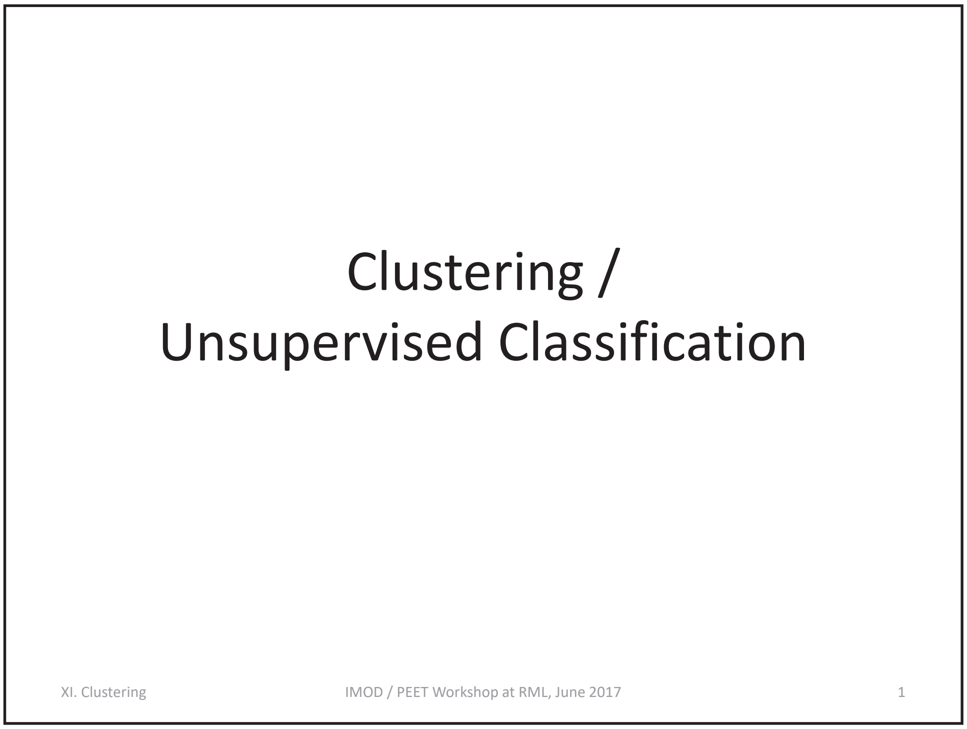
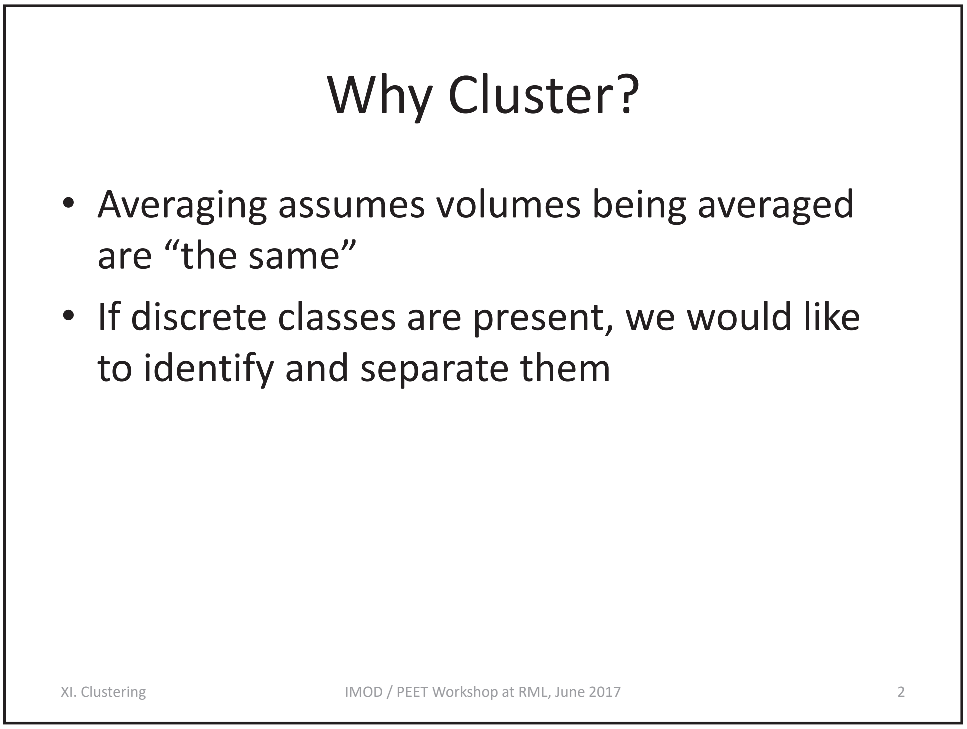
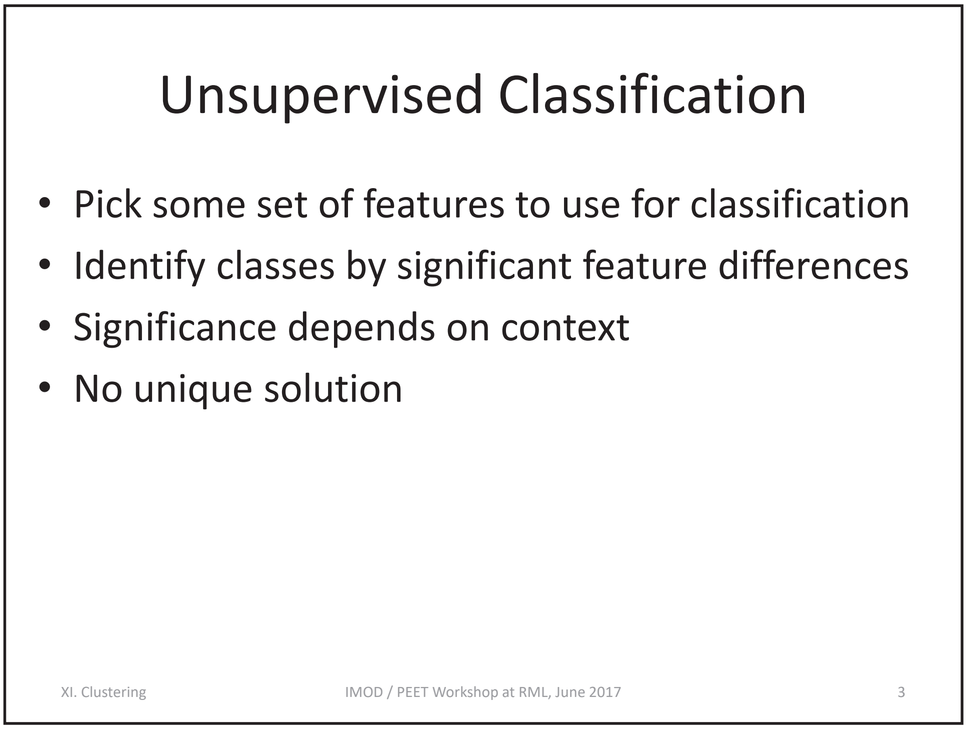


# Clustering / Unsupervised Classification



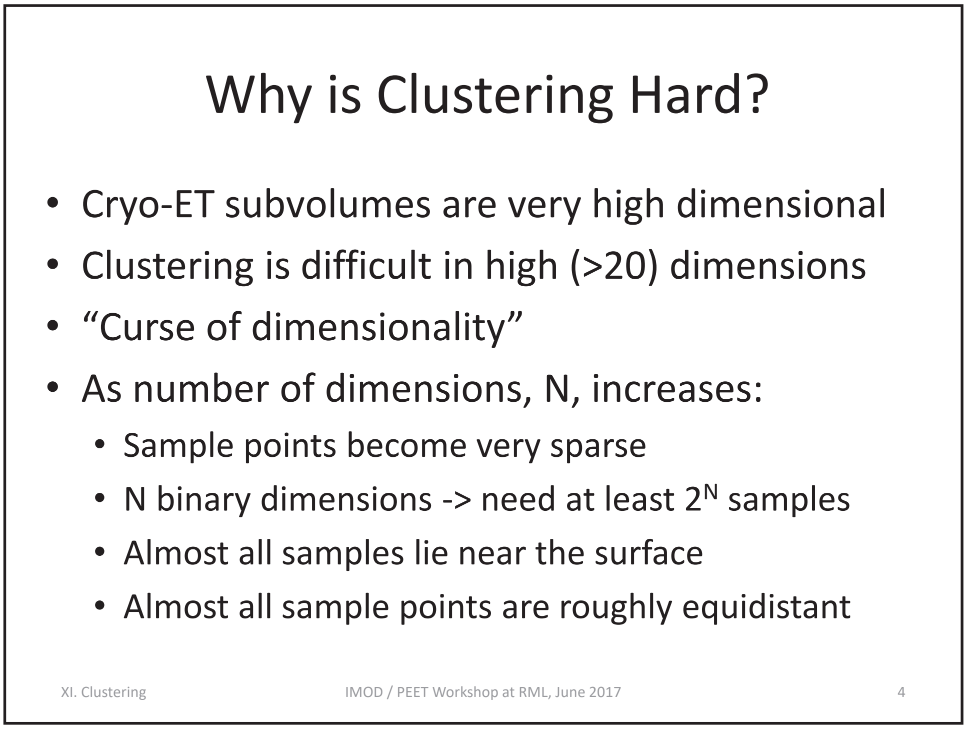
## Why Cluster?

- Averaging assumes volumes being averaged are “the same”
- If discrete classes are present, we would like to identify and separate them



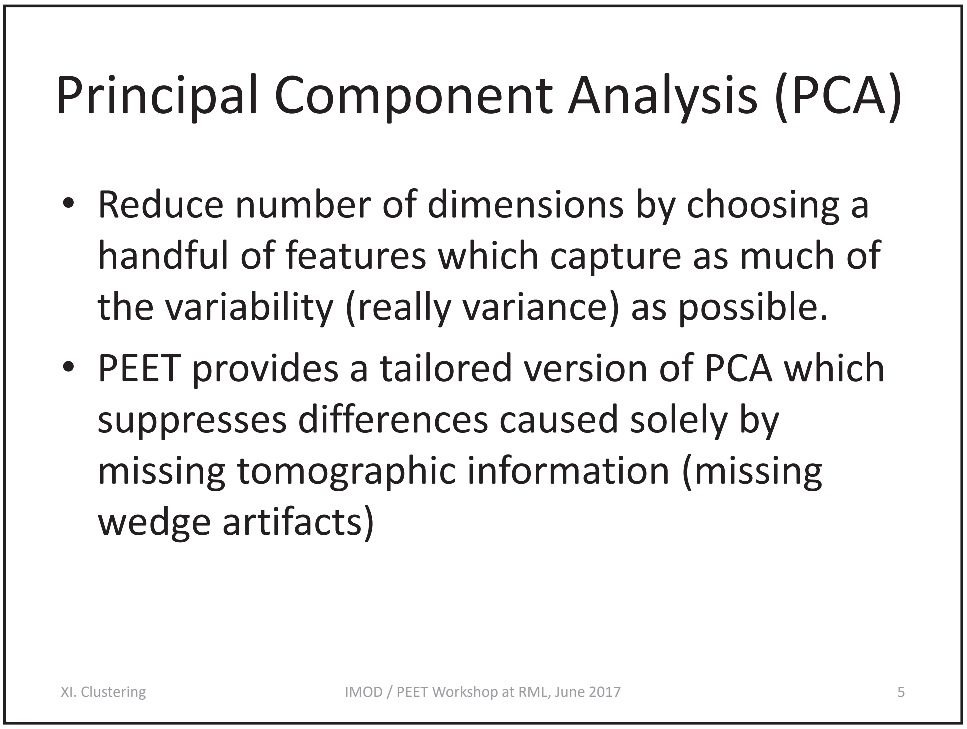
## Unsupervised Classification

- Pick some set of features to use for classification
- Identify classes by significant feature differences
- Significance depends on context
- No unique solution



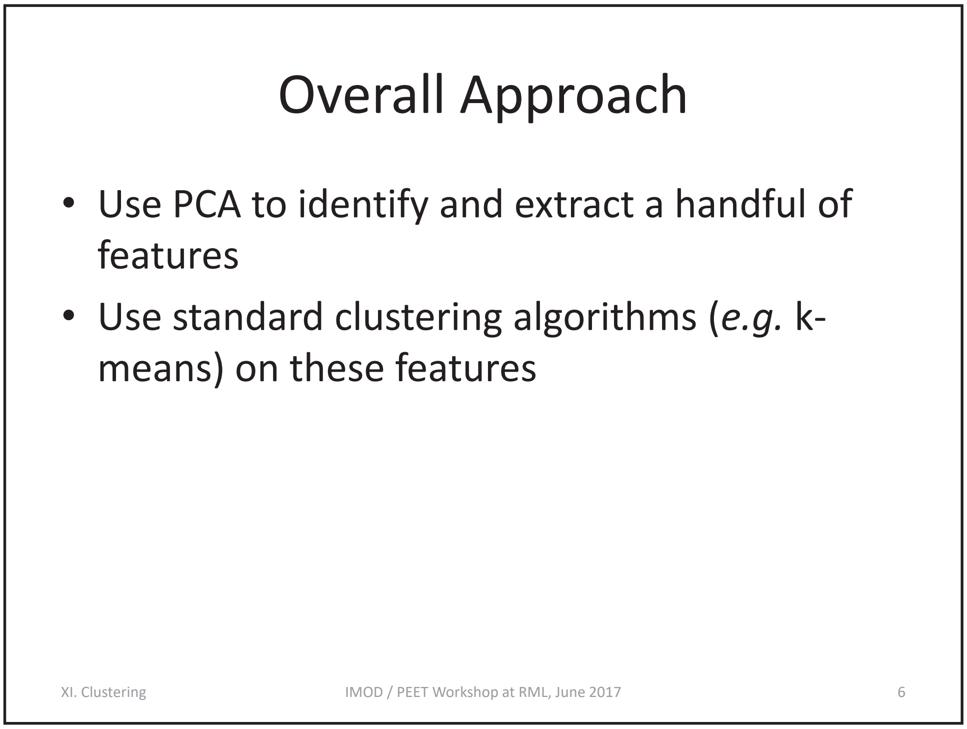
## Why is Clustering Hard?

- Cryo-ET subvolumes are very high dimensional
- Clustering is difficult in high (>20) dimensions
- “Curse of dimensionality”
- As number of dimensions, N, increases:
  - Sample points become very sparse
  - N binary dimensions -> need at least $2^N$ samples
  - Almost all samples lie near the surface
  - Almost all sample points are roughly equidistant



## Principal Component Analysis (PCA)

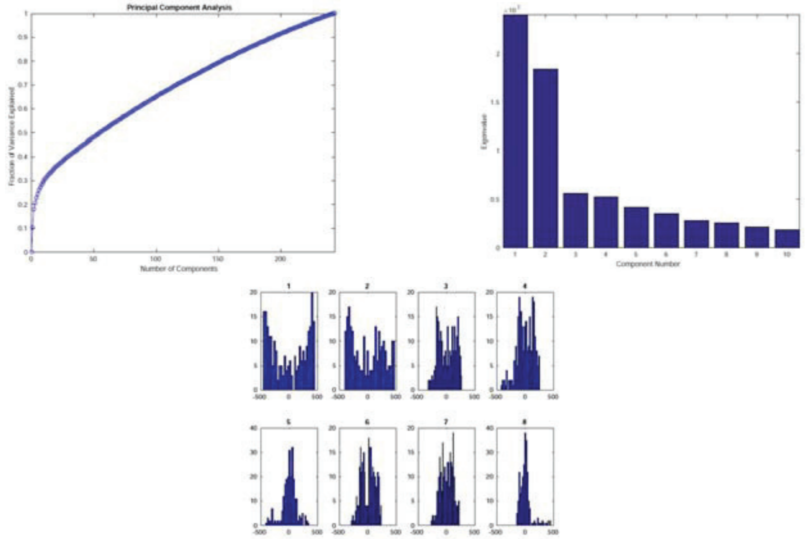
- Reduce number of dimensions by choosing a handful of features which capture as much of the variability (really variance) as possible.
- PEET provides a tailored version of PCA which suppresses differences caused solely by missing tomographic information (missing wedge artifacts)



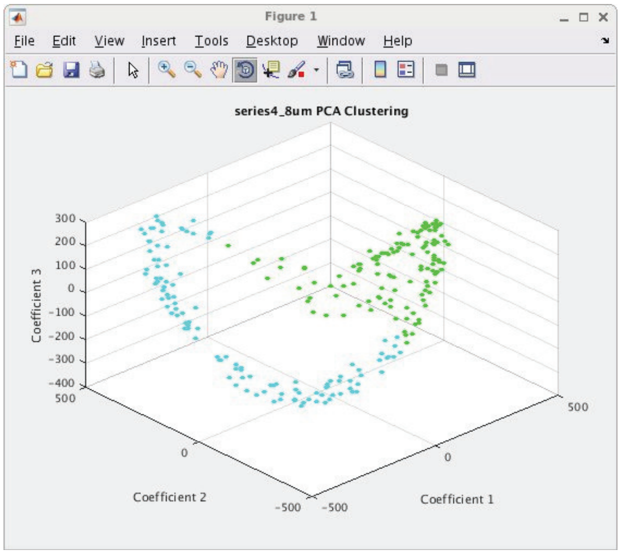
## Overall Approach

- Use PCA to identify and extract a handful of features
- Use standard clustering algorithms (*e.g.* k-means) on these features

# Sample PEET PCA Output



# MT 1st Search Clustering

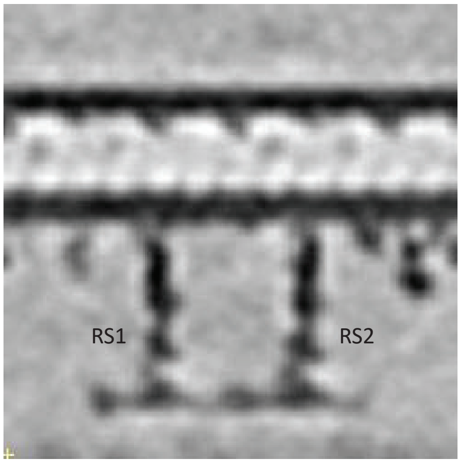


## MT 1$^{st}$ Search Clustering Interpretation

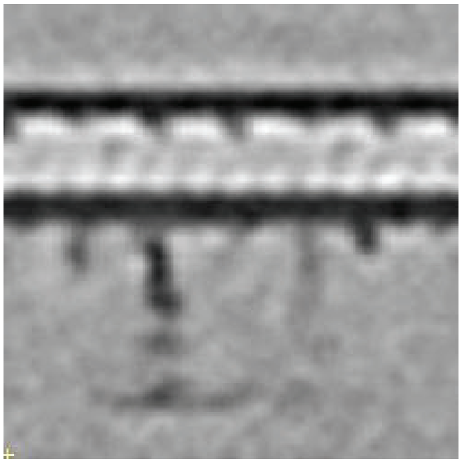
- Statistically significant but probably not due to real sample heterogeneity
- Missing wedge artifacts affect alignment which affects clustering
- Variation with axial orientation

## Chlamydomonas Axonemes

(from D. Nicastro and T. Heuser)



pWT



6E6

## Chlamydomonas Axonemes

- RS2 density greatly reduced in 6E6 mutant
- Is this due to flexibility or heterogeneity?
  - Maybe all mutants have RS2, but configuration is variable
  - Maybe some mutants completely lack RS2
- We will explore this case in the lab exercise

# Questions?