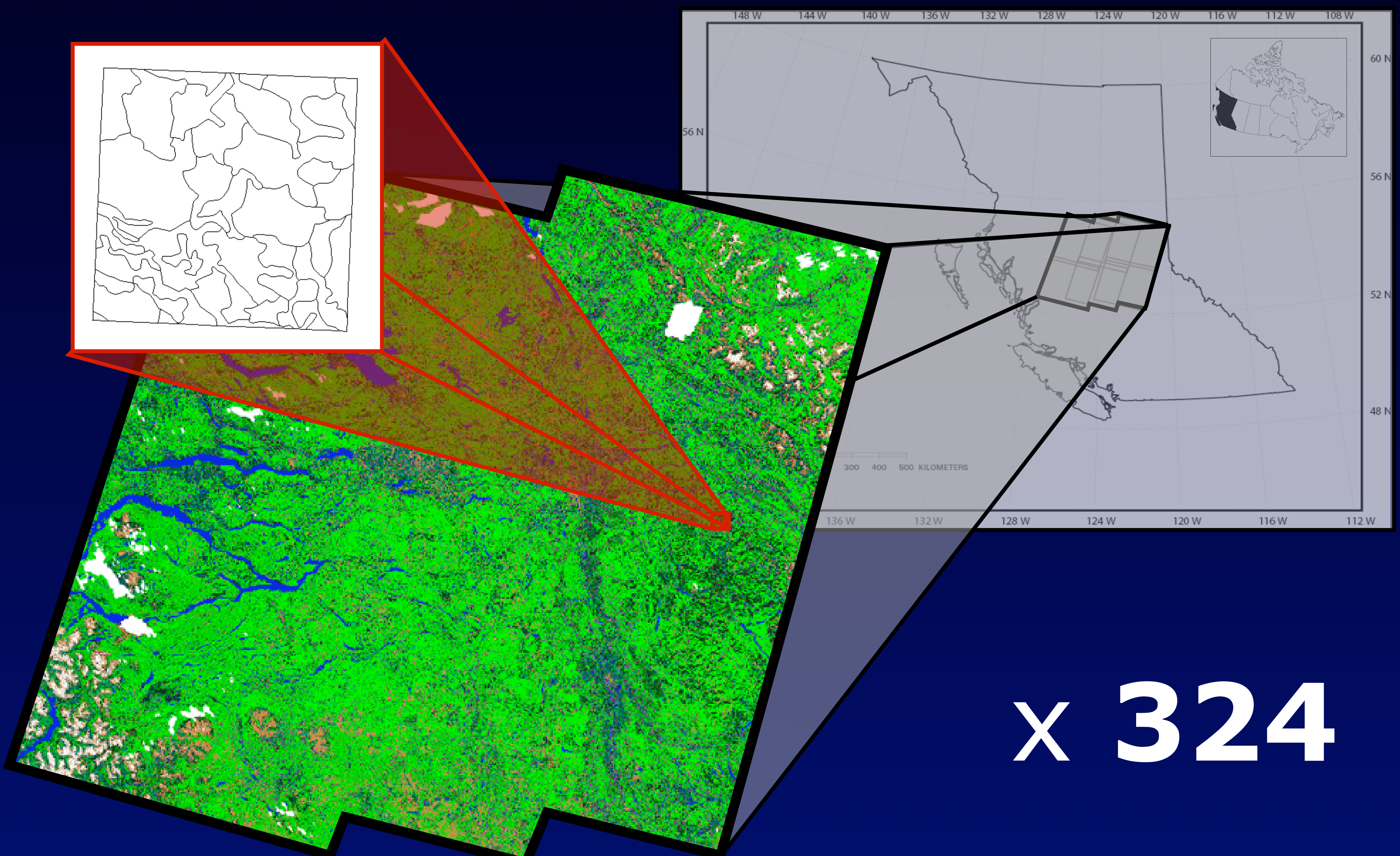
- > 1200 Landsat images
- More than 450 images with greater than 10% forest cover
- circa 2000 imagery
- For completion in 2005/06
- Hyperclustering and labeling; 6 optical channels, plus intra-pixel pan variance

National Forest Inventory

- National systematic sample
- Rationale – standardization nationally
- Sample units are 2 x 2 km photo plots on a 20 km grid
- Ground Plots
  - many attributes
  - including DOM and soil C



Demonstration: Prince George, BC Forest Classification



x **324**

DATA DESCRIPTION SUMMARY

- area proportions: discrepancies
  - missing data
  - barren land
  - wetlands
  - conifer bias
  - density mismatch

	water	rock	shrub	broadl	conif	mixedw
water	4434	1312	1170	162	2695	347
rock	1986	3695	6165	957	19385	2094
shrub	2554	2256	6165	12919	55088	12606
broad	248	8664	3494	22381	70210	21043
conif	1100	6757	3622	13739	77% 69882	42782
mixed	459	2406	9491	8429	53198	16% [illegible]

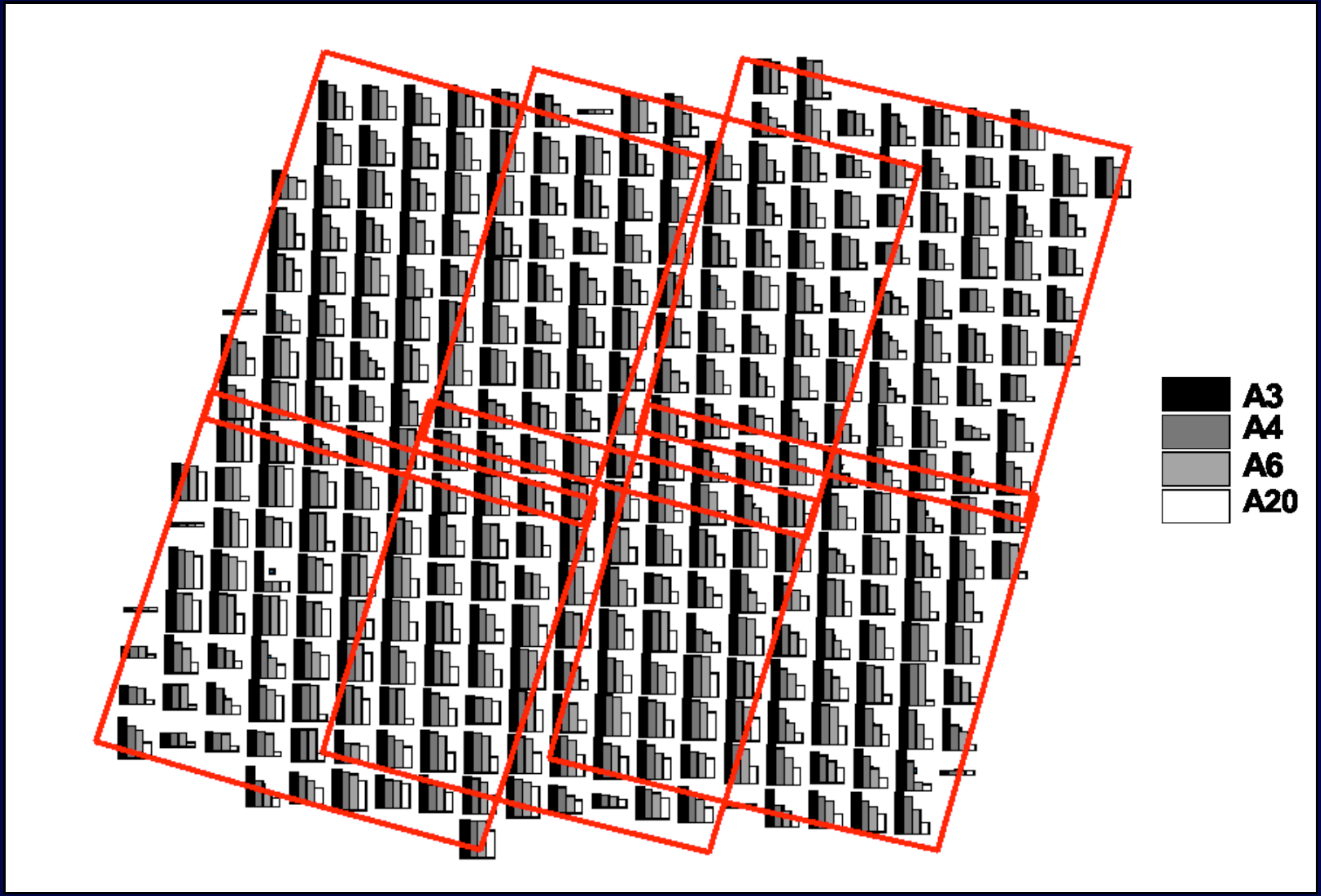
- OVERALL MATCH:

```
AGG3:  91.2%
AGG4:  79.4%
AGG6:  64.1%
AGG20: 26.1%
```

PGSA %	NFI	EOSD
NODATA	0.0	3.0
SHADOW	0.0	0.0
SNOW/ICE	0.0	1.0
ROCK	0.0	0.0
EXP.LAND	1.0	7.2
WATER	3.8	4.0
SHRUB-TALL	2.6	0.6
SHRUB-LOW	6.1	5.7
HERB	1.6	5.3
BRYOIDS	0.0	0.0
WETLAND-TREED	1.2	0.0
WETLAND-	0.0	0.0
WETLAND-HERB	2.3	0.0
CONIFER-DENSE	8.1	28.3
CONIFER-OPEN	51.0	25.3
CONIFER-	7.0	2.5
BROADL-DENSE	0.9	3.3
BROADL-OPEN	3.2	1.4
BROADL-SPARSE	0.1	6.1
MIXEDW-DENSE	0.5	5.2
MIXEDW-OPEN	3.4	0.8
MIXEDW-SPARSE	1.4	0.0

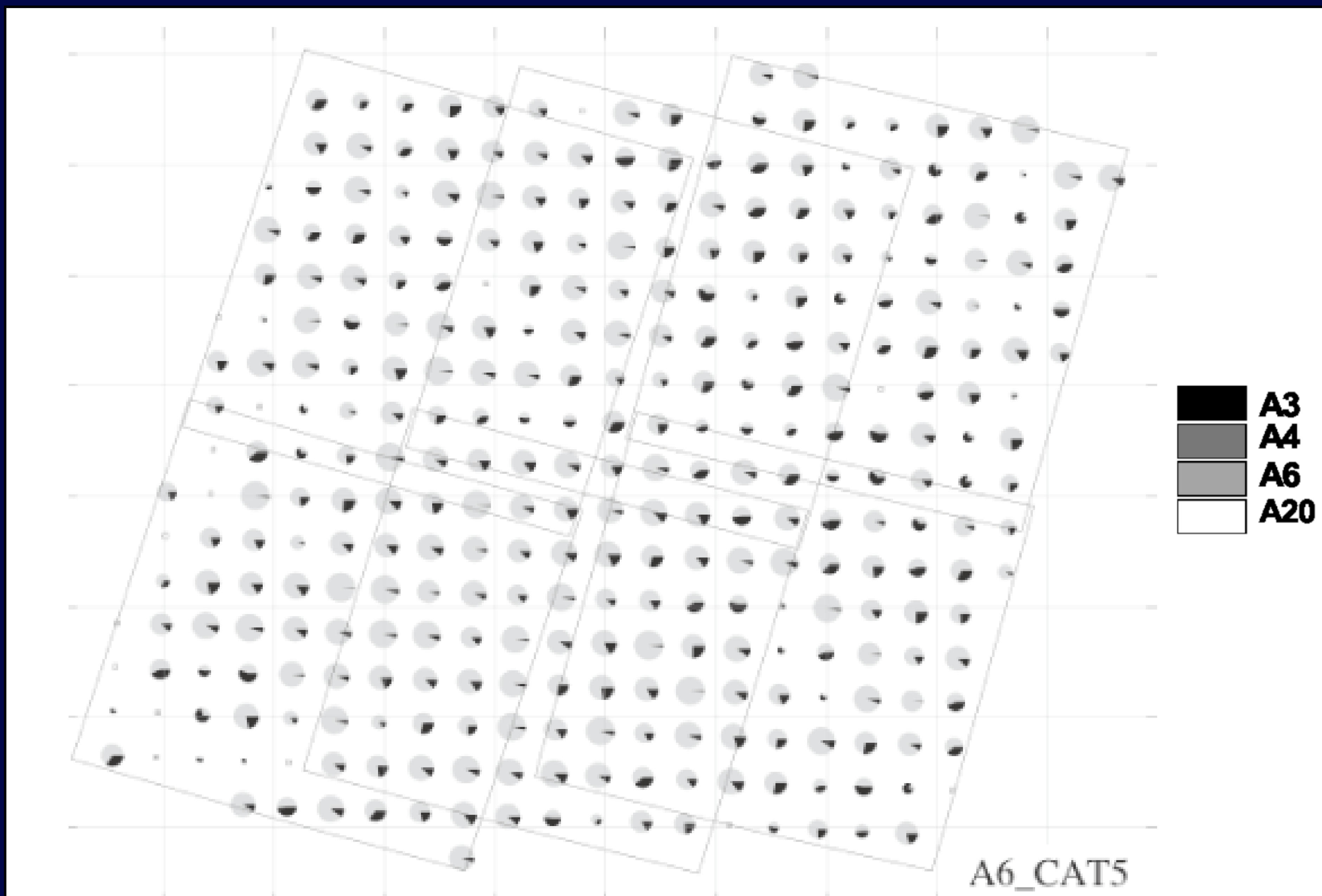
GEOGRAPHICAL VARIATION

- where is the mismatch? (...and is it "well mixed"?)
  - **overall distribution of coincidence across aggregation levels**



GEOGRAPHICAL VARIATION

- where is the mismatch? (...and is it "well mixed"?)
  - **overall distribution of coincidence across aggregation levels**
  - **coincidence by individual categories (coniferous)**

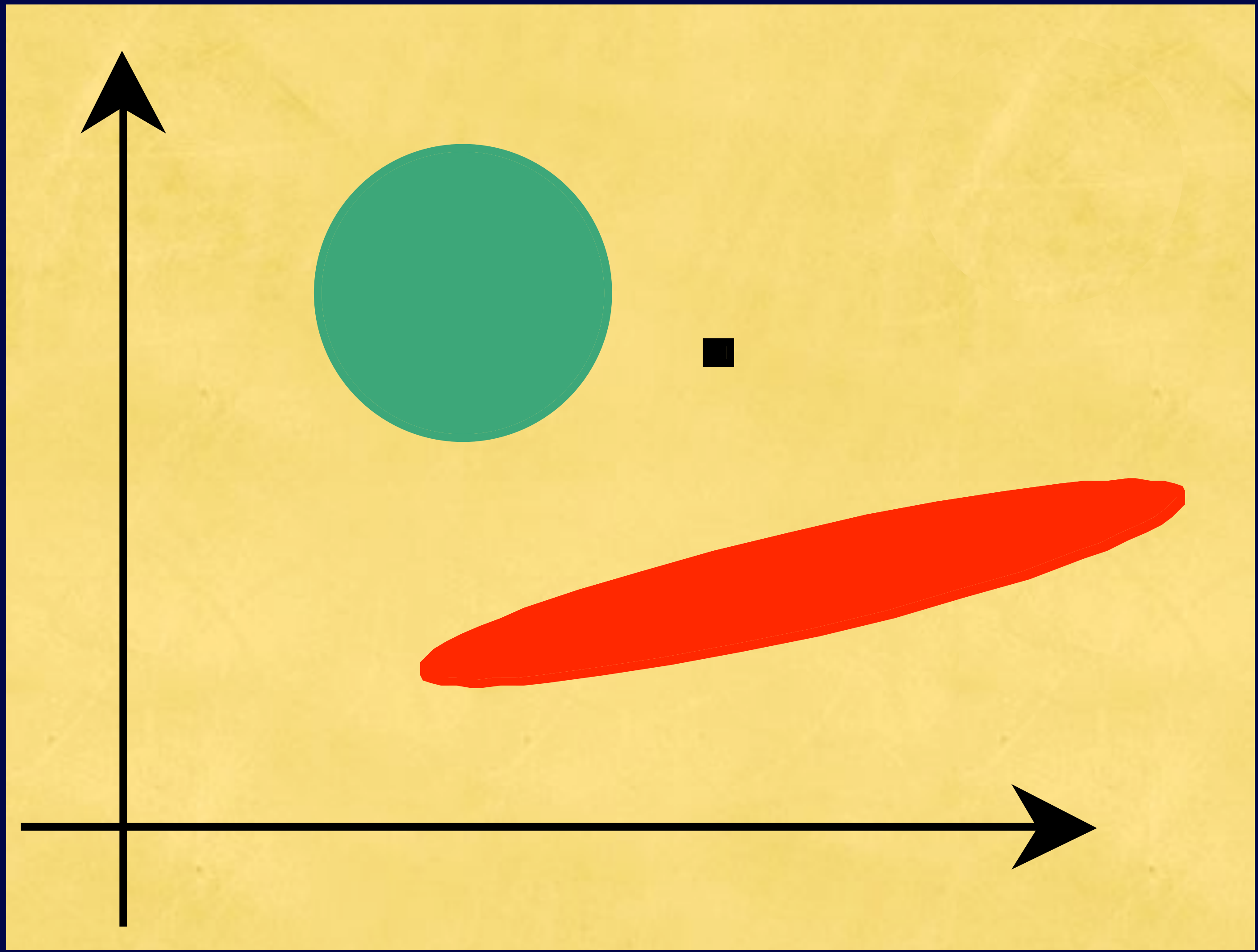


Instead...

Mahalanobis Distance

$$P(X|c) = \log(\det|V_c|) + (X-m_c)T*V_c - 1*(X-m_c)$$

- $P(X|c)$ = likelihood of a pixel belonging to class
- V_c = variance-covariance matrix
- $X-m_c$ = distance between the pixels and the cluster centroids
- classification provides m_c and v_c

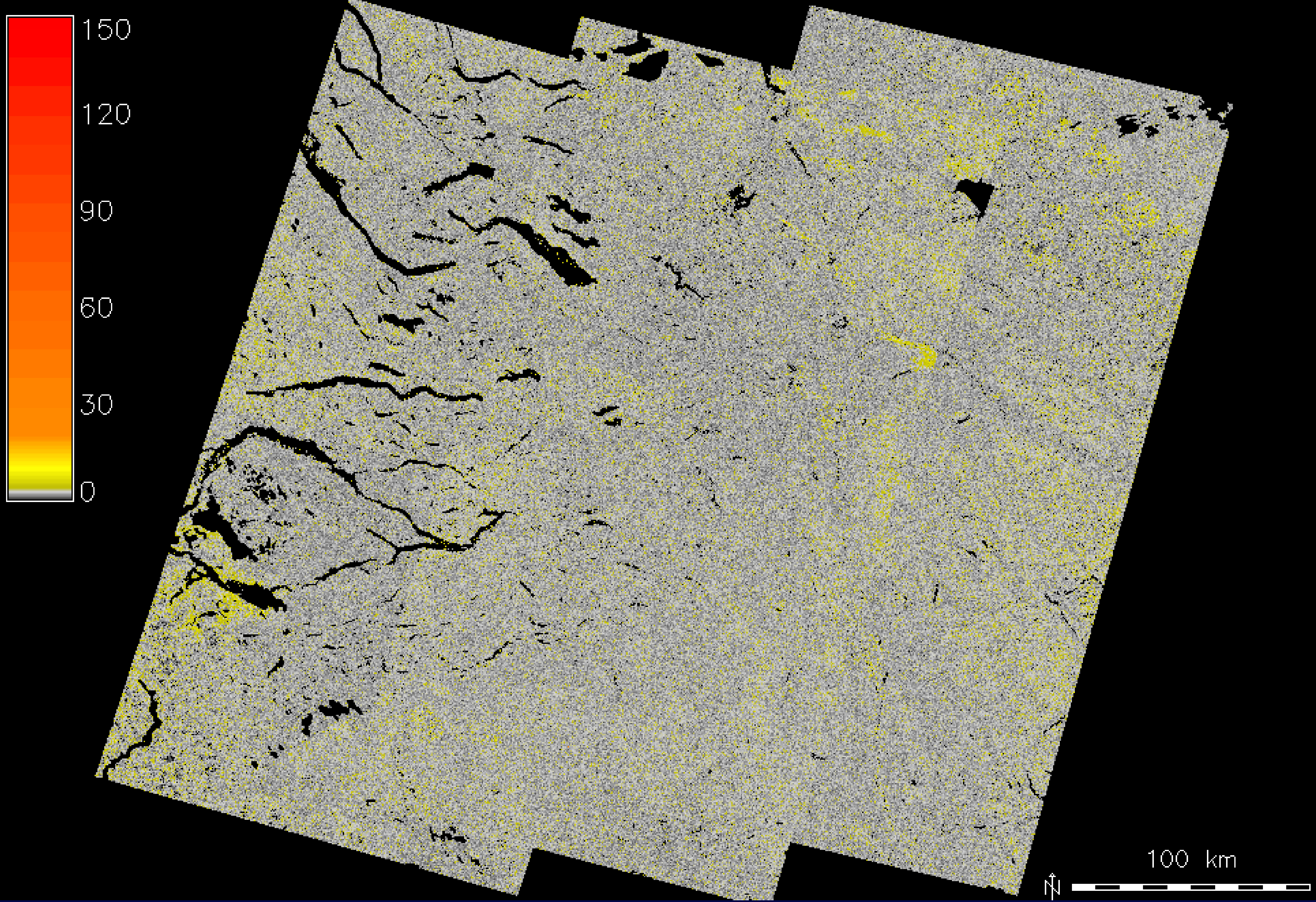


Start with standardized distance ($X\text{-}m_c$)

How far to the closest cluster?

$$Dist = \sqrt{\left(\frac{X - m_1}{v_1}\right)^2 + \ldots + \left(\frac{X - m_n}{v_n}\right)^2} \quad n = \#\, classes$$

```
r.mapcalc cluststddist = sqrt(
     exp((tm1 - cluster.tm1avg) / cluster.tm1stddev, 2)
   + exp((tm2 - cluster.tm2avg) / cluster.tm2stddev, 2)
   + exp((tm3 - cluster.tm3avg) / cluster.tm3stddev, 2)
   + exp((tm4 - cluster.tm4avg) / cluster.tm4stddev, 2)
   + exp((tm5 - cluster.tm5avg) / cluster.tm5stddev, 2)
   + exp((tm7 - cluster.tm7avg) / cluster.tm7stddev, 2)
   + exp((texture - cluster.textureavg) / cluster.texturestddev, 2)
   )
```

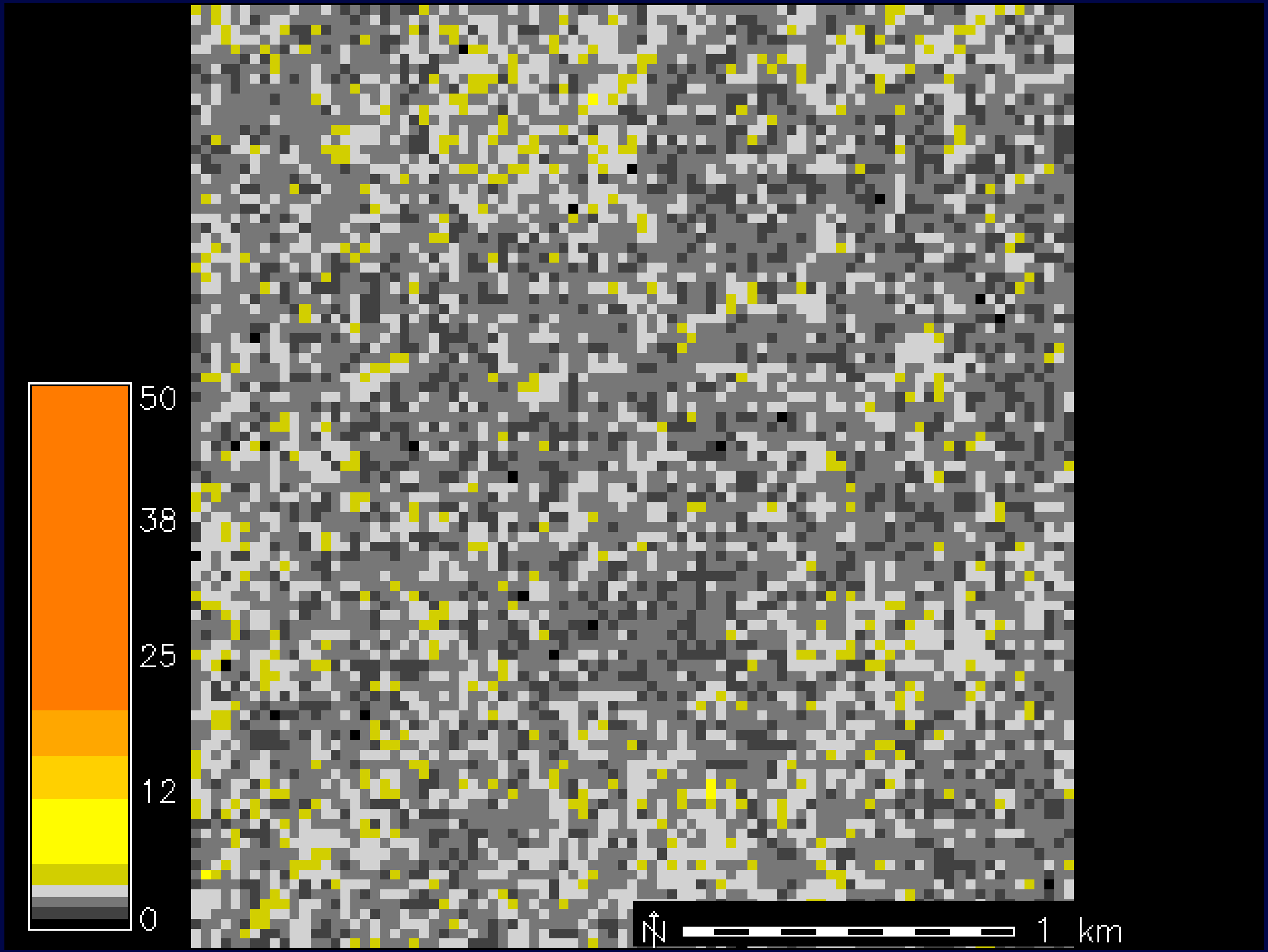


Misclassifications in areas with larger distances ?

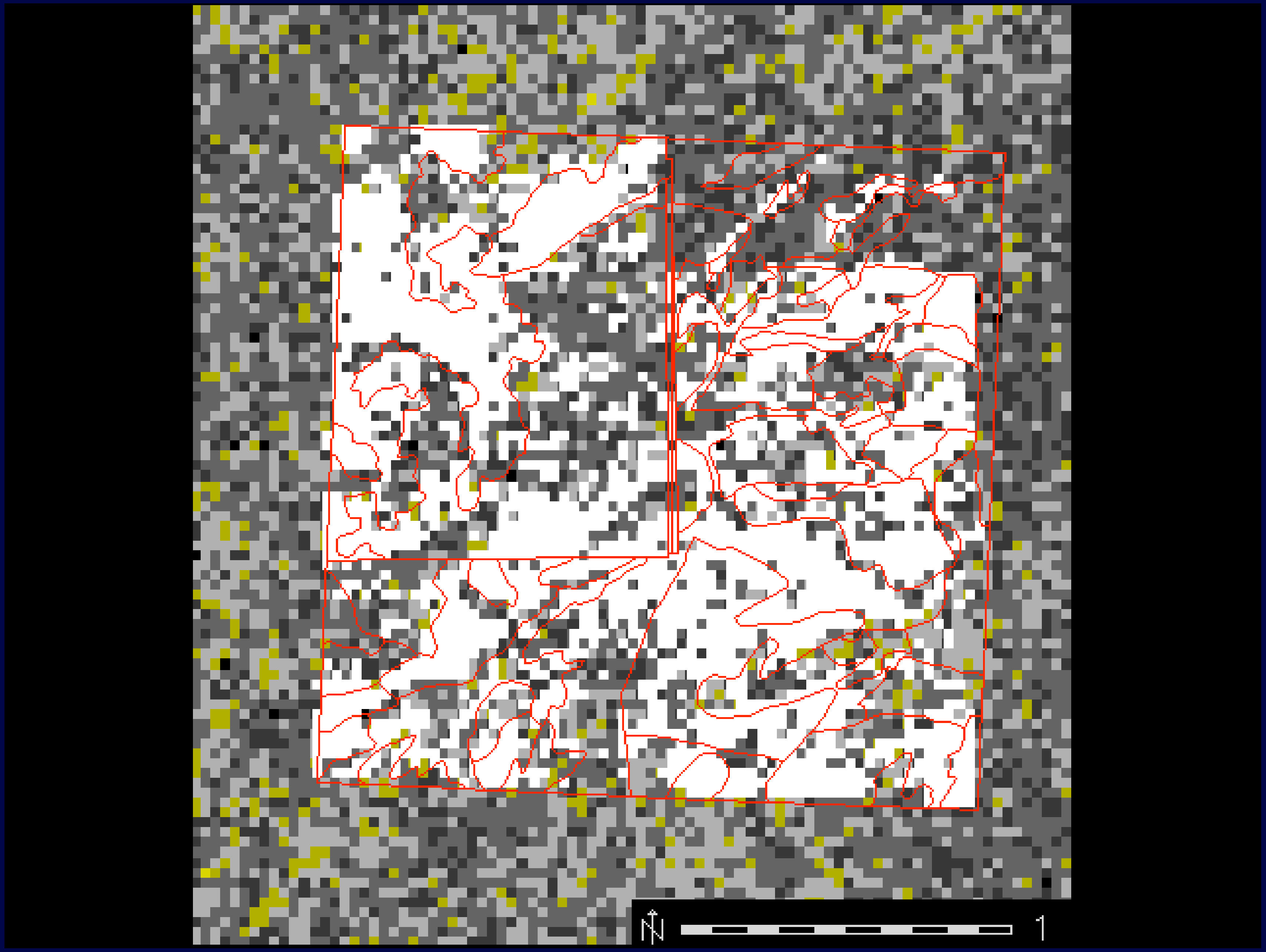
A detailed look:

A detailed look:

A detailed look:

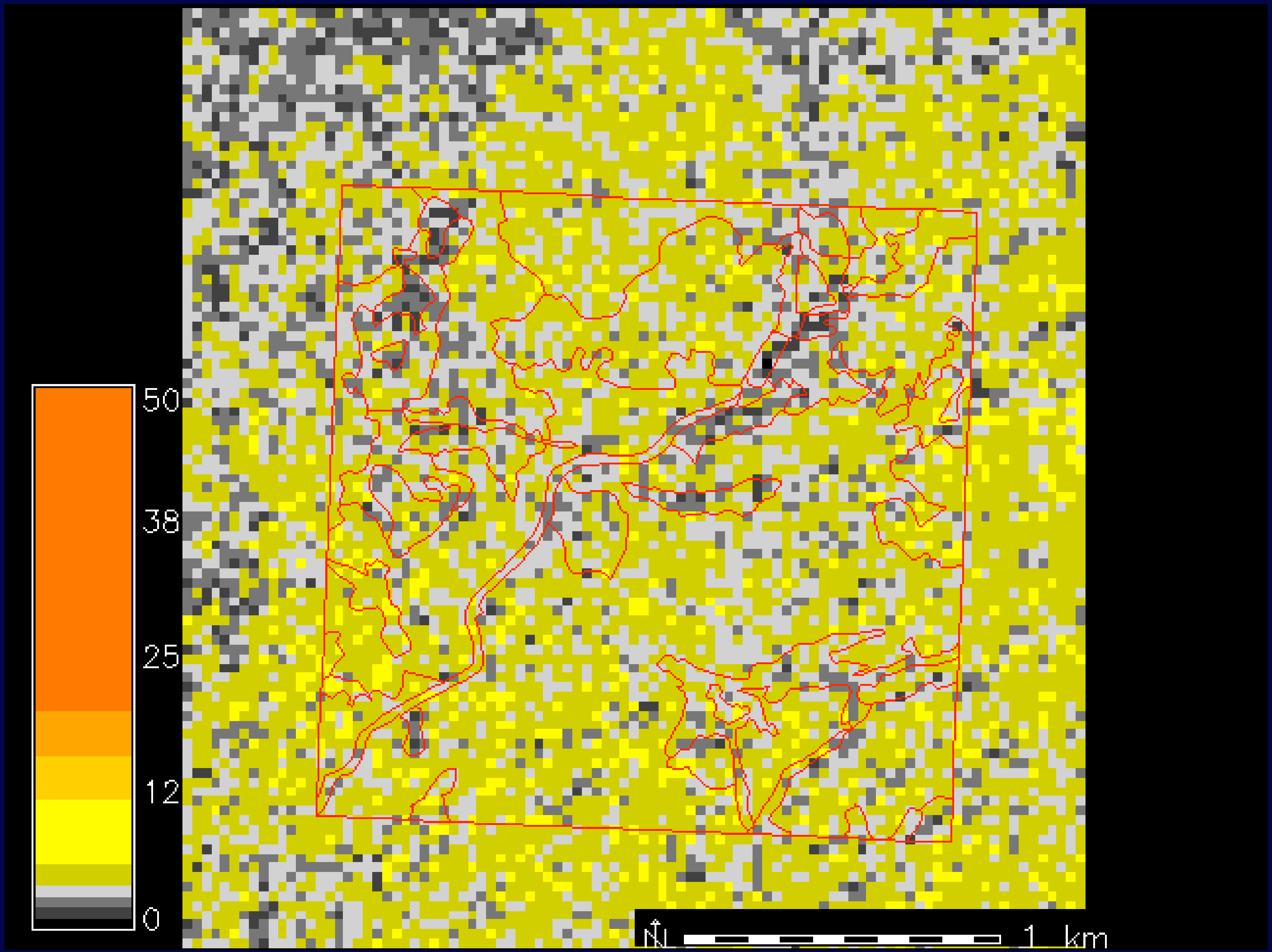


A detailed look:

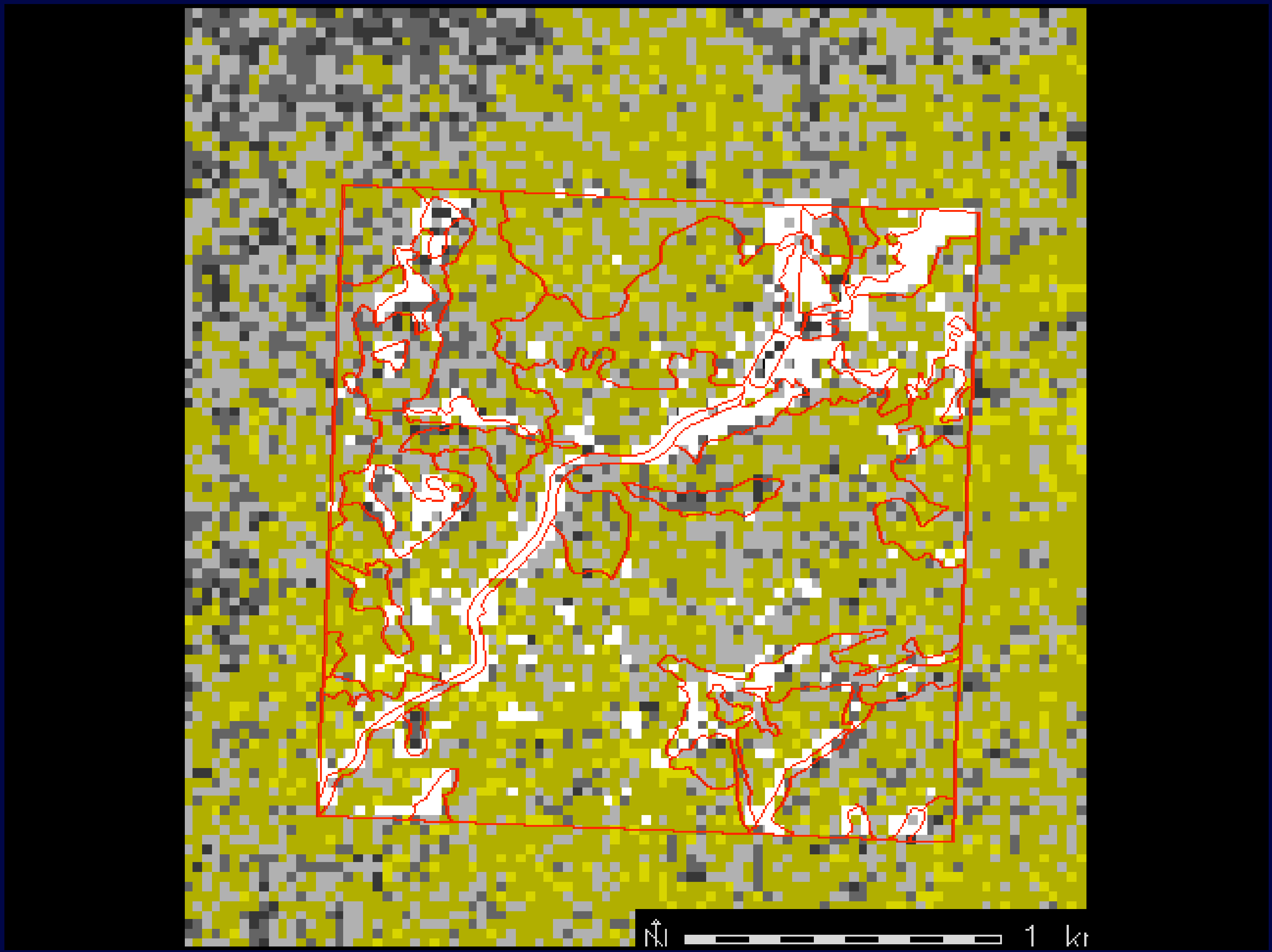


A detailed look:

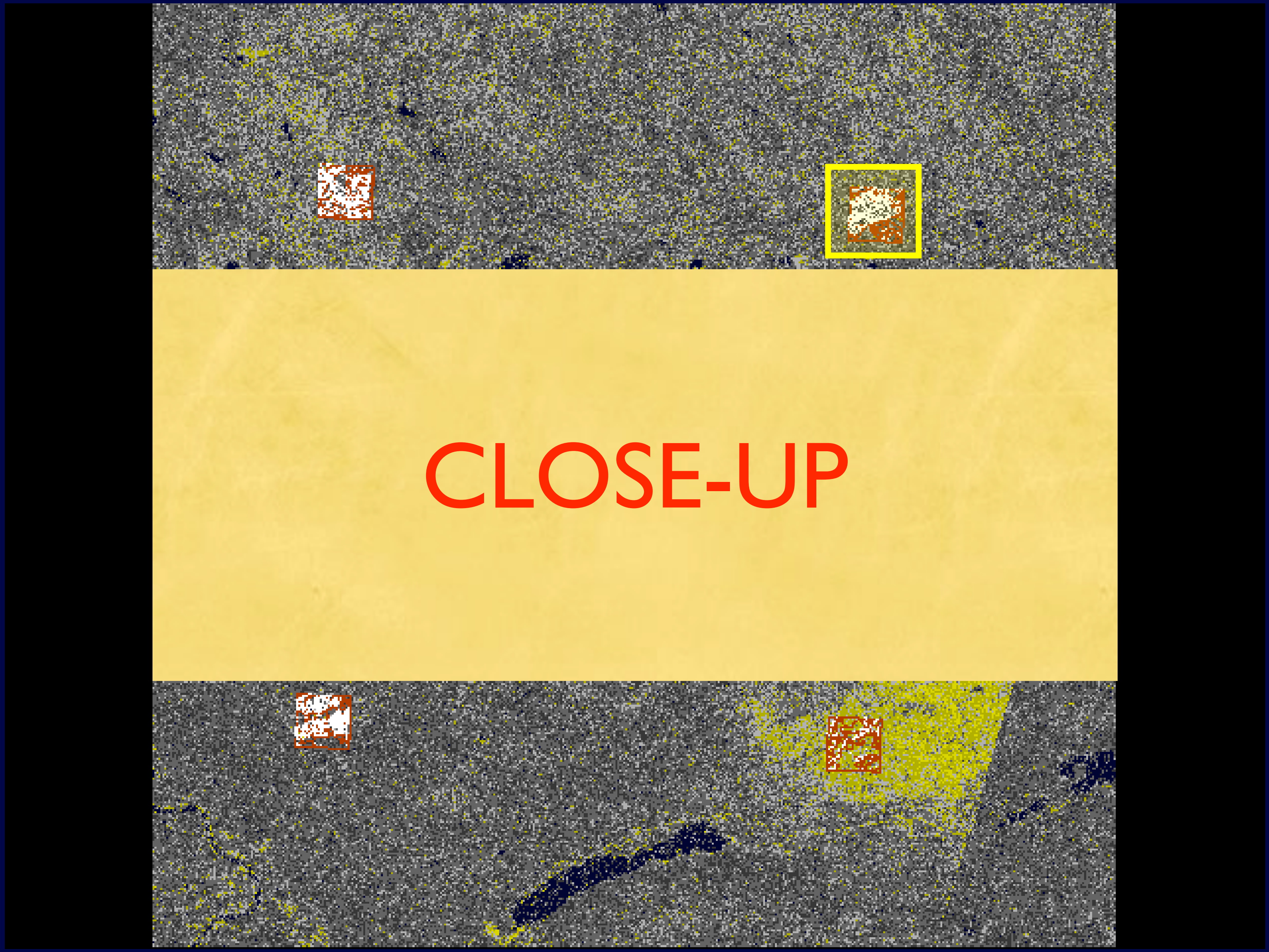
A detailed look:



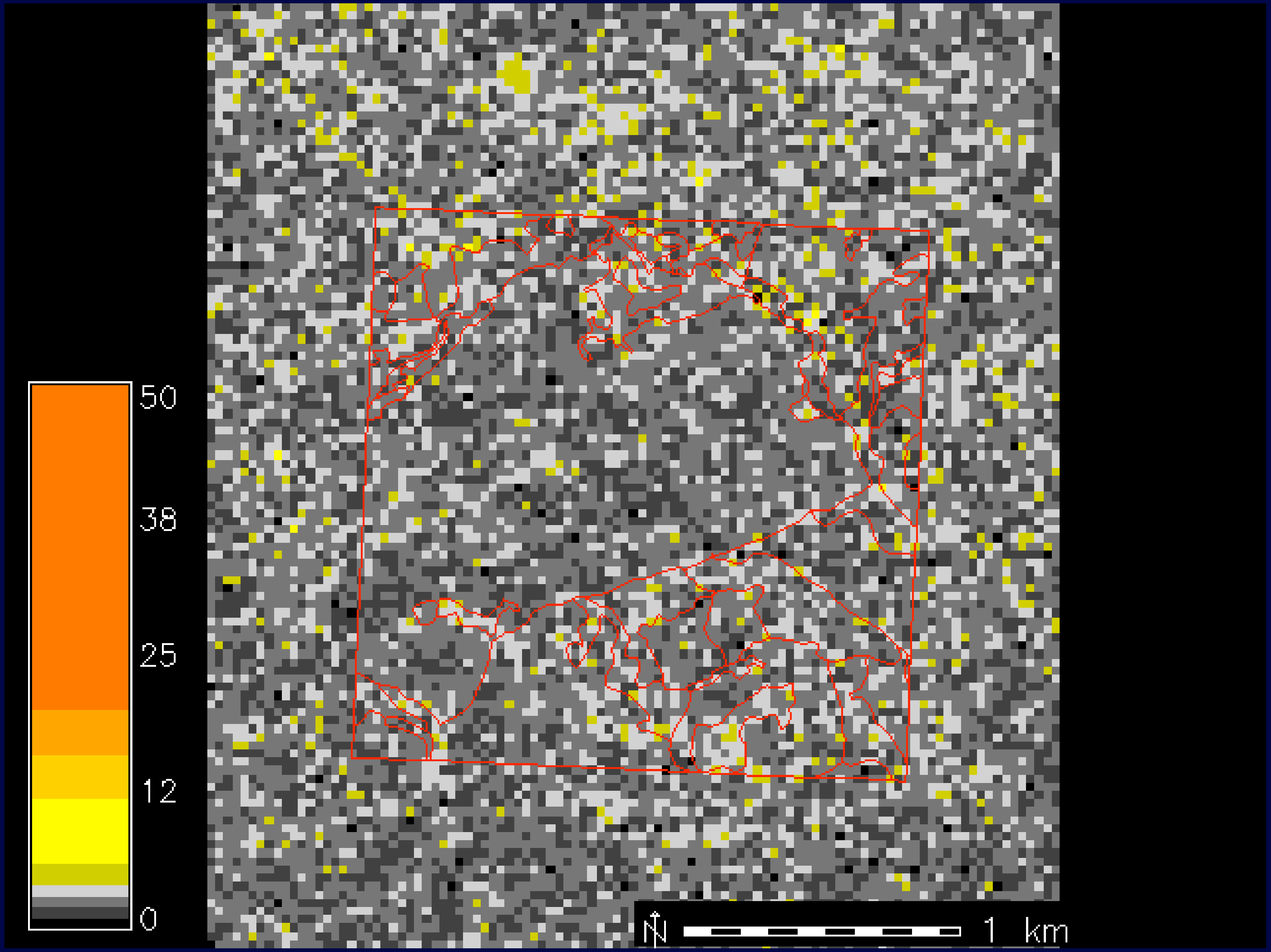
A detailed look:



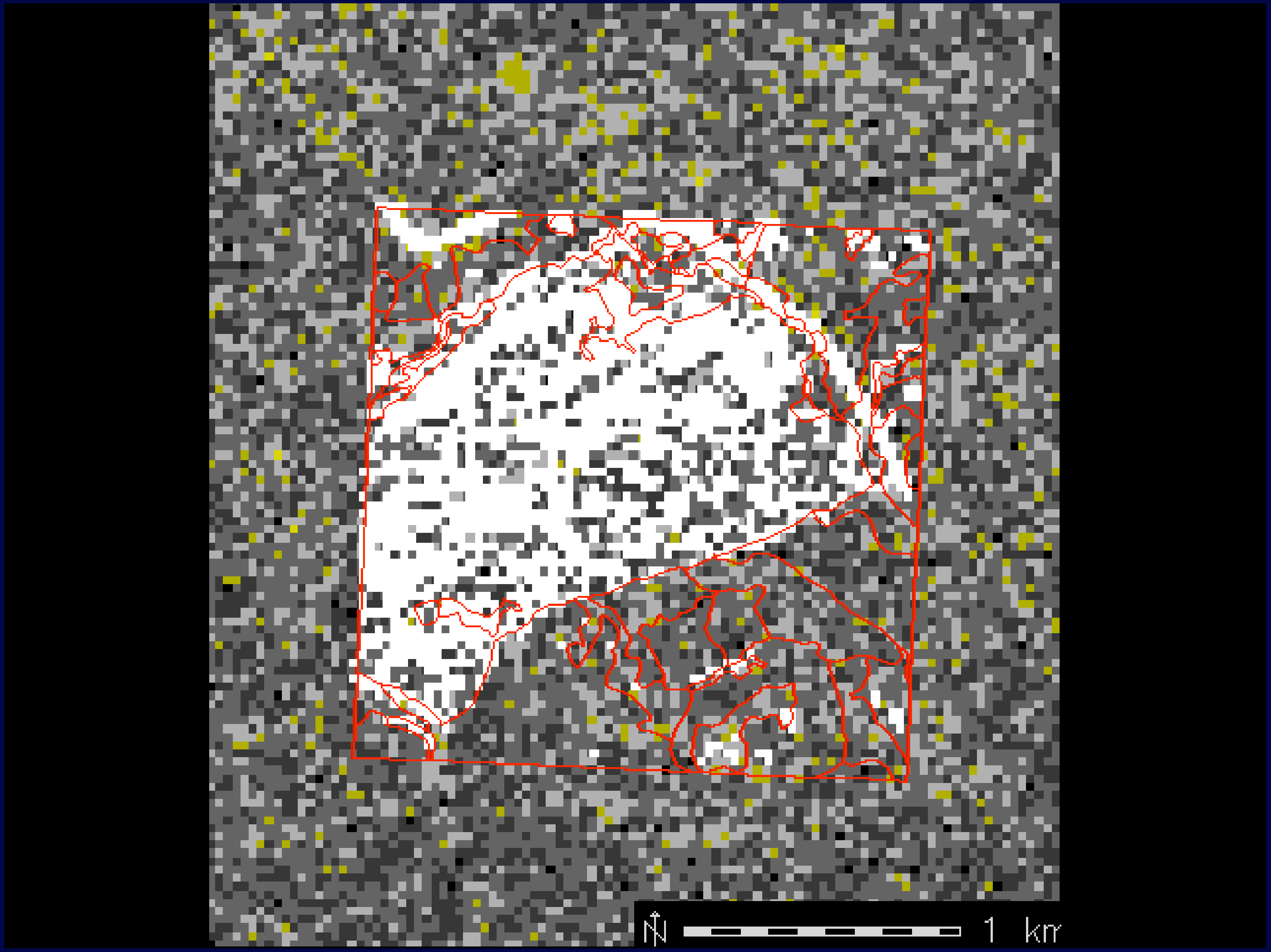
A detailed look:



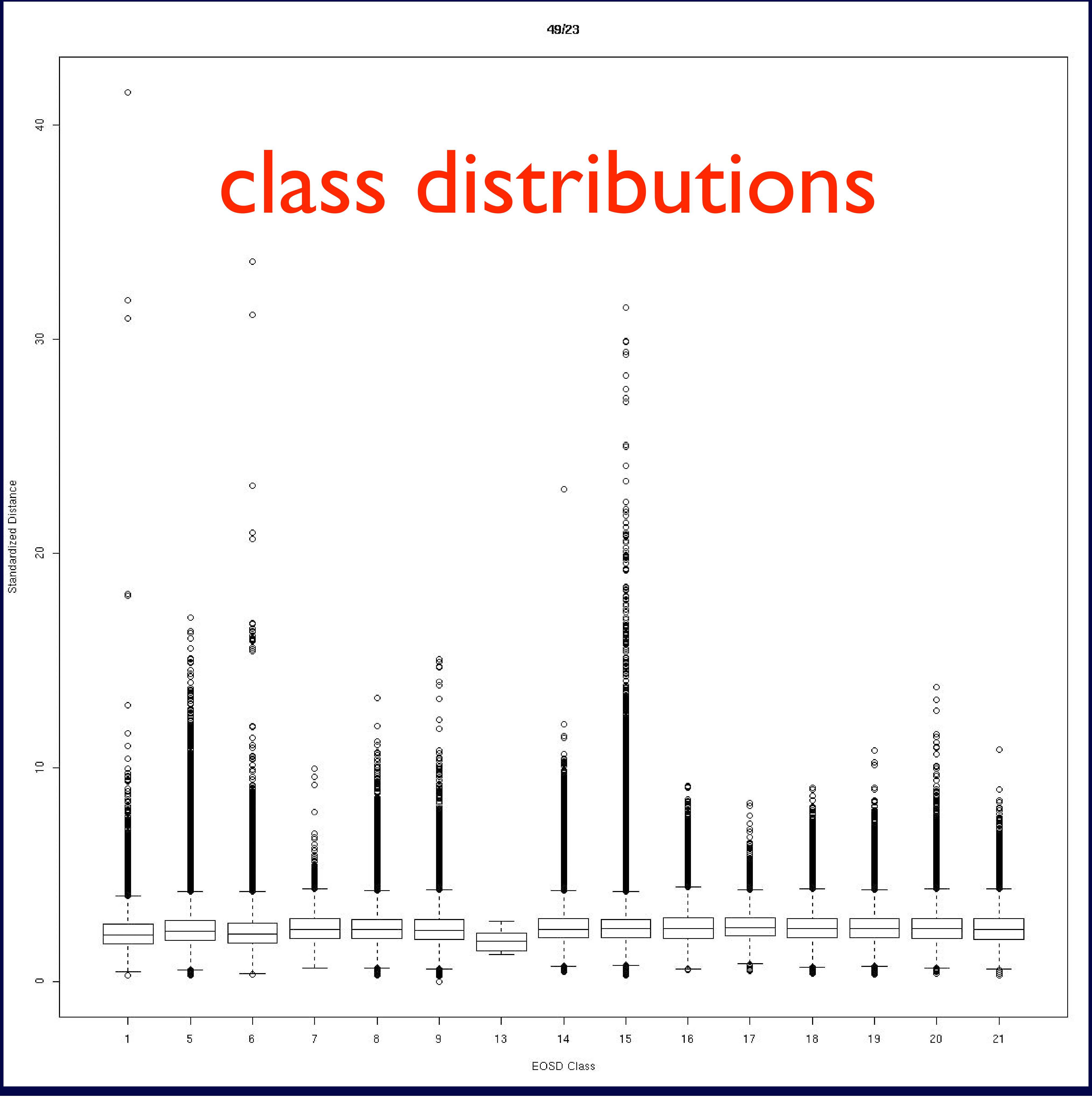
A detailed look:



A detailed look:



class distributions

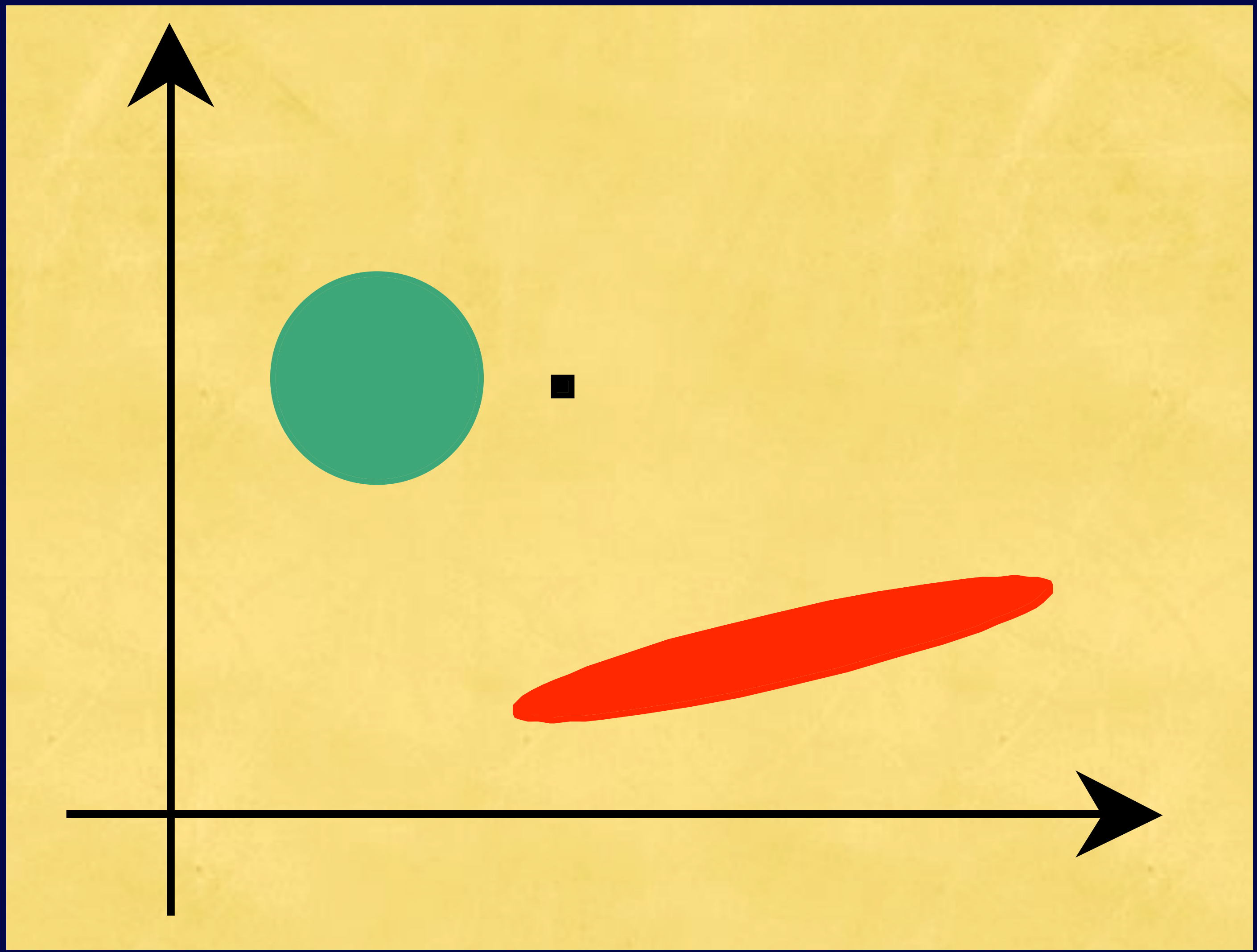


The new stuff:

- how far to NEXT closest cluster...
- relative distances, statistical significance
- R code, GRASS module / PCI output, GRASS native + ... ?
- Future?

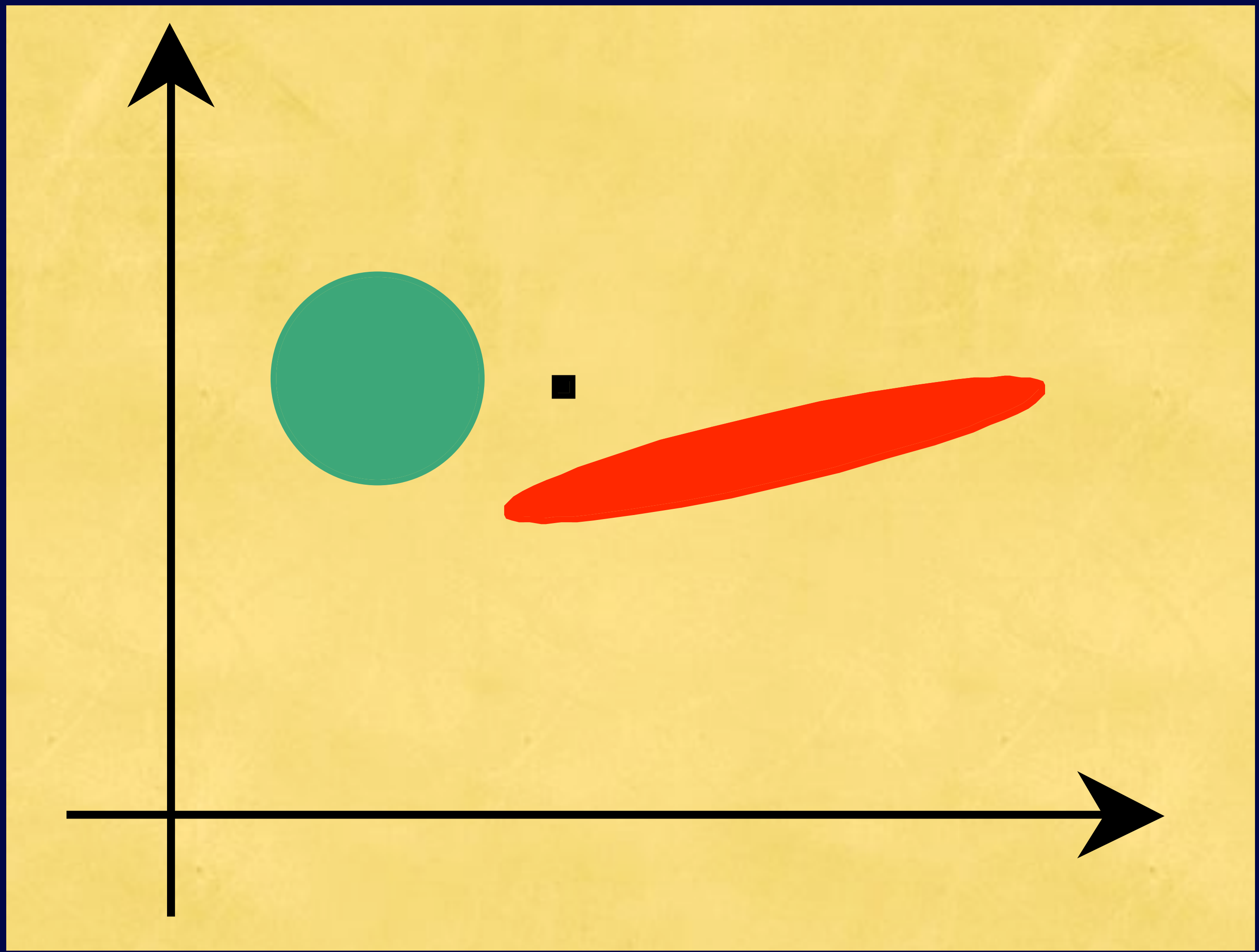
“Second closest” clusters

- instead of just calculating distance to “final” cluster, check distance of each pixel to EVERY cluster, and sort
- makes a difference if first and second distances are similar or contrasting

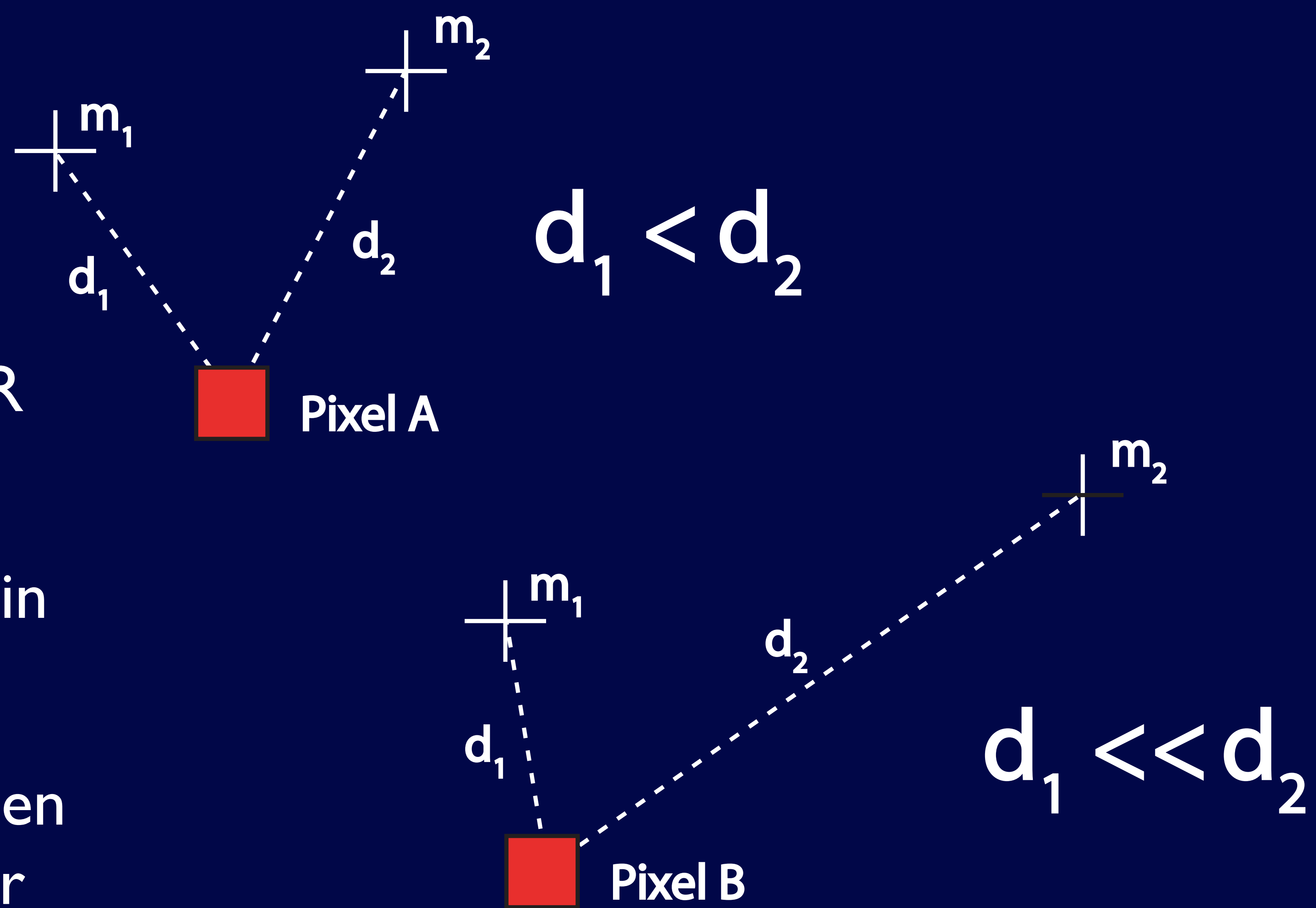


“Second closest” clusters

- instead of just calculating distance to “final” cluster, check distance of each pixel to EVERY cluster, and sort
- makes a difference if first and second distances are similar or contrasting



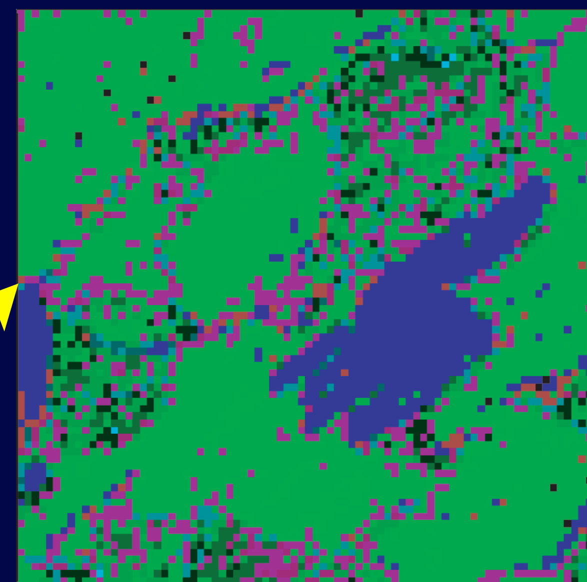
GRASS/R link to the rescue



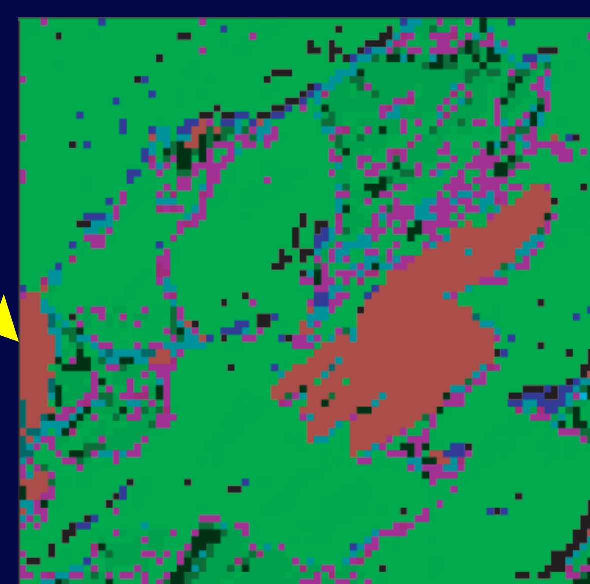
- GRASS5.4/R (original R grass package)
- 80x80 regions chosen in GRASS, read in to R
- determine ratio between first and second cluster distances

```
args(secondcluster)
function (inputclass = classes$classes, outclass = "secondclass",
    nlayers = 4, clusclas = "clusclastbl", cluscentres = "4822",
    clustermaps = clusters, nclusters = 241, nbands = 7, tmdata = tm4822,
    nrows = 80, ncols = 80, stddist = TRUE, verbose = TRUE)
```

Second Closest

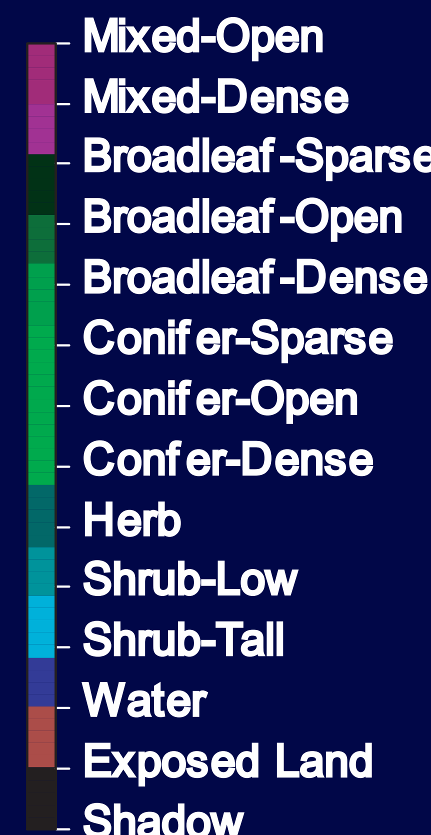


80x80 image subsets processed in R



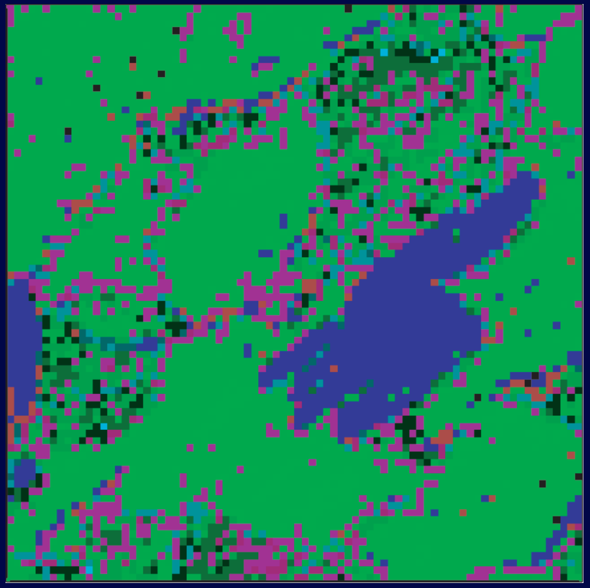
Closest

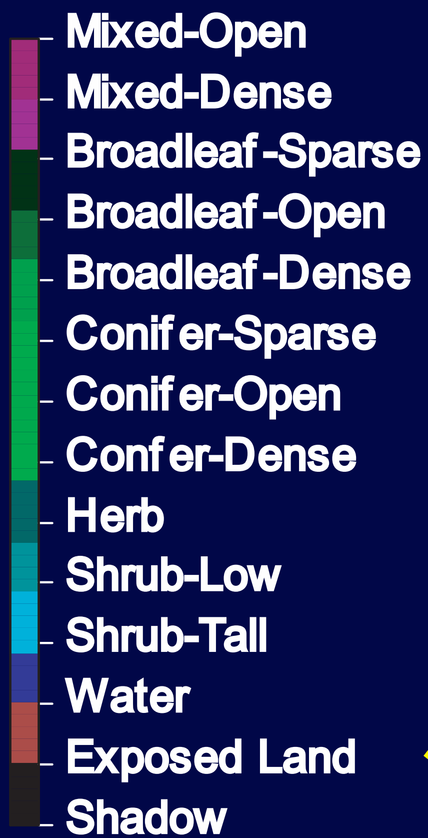
(& distances)

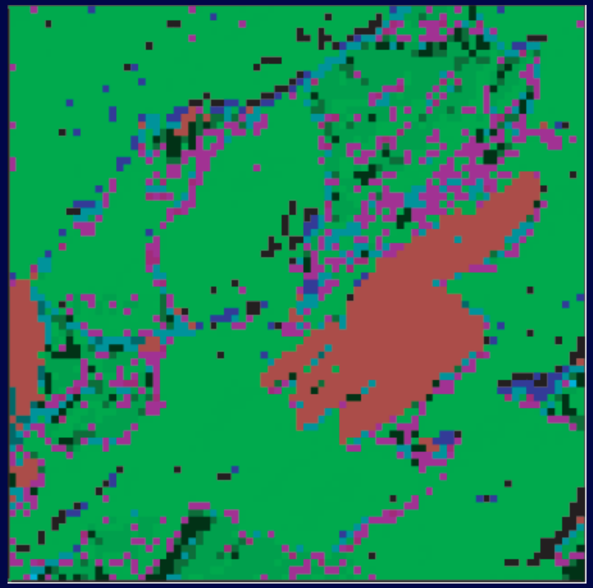


Potentially confused classes

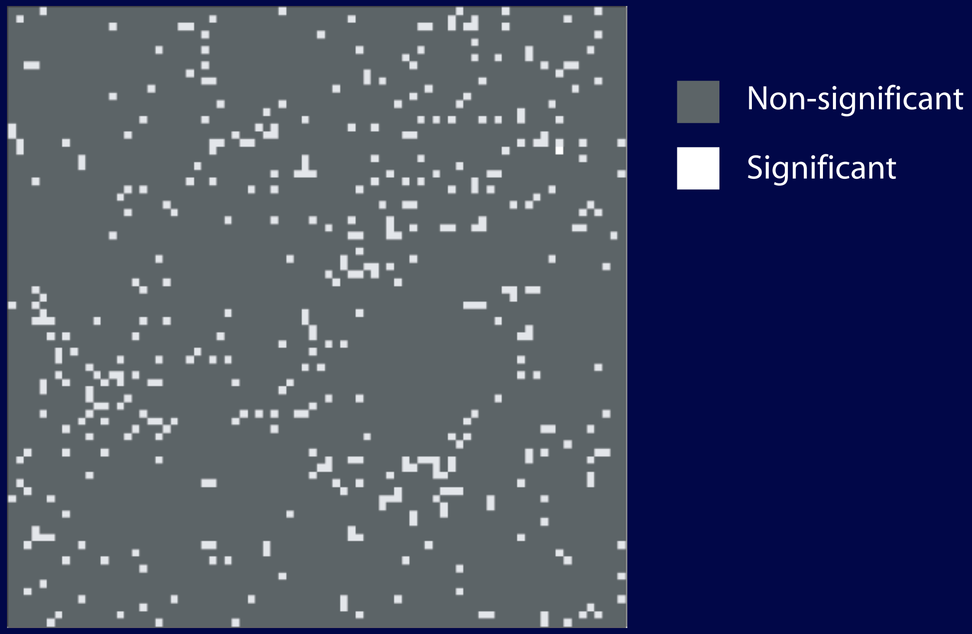
Second Closest



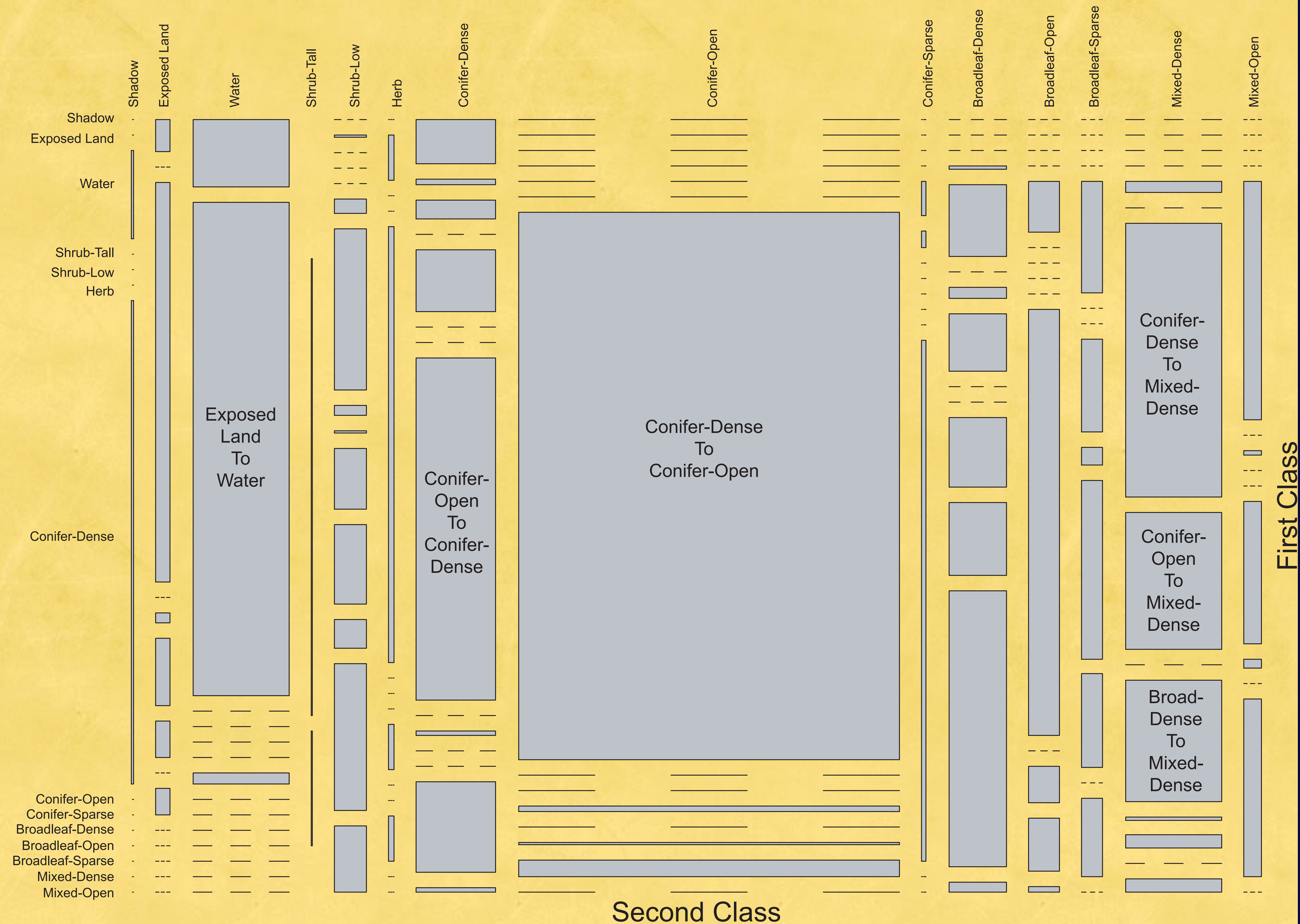




Closest



Difference Ratio
Z-score significance


Shadow
Exposed Land
Water
Shrub-Tall
Shrub-Low
Herb
Conifer-Dense
Conifer-Open
Conifer-Sparse
Broadleaf-Dense
Broadleaf-Open
Broadleaf-Sparse
Mixed-Dense
Mixed-Open
Exposed Land To Water
Conifer-Open To Conifer-Dense
Conifer-Dense To Conifer-Open
Conifer-Dense To Mixed-Dense
Conifer-Open To Mixed-Dense
Broad-Dense To Mixed-Dense
Second Class
First Class

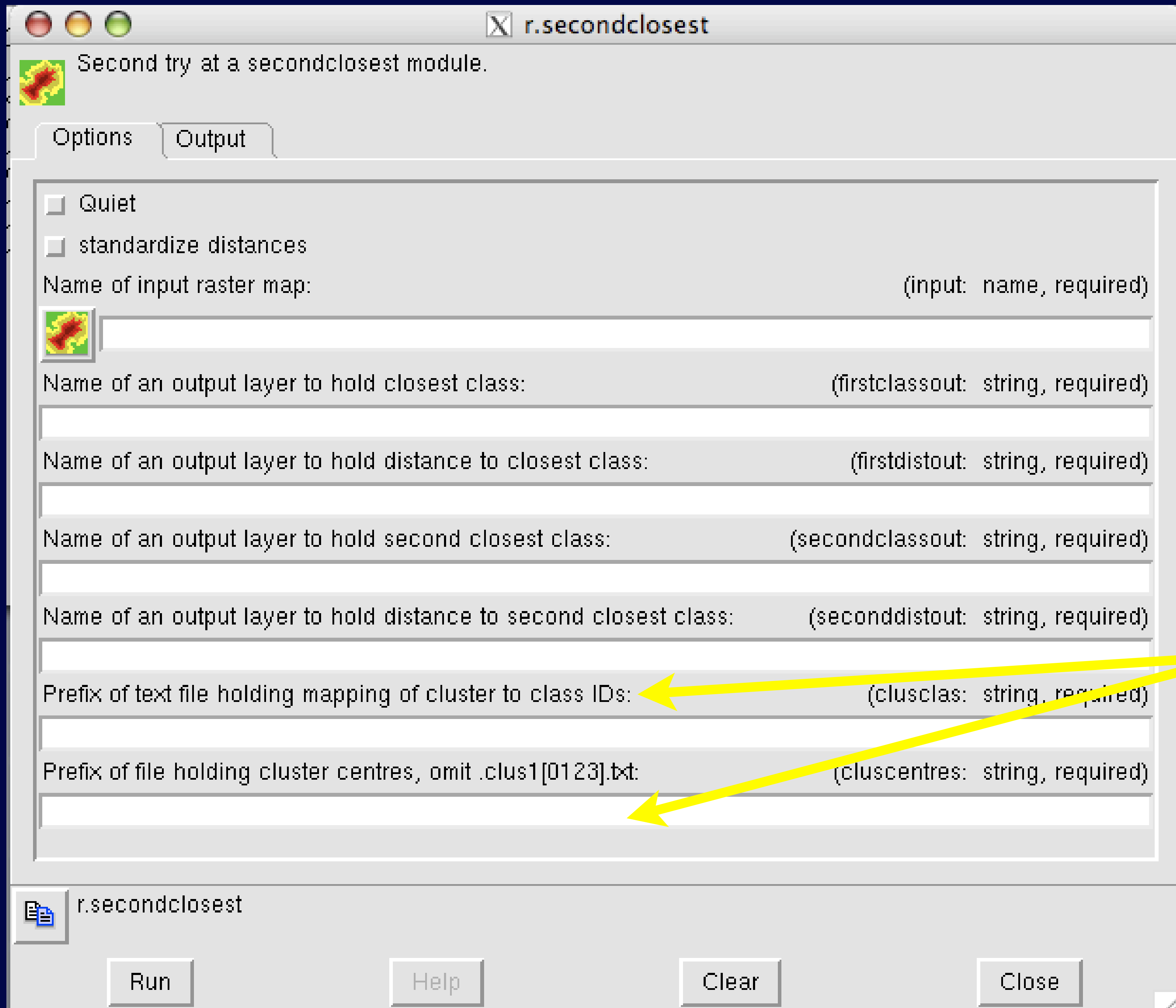
What we learned...

- Shrub-low is the most frequently confused class (with completely different thematic groups)
- within forest types, density is easily confused
- technique provides a MAP (contrast to global measures) of significance of thematic confusion
- exploratory tools in R provide rich environment to identify both systematic and spatial anomalies in classification dataset
- data volume in typical remote sensing applications a challenge for R using a monolithic approach... need alternatives...

Feasibility shown, now what?

- some options:
  - this is all per-pixel (spatially "independent"), so could chop up and put back together; still a time issue
  - could sample
  - could write a GRASS module
- r.secondclosest

GRASS module



PCI/CFS kludges

Coming soon...

- have started work on cleaned up code that is not tied to this data flow (probably i.clusterdists or ... ?)
  - remove the #$%^@#$ "layered data" handling !
  - remove the "shadow correction" checking (*)
  - replace parsing of PCI cluster stats table with use of GRASS signatures
- extensions of the approach: fuzzy logic membership functions, ...

Conclusions

- For algorithm development, R and the GRASS/R links (& friends) provide a good environment to test ideas (especially for those of us with weak C skills)
- some tasks challenge R's interpreted, in memory approach; R is not intended to be a GIS (even with spgrass6!)
- getting the algorithm right in R allowed for faster development of a GRASS module
- the cluster distances approach provides an effective evaluation of per-pixel classification confidence