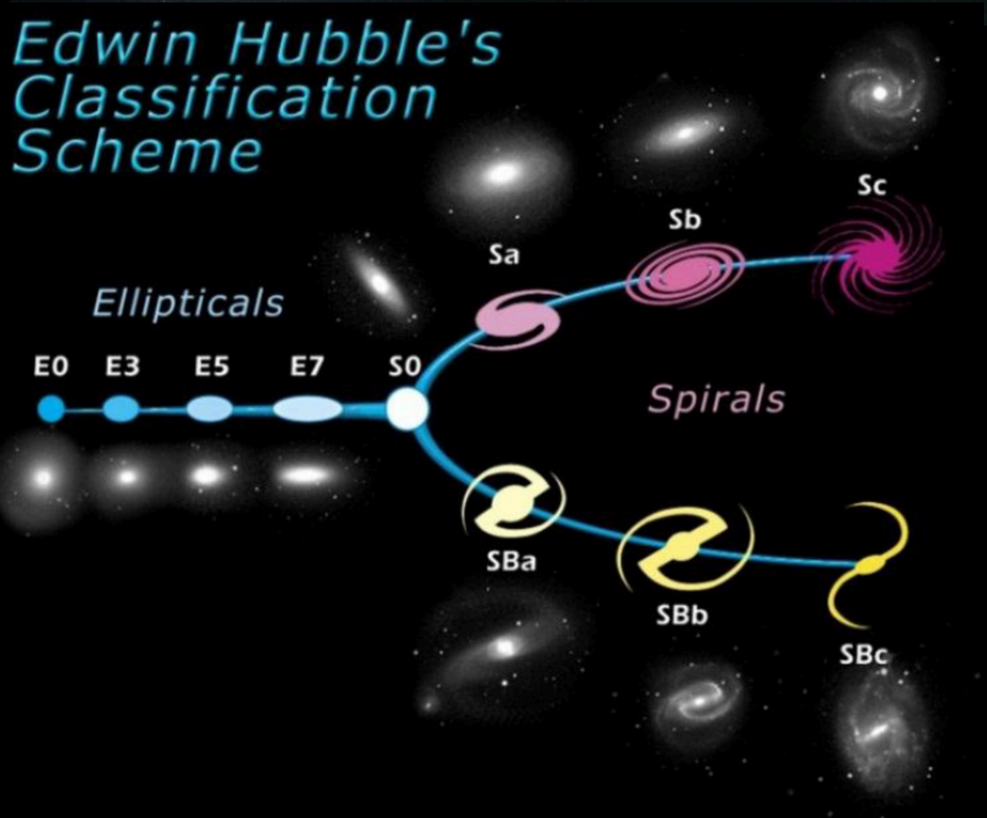
AN ANALYSIS OF LARGE SPECTRAL AND PHOTOMETRIC DATASETS

GALAXY MORPHOLOGICAL CLASSIFICATION

Son pham
4/28/22

OVERVIEW

- Produce a *fast and effective* model for galaxy classification
  - read sky survey data through SQL queries
  - preprocess and integrate heterogeneous data
  - build quick and accurate machine learning classification models
- 2 galaxy classes:
  - spiral
    - flat stellar and gaseous disk with halo of dark stars, gas, and dark matter
    - high interstellar densities
  - elliptical
    - massive galaxies formed from mergers of spiral galaxies
    - low densities - older stars

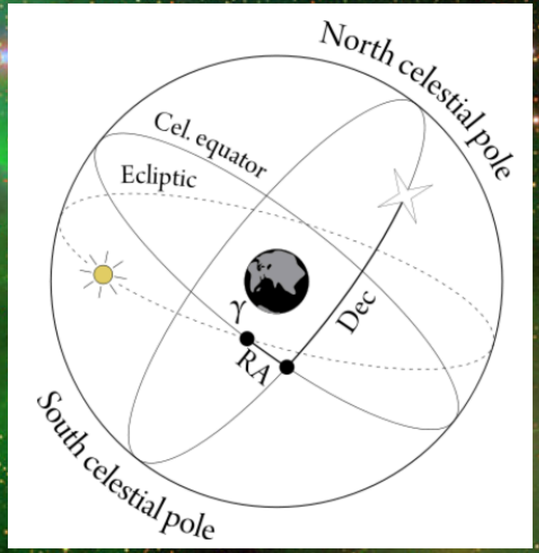


UPDATES

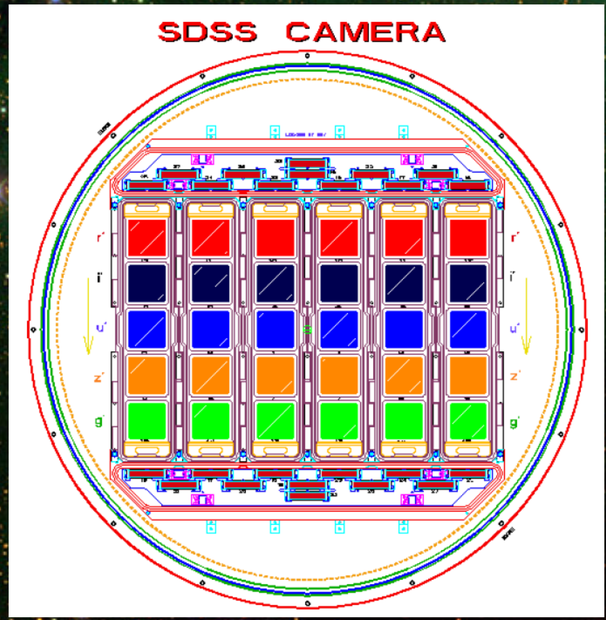
- converted csv's to feather files
  - reading full dataset (4.3M objects):
    - csv: 8.069 s
    - feather: 0.556 s
  - reading processed data:
    - csv: 0.45 s
    - feather: 0.052 s

PRE-PROCESSING

- convert and compute distances using celestial (ra,dec) coordinates
- Perform kd-tree crossmatching on SDSS and GalaxyZoo datasets
- Remove non-galaxy data objects
- Convert galaxy labels: spiral=0, elliptical=1
- compute u-g, g-r, r-i, i-z color filters (4D SDSS color-color space)
- compute average 4th moments of color intensities

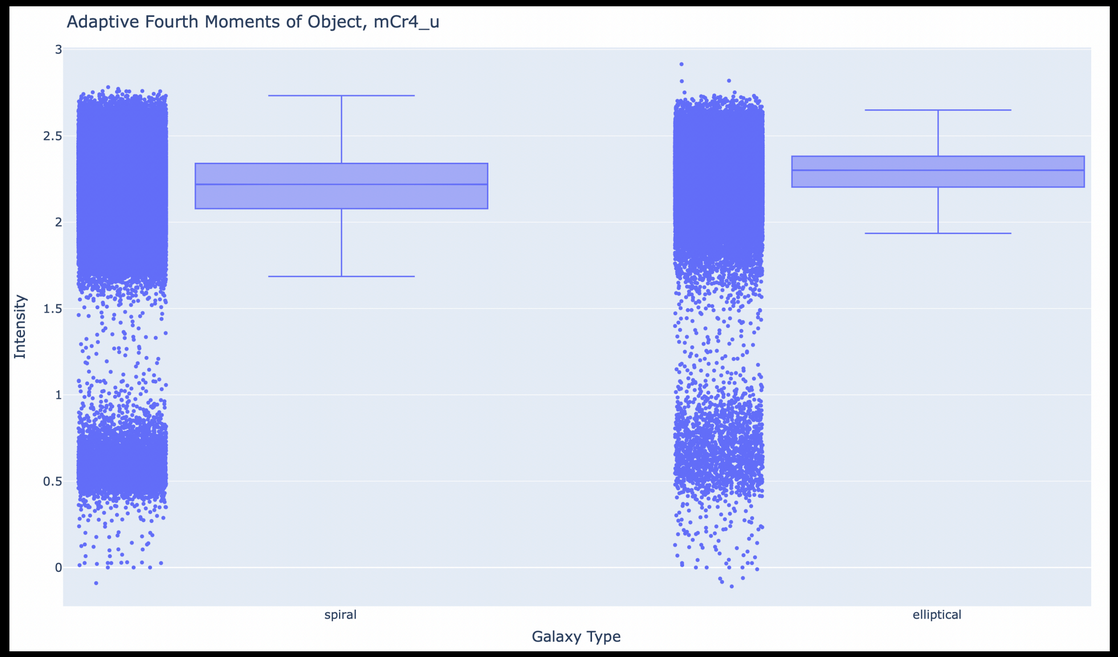


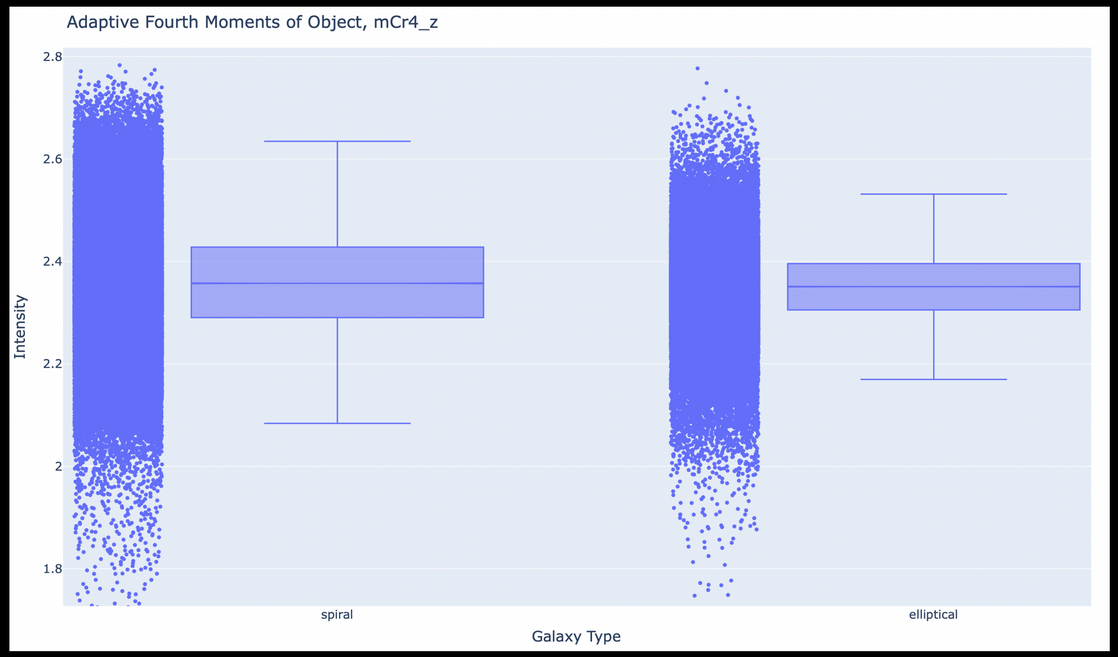
The Distribution, Classification, and Color Evolution of Galaxies Since z~0.5 in the Sloan Digital Sky Survey



STATISTICAL ANALYSIS

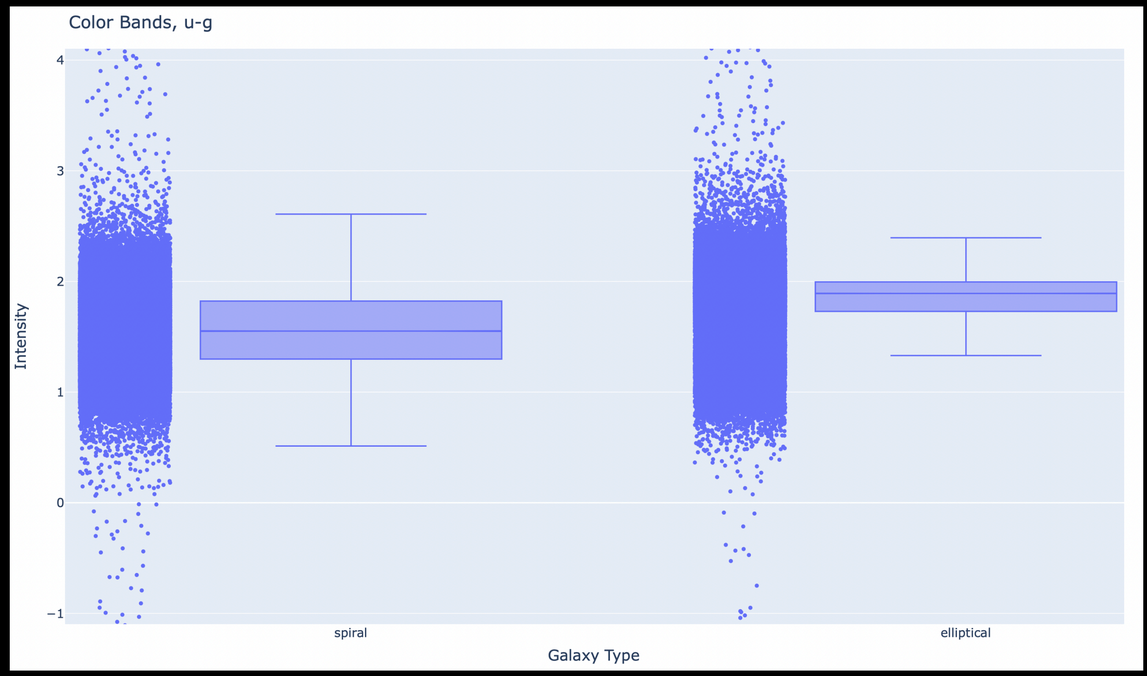
- adaptive fourth moments of object
  - attribute for each color-band filter
  - represents color intensity adapted to the shape and size of the object
  - u, g - similar pattern
    - higher intensities relative to spirals
  - r, i, z - similar pattern
    - tighter clusters for elliptical galaxies





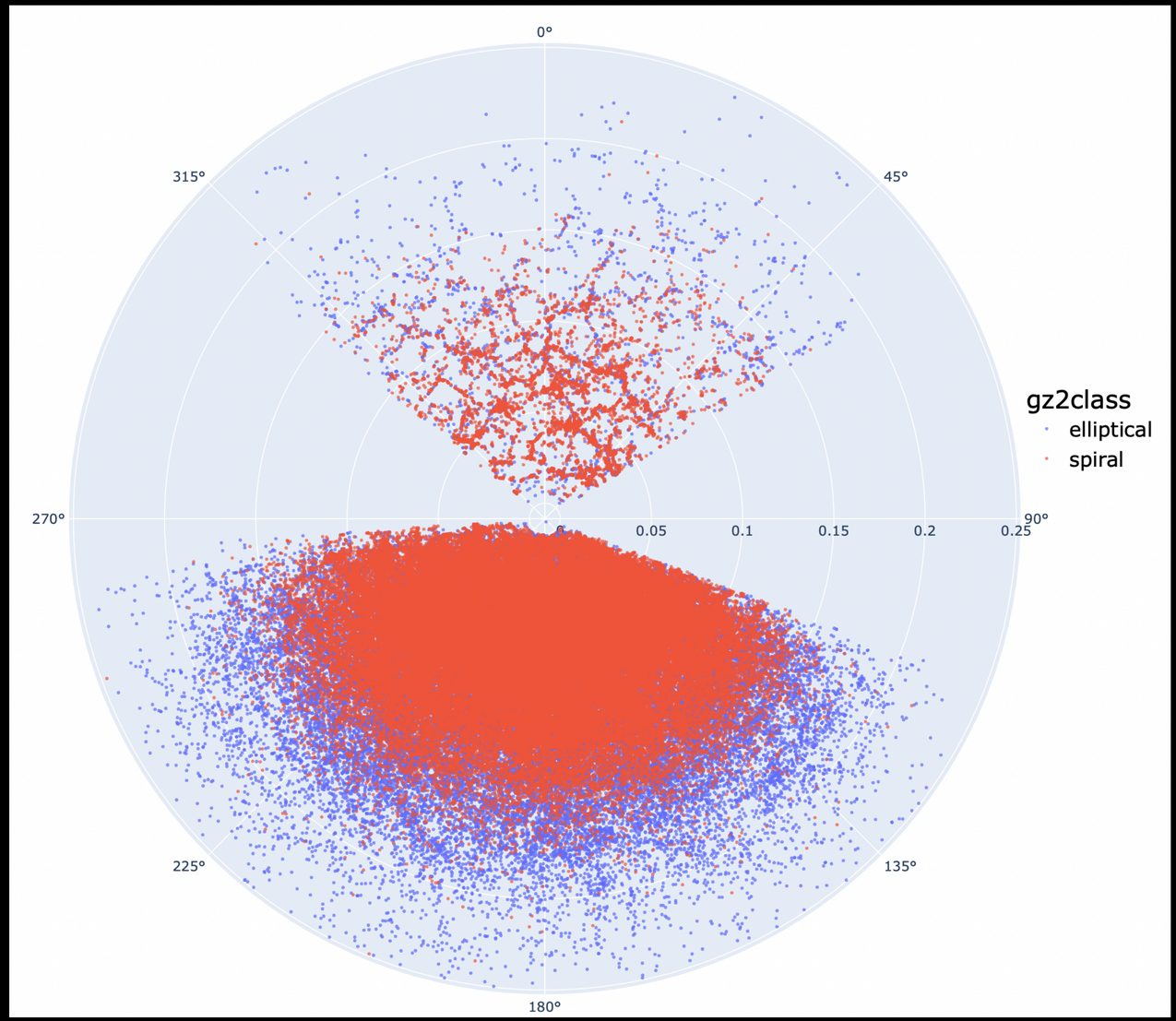
STATISTICAL ANALYSIS

- color bands
  - u-g, g-i, i-r, r-z
  - 4D representation of color distribution
  - all dimensions show similar pattern



STATISTICAL ANALYSIS

- redshift
  - increase in wavelength of an object as it travels away from a reference point
  - redshift vs. right ascension
    - redshift of 0 to 0.25
    - 360 deg of right ascension
  - most spiral galaxies have lower redshift compared to elliptical galaxies



MACHINE LEARNING

- Build several models for speed and accuracy/precision comparison
  - Ensembles: random forest, adaboost, stochastic gradient boosting
  - State vector machines
  - K-nearest neighbors
  - logistic regression with L2
  - Multi-layer Perceptron NN

Random Forest

# Of Trees	F1 (228K Objects)	Speed (228K)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.7303	9.78	0.7442	4.10	0.7498	1.98	0.7761	0.98
5	0.7805	32.23	0.8044	13.09	0.8073	6.12	0.8252	2.89
10	0.8082	54.23	0.8322	22.30	0.8343	10.08	0.8505	4.85
15	0.8178	71.60	0.8392	29.00	0.8413	13.53	0.8571	6.24
20	0.8203	82.51	0.8385	33.42	0.8417	15.17	0.8595	6.82
30	0.8205	90.19	0.8398	35.61	0.8410	15.89	0.8563	6.98
50	0.8210	89.98	0.8395	35.63	0.8390	15.79	0.8549	6.93

AdaBoost

# Of Learners	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.7267	2.72	0.7495	1.19	0.7629	0.58	0.7871	0.31
5	0.7431	5.19	0.7688	2.27	0.7896	1.08	0.7944	0.53
10	0.7666	8.46	0.7875	3.50	0.7964	1.71	0.8044	0.82
20	0.7786	14.56	0.7967	6.07	0.8077	2.93	0.8152	1.41
50	0.7912	33.04	0.8093	13.52	0.8171	6.52	0.8370	3.14
100	0.7988	62.84	0.8157	25.77	0.8245	12.74	0.8458	6.01
200	0.8040	121.54	0.8213	49.33	0.8287	24.59	0.8477	11.63
500	0.8078	293.48	0.8257	120.40	0.8366	58.97	0.8507	27.78

MACHINE LEARNING

- Build several models for speed and accuracy/precision comparison
  - Ensembles: random forest, adaboost, stochastic gradient boosting
  - State vector machines
  - K-nearest neighbors
  - logistic regression with L2
  - Multi-layer Perceptron NN

Stochastic Gradient Boost

# Of Estimators	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.3646	3.81	0.4028	1.62	0.4030	0.80	0.4104	0.38
5	0.7471	10.41	0.7002	4.32	0.7111	2.01	0.7119	0.96
10	0.7804	18.69	0.7824	7.72	0.7692	3.63	0.7956	1.70
20	0.7893	34.94	0.8074	14.27	0.8066	6.54	0.8260	3.08
50	0.8021	83.73	0.8226	33.76	0.8260	16.19	0.8498	7.52
100	0.8107	162.53	0.8293	66.33	0.8360	31.99	0.8599	14.44
200	0.8169	318.45	0.8353	128.69	0.8390	62.66	0.8634	29.11
500	0.8218	774.50	0.8404	316.01	0.8454	151.51	0.8651	70.97

State Vector Machines

Kernel Type	F1 (228K Objects)
Linear	0.7581
Poly	0.7416
Rbf	0.7340

MACHINE LEARNING

- Build several models for speed and accuracy/precision comparison
  - Ensembles: random forest, adaboost, stochastic gradient boosting
  - State vector machines
  - K-Nearest neighbors
  - logistic regression with L2
  - Multi-layer Perceptron NN

K-Nearest Neighbors

# Of Neighbors	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.7060	131.96	0.7303	24.73	0.7249	7.07	0.7295	1.77
5	0.7571	152.52	0.7791	34.61	0.7803	8.81	0.7866	2.07
10	0.7656	152.52	0.7860	34.64	0.7852	9.07	0.7881	2.01
20	0.7735	152.52	0.7919	34.49	0.7930	8.91	0.7973	2.06
30	0.7747	152.55	0.7942	34.54	0.7952	9.05	0.8023	2.06
50	0.7757	152.56	0.7944	34.70	0.7956	8.89	0.8022	2.08

MACHINE LEARNING

- Build several models for speed and accuracy/precision comparison
  - Ensembles: random forest, adaboost, stochastic gradient boosting
  - State vector machines
  - K-Nearest neighbors
  - logistic regression with L2
  - Multi-layer Perceptron NN

Logistic Regression

Penalty Type	F1 Score
L2	0.7331

Multi-Layer Perceptron NN

# Of Layers	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
(1,1)	0.7853	11.27	0.8078	8.70	0.8141	6.42	0.8217	3.20
(2,2)	0.7962	12.48	0.8167	8.14	0.8145	7.07	0.8143	2.04
(5,2)	0.8118	16.81	0.8182	8.96	0.8267	6.62	0.8340	2.52
(5,5)	0.8094	15.97	0.8310	12.60	0.8408	4.57	0.8545	4.12
(2,2,2)	0.7859	33.54	0.8082	15.14	0.8133	10.22	0.8169	3.34
(4,4,4)	0.8099	19.06	0.8318	11.11	0.8380	6.44	0.8386	3.77
(8,8,8)	0.8167	17.56	0.8365	13.78	0.8414	6.55	0.8535	3.49
(2,2,2,2)	0.7976	24.31	0.8142	5.42	0.8178	2.35	0.8298	1.94
(5,5,5,5)	0.8100	32.56	0.8315	11.69	0.8396	6.03	0.8559	3.47
(8,8,8,8)	0.8154	23.34	0.8348	14.87	0.8417	9.08	0.8557	3.29

FUTURE WORK

- Add irregular galaxies class labels
- Add Hubble classification scheme labels
- Add non-galaxies data objects (use full size dataset)
- improve cross-matching algorithm
  - hash functions, indexing
- use self-supervised learning and convolutional neural networks to classify images