Learning to Classify Galaxy Shapes using the EM Algorithm

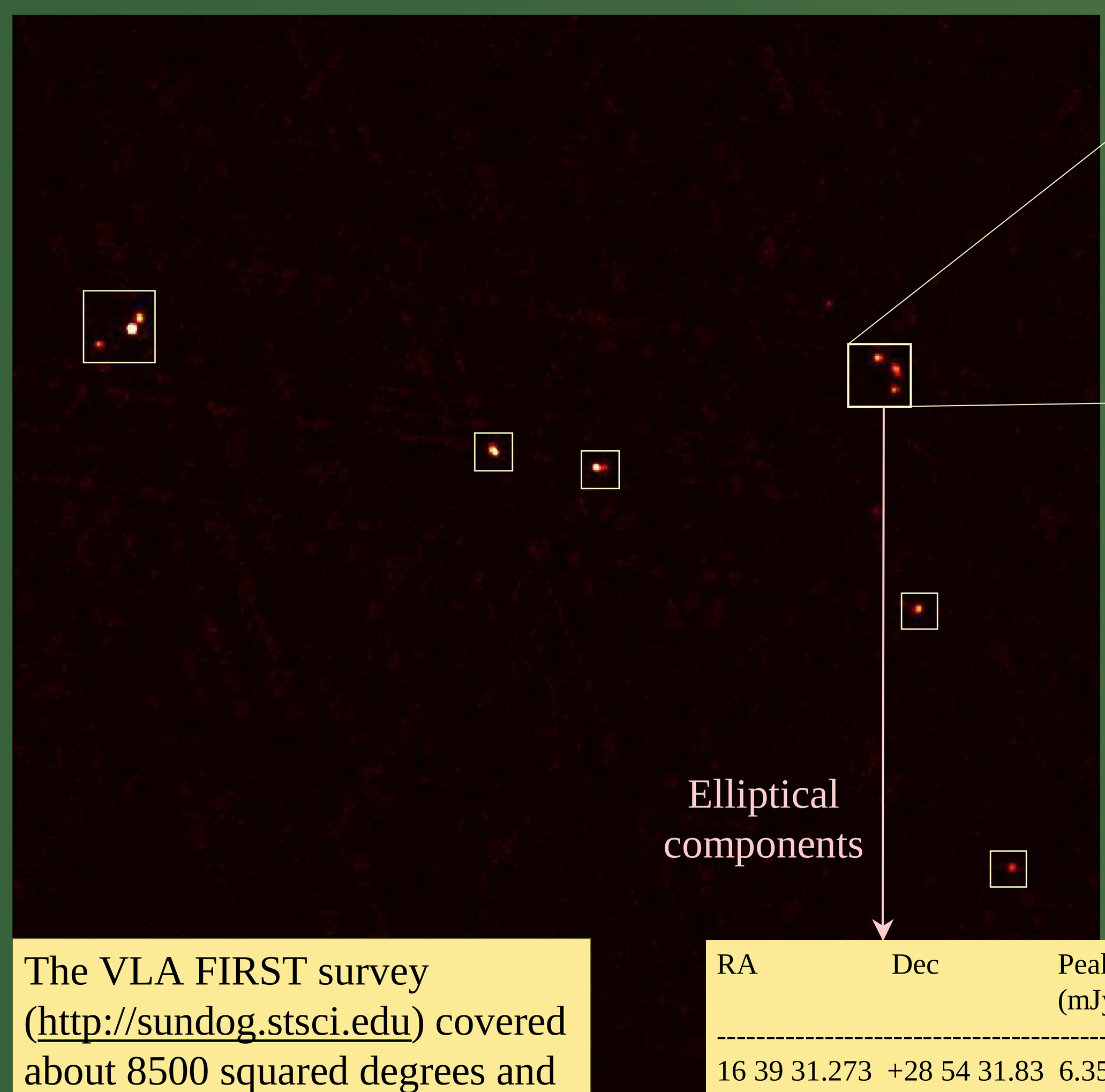
Sergey Kirshner[1], Igor V. Cadez[2], Padhraic Smyth[1], Chandrika Kamath[3]

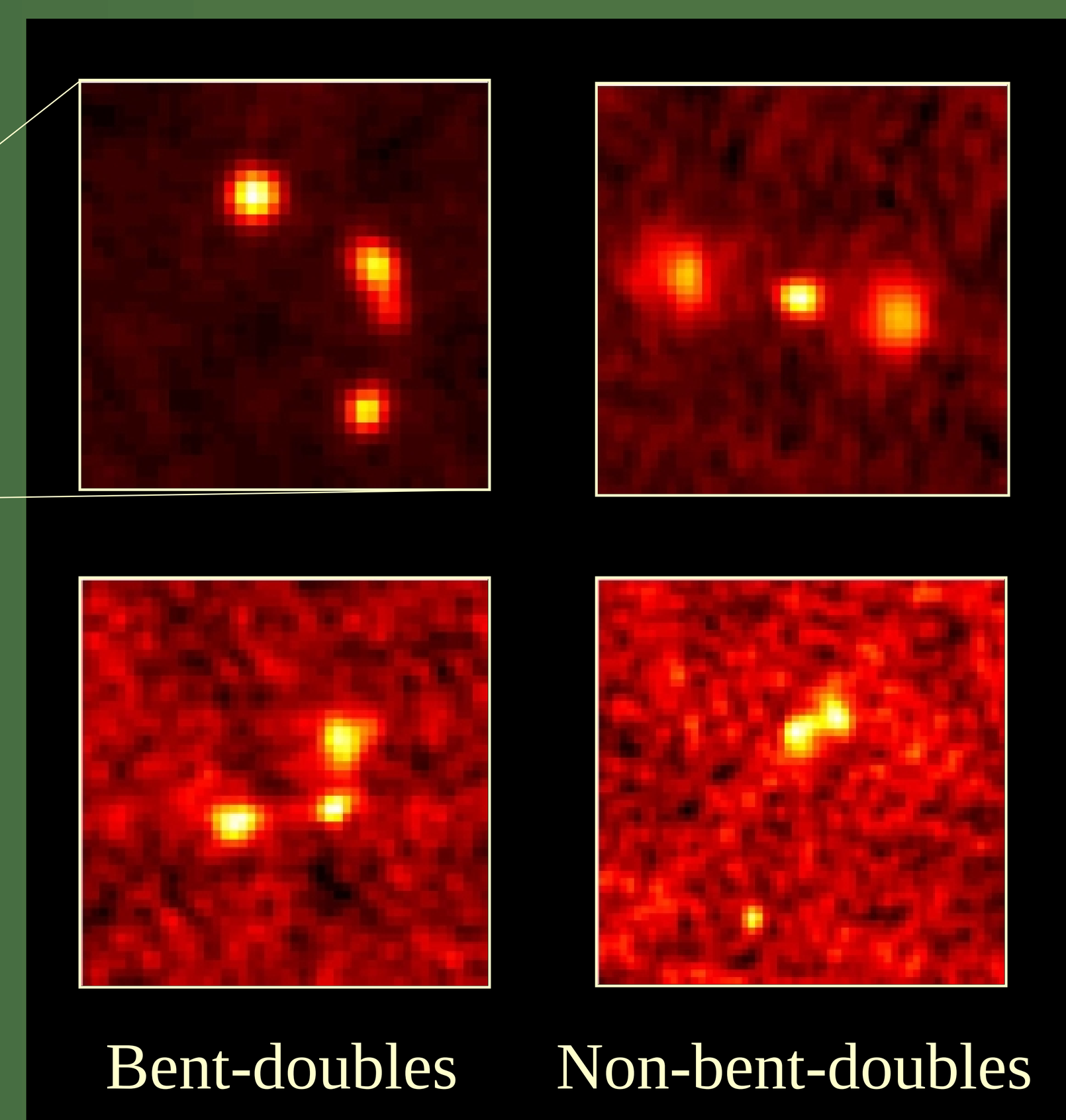
[1] Information and Computer Science, University of California, Irvine, CA 92697-3425, skirshne@ics.uci.edu, smyth@ics.uci.edu

[2] Sparta Inc., 23382 Mill Creek Drive #100, Laguna Hills, CA 92653, icadez@ics.uci.edu

[3] Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551, kamath2@llnl.gov

We describe an application of probabilistic learning to the problem of automatic identification of astronomical radio sources with a bent-double morphology. Calculation of object features requires identifying an ordering on the "corners" of the object of interest. We describe an approach for automatic ordering and calculation of features and demonstrate how classes of objects can be learned in both a supervised and unsupervised manner.





The VLA FIRST survey (http://sundog.stsci.edu) covered about 8500 squared degrees and identified about 7.7×10^5 sources.

RA	Dec	Peak Flux (mJy/bm)	Integrated Flux (mJy)	RMS (mJy/b)	Major Axis (arcsec)	Minor Axis (arcsec)	Position Angle (degrees)
16 39 31.273	+28 54 31.83	6.35	14.40	0.134	11.03	6.00	20.5
16 39 31.399	+28 54 9.36	6.76	7.93	0.134	6.25	5.48	153.3
16 39 32.800	+28 54 44.58	8.97	14.12	0.135	6.83	6.72	0.1

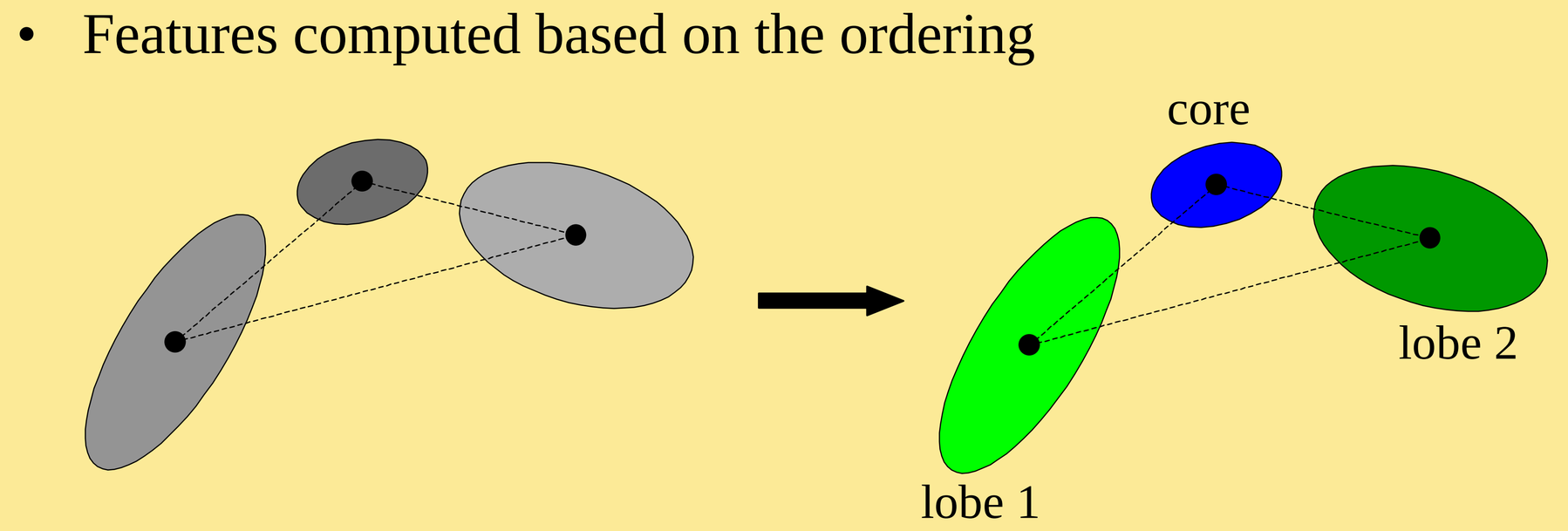
- Set labeled by scientists
  - 128 bent-doubles
  - 22 non-bent-doubles
- Objective: to classify new configurations

Orderings and Features

- Build probabilistic model for bent-doubles
- Features based on symmetry
- Need to order components to calculate features
- Compute features given the ordering:
  - angles
  - side ratios
  - peak intensity ratios
- Assume independence of features given the ordering
- Assume pairwise independence for each feature
- Incorporate constraints for features
  - angles add up to π
  - product of ratios is 1

How to Compute Features

- Features computed based on the ordering



- One value of feature per component per ordering
  - angle(●) = … angle(●) = … angle(●) = …
  - side ratio(●) = … side ratio(●) = … side ratio(●) = …
  - intensity ratio(●) = peak(●)/peak(●) intensity ratio(●) = peak(●)/peak(●) intensity ratio(●) = peak(●)/peak(●)

P(features|) = P(angles|)×P(side ratios|)×P(intensity ratios|)

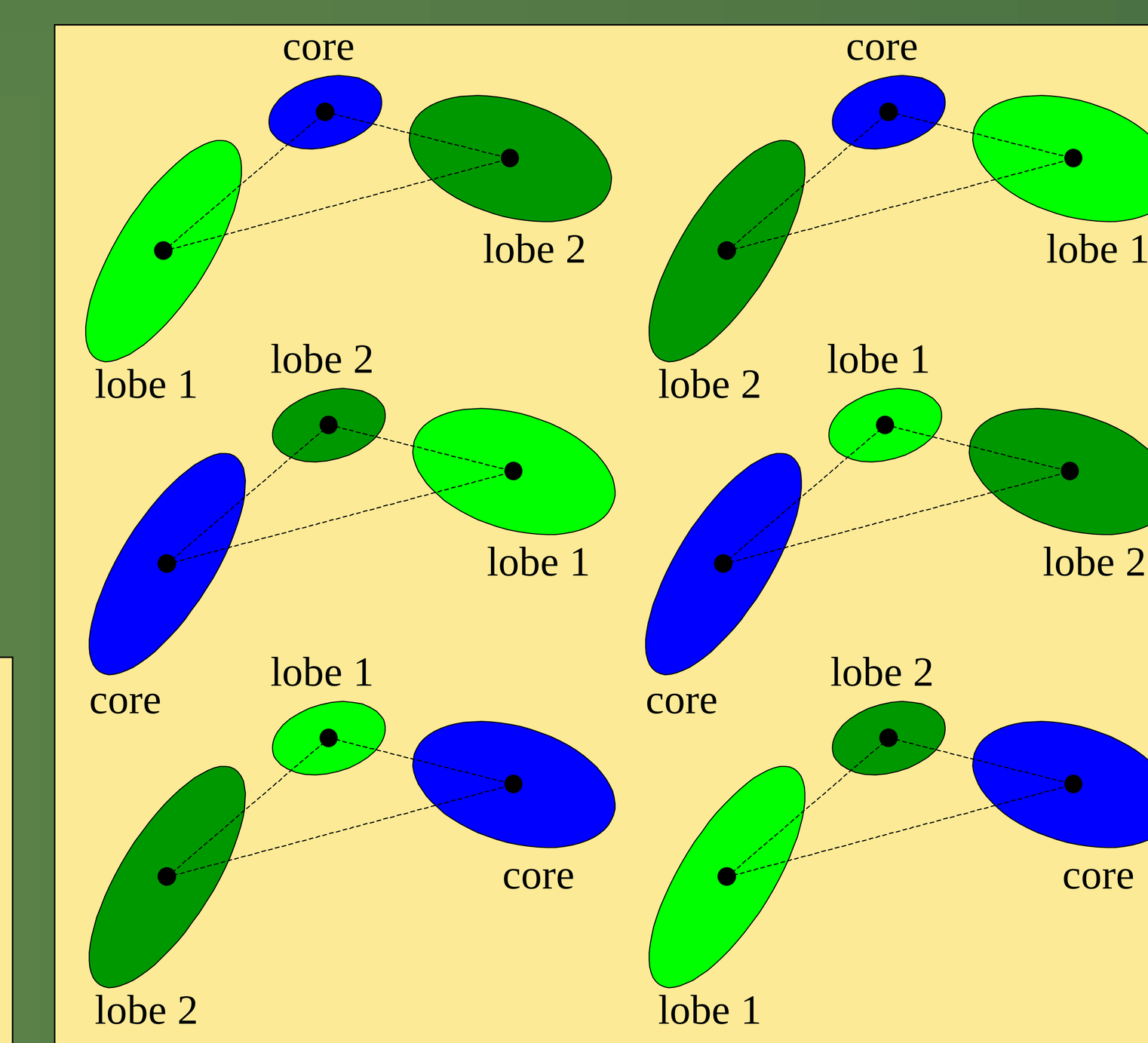
P(angles|) = P(angle(● |), angle(● |), angle(● |)) =
= P(angle(● |)|angle(● |), angle(● |))×
×P(angle(● |)|angle(● |))×P(angle(● |)) =
= P(angle(● |))× P(angle(● |))

P(side ratios|) = P(side ratio(● |))×P(side ratio(● |))

P(intensity ratios|) = P(intensity ratio(● |))×P(intensity ratio(● |))

Two Problems

- There are six possible orderings – which one is best? Proper orderings for bent-doubles in the data may not be known.
- Non-bent-doubles do not have a correct ordering since they may not exhibit any symmetry that the features are designed to capture. Need to come up with a scheme to be able to assign a probability score to all configurations.



Unscrambling Problem

Examples	Components p_1	p_2	…	p_M
Object 1	e_{11}	e_{12}	…	e_{1M}
Object 2	e_{21}	e_{22}	…	e_{2M}
Object 3	e_{31}	e_{32}	…	e_{3M}
Object 4	e_{41}	e_{42}	…	e_{4M}
……	…	…	…	…
Object N	e_{N1}	e_{N2}	…	e_{NM}

Scramble function →

Object 1	$e_{1?}$	$e_{1?}$	…	$e_{1?}$
Object 2	$e_{2?}$	$e_{2?}$	…	$e_{2?}$
Object 3	$e_{3?}$	$e_{3?}$	…	$e_{3?}$
Object 4	$e_{4?}$	$e_{4?}$	…	$e_{4?}$
……	…	…	…	…
Object N	$e_{N?}$	$e_{N?}$	…	$e_{N?}$

- Components correspond to vectors of features calculated for elliptical components.
- How to unscramble the components?
- How to learn parameters for distributions $p_1, \dots, p_M$ (or their joint distribution) and the scramble function?

Mixture of Orderings

- Reformulate probability of a configuration as a mixture over possible orderings

P( ) = P(features|)×P( ) + … + P(features|)×P( )

Parameterization

- Parametrize $P_{●}$ and $P_{●}$ for all features as Gaussians
  - use transformed features
- Want to learn the parameters of the model for bent-doubles to fit the available data best
  - maximize the likelihood of the data:

$$L = \prod_{e\in \text{bent-doubles}} P(e) = \prod_{e\in \text{bent-doubles}} \sum_{o\in \text{orderings}} P(\text{features}(e)|o)\times P(o)$$

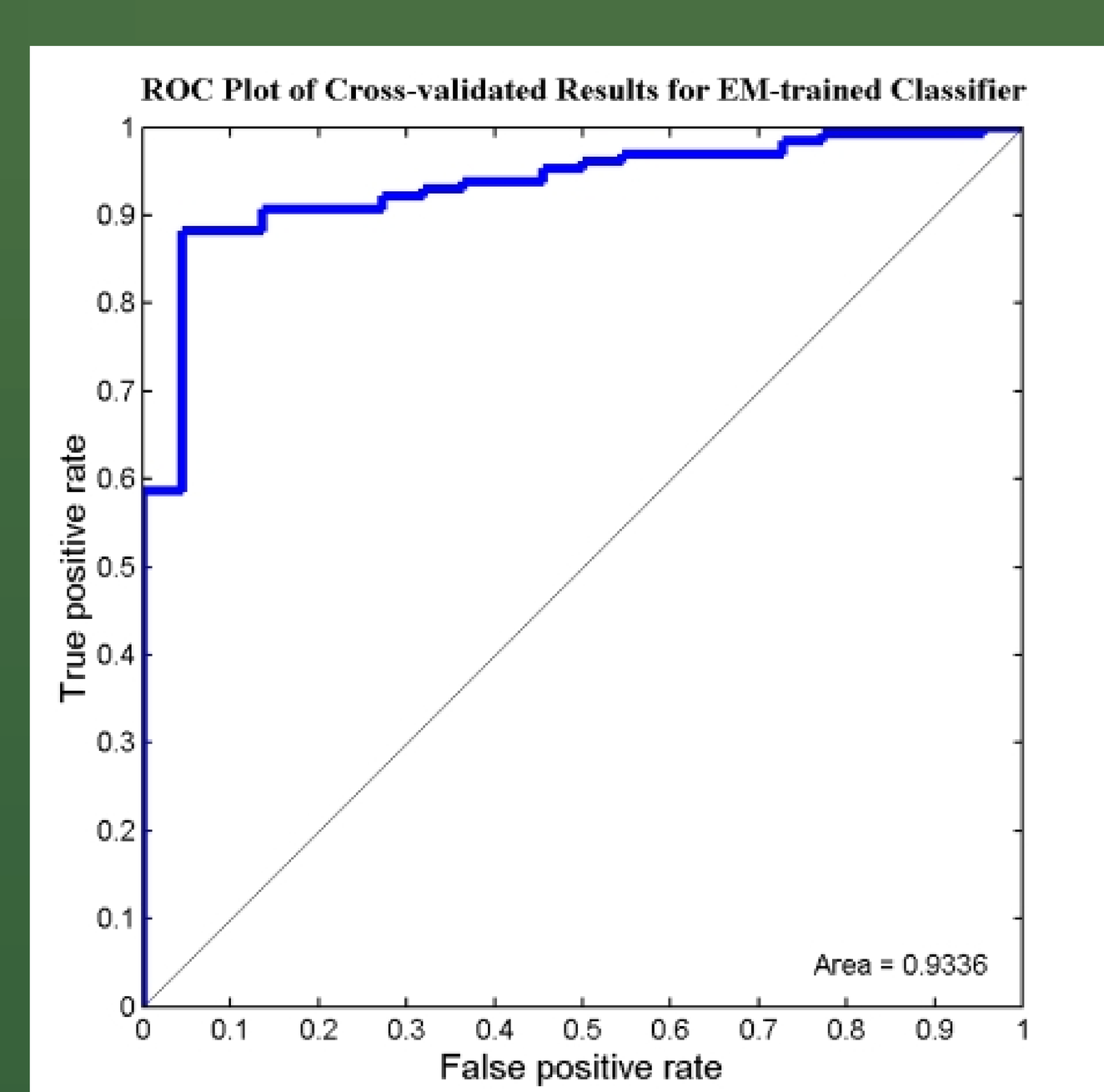
  - sum defined in Mixture of Orderings
  - consider distribution for orderings uniform
  - learn the parameters of the Gaussians (means and variances)
- Use Expectation-Maximization algorithm to learn the parameters
- Once the parameters known, can calculate probability given the model for any configuration

Evaluation

- Use receiver operating characteristics (ROC) curve plot for evaluation
  - rate of false positives vs true positives
- Build ROC curve
  - for each configuration in the set
    - train model on all other bent-doubles in the set
    - calculate the score of the configuration given the model
  - rank the examples by their scores
  - for each threshold between consecutive scores, calculate how many true positives and false positives one would get
  - plot as a ratio to all positives
- Use area below the curve to compare methods
- Can be converted into classifier by setting a threshold for positives

Expectation Maximization (EM)

- Iterative algorithm changing the parameters to increase the log-likelihood of the data in two steps per iteration
- Initialize – can select parameters at random
- E-step – calculate probability distribution
  - for each example in the training set, calculate the probability of each possible ordering given the current model
- M-step – maximize the expected log-likelihood
  - select parameters for the new model to maximize the expected log-likelihood of data and orderings according to the distribution from E-step
- Repeat until convergence reached
- Convergence guaranteed; possibly, local maximum



ROC Plot

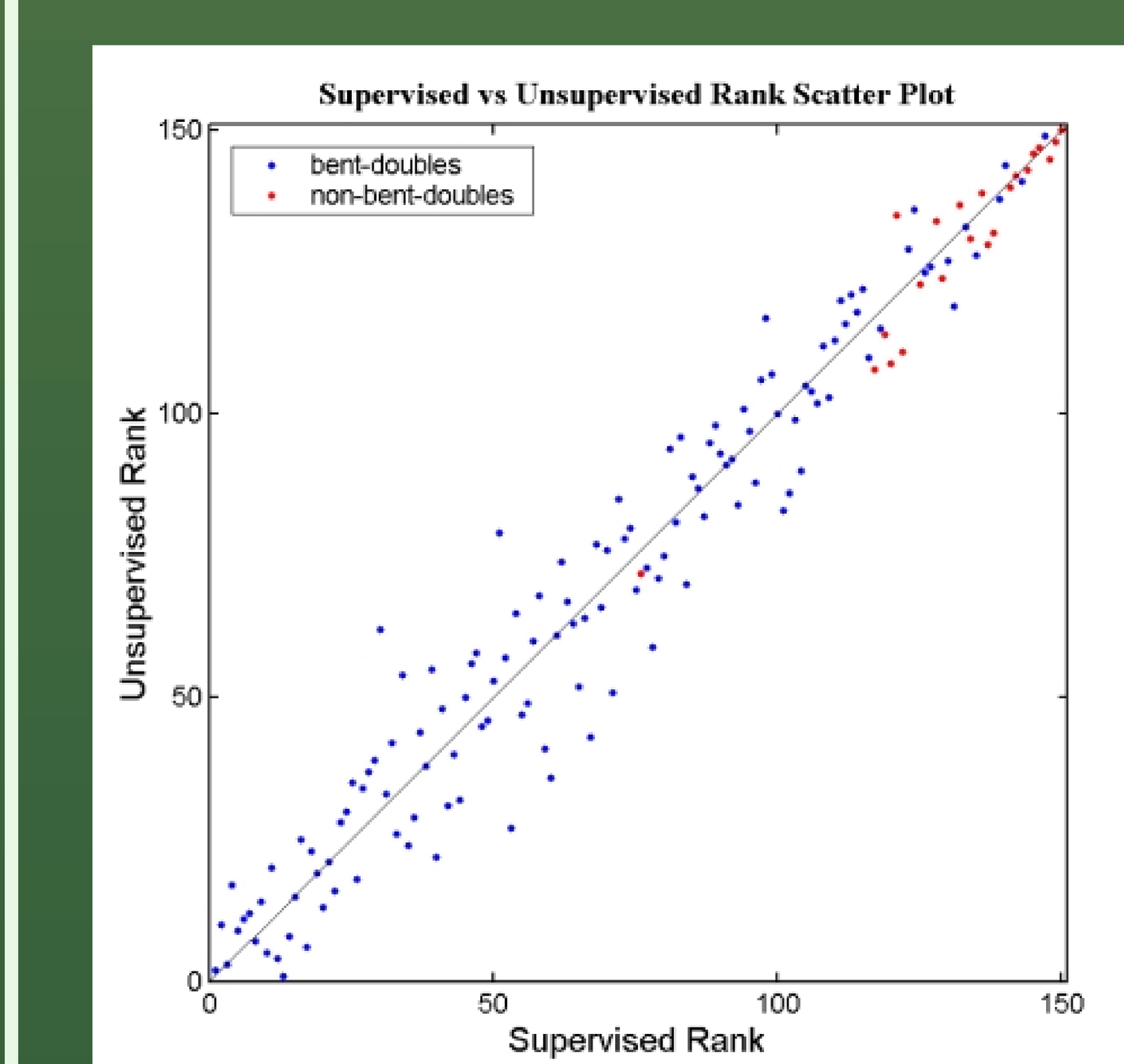
Random assignments would yield an ROC area of 0.5, best possible is 1.0, EM-trained model achieves 0.9336.

Motivation for Unsupervised Learning

- Only a very small fraction of the large data set is labeled
- Indicates how much class information is contained in the features
- Can be used for automated discovery of sub-classes of galaxies

Unsupervised Approach

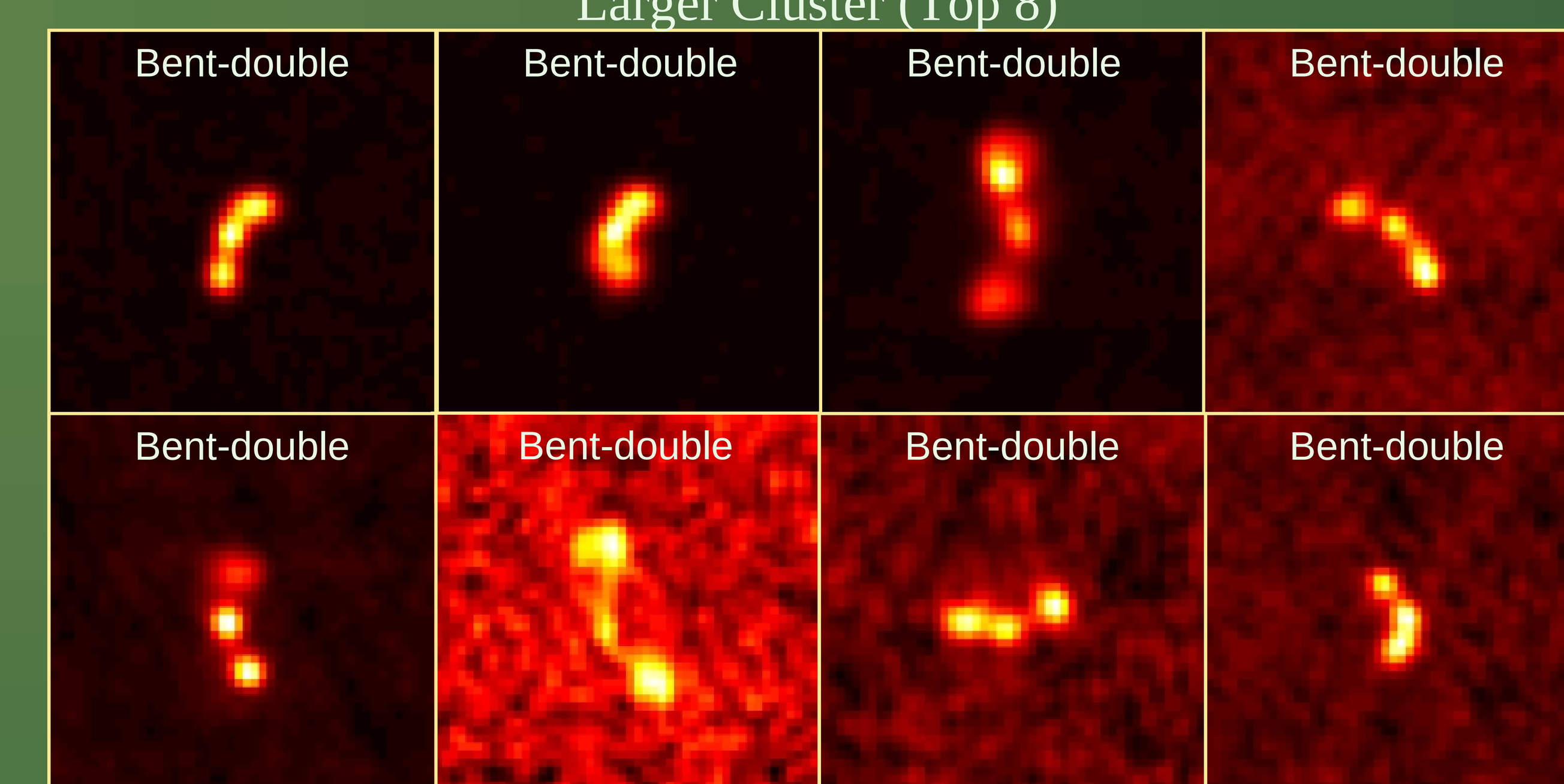
- Disregard the labels and probabilistically cluster the data using EM algorithm
- Parameters:
  - probability for each class
  - probabilities of each orientation per class – uniform
  - parameters for Gaussians for P_a and P_a per orientation and class
- Use all available examples (all 150)
- Ideally, for 2 clusters, want one cluster to correspond to bent-double, another to non-bent-doubles



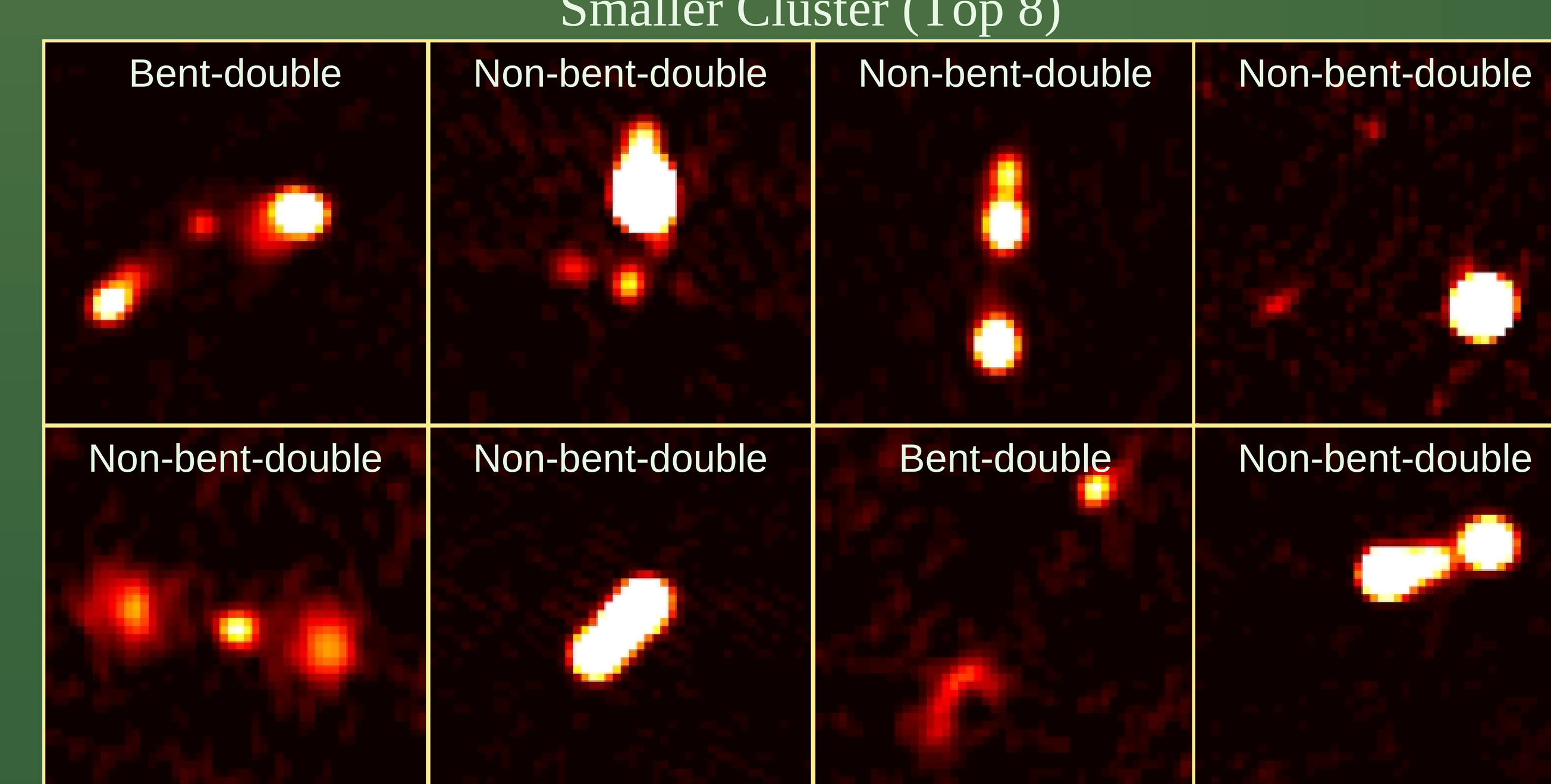
Scatter Plot

The plot shows the rankings of configurations according to the supervised method and to the larger cluster of the unsupervised method. Rankings appear to be strongly correlated leading to believe that unsupervised method can produce results similar to the supervised.

Larger Cluster (Top 8)



Smaller Cluster (Top 8)



2-Cluster Results

- 89 in the larger cluster; 61 in the smaller cluster
- 88 out of 89 examples in the larger cluster are bent-doubles
  - Bent-double cluster as hoped
- Top 8 configurations for each cluster
  - larger cluster (left, top)
  - smaller cluster (left, bottom)

Conclusions and Future Directions

- Method potentially accurate enough to be used by astronomers in screening candidate galaxies from large catalogs
- Strong structure in the data as indicated by the unsupervised algorithm
- Tests on larger set (underway)
- Compare with other methods (underway)
- Generalize to different number of components (2, 4+) and to missing components
- Generalize to weighted misclassification function
  - Worse to classify a bent-double as a non-bent-double than vice versa
- Consider other features and parametrizations