Galaxy Morphology

Archa Jain, Alexandre Gauthier, Emil Noordeh

Background

- Classify galaxies by shape
- Astronomers have billions of pictures of galaxies
- Need machine learning to classify galaxies to learn about distribution of shapes
- Galaxy Zoo project provides crowdsourced training data
- Main Categories: **Spiral, Elliptical, Other**

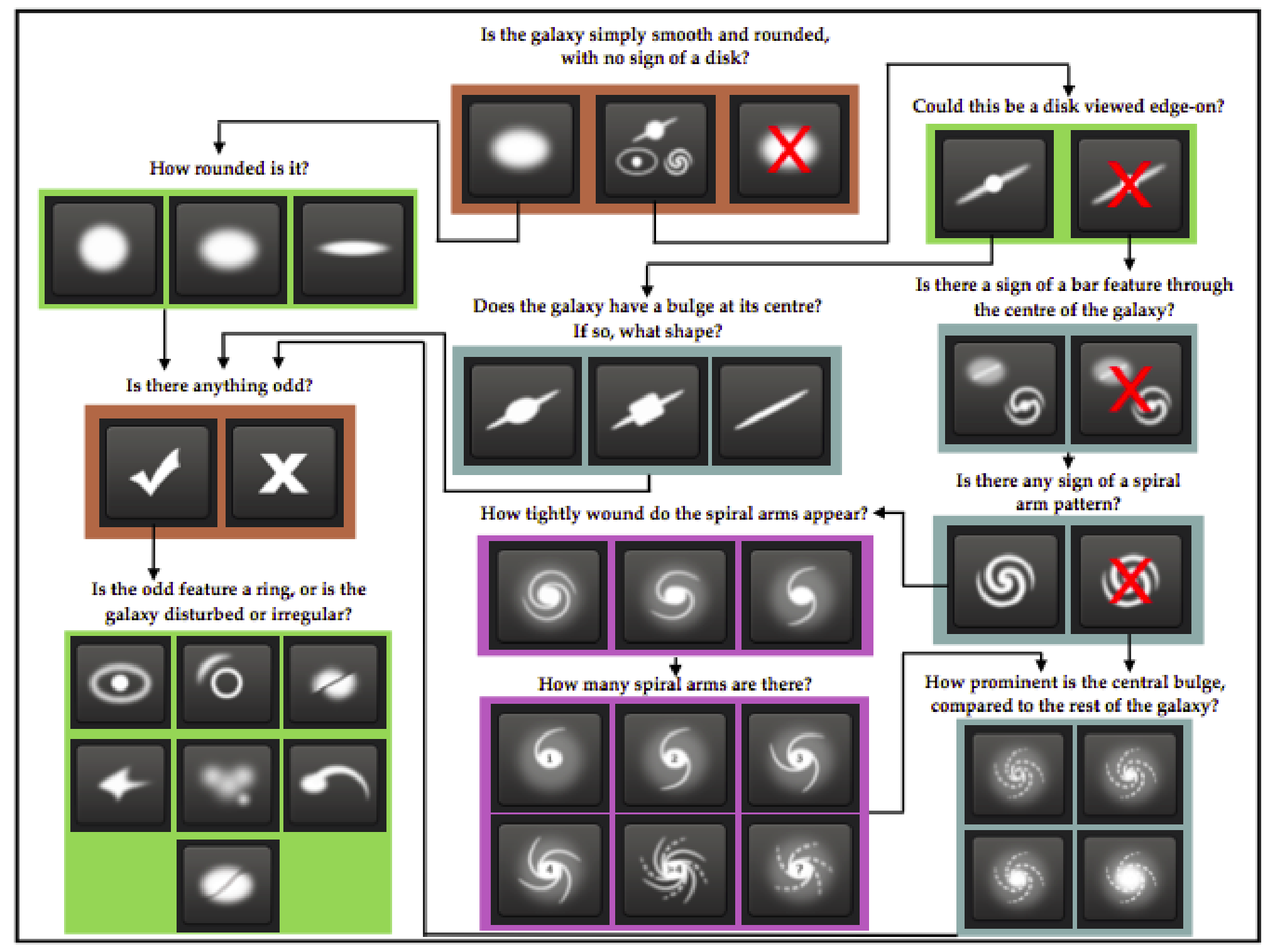
Is the galaxy simply smooth and rounded, with no sign of a disk?
How rounded is it?
Could this be a disk viewed edge-on?
Is there a sign of a bar feature through the centre of the galaxy?
Does the galaxy have a bulge at its centre? If so, what shape?
Is there anything odd?
Is there any sign of a spiral arm pattern?
How tightly wound do the spiral arms appear?
Is the odd feature a ring, or is the galaxy disturbed or irregular?
How many spiral arms are there?
How prominent is the central bulge, compared to the rest of the galaxy?

Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Fig. 1: Decision tree for questions asked of Galaxy Zoo participants [1]

Preprocessing

Compress to 125 features

1. **Crop** to 160x160 - with a fixed window selected to minimize loss of information
2. **Subtract background** - find boundaries of all non background artifact and erase all but the center one
3. **Rotate** to align galaxies vertically - calculate the angle of the galaxy and rotate
4. Use **PCA** to reduce images to 125 features

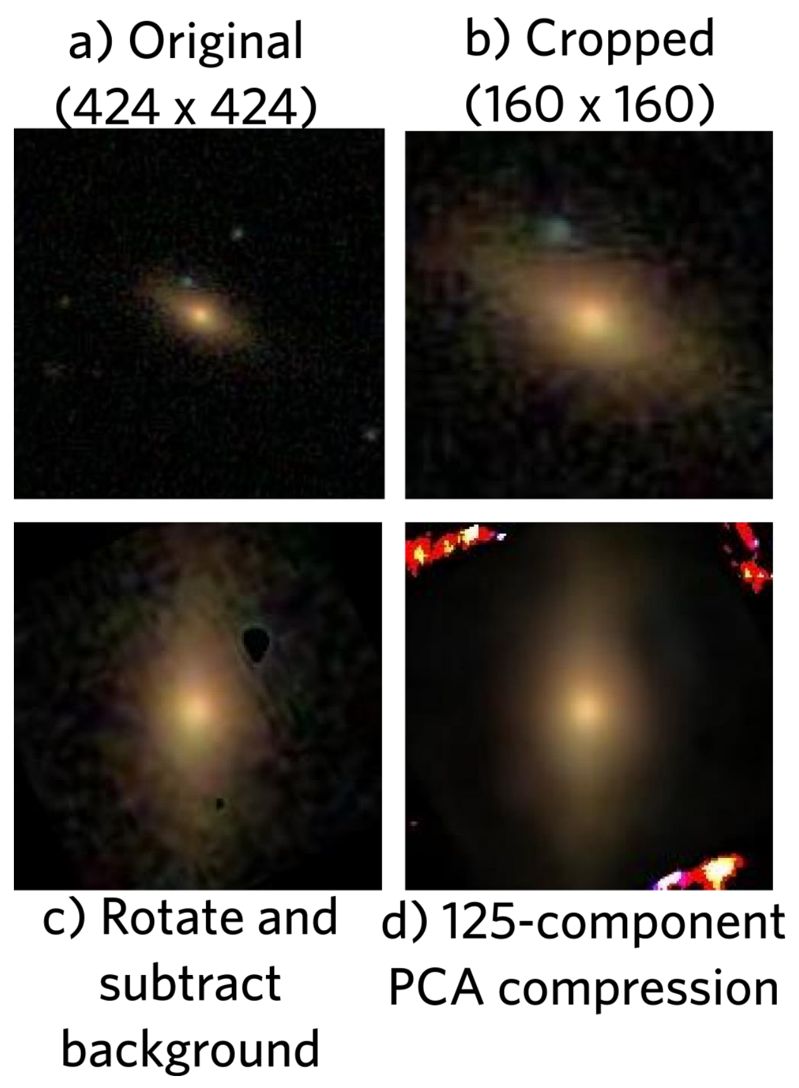
a) Original (424 x 424)
b) Cropped (160 x 160)
c) Rotate and subtract background
d) 125-component PCA compression

Fig. 2: Preprocessing

Dataset

- Images obtained from GalaxyZoo Kaggle dataset [2]
- 61,578 images : 424x424x3(RGB) - 539,328 features
- Each image is centered around a single galaxy.
- Labels are a list of probabilities of answering a certain question from the decision tree in Fig 1.

Clustering

Gain insights into overall distribution using speedy
K Means Clustering

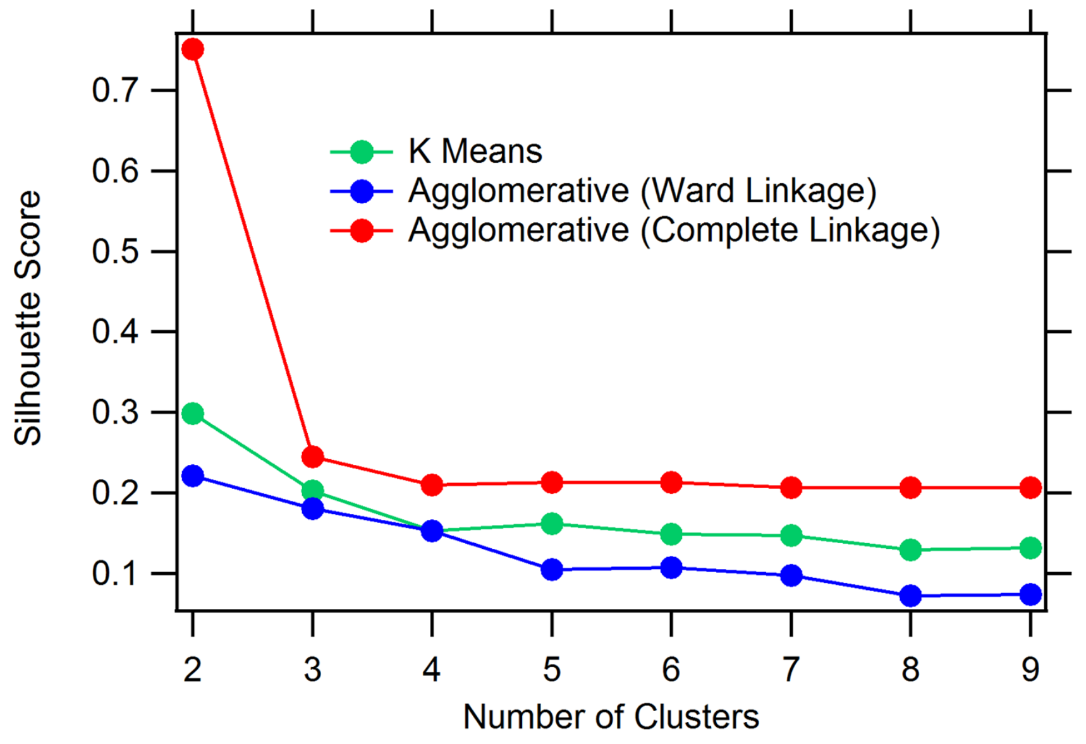

Fig. 3: Silhouette scores. Note peak at n=2 for complete linkage is anomalous; all but one galaxy in same cluster.

Classification

- Generated labels for the following categories - Spiral, Elliptical, Disc, Round, Other (for odd / stars etc.), generated if the corresponding probabilities were comparably high.
- Performed multiclass classification with - One-VS-All classifier with SVM with an RBF Kernel, Decision Tree, Random Forest, AdaBoost Classifier, selected Random Forest with Cross Validation, K nearest neighbors.

Regression

- Label generation for classification was noisy - since a lot of probabilities for some questions are around ~0.5, making it hard to select a label.
- Trained a Decision Tree regressor (with 200 estimators) for each of the probabilities for the above mentioned five classes.

Acknowledgements

We would like to thank Professors Ng and Duchi, the course staff (especially Nihit Desai), for their feedback, help and support.

Results and Discussion

- Classification precision high, except for confusion between 'Spiral' and 'Other'
- Accuracy went up by adding 2 additional classes as informed by the data
- Confusion was caused by low probability labels; directly predicting probabilities led to higher accuracy
- Clustering indicated the data is not very cleanly separable (also indicated by label confusion).
- Visualization revealed 2 dominant dimensions:

Eccentricity **Brightness**

	Disc	Spiral	Ellipse	Round	Other
Disc	**620**	69	102	3	56
Spiral	35	**1524**	156	83	184
Ellipse	48	213	**1299**	158	143
Round	0	131	149	**1241**	67
Other	48	790	366	212	**303**

Fig. 4: Confusion Matrix (X: true, Y: pred.)

Category	Accuracy
Spiral	0.928
Ellipse	0.926
Disc	0.962
Round	0.949
Other	0.937

Fig. 5: Regression Accuracy

Applying 2-component PCA compression to 125-dimensional data after clustering

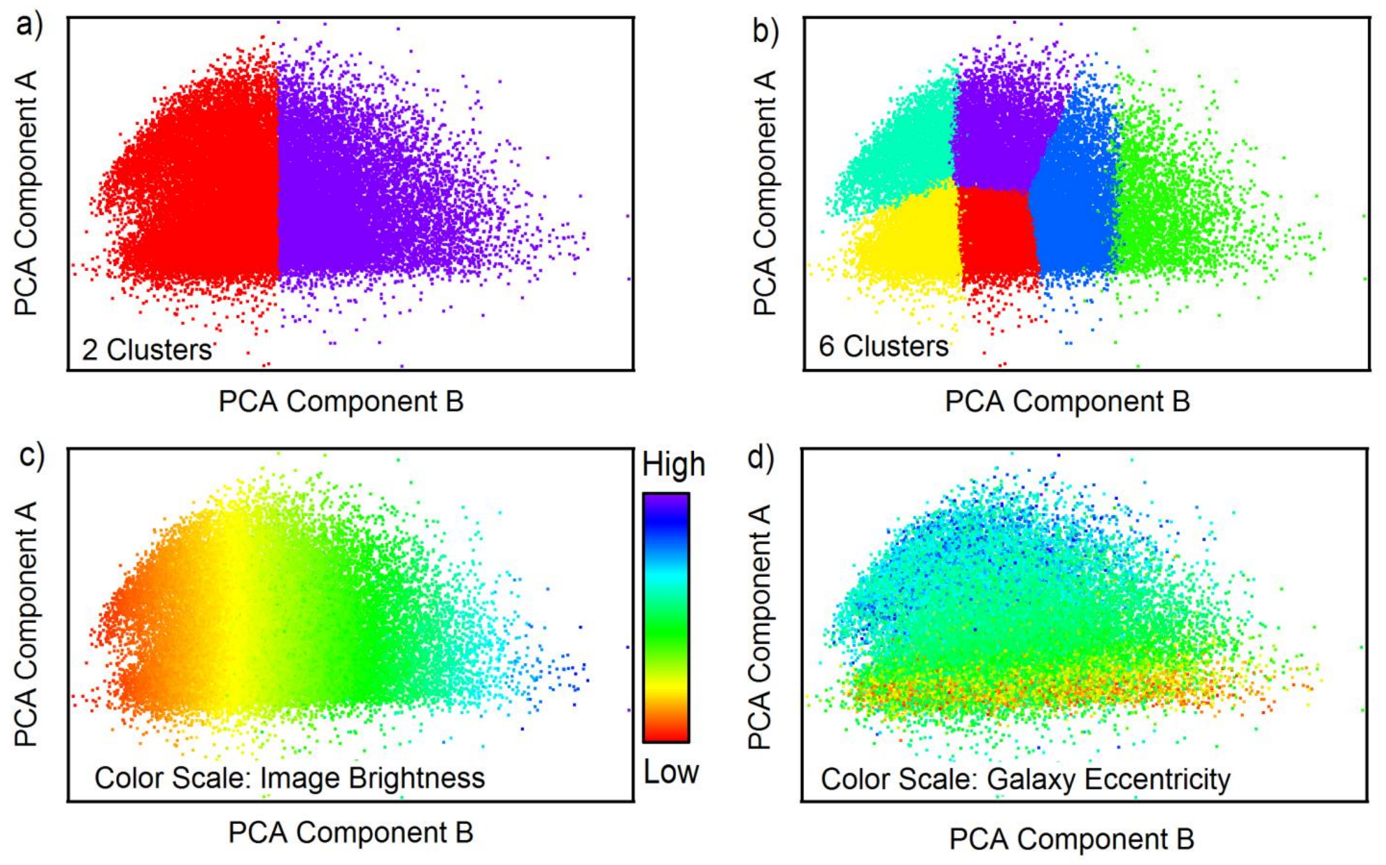

Fig. 6: 40,000 galaxies plotted in the space of the two-dimensional PCA. (a) and (b) show the distribution of K Means clusters for n=2 and n=6, respectively. (c) colors each galaxy by the average brightness of the image, and (d) colors based upon the eccentricity (un-roundedness) of the galaxy as determined by Galaxy Zoo users.

References

[1] arXiv:1308.3496
[2] "Galaxy Zoo – The Galaxy Challenge." *Data – Galaxy Zoo – The Galaxy Challenge | Kaggle.* N.p., n.d. Web. 11 Dec. 2016