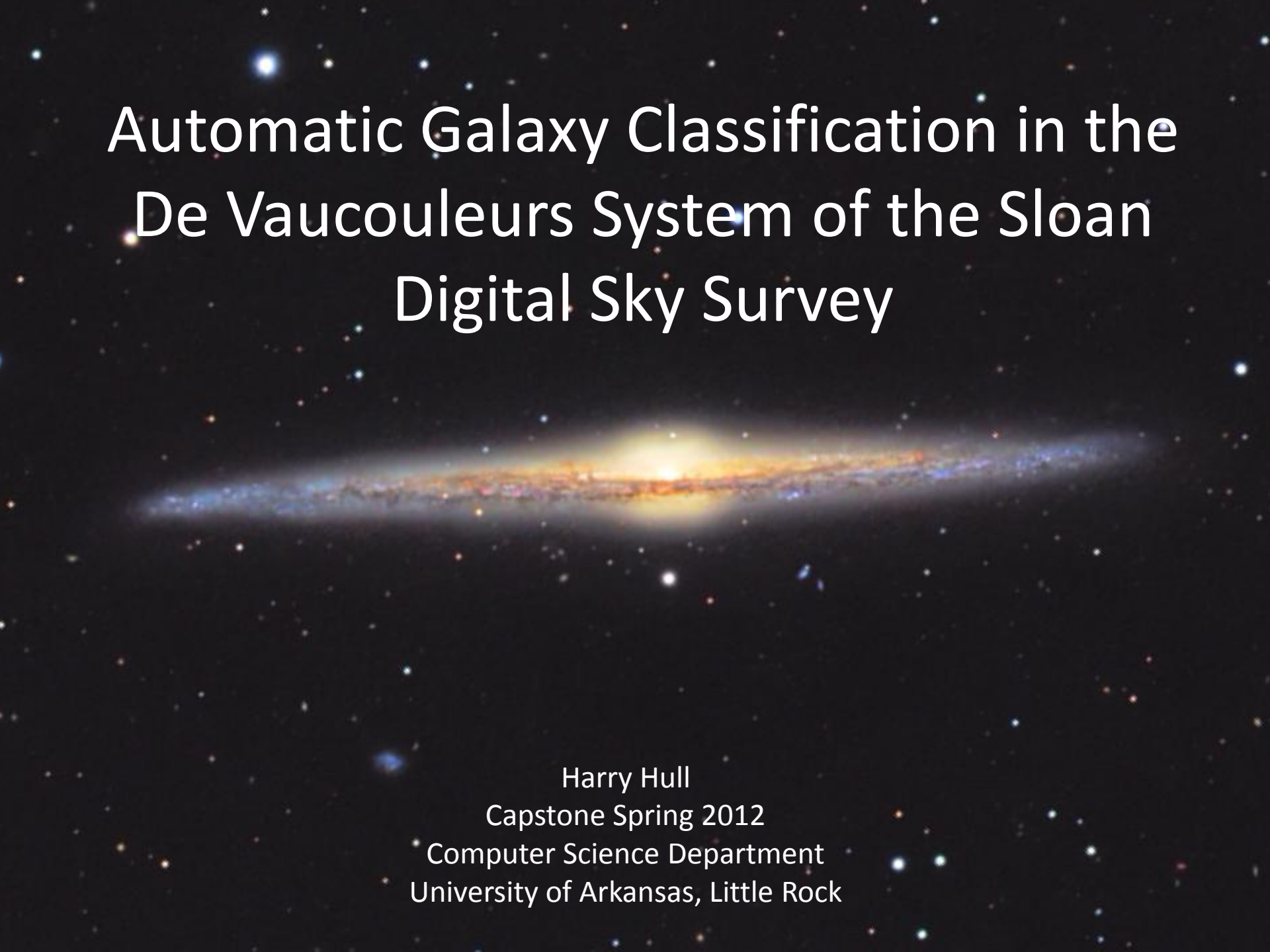
Automatic Galaxy Classification in the De Vaucouleurs System of the Sloan Digital Sky Survey

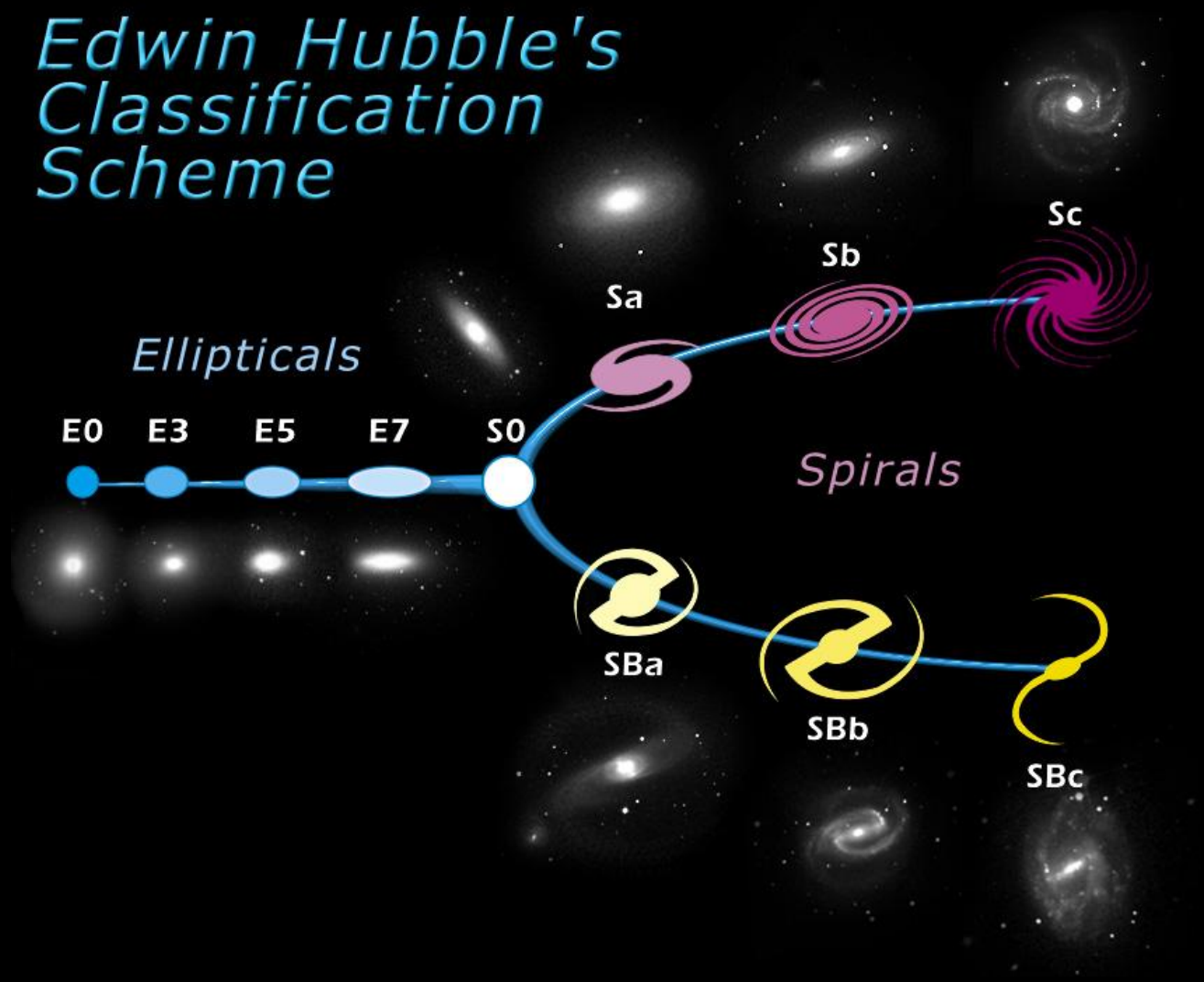


Harry Hull
Capstone Spring 2012
Computer Science Department
University of Arkansas, Little Rock

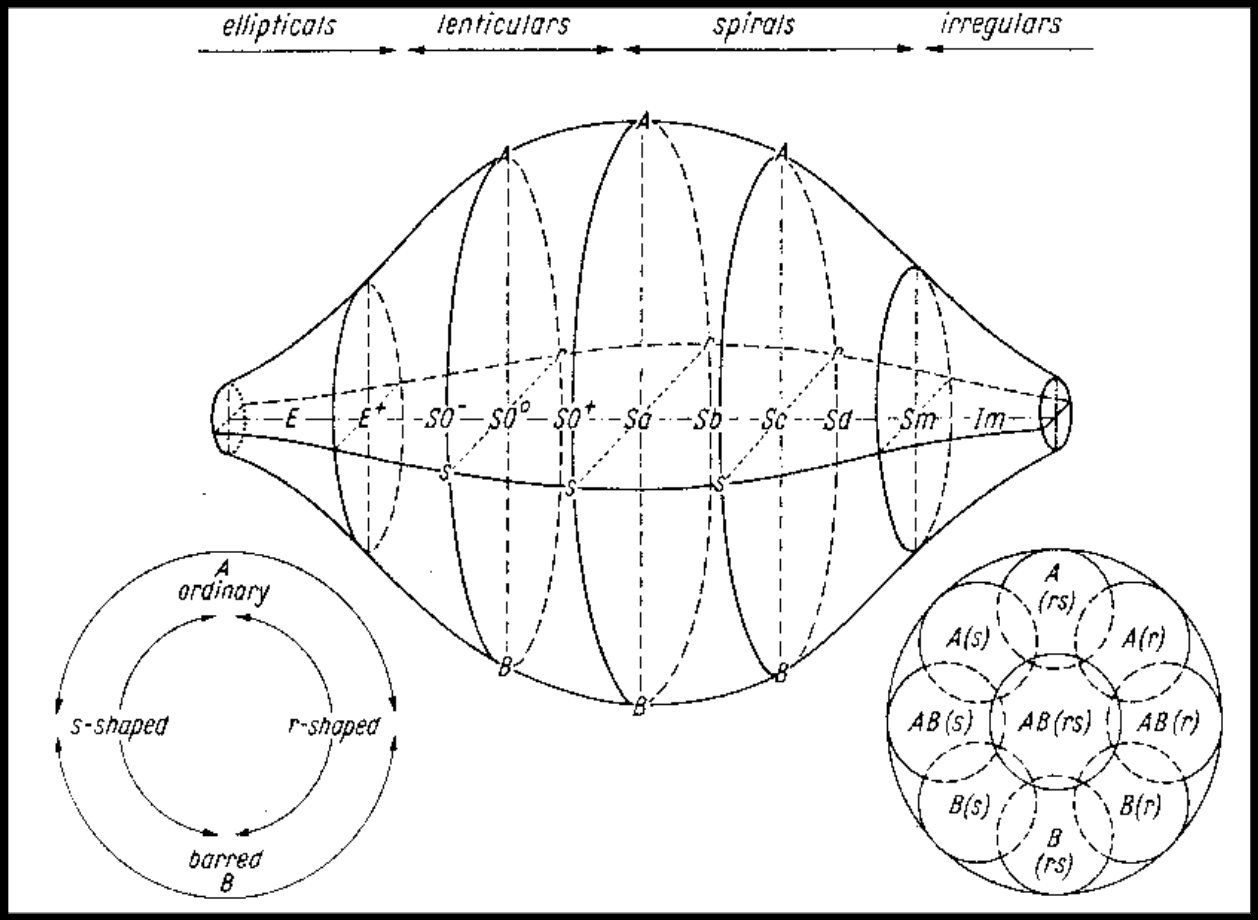
Project Goal

Build a tool to classify galaxies from the Sloan Digital Sky Survey (SDSS) into the 17 class De Vaucouleurs System

Hubble Tuning Fork Classification



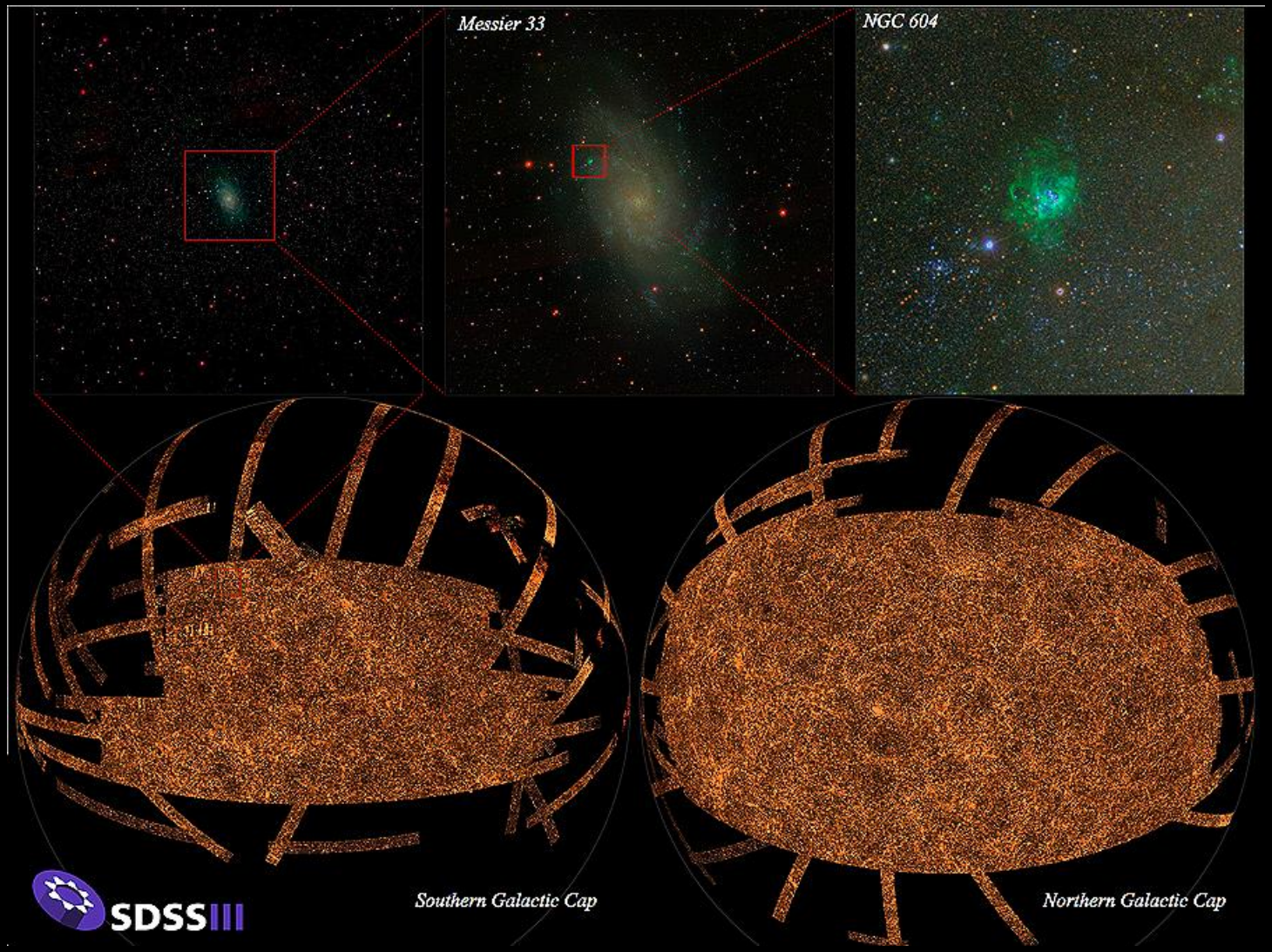
De Vaucouleurs Classification System



Numerical Hubble stage

Hubble stage T	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
de Vaucouleurs class[5]	cE	E	E^+	$S0^-$	$S0^0$	$S0^+$	S0/a	Sa	Sab	Sb	Sbc	Sc	Scd	Sd	Sdm	Sm	Im	
approximate Hubble class[6]	E			S0			S0/a	Sa	Sa-b	Sb	Sb-c	Sc			Sc-Irr	Irr I		

Sloan Digital Sky Survey



What currently exists for automatic galaxy classification?

- Galaxy Zoo is classifying galaxies via crowdsourcing (mechanical turk style).
- No widely used automatic software for classification.

Strategy

- Training Data?
  - Need a large dataset of pre-classified De Vaucouleurs galaxies
- Storage Limitations?
  - SDSS R7 is around 16 TB uncompressed.
- Machine Learning technique to classify?

Training Data

- EFIGI Database (http://www.astromatic.net/projects/efigi)
  - ~ 4,000 De Vaucouleurs preclassified galaxies.

Storage Limitations

- Instead of downloading the entire SDSS I decide to use a galaxy catalog that lists galaxy coordinates.

- I used the NYU Value-Added Galaxy Catalog (NYU-VAGC) http://sdss.physics.nyu.edu/vagc/

NYU Value-Added Galaxy Catalog

- FITS files that contain the Equatorial Coordinates of around 2 million galaxies from the SDSS.

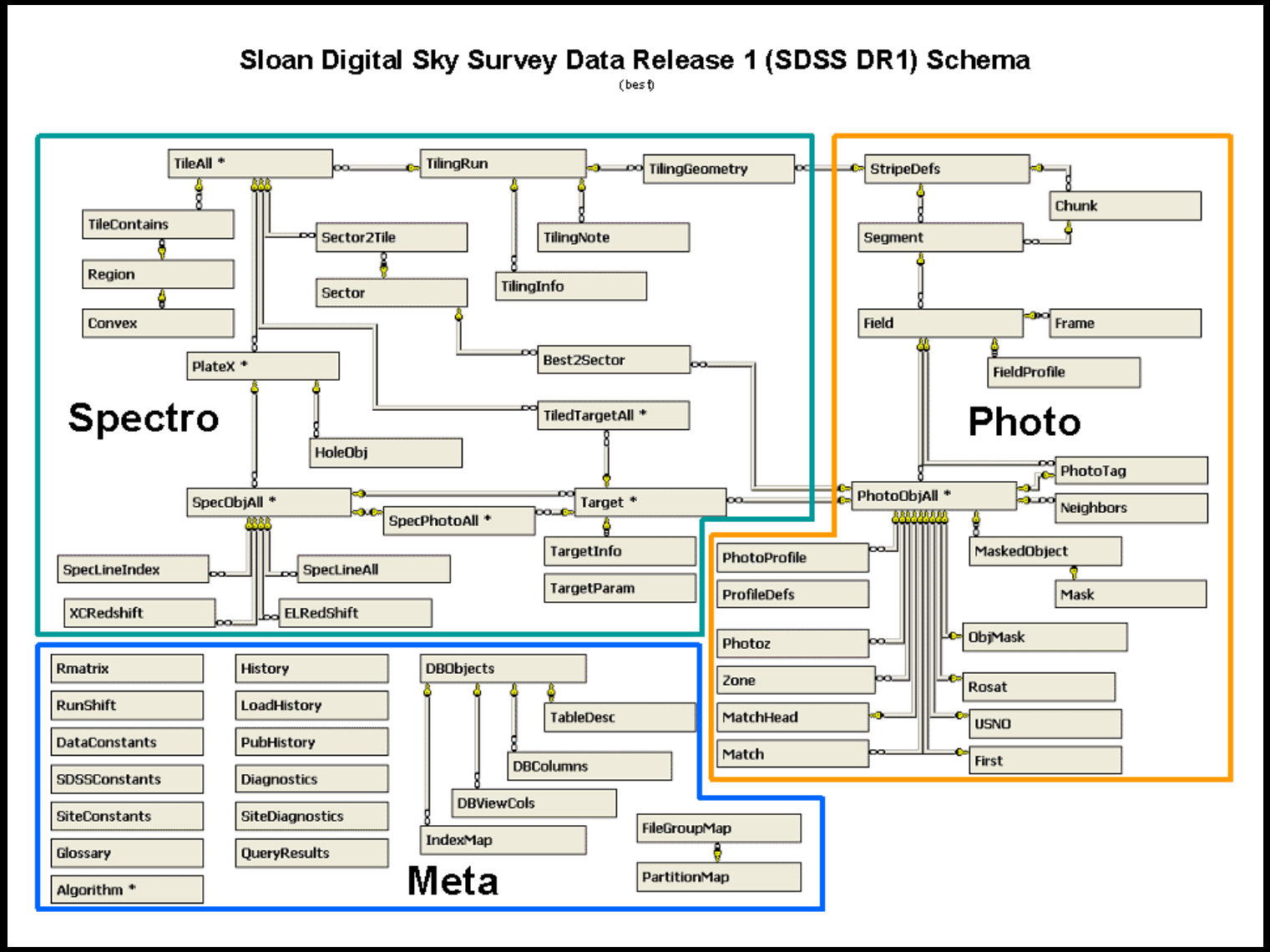
- Files contain other astronomical data but no classifications.

Obtaining the galaxies

- I used the SDSS SkyServer Database to grab images around the EQ. coords from NYU-VAGC.

- I used 30,000 galaxies for the purposes of this project.

SDSS SkyServer



Machine Learning Strategy

- The following were tested and used:
  - K-nearest neighbor
  - Artificial Neural Network
  - M5P Regression Tree

- M5P produced optimal results with a 10-Fold Cross Validation resulting in approximately 1.88 RMSE (RMSE of 1.8 ± 0.3)

Galaxy Tool Design

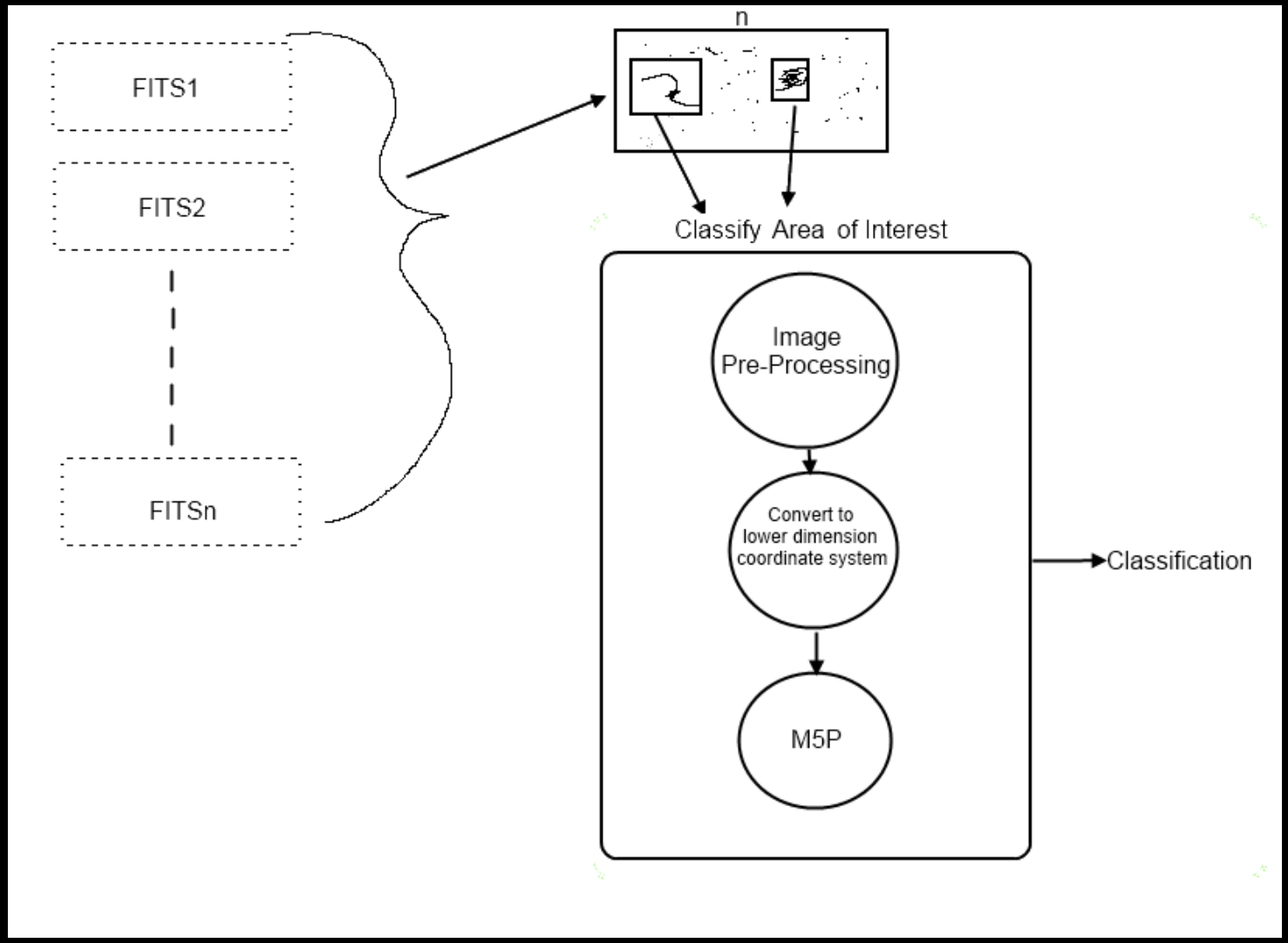


Image Pre-processing

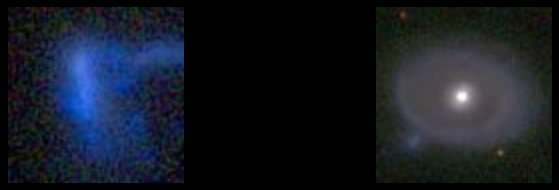
- Rotate image to level all galaxies same direction
- Stretch image so galaxy fills the picture

M5P Tree inputs

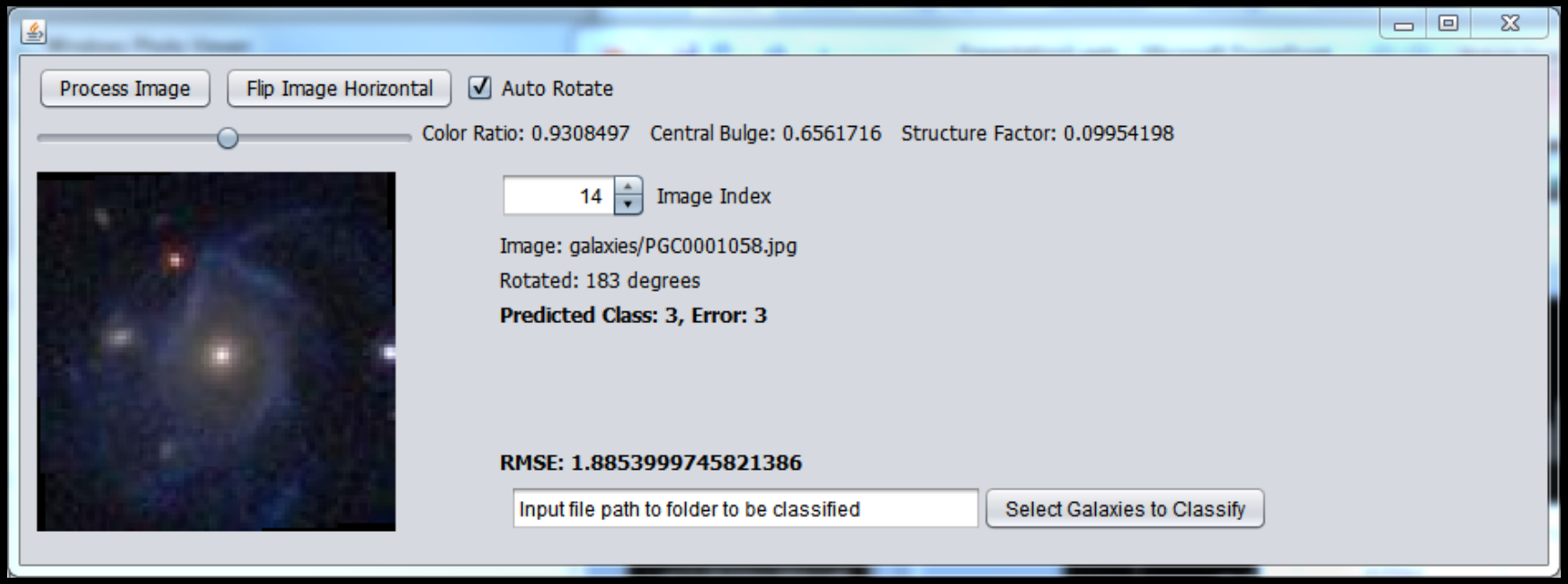
- SVD reduced R, G, B arrays
- Central Bulge Factor

- Chirality Factor

- Consistency Factor

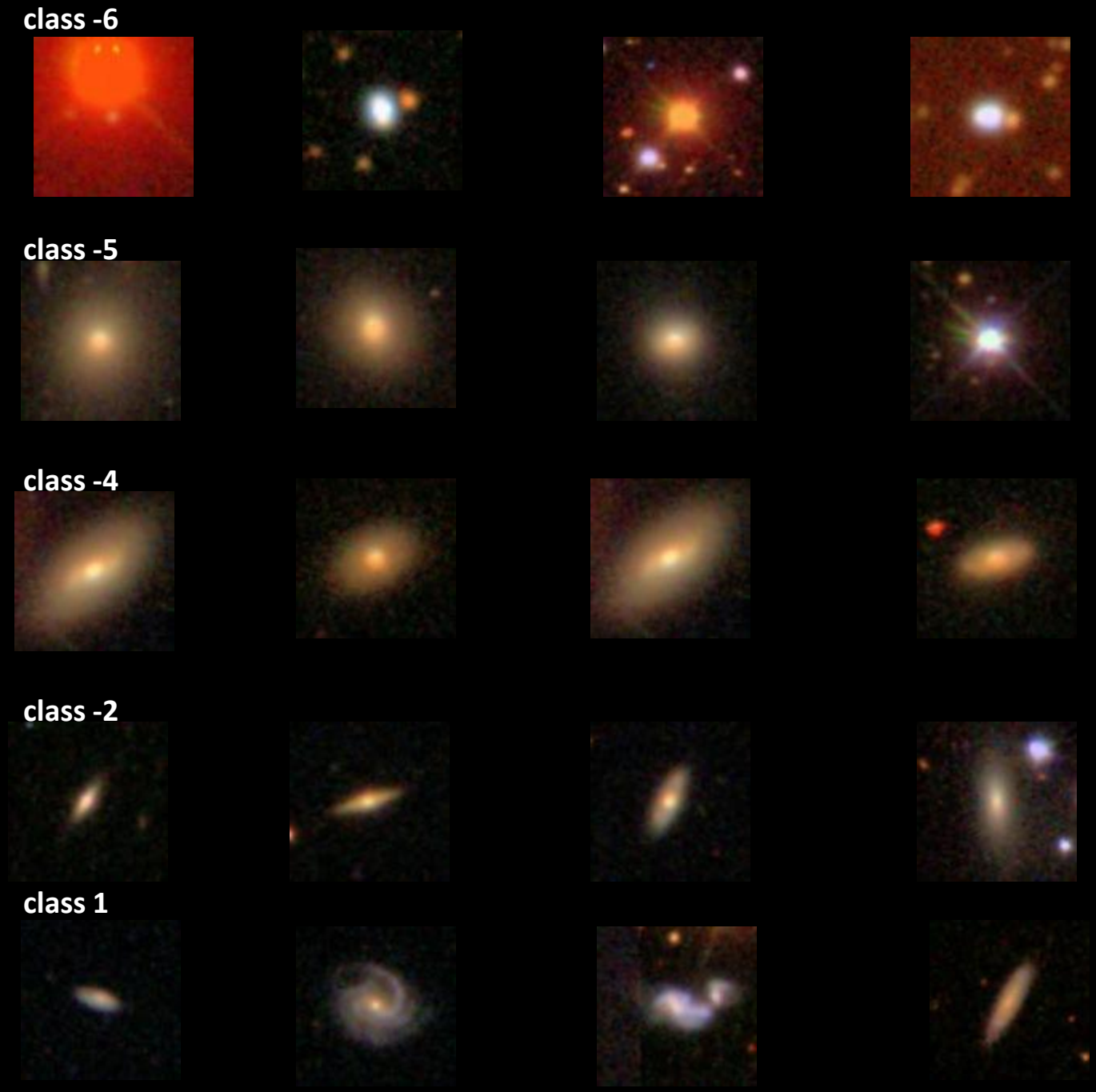


Galaxy Tool Interface



Results

- Ellipse Total: 1195
- Ellipse Percent: 4.5652506112469435
- Sprial Total: 22858
- Sprial Percent: 87.32426650366747
- Irregular Total 2123
- Irregular Percent 8.110482885085576


class -6
class -5
class -4
class -2
class 1

Future Work

- Use consistency function and some lower bound to create a ‘Gold standard’ and other tiers of categories.
- Finish the rest of the SDSS.
- Implement / improve galaxy parsing from the RAW data to help add to existing galaxy cataloges.

Special Thanks

- NYU-VAGC (Eq. Coords FITs Files)
- SDSS (The Testing Data)
- EFIGI (The Training Data)
- WEKA Machine Learning Library (M5P)
- JAMA Java Matrix package (SVD)
- Dr. Marc Seigar – UALR Astronomy Dept
- Dr. Keith Bush – UALR Computer Science Dept
- Craig Williams – UALR Comp Science Graduate