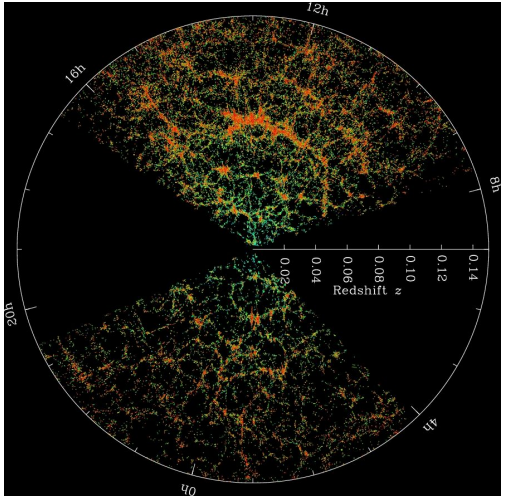
Galaxy Classification with Decision Tree & Random Forest

Tintin Nguyen
PHYS 305 Final Project
University of Arizona
April 28, 2022

INTRODUCTION: SDSS GALAXY OBSERVATIONS

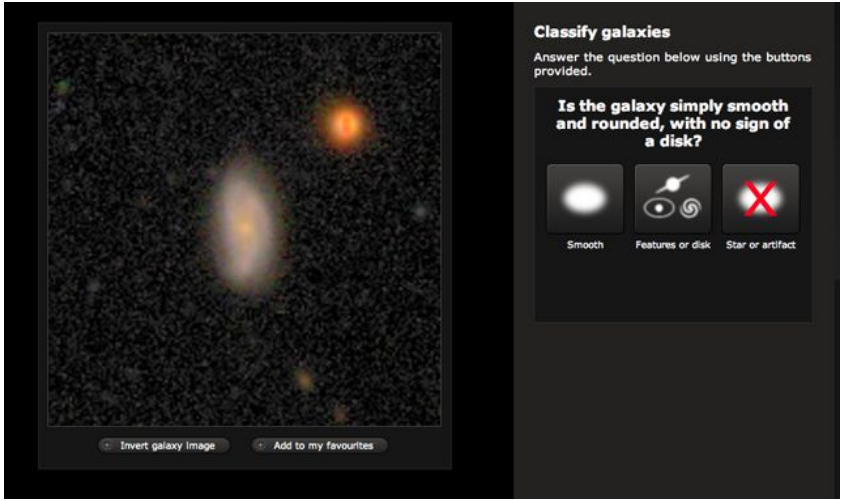


Sloan Digital Sky Survey (SDSS)

From 1M+ galaxy observations:

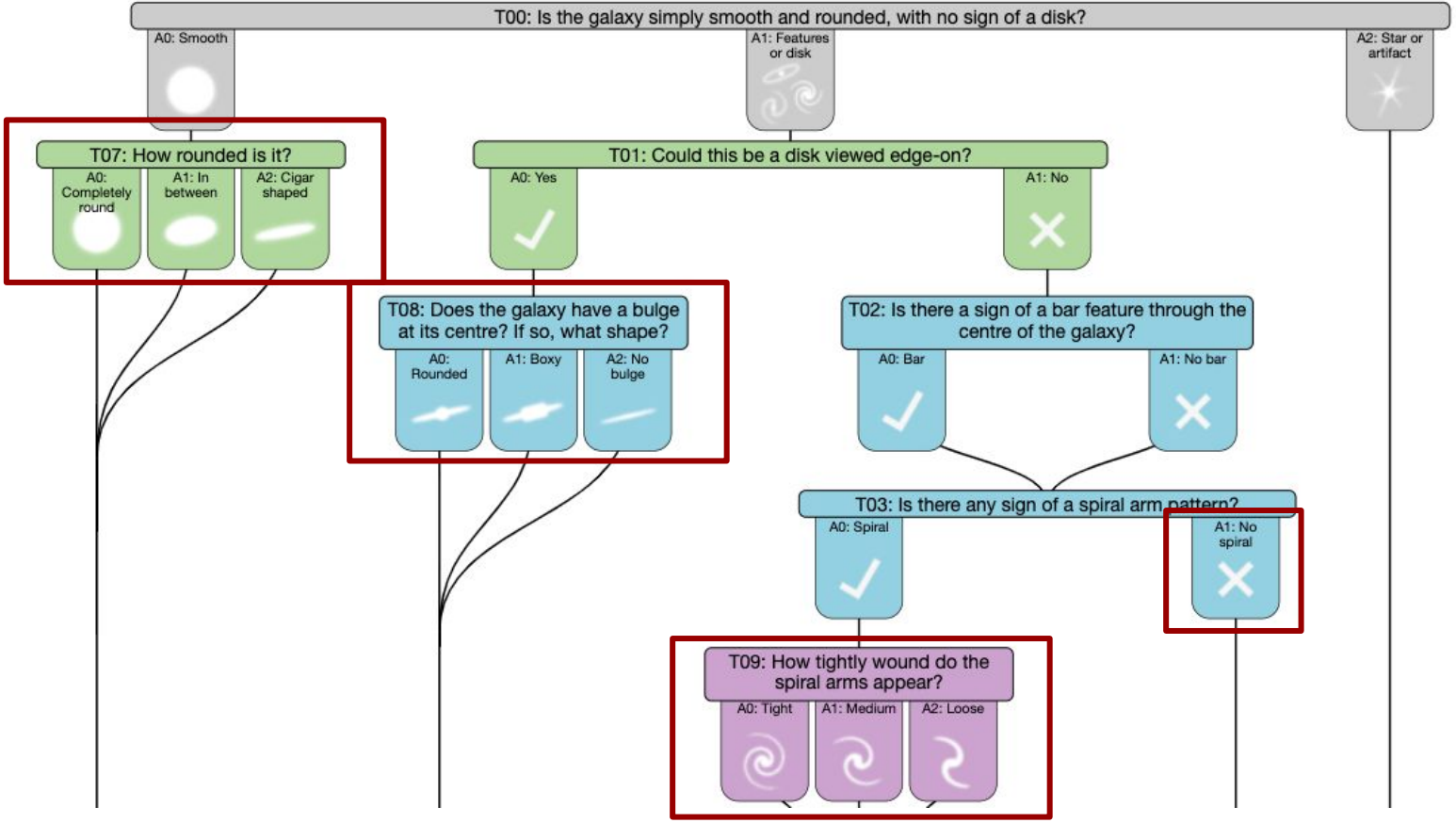
- ★ **Origin and evolution of galaxies**
  - Star formation zones
  - Ionized gas in galaxy center
- ★ **Large-scale structure**
  - Cosmic inflation
  - Dark matter
  - Dark energy

INTRODUCTION: GALAXY ZOO

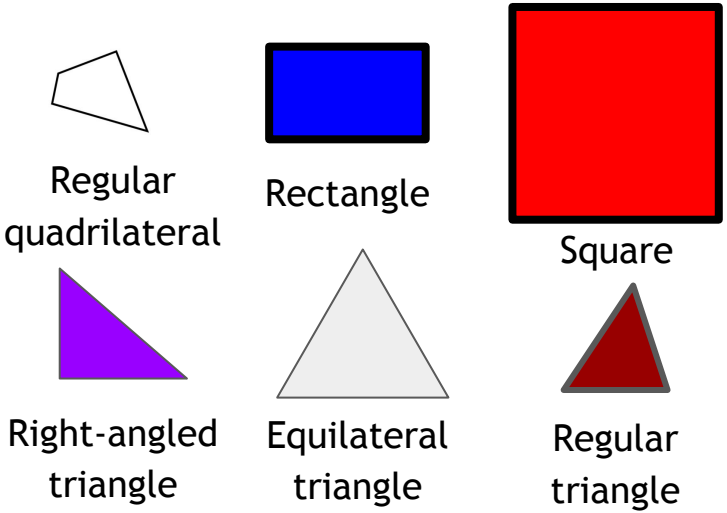


- ★ **PROBLEM:** Too much data!!!
- ★ **SOLUTION:** Volunteer helps classify galaxies
  - Types of galaxies with consistent majority vote are reliable
  - Immense dataset to train machine learning galaxy classifications

INTRODUCTION: GALAXY CLASSIFICATION

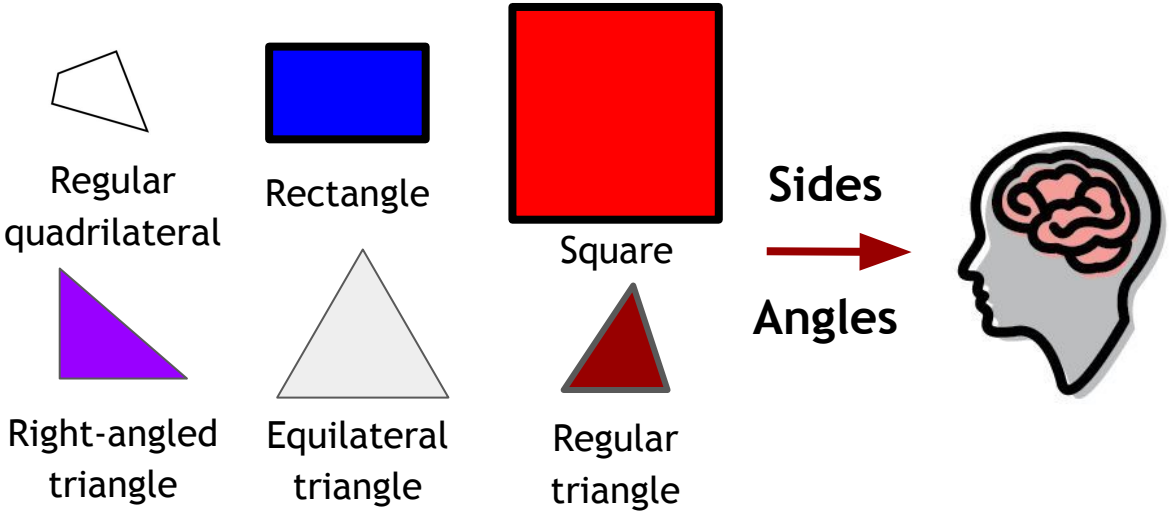


METHOD: DECISION TREE



TRAINING DATASET

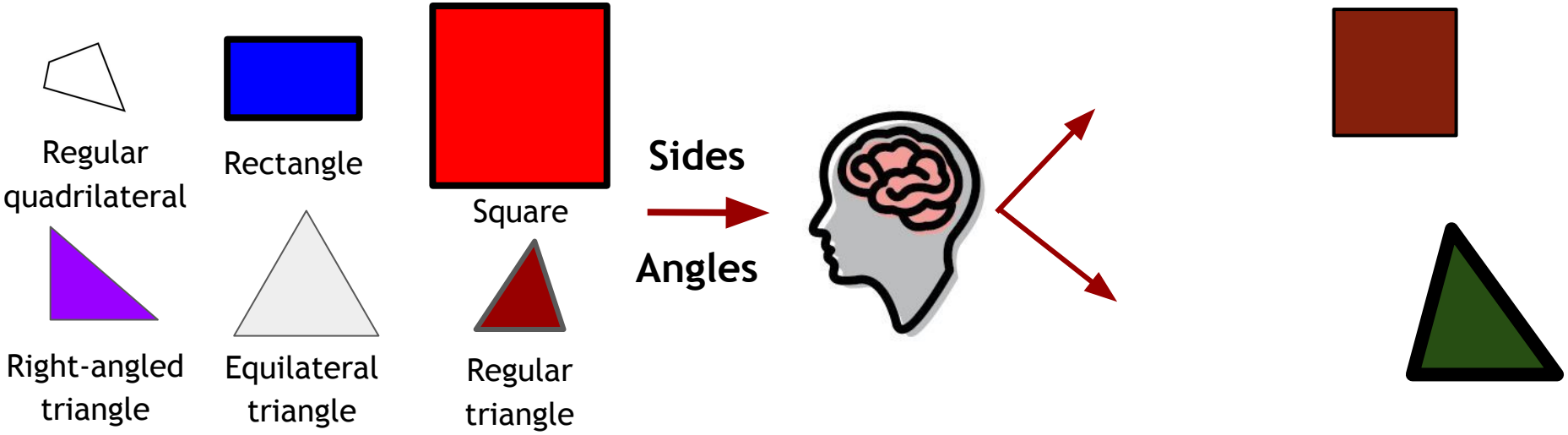
METHOD: DECISION TREE



TRAINING DATASET

DECISION TREE

METHOD: DECISION TREE

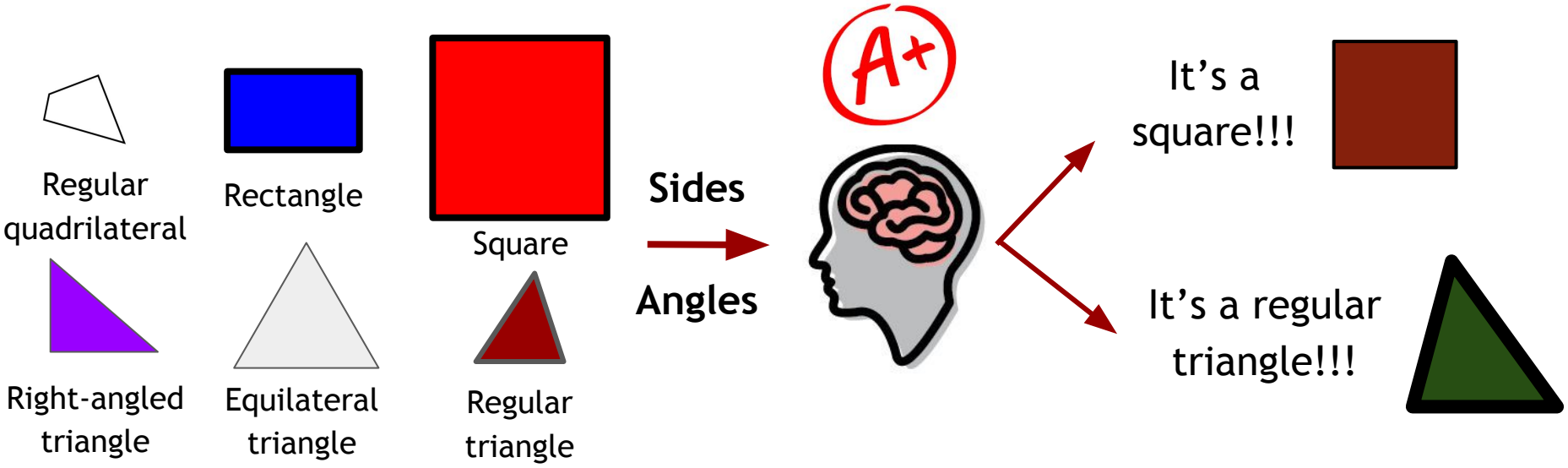


TRAINING DATASET

DECISION TREE

TESTING DATASET

METHOD: DECISION TREE

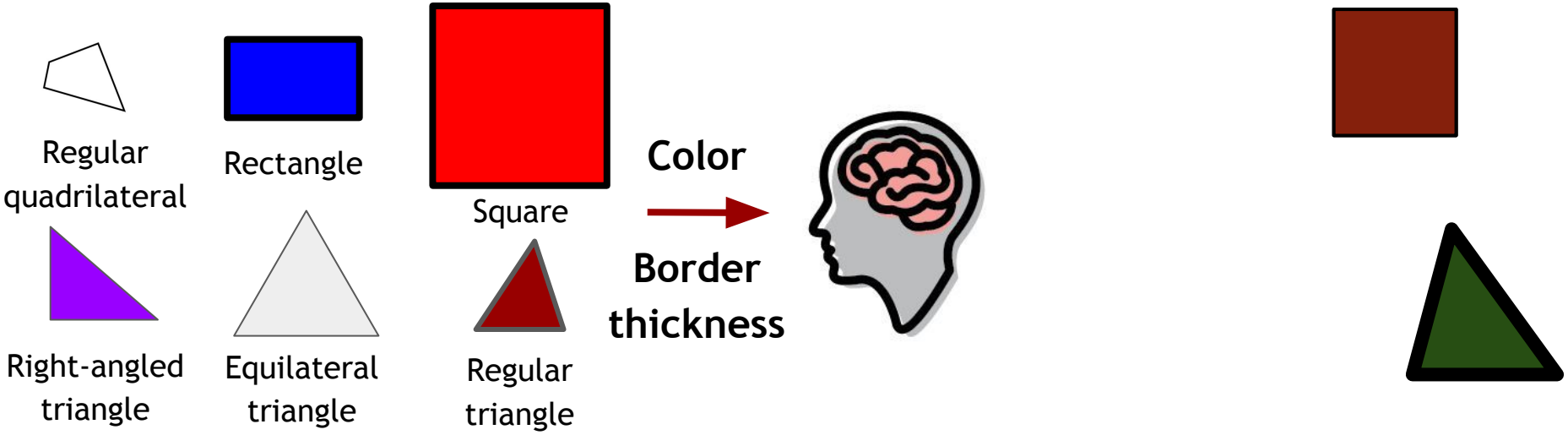


TRAINING DATASET

DECISION TREE

TESTING DATASET

METHOD: DECISION TREE



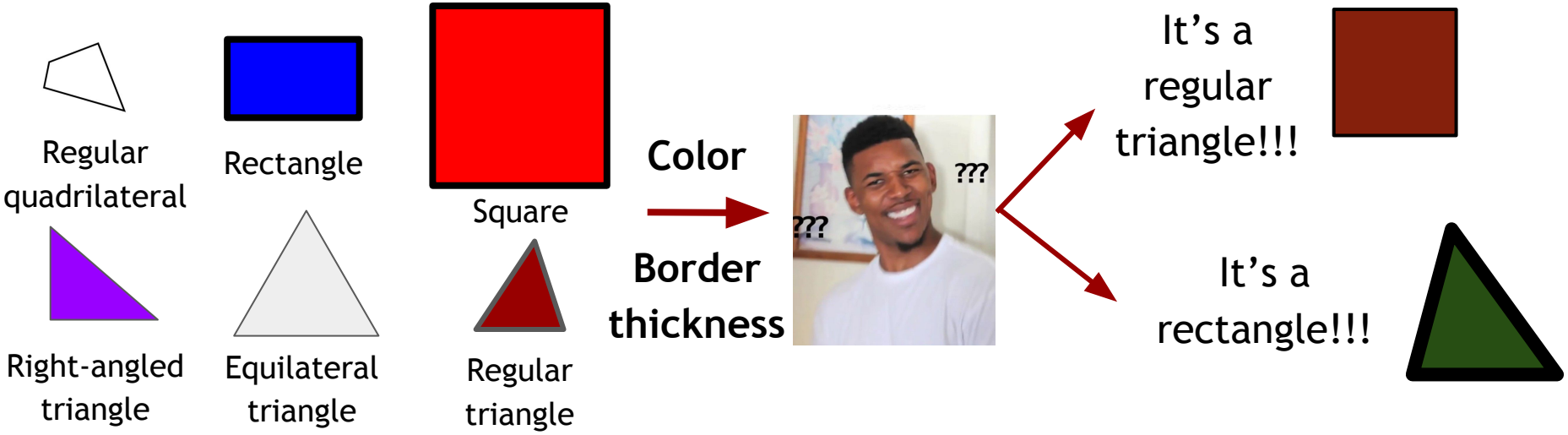
TRAINING DATASET

DECISION TREE

TESTING DATASET

METHOD: DECISION TREE

EXTRACTING RIGHT FEATURES ARE IMPORTANT!

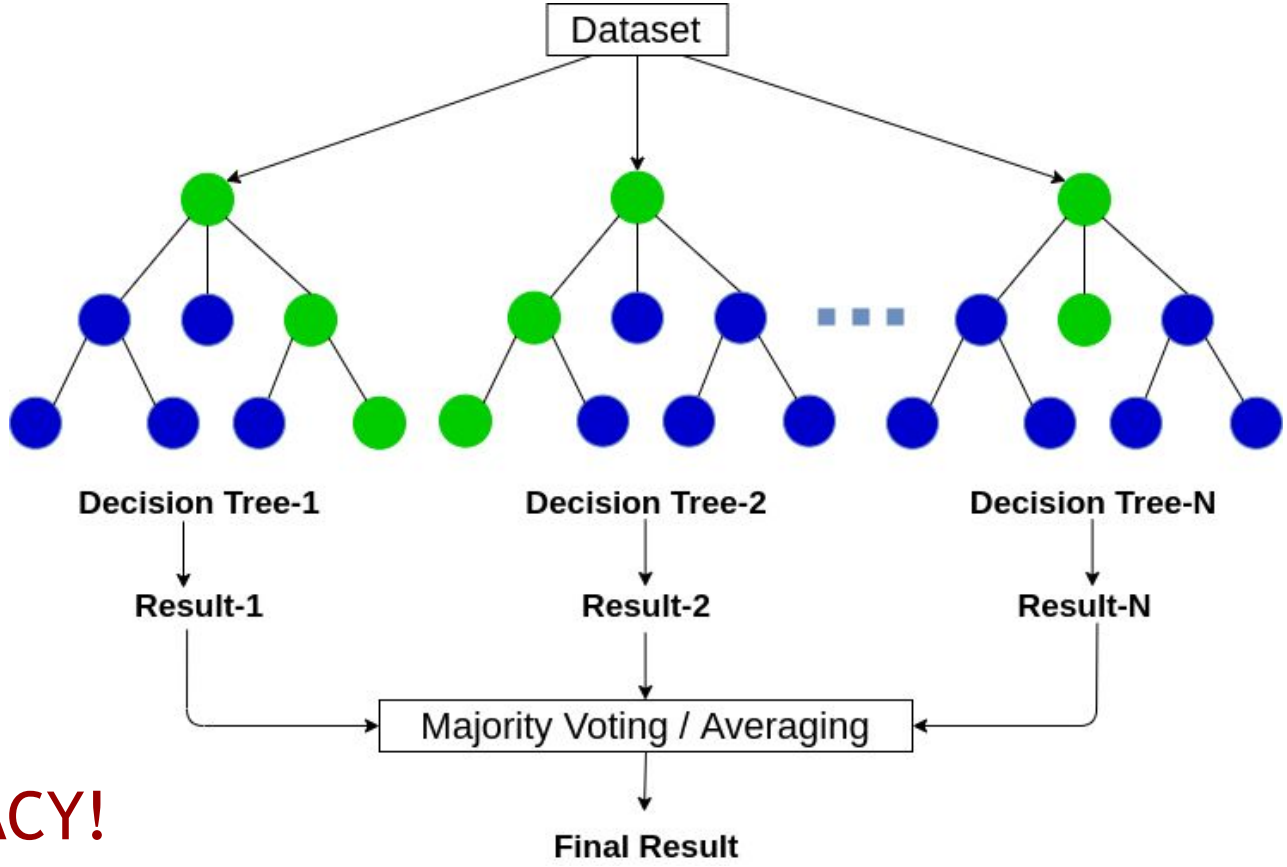


TRAINING DATASET

DECISION TREE

TESTING DATASET

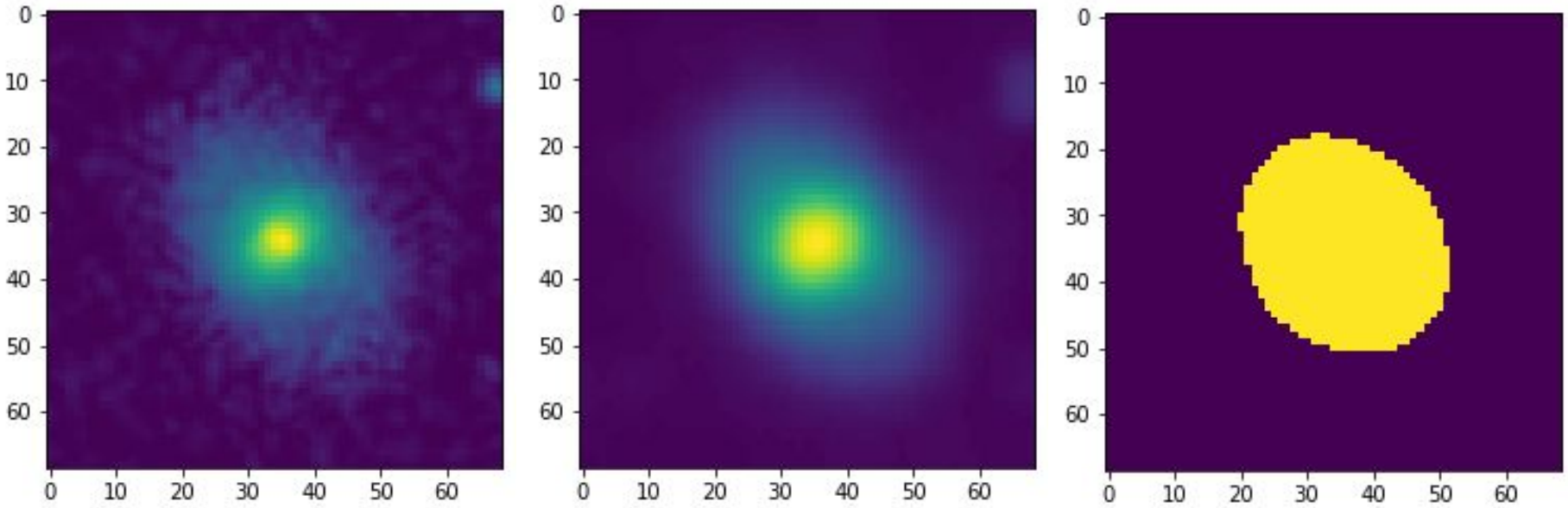
METHOD: RANDOM FOREST



DEMOCRACY!

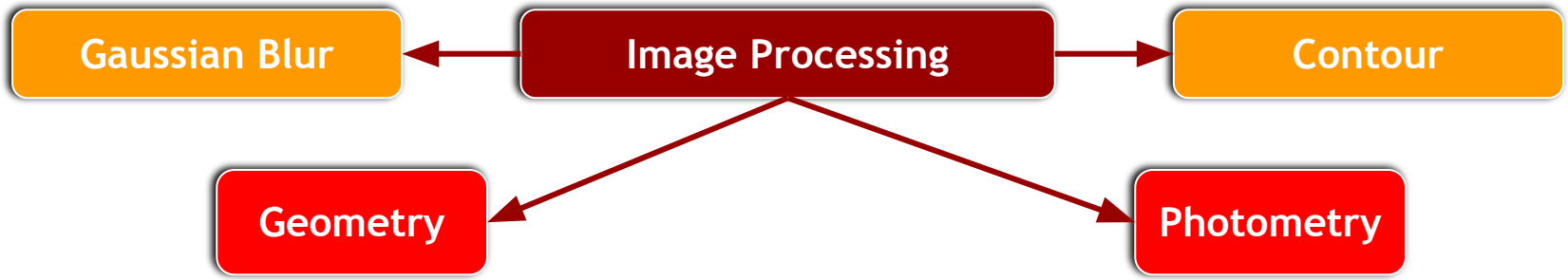
METHOD: IMAGE PROCESSING

Image Processing

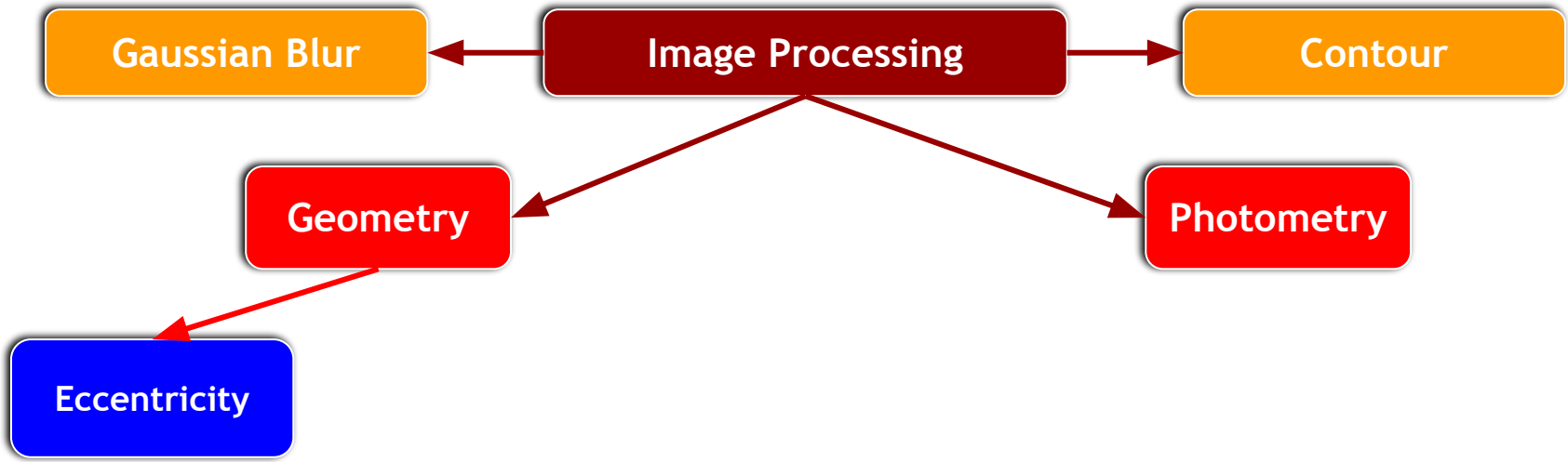


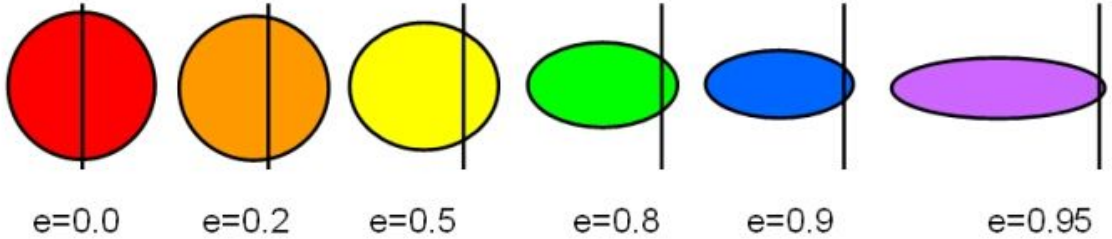
Raw image → Apply Gaussian Blur → Draw contour

METHOD: EXTRACTING GALAXY FEATURES

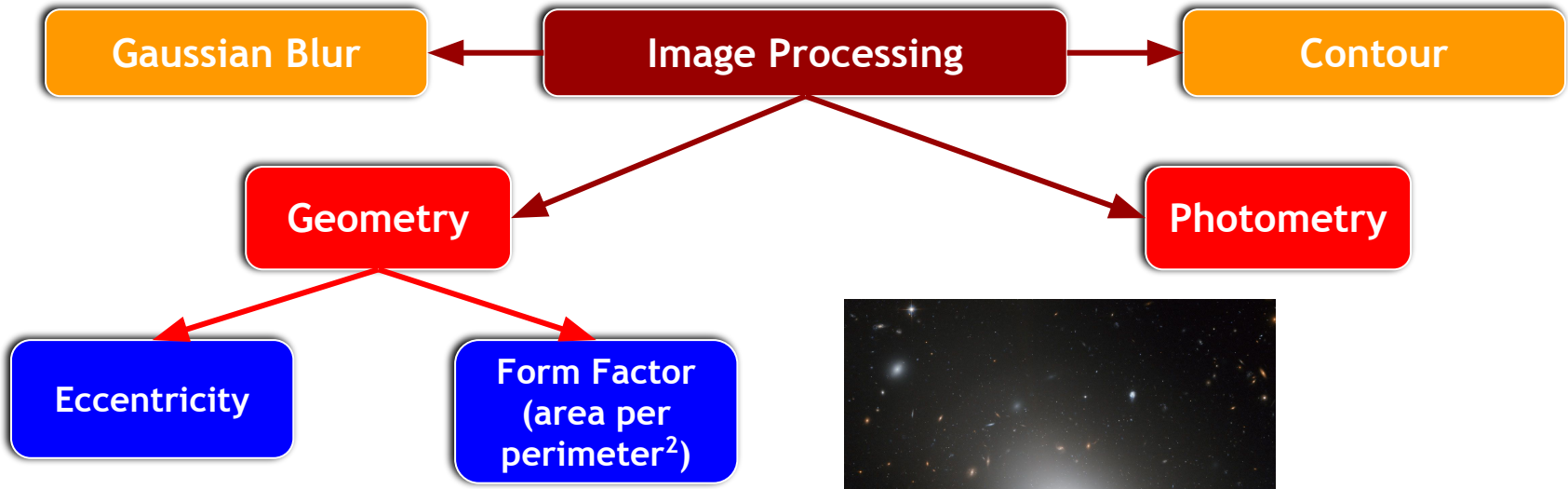


METHOD: EXTRACTING GALAXY FEATURES

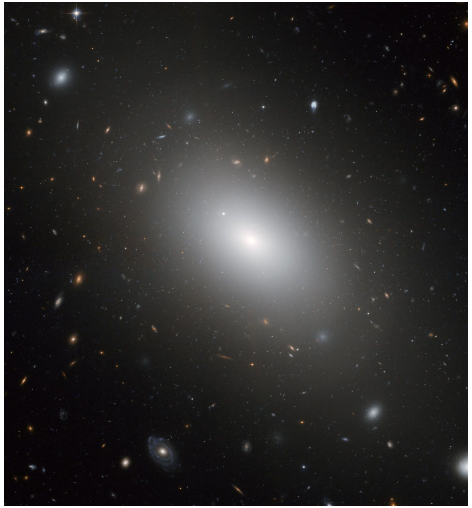




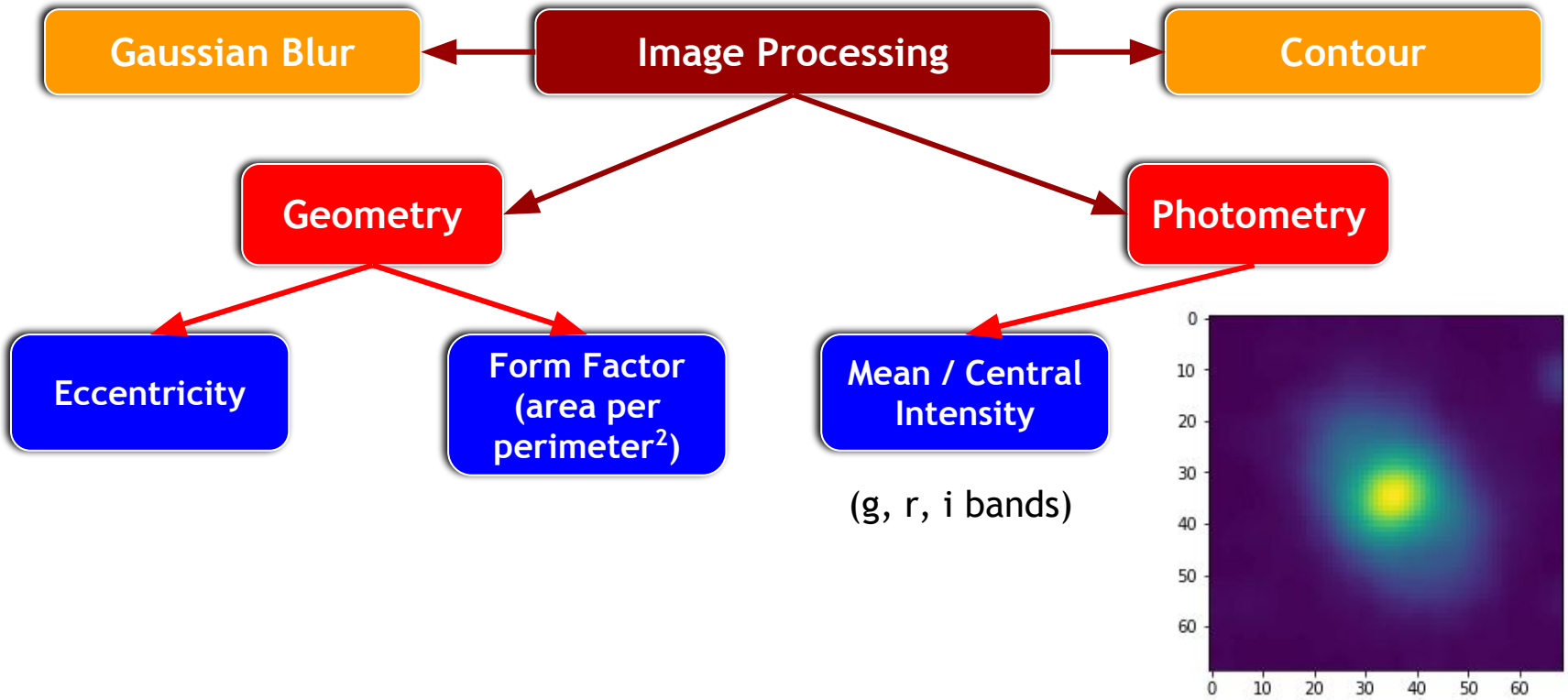
METHOD: EXTRACTING GALAXY FEATURES



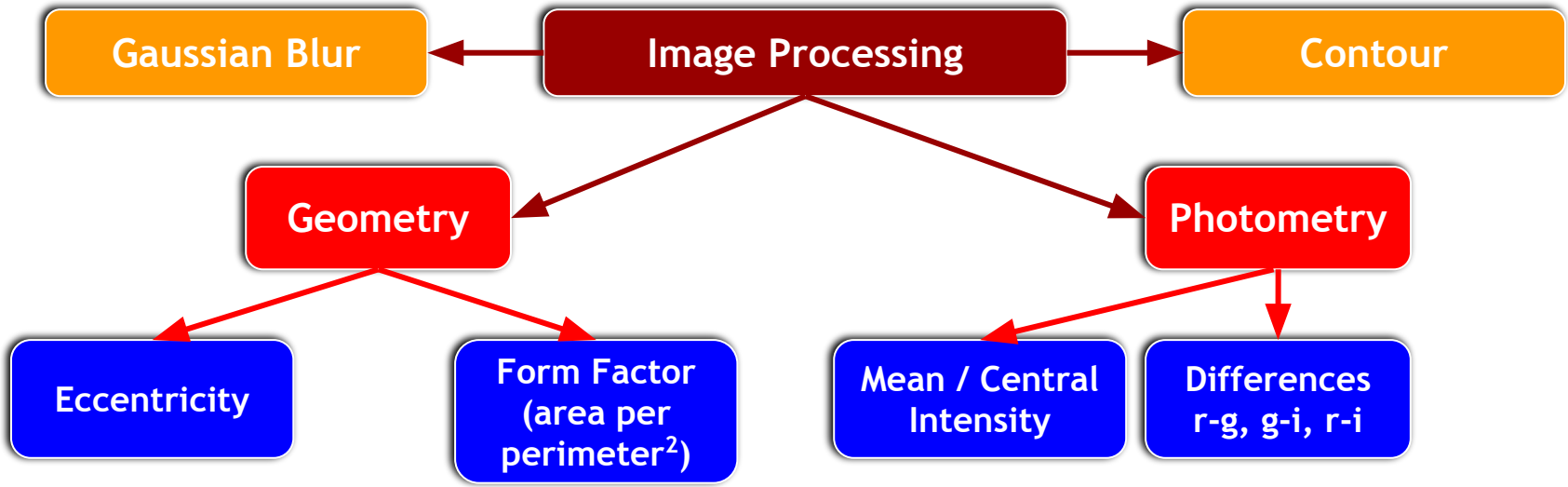
Elliptical galaxy



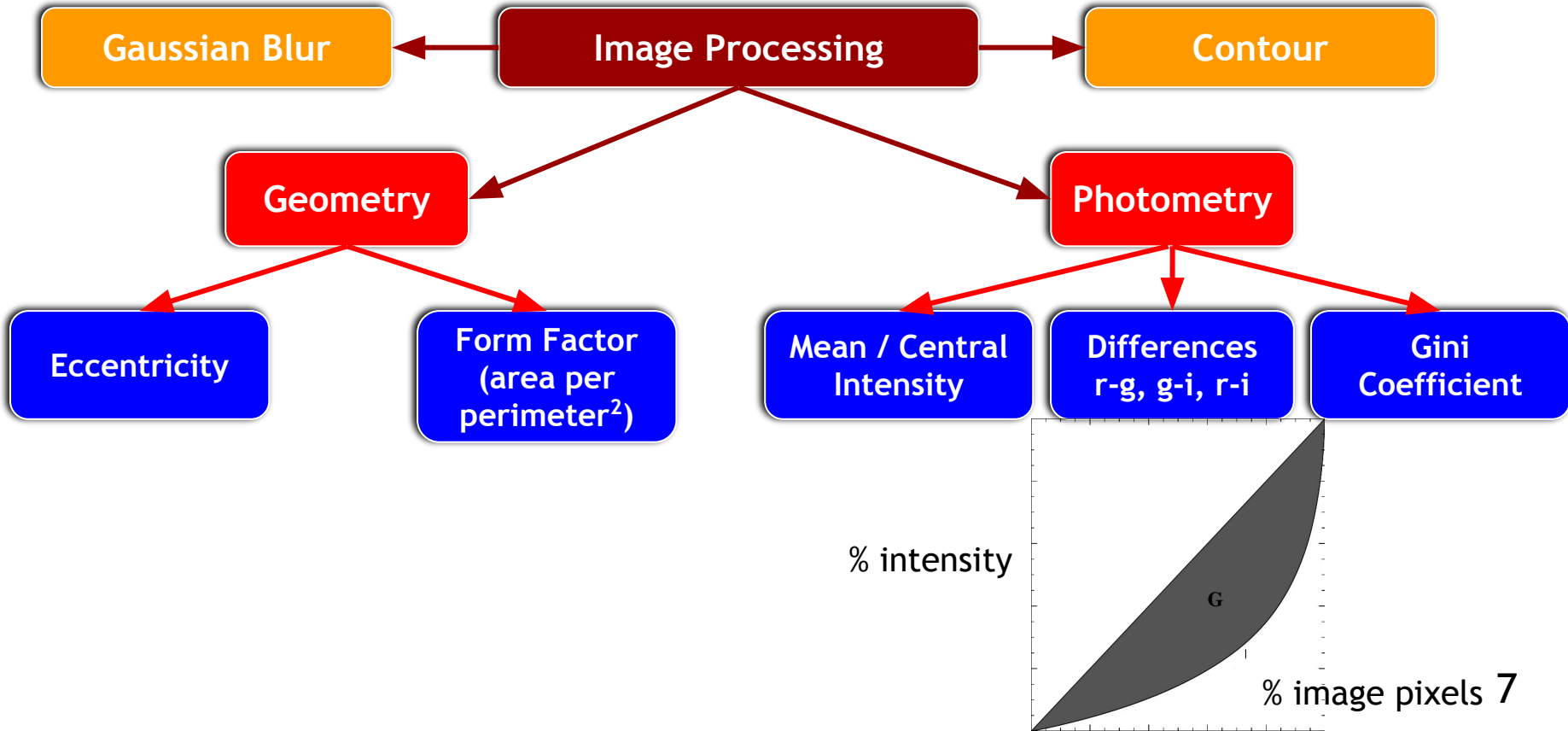
METHOD: EXTRACTING GALAXY FEATURES



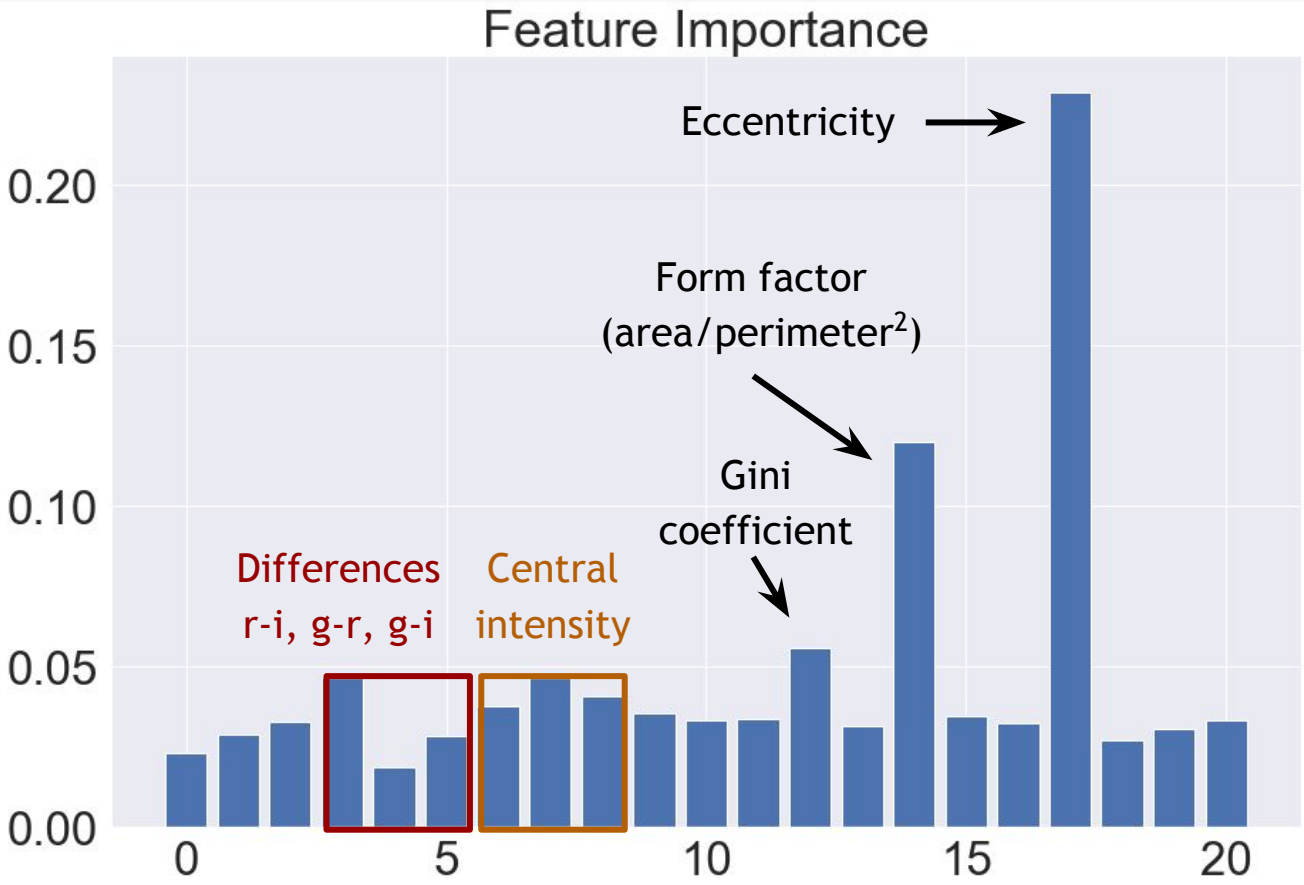
METHOD: EXTRACTING GALAXY FEATURES



METHOD: EXTRACTING GALAXY FEATURES

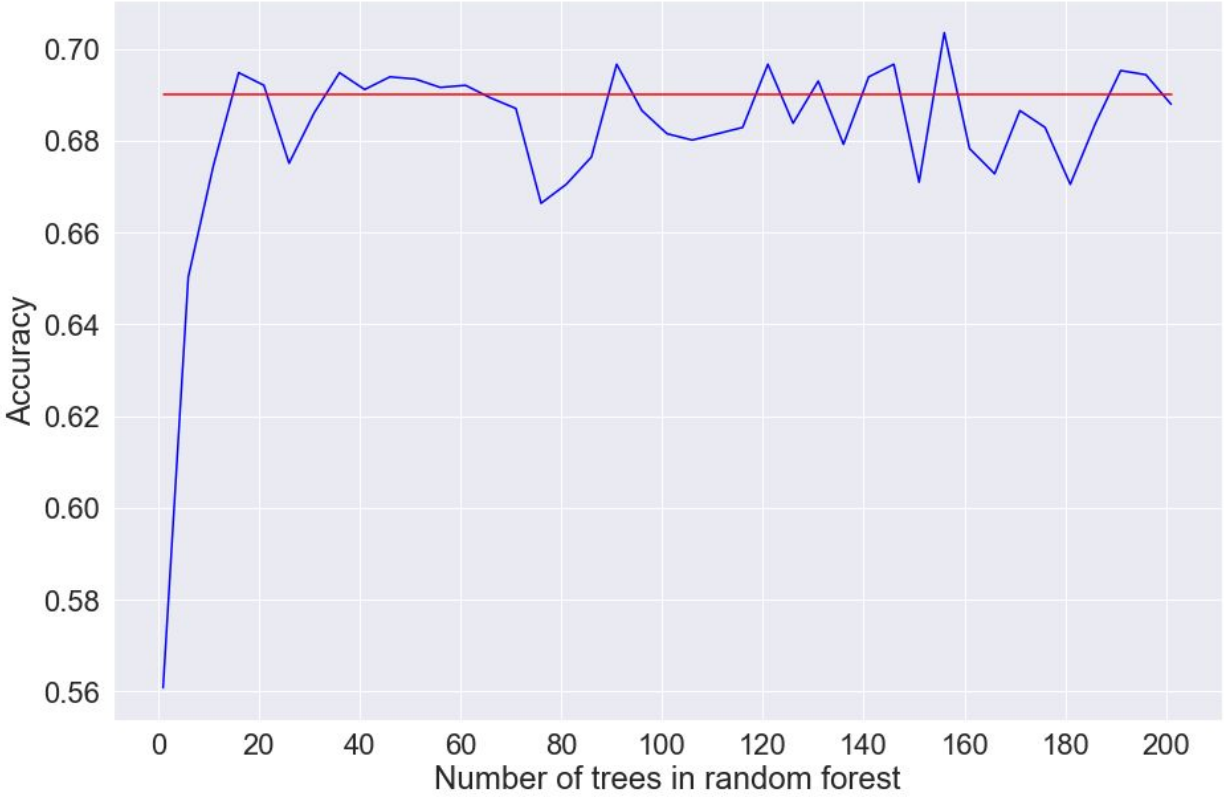


RESULTS: IMPORTANT FEATURES



RESULTS: DECISION TREE vs RANDOM FOREST

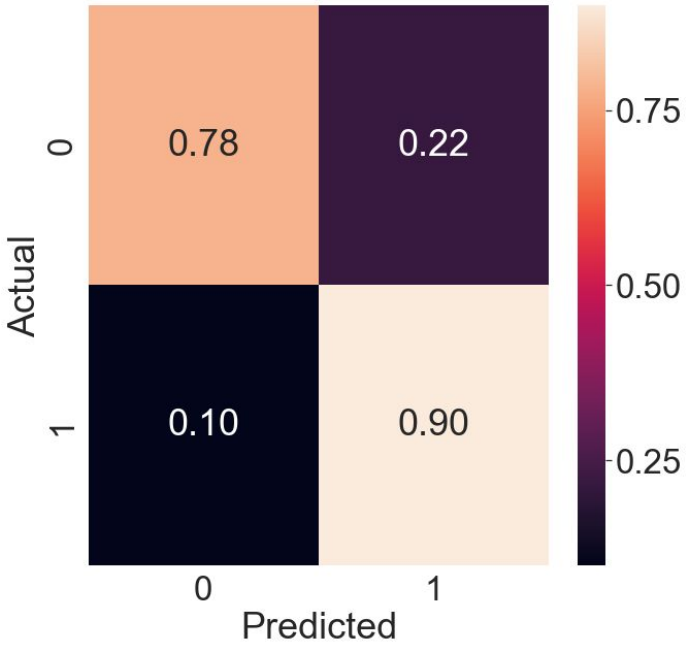
- ★ Decision tree: ~56%
- ★ Tuned random forest: ~69%



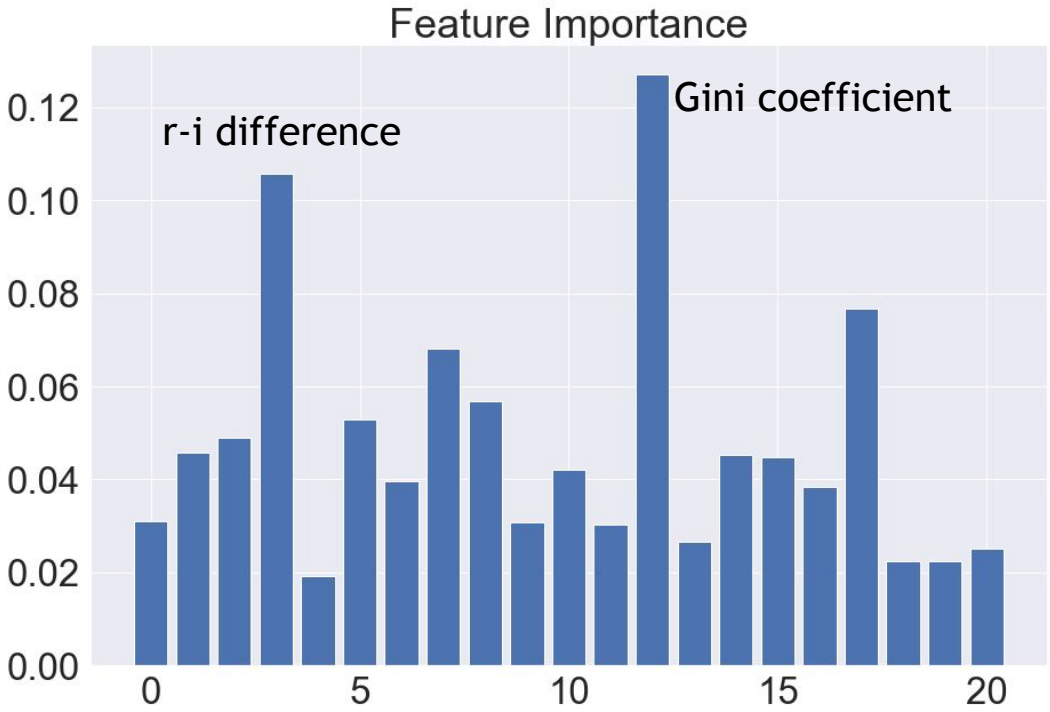
ANALYSIS

Disk (0) or smooth (1)?

Accuracy: ~85%

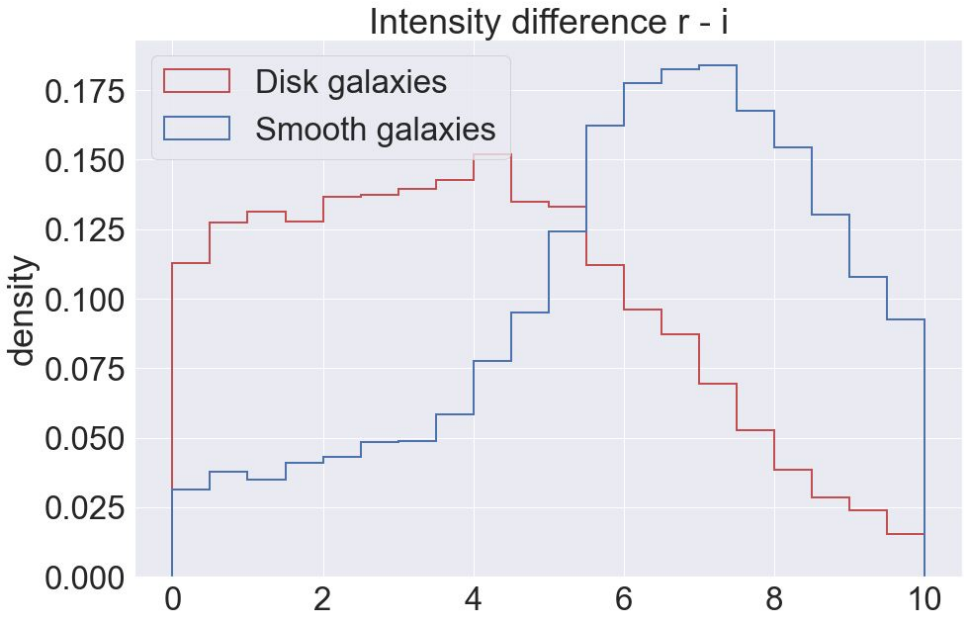


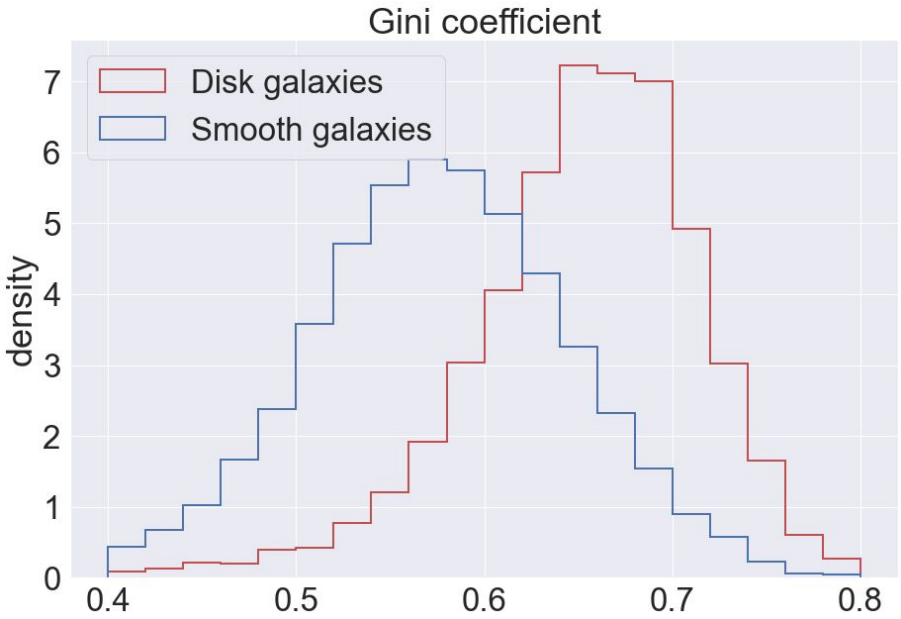
Confusion Matrix



ANALYSIS

Disk (0) or smooth (1)?

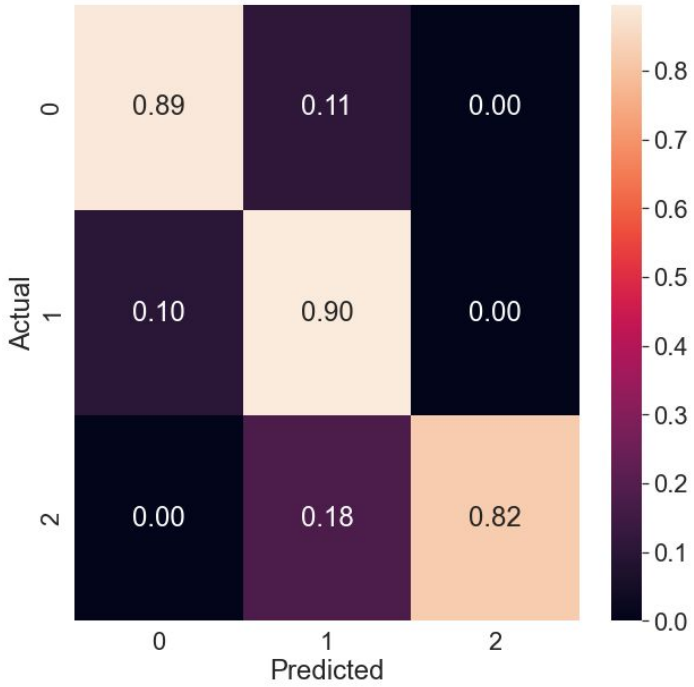


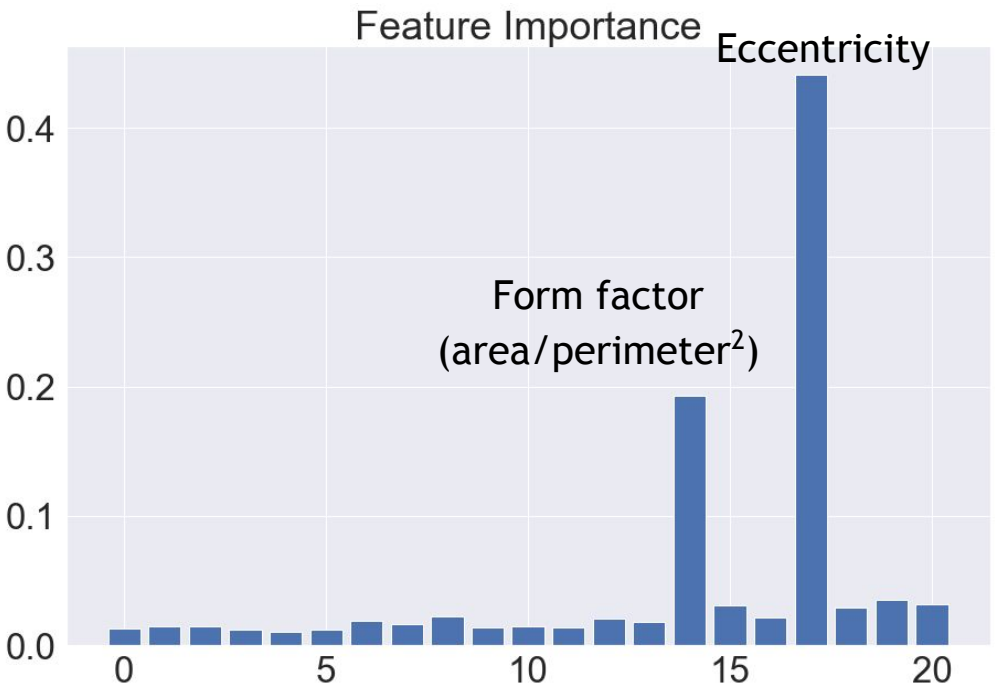


ANALYSIS

SMOOTH - How rounded? Completely round (0), in-between (1), or cigar-shaped (2)?

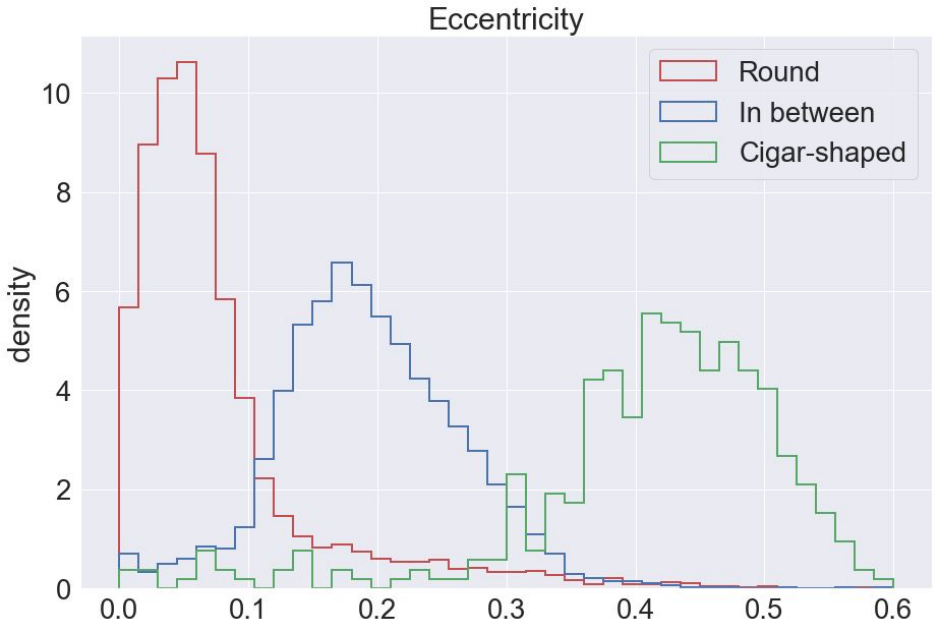
Accuracy: ~89%

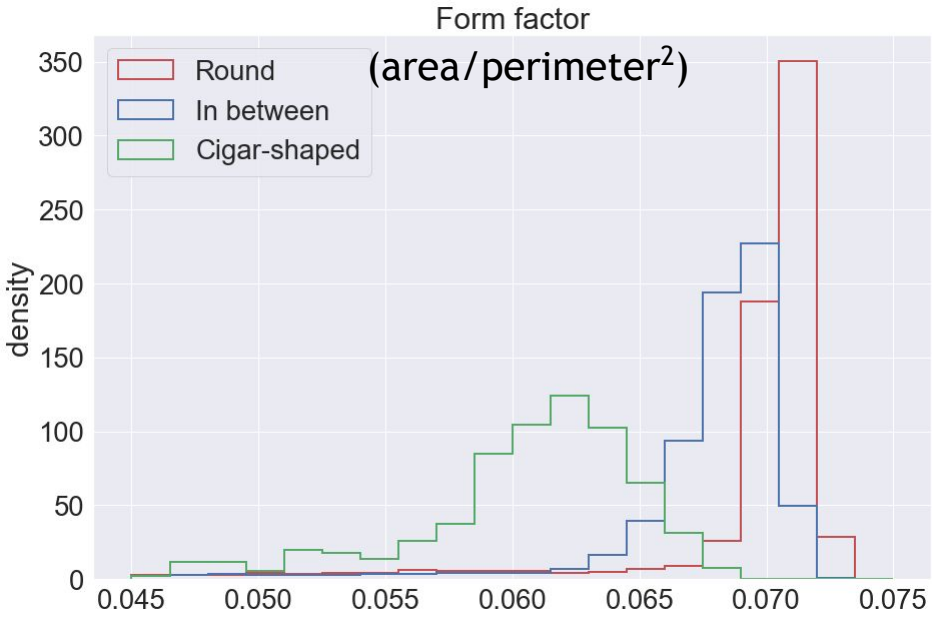




ANALYSIS

SMOOTH - How rounded? Completely round (0), in-between (1), or cigar-shaped (2)?

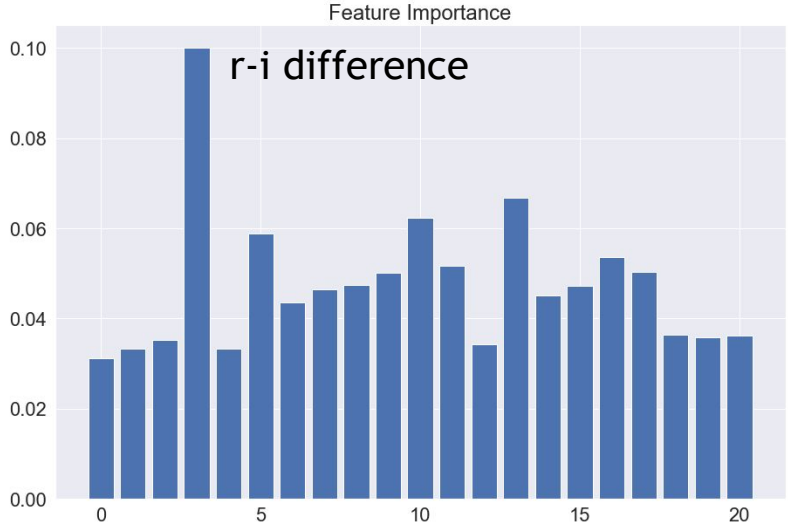


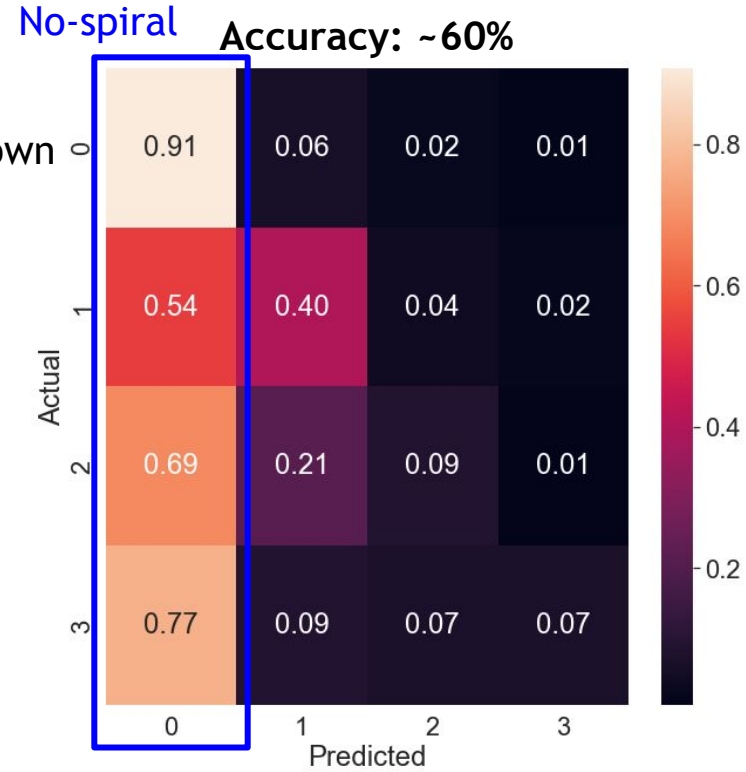


ANALYSIS

DISK + FACE ON - Spiral?: No (0), tight (1), medium (2) or loose (3)?

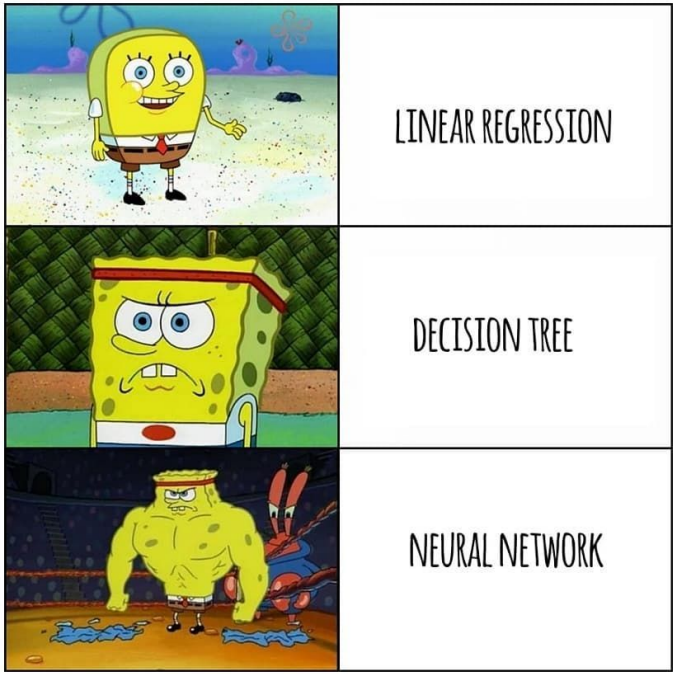
- ★ No-spiral galaxies form the majority of this subgroup
- ★ Algorithm is biased towards predicting that an unknown galaxy belongs to the majority class (no-spiral)





CONCLUSION AND FUTURE WORK

- ★ Random forests (~69%) outperform decision trees (~56%)
- ★ Great prediction for the following cases:
  - Disk vs Smooth (~85%)
  - Roundedness of smooth galaxies (~89%)
- ★ Terrible prediction (~60%) for spiral galaxies classification with bias towards the majority type
- ★ Possible next steps:
  - Minimize bias towards the majority type
  - Find better classification features for spiral galaxies
  - Compare performance with neural networks



REFERENCES / ACKNOWLEDGEMENTS

PAPERS

S. Goderya and S. Lolling, 2001, "Morphological Classification of Galaxies Using Computer Vision and Artificial Neural Networks: A Computational Scheme"

J. Lotz et al., 2004, A New Non-Parametric Approach to Galaxy Morphological Classification"

K. Willett et al., 2013, "Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey"

M. Walmsley et al., 2021, "Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies"

TUTORIALS / CODE REFERENCES

OpenCV: Image Processing, Contour Features

Towards Data Science: Decision Tree, Random Forest, Hyperparameter Tuning

Scikit-learn and OpenCV documentation

ACKNOWLEDGEMENTS: **SDSS** for the galaxy images, **Galaxy Zoo and the volunteer citizen scientists** for the training and testing datasets, **AstroNN** for sampling the Galaxy Zoo dataset for educational purposes, **Kaggle Galaxy Zoo Challenge contestants** for sharing their ideas, and **Prof. Johns** for valuable suggestions and feedbacks