Galaxy "*visual*" morphologies with deep-learning

Marc Huertas-Company

Machine Learning for Astronomy - CEA - 22/01/16

Why (visually) classifying galaxies?





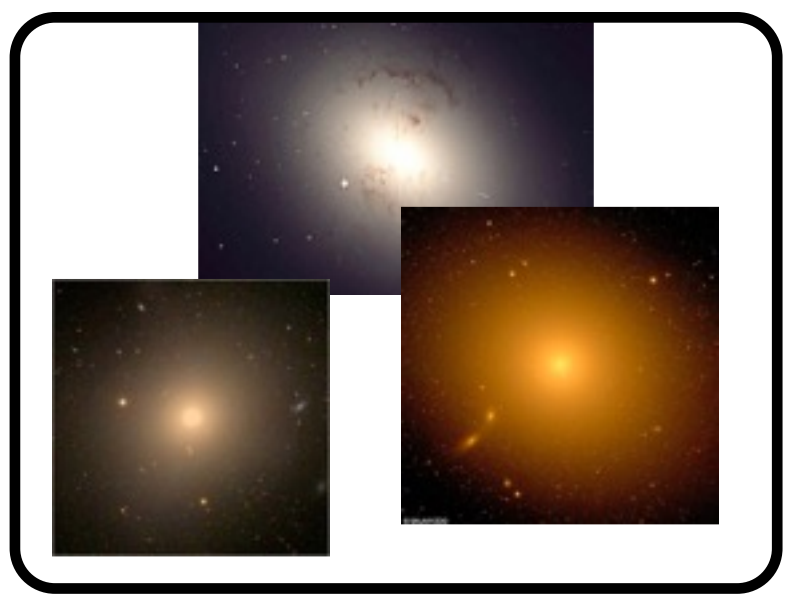

Complex problem

BOX I

BOX II

BOX I

BOX II

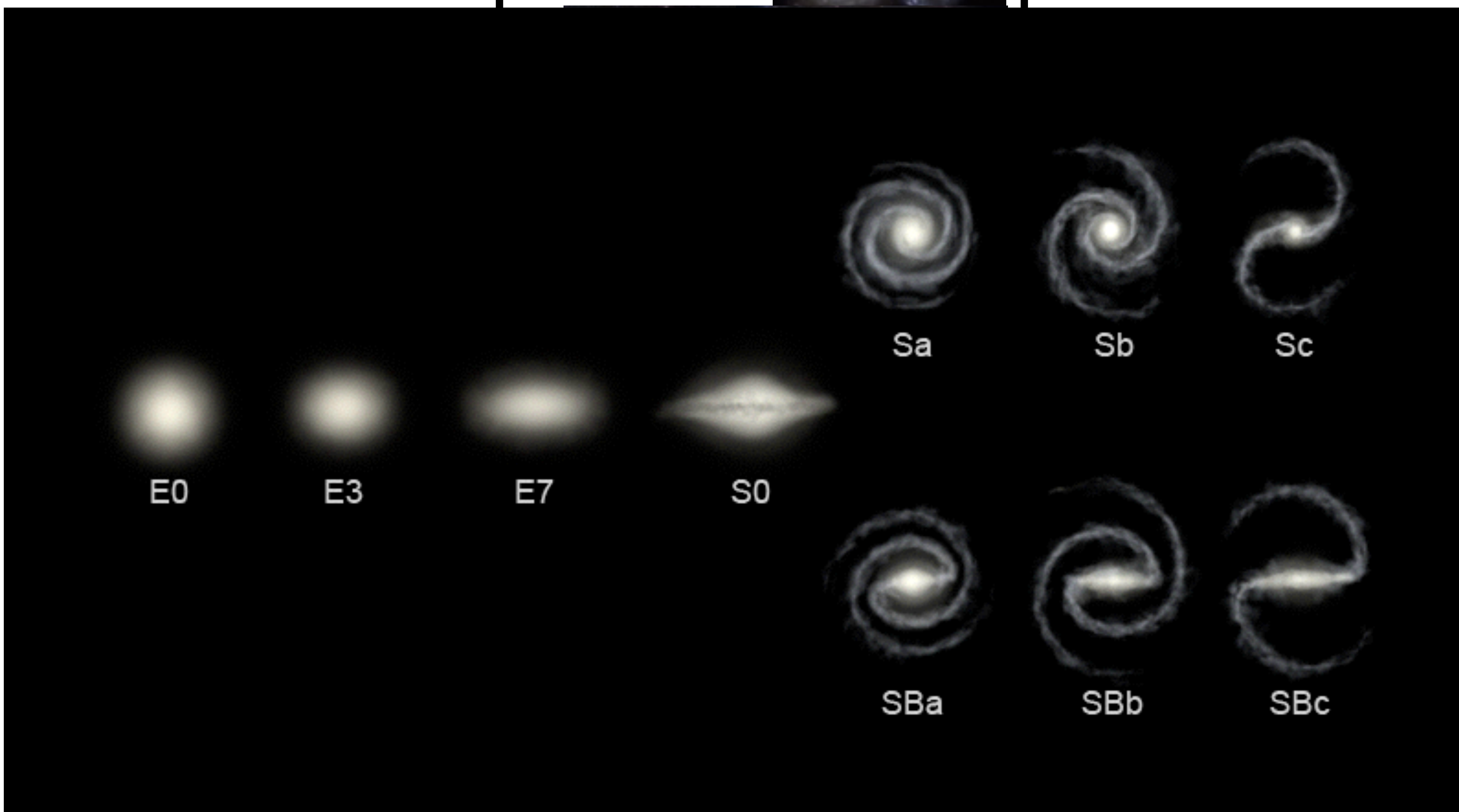


BOX I

BOX II

“Objects in the same box experienced the same physics”

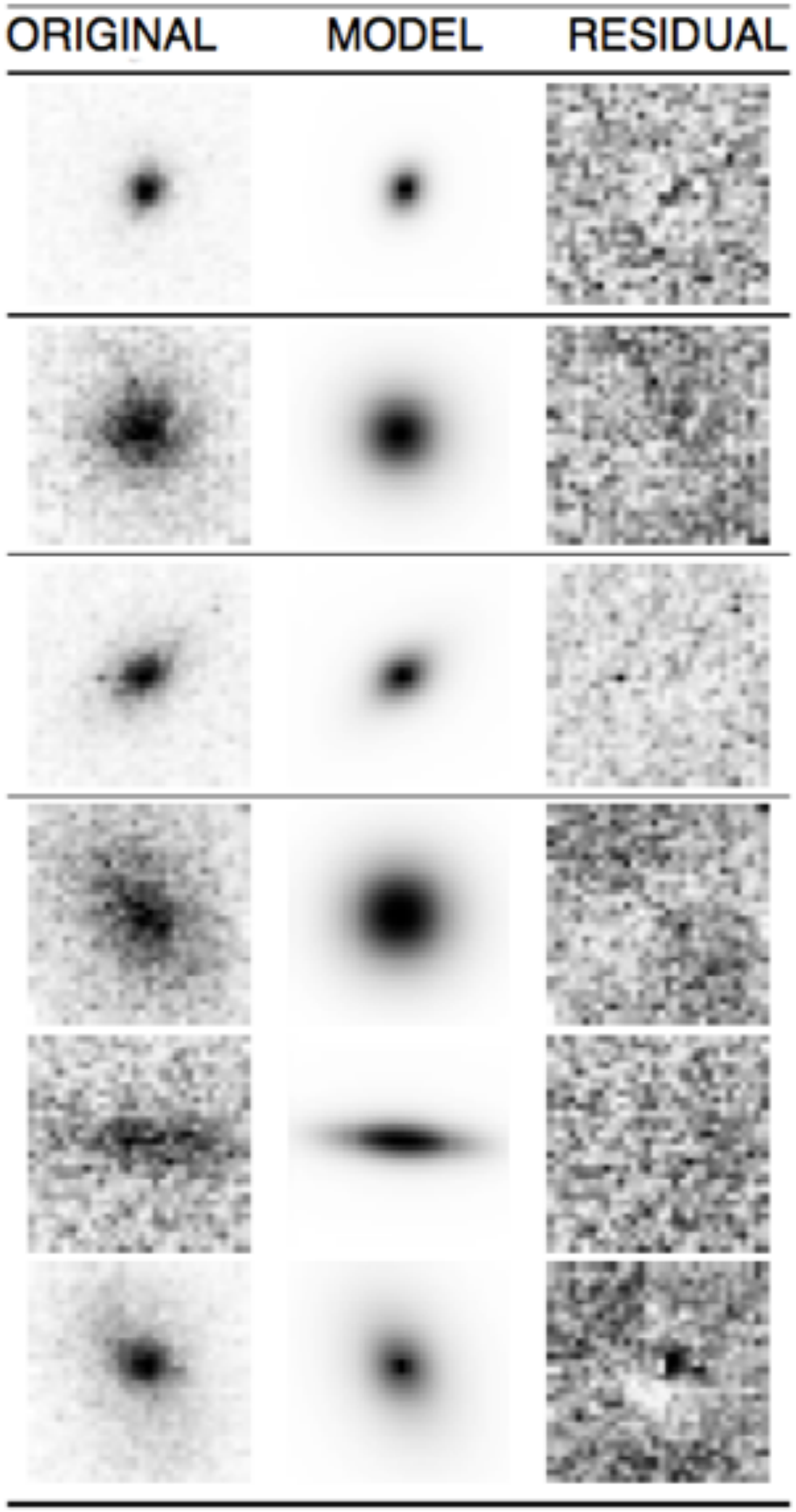
BOX 1



n the
x
ted
e
,

Parametric description

- **Parametric** - **Fit of an analytic profile** (Sersic, Exp …) - e.g Peng+02, Barden+12 etc..
    - +: quantitative, structural information
    - -: **Model assumption.** bad model for many galaxies, bad match with Hubble sequence



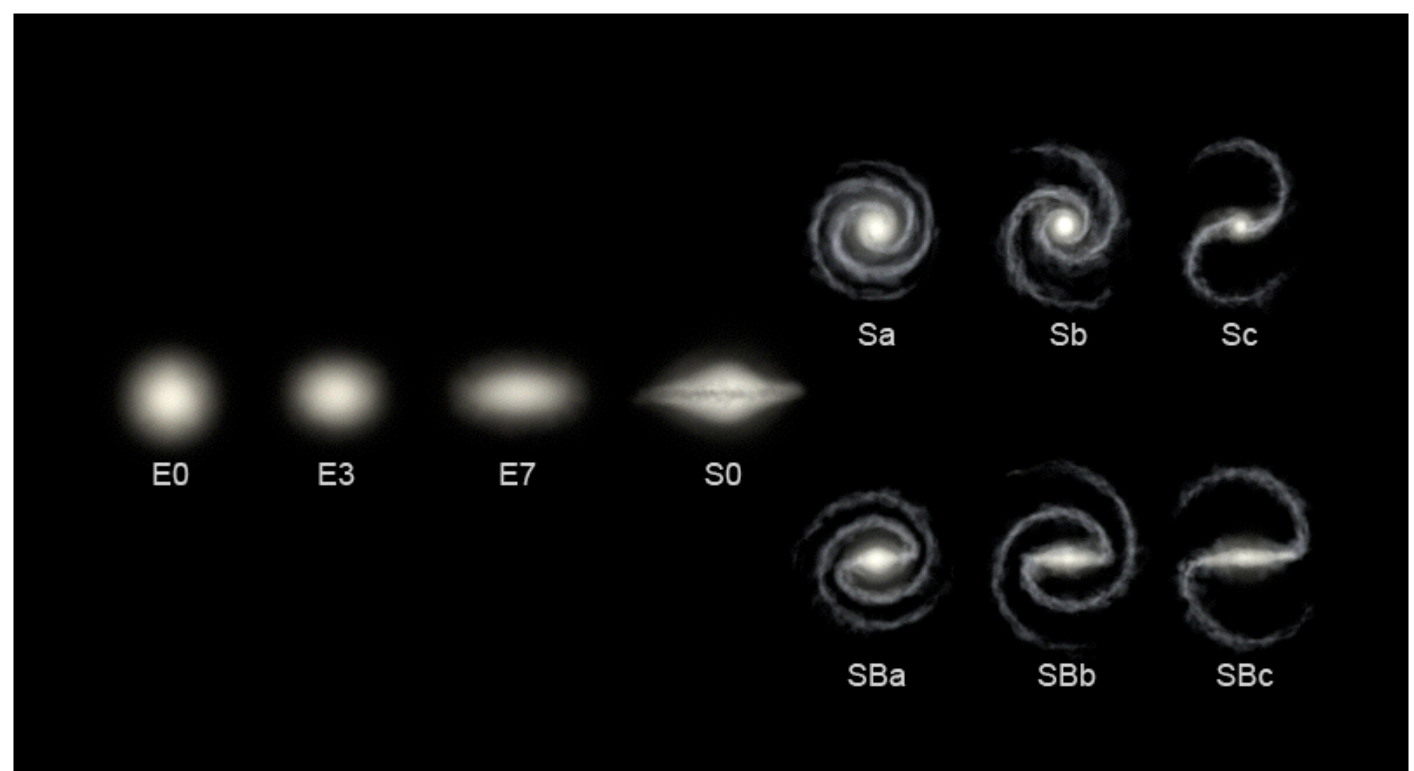
MHC+07

Going beyond Sersic/BD decompositions is highly desirable

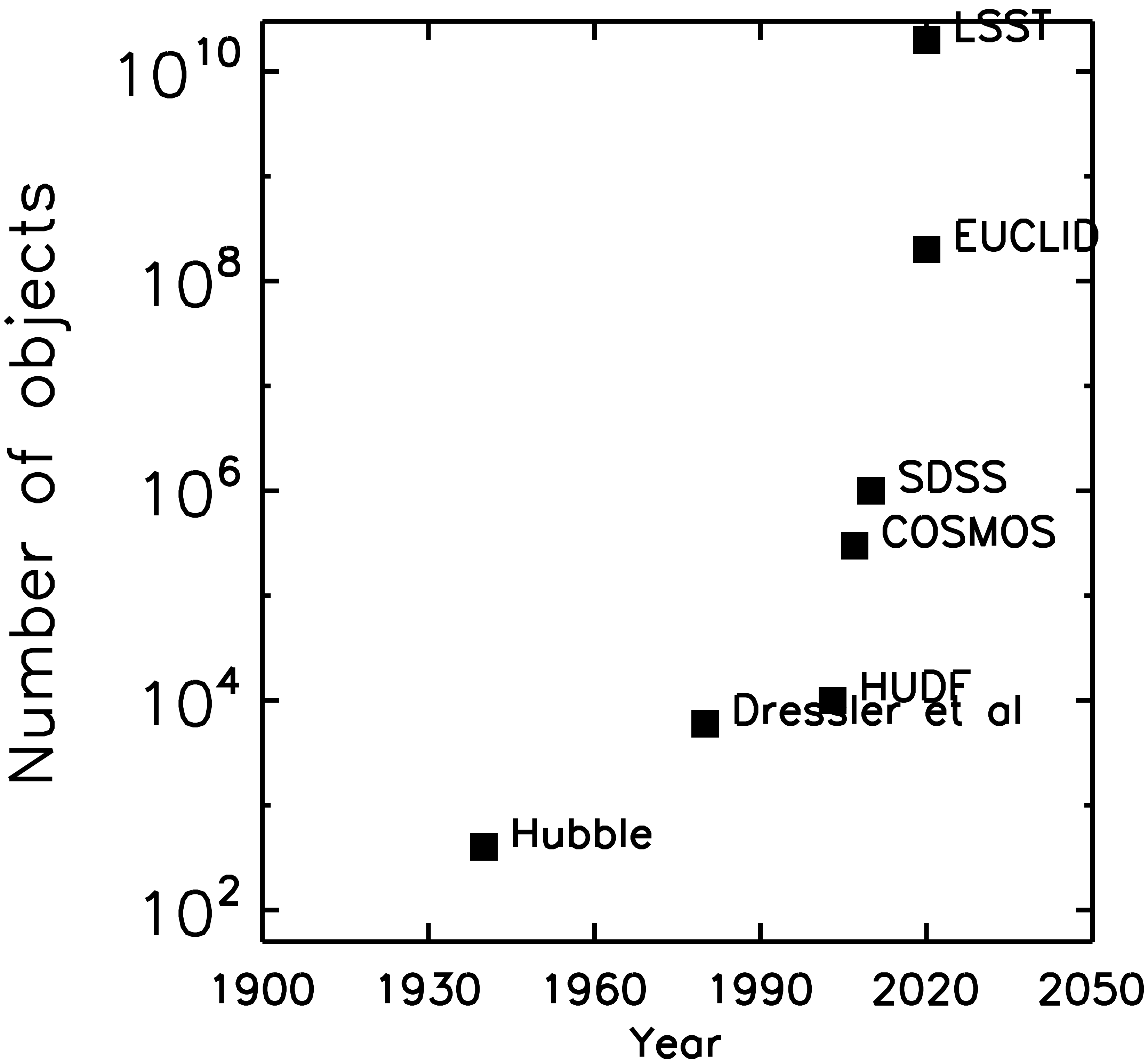
- GZOO: citizen science based visual classification of SDSS, CANDELS etc…
  - 314 papers, > 3000 citations
- All big surveys have visual classifications - COSMOS, CANDELS etc..

Going beyond Sersic/BD decompositions is highly desirable

- GZOO: citizen science based visual classification of SDSS, CANDELS etc…
  - 314 papers, > 3000 citations
- All big surveys have visual classifications - COSMOS, CANDELS etc..

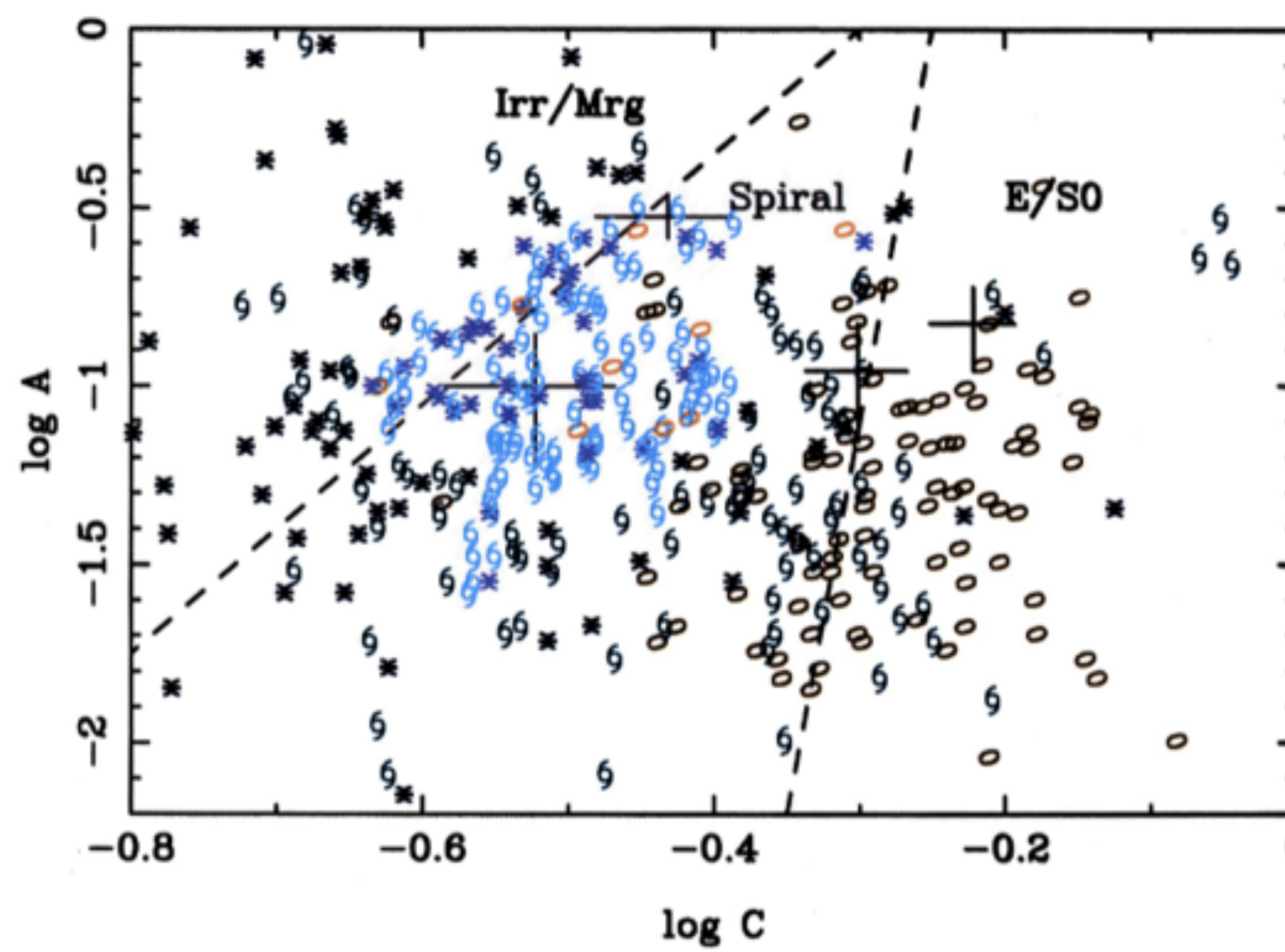


Combination of machine intelligence with human-based inspections is needed



Automated morphologies

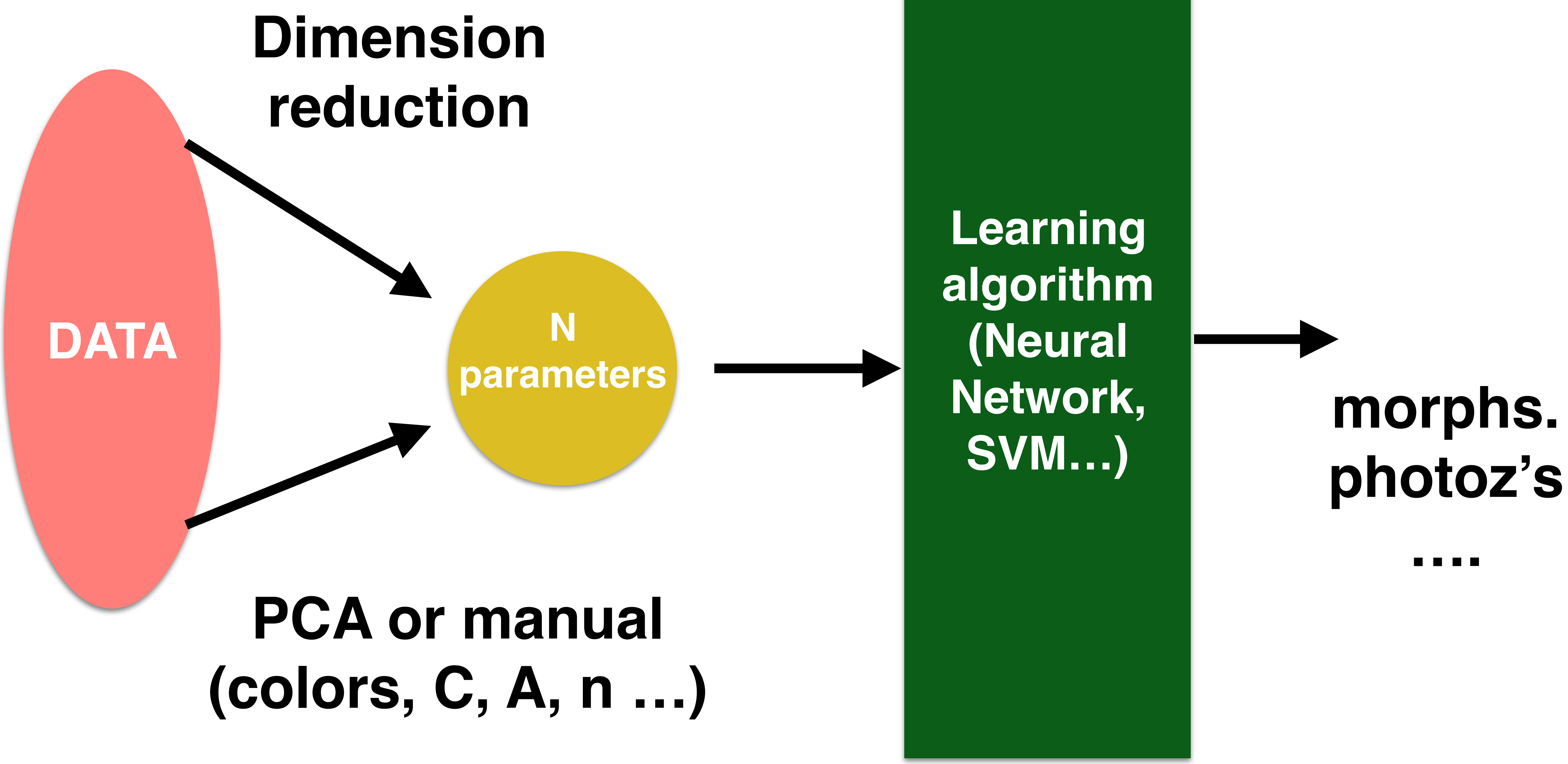
Non-parametric - measure **features** on the images without model assumption and establish correlations between values and morphological types **(C,A,M20,Gini…)**



Abraham+96

- +: not model dependent
- -: not quantitative

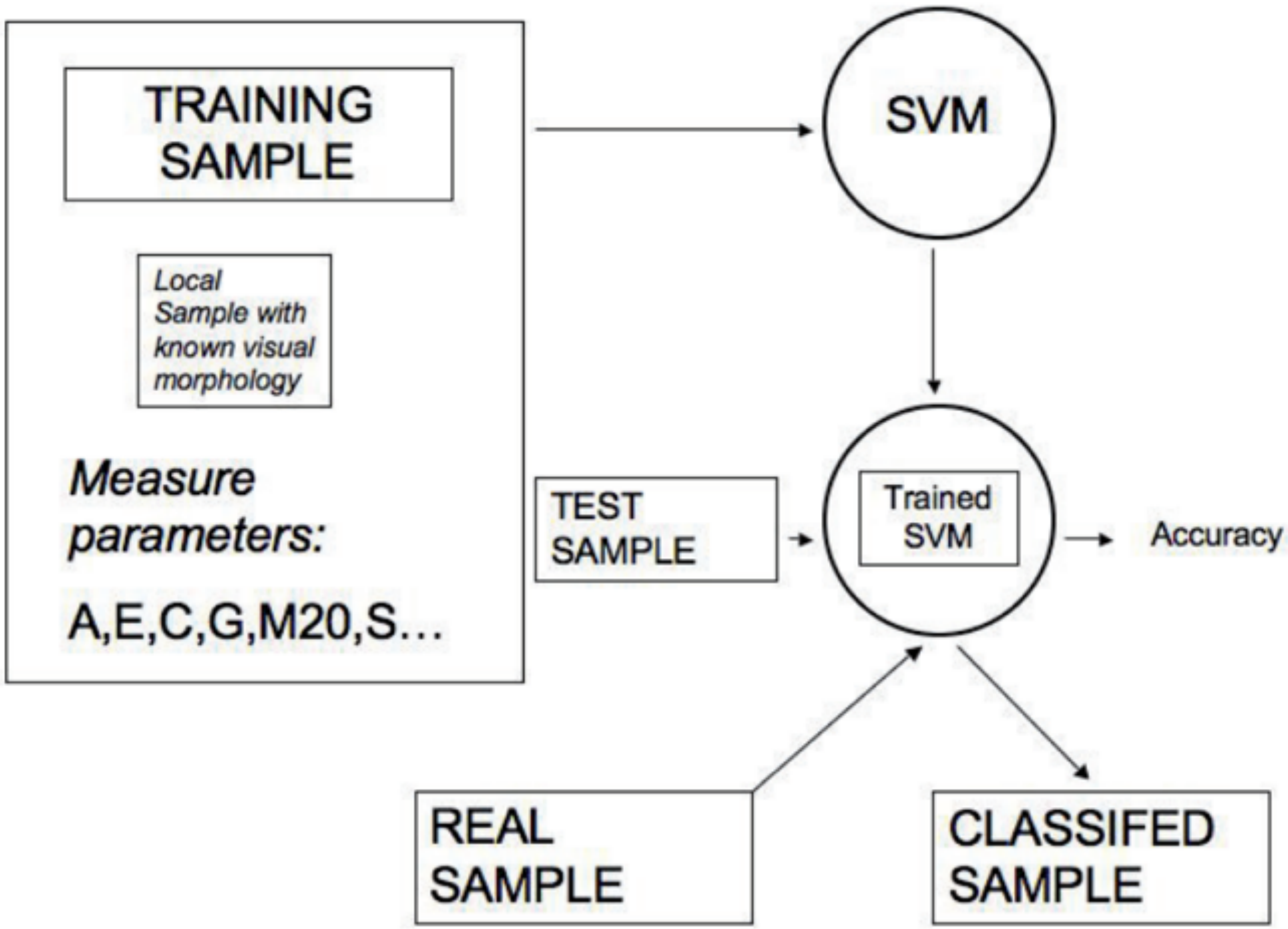
FIRST ML TESTS: LINEAR REDUCTION OF DIMENSION + CLASSIFIER



MHC+08,09,11

http://gepicom04.obspm.fr/galSVM/Home.html

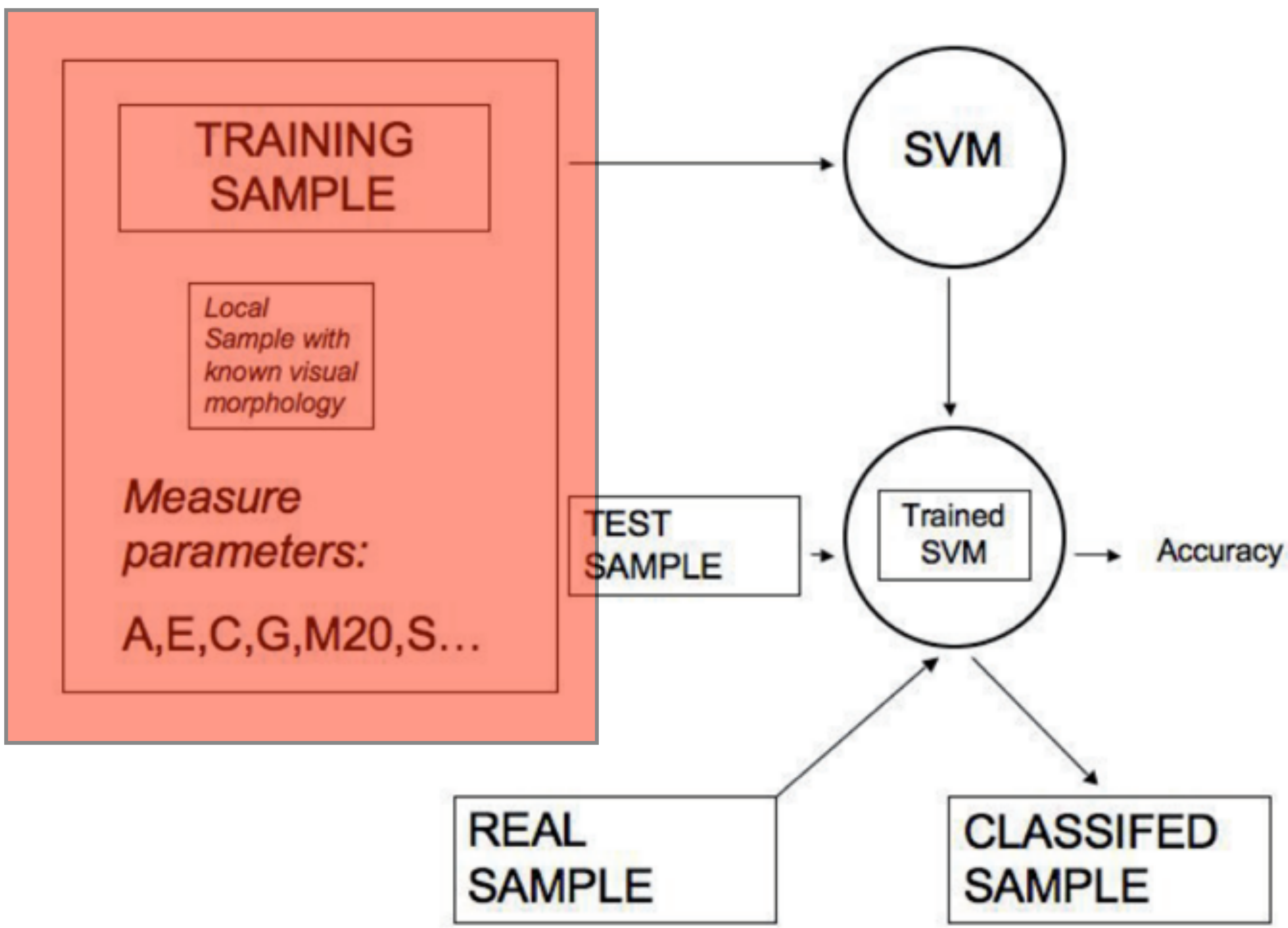
TEST OF MACHINE MACHINE LEARNING ALGORITHMS



TEST OF MACHINE MACHINE LEARNING ALGORITHMS

MHC+08,09,11

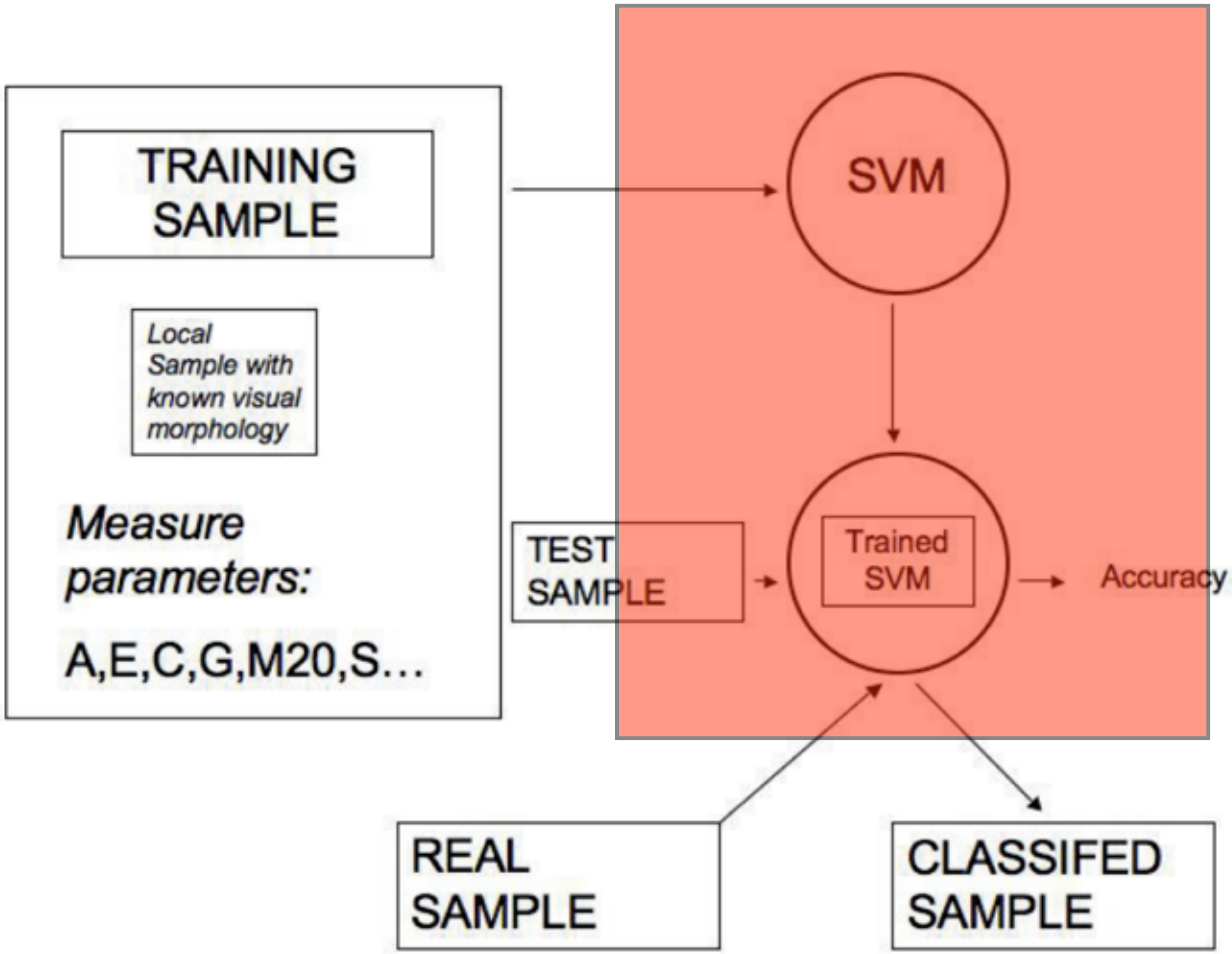
http://gepicom04.obspm.fr/galSVM/Home.html

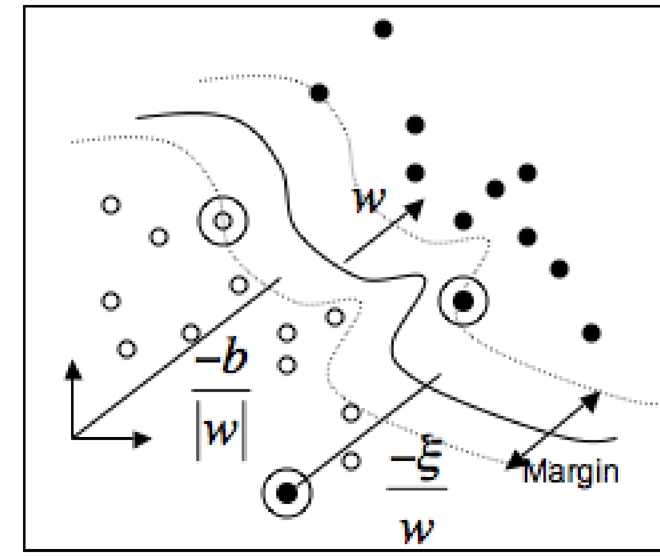


TEST OF MACHINE MACHINE LEARNING ALGORITHMS

MHC+08,09,11

http://gepicom04.obspm.fr/galSVM/Home.html

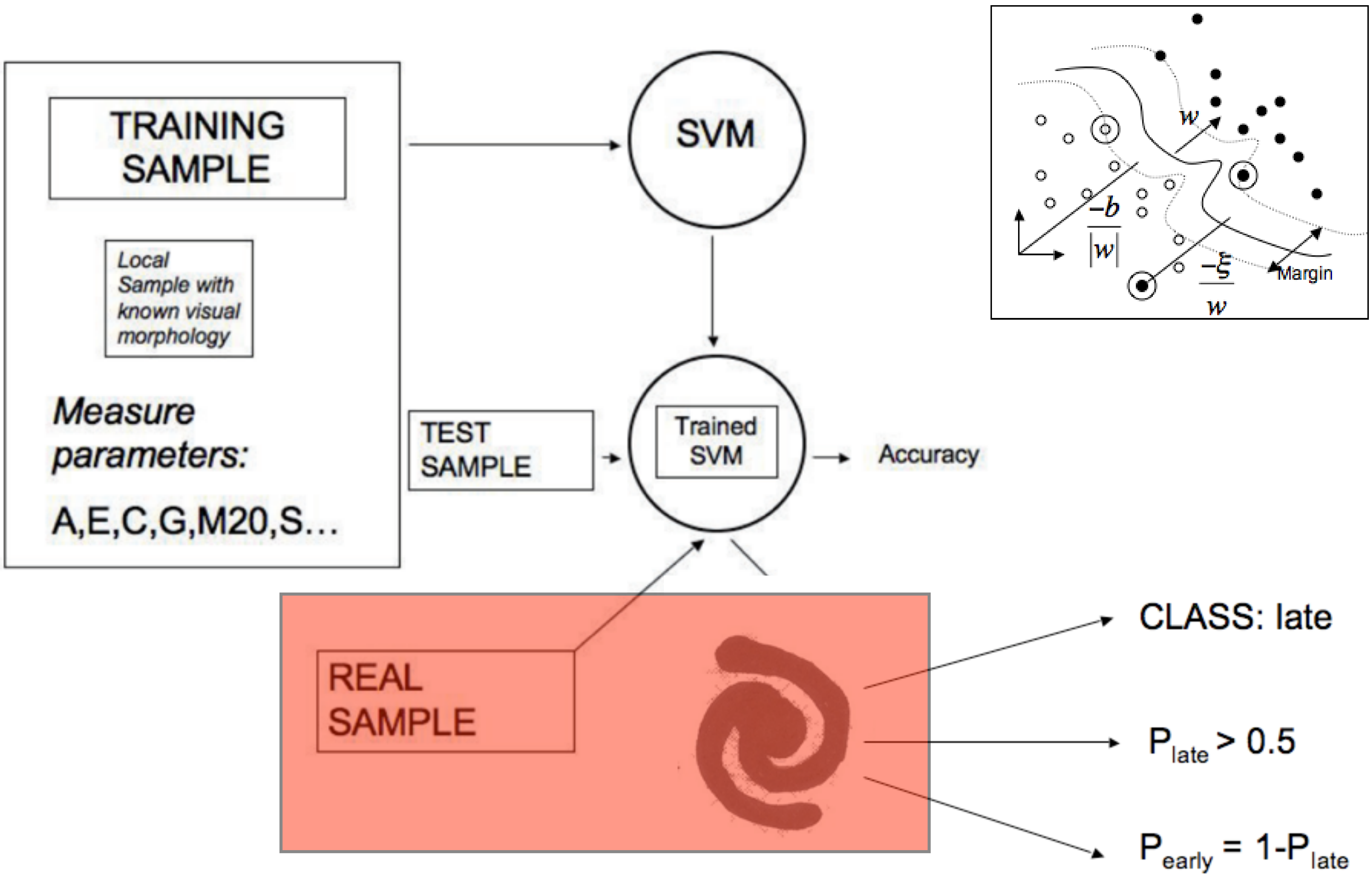




TEST OF MACHINE MACHINE LEARNING ALGORITHMS

MHC+08,09,11

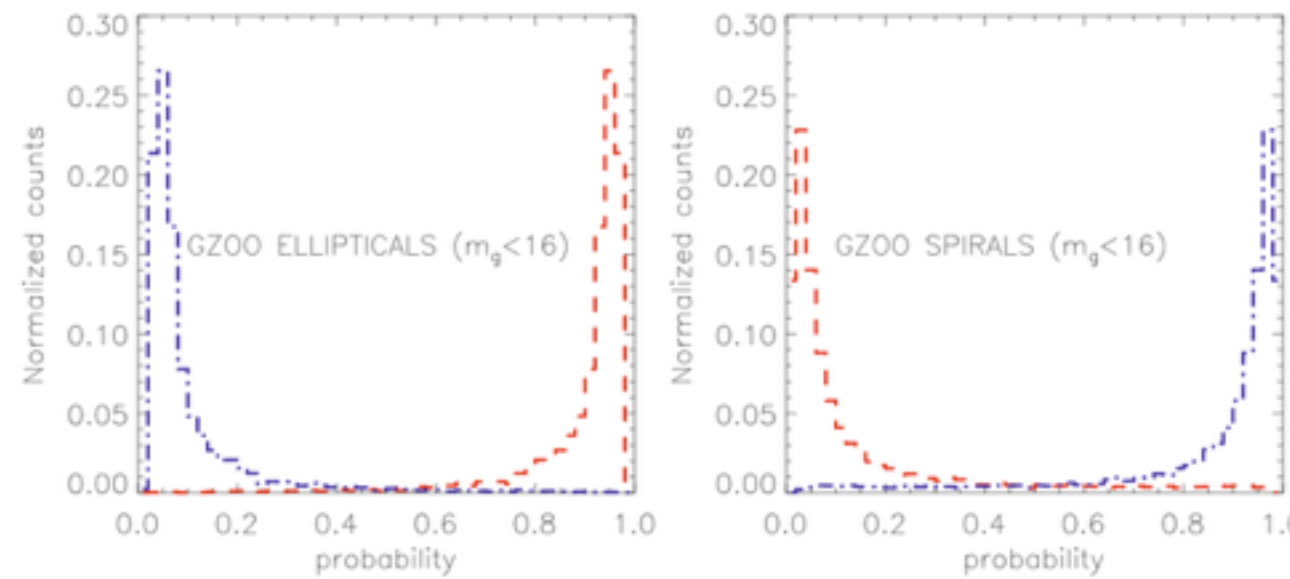
http://gepicom04.obspm.fr/galSVM/Home.html

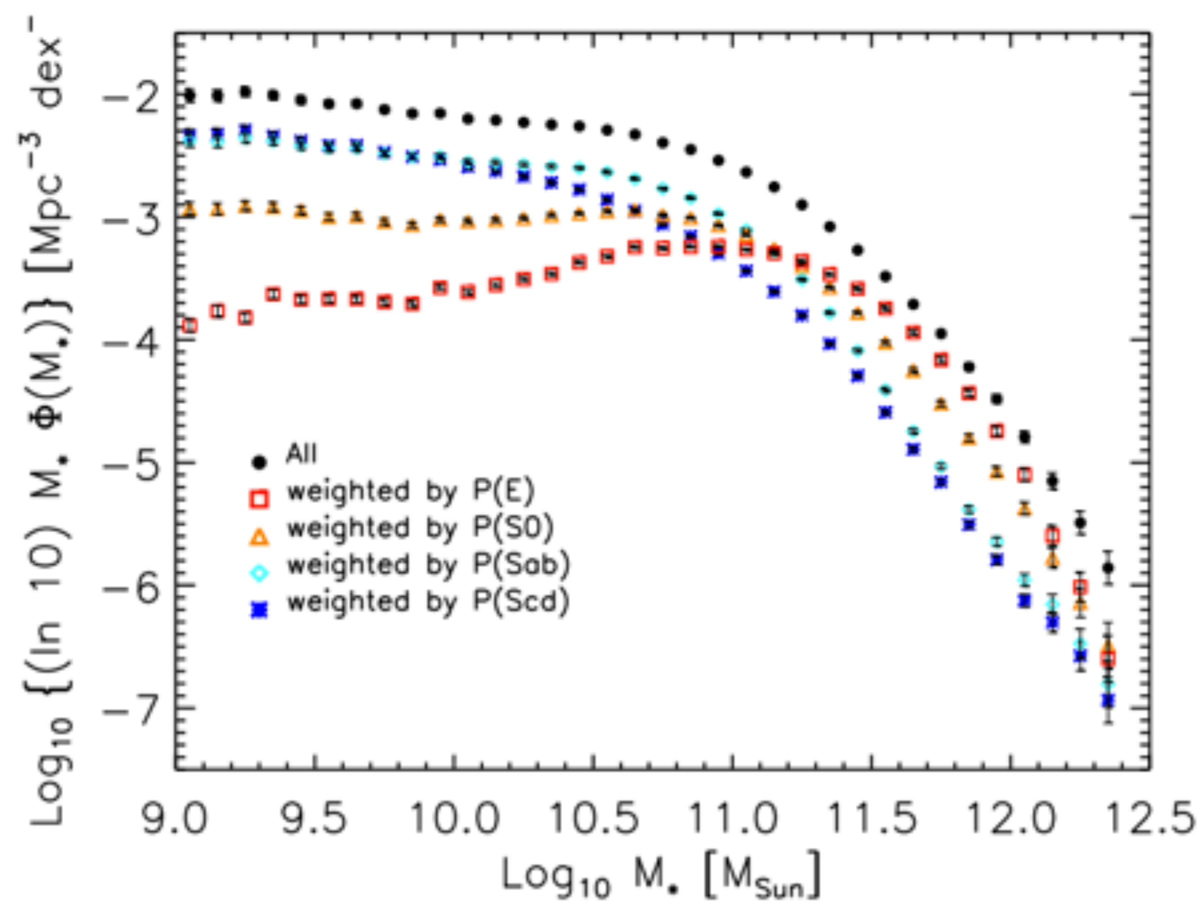


SOME EXAMPLES

- **z~0 - SDSS - Bayesian classification of ~1e6 galaxies -** *gepicom04.obspm.fr/**sdss_morphology/M***

MHC+11



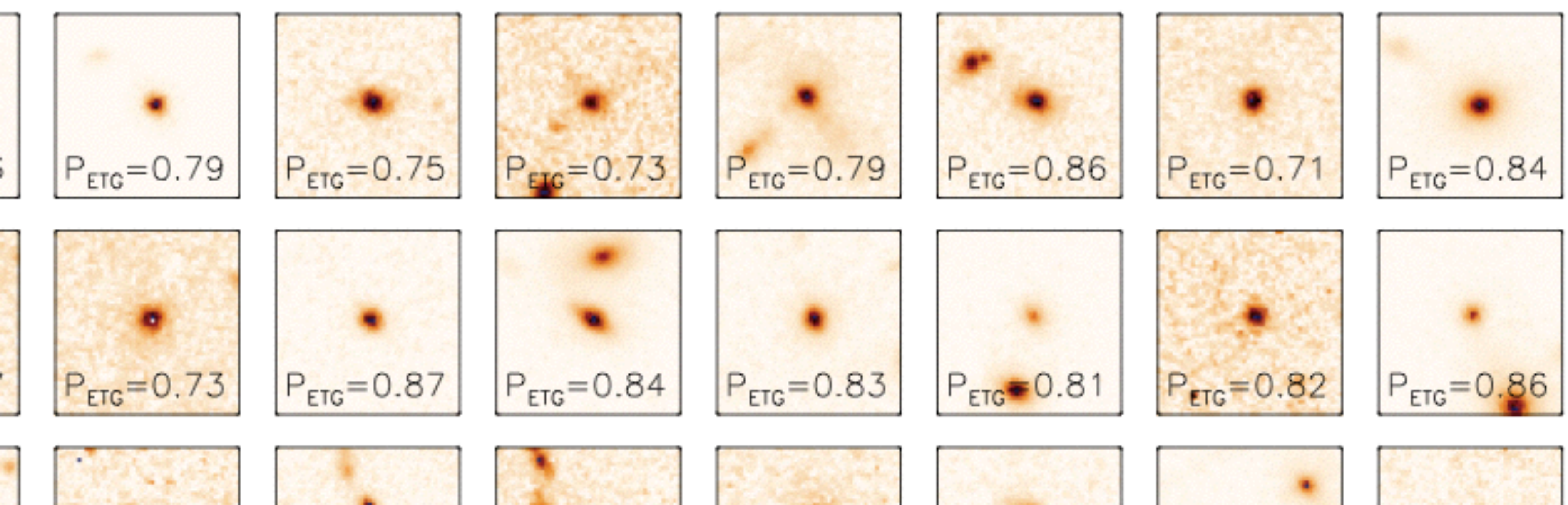


Bernardi+13b

SOME EXAMPLES

z>1

WFC3/ERS

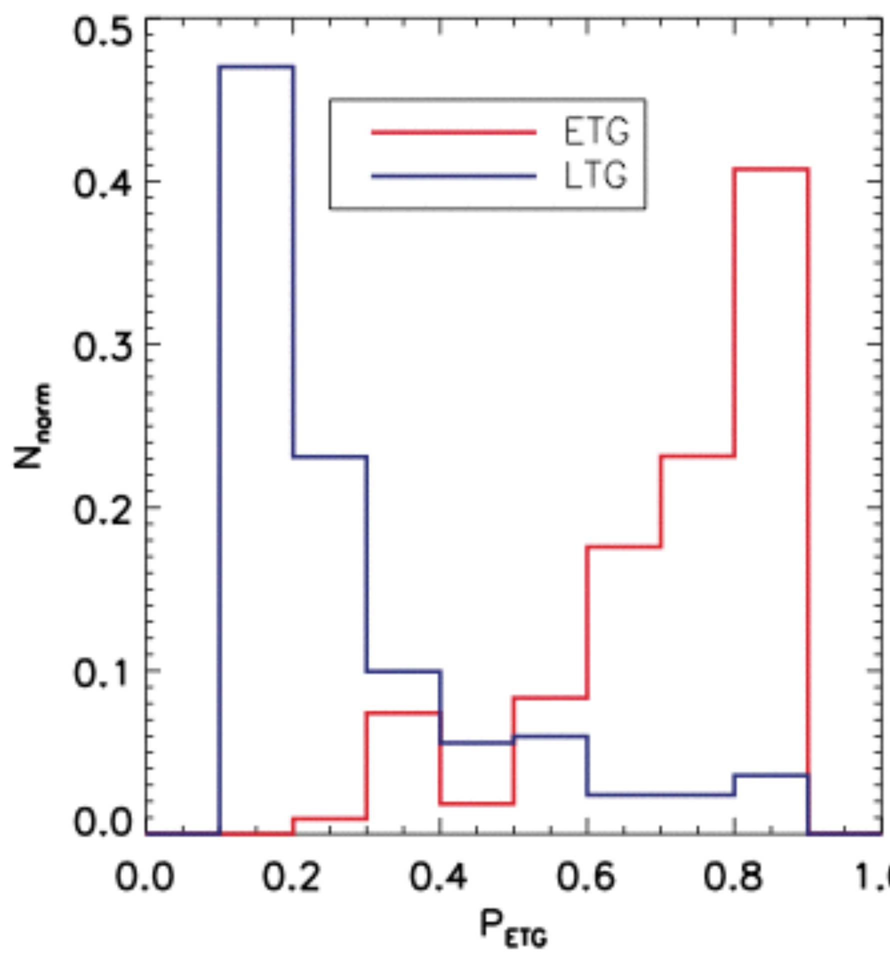


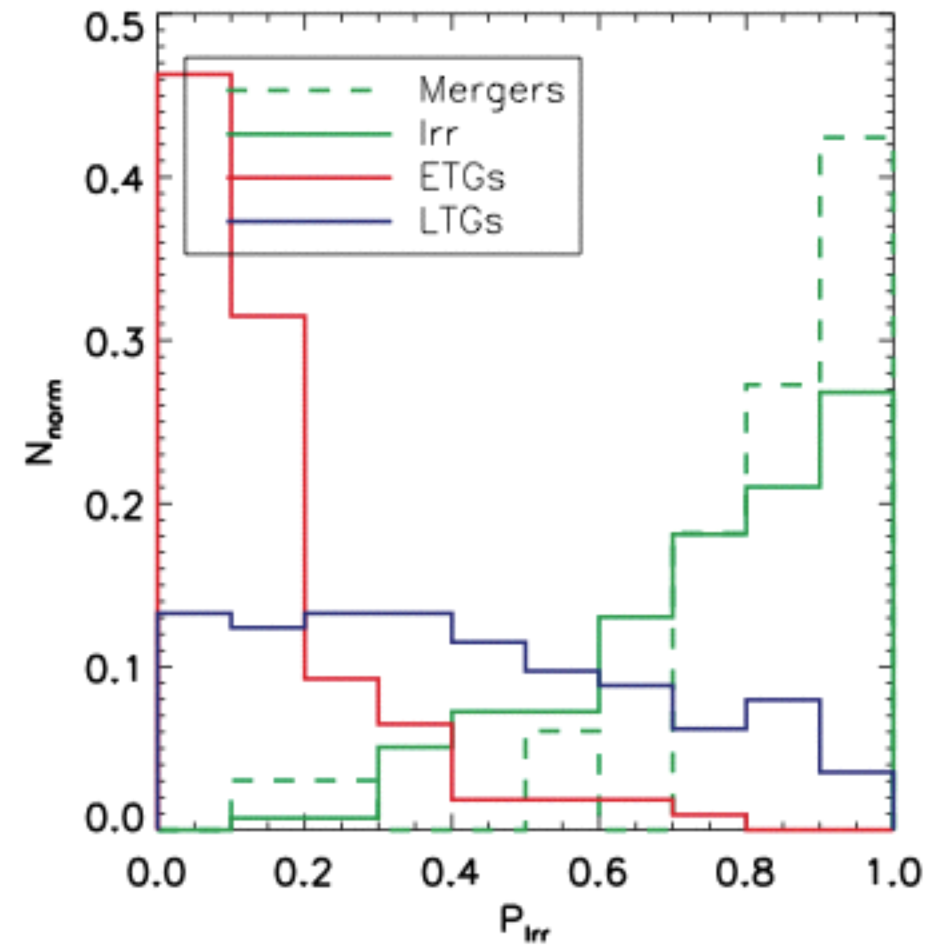
MHC+14a

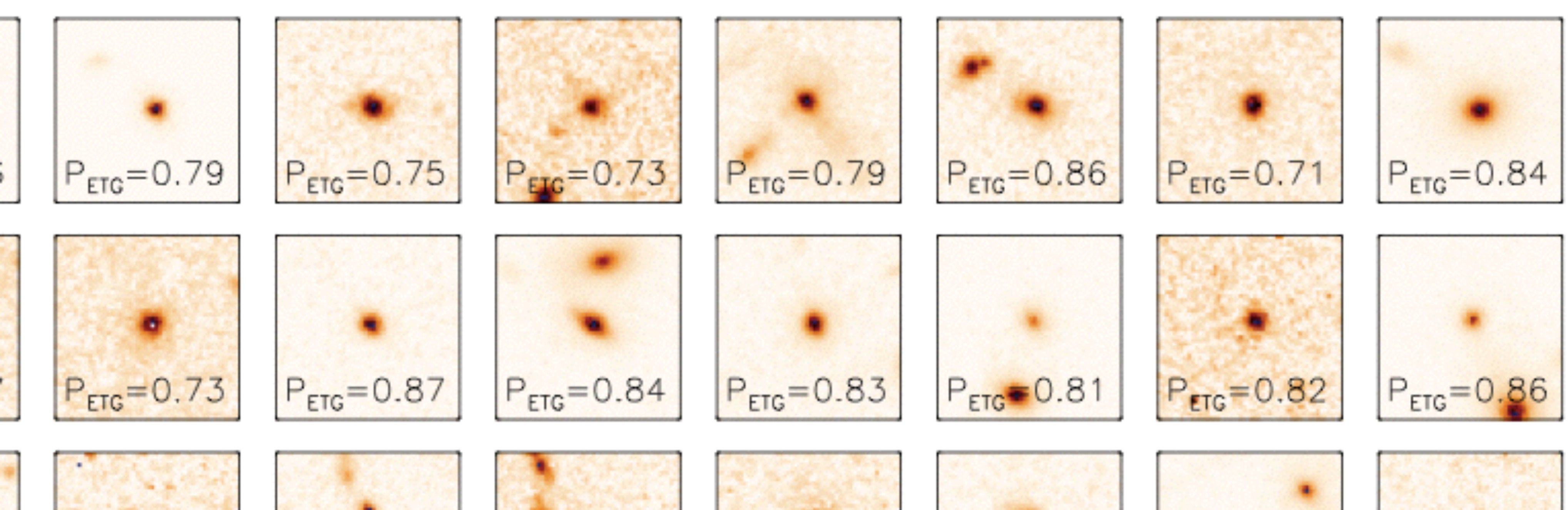
SOME EXAMPLES

z>1

WFC3/ERS







MHC+14a

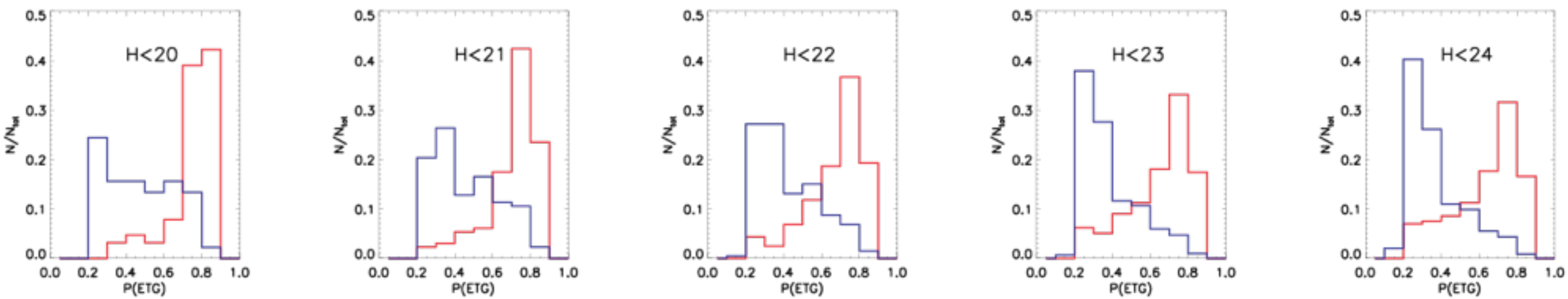
20-30% contamination in a sample of ETGs at z>1

Limitations

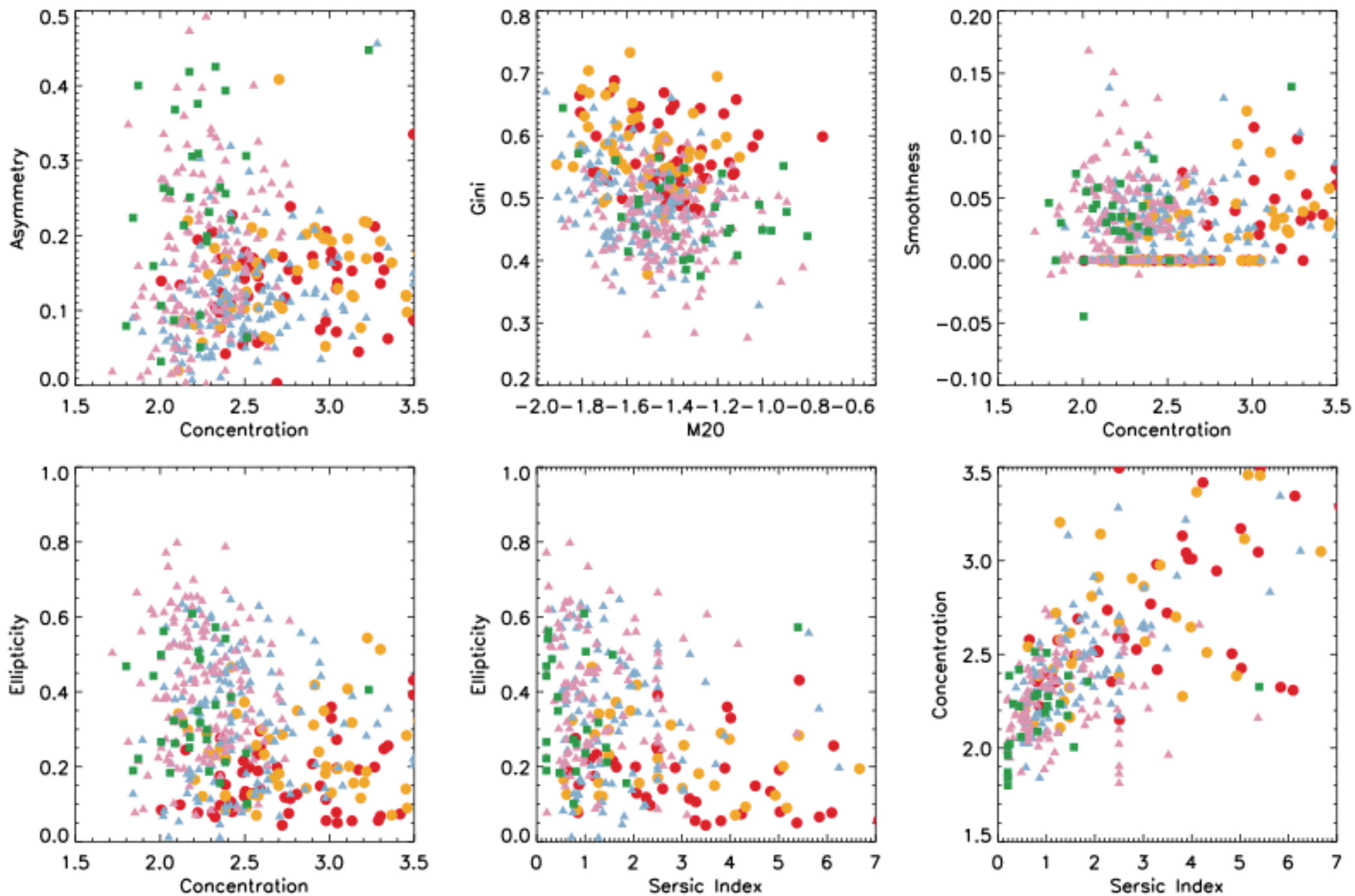
P_{thresh}	$P^{ERS}(ERS)$	$C^{ERS}(ERS)$	$P^{SDSS}(ERS)$	$C^{SDSS}(ERS)$
ETGs				
0.3	53.68	99.03	48.80	89.71
0.4	62.50	92.23	56.61	78.68
0.5	70.45	90.29	66.42	66.91
0.6	78.70	82.52	71.56	57.35
0.7	80.00	66.02	77.11	47.06
0.8	83.02	42.72	85.96	36.03

MHC+14a

EUCLID



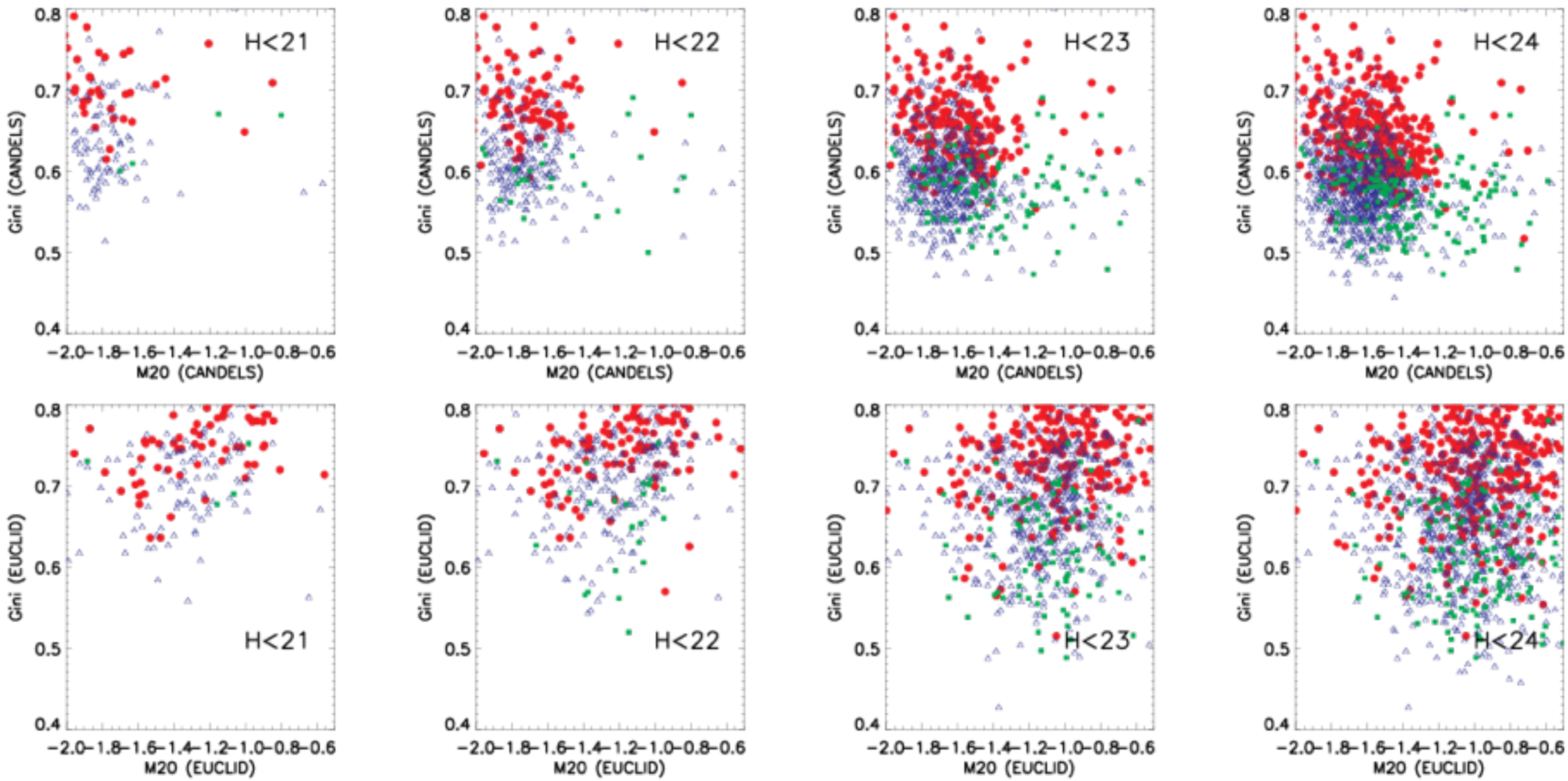
Good choice of features?

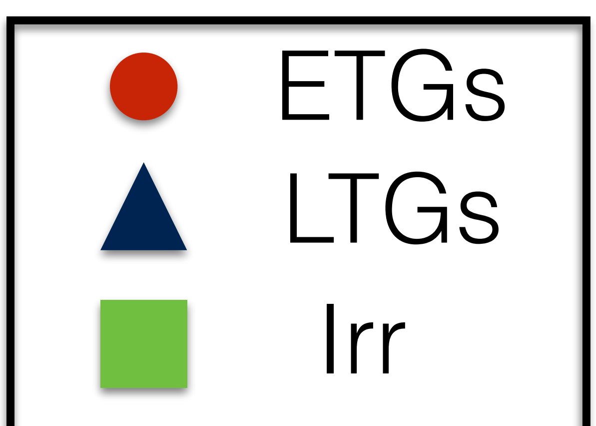


MHC+14a

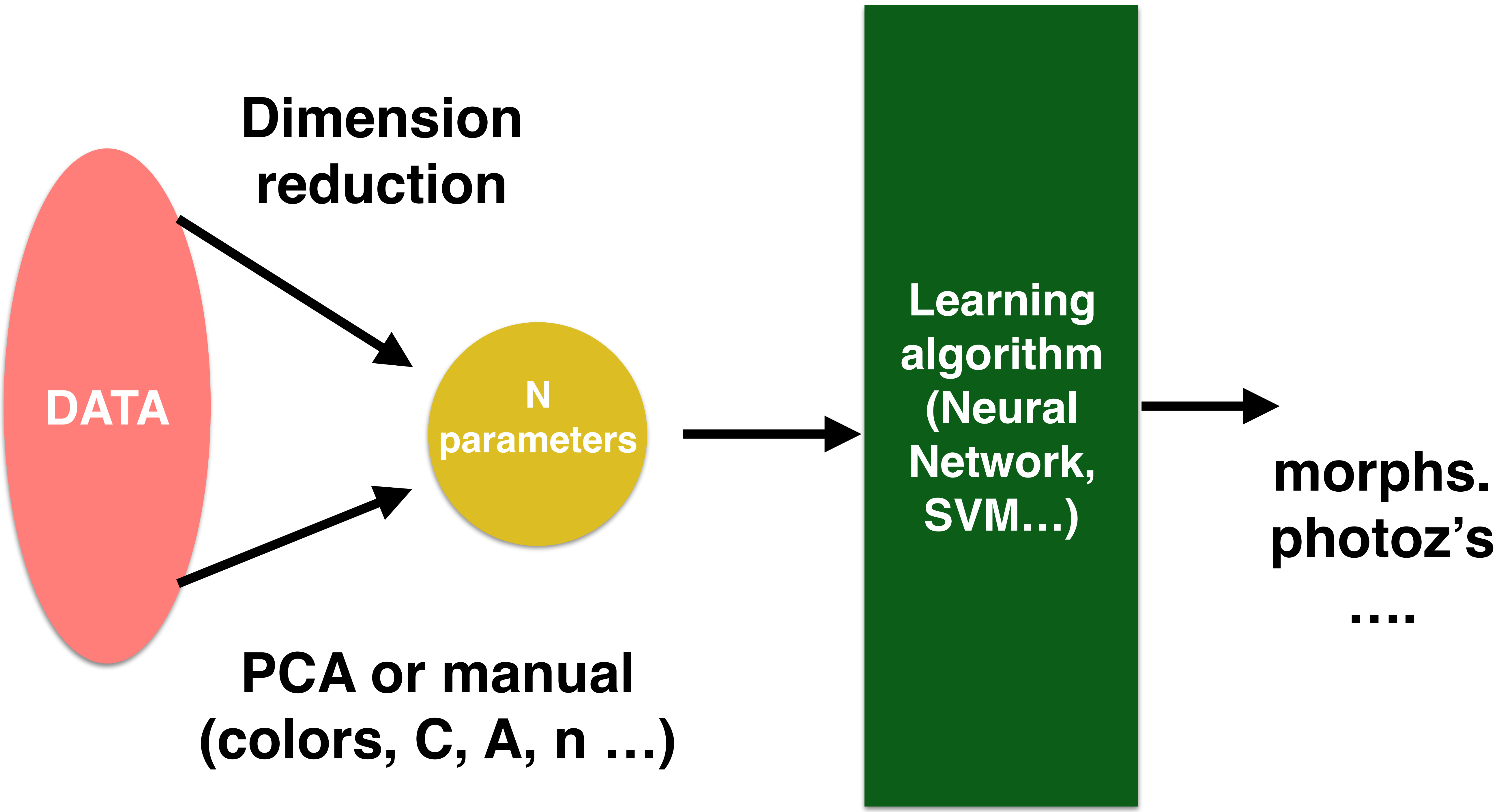
LEARN THE FEATURES!

Gini-M20 plane (EUCLID emulated images)

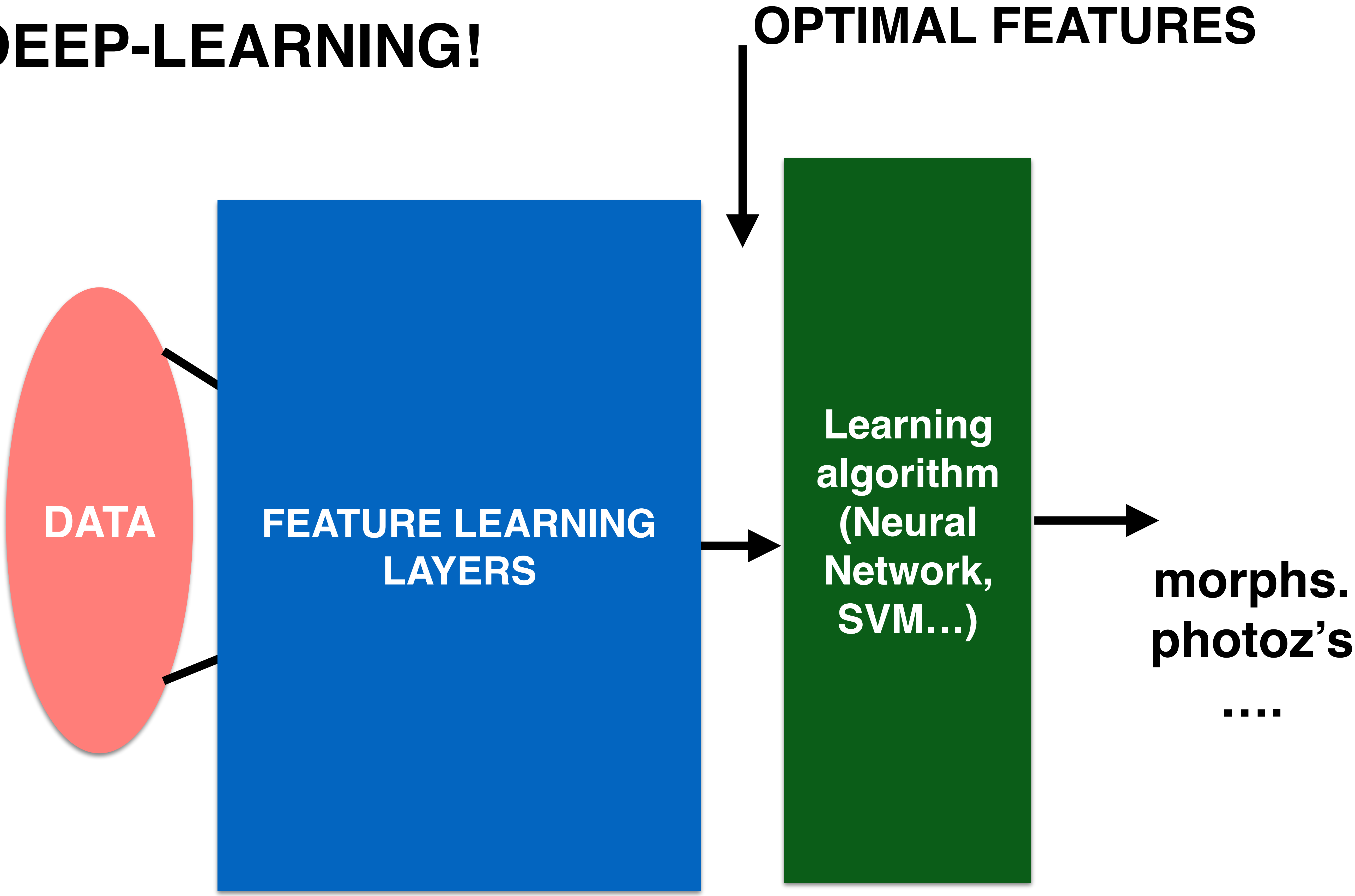




Very noise/resolution dependent…


Dimension
reduction
DATA
N
parameters
Learning
algorithm
(Neural
Network,
SVM…)
morphs.
photoz's
….
PCA or manual
(colors, C, A, n …)

DEEP-LEARNING!



GZOO2 decision tree

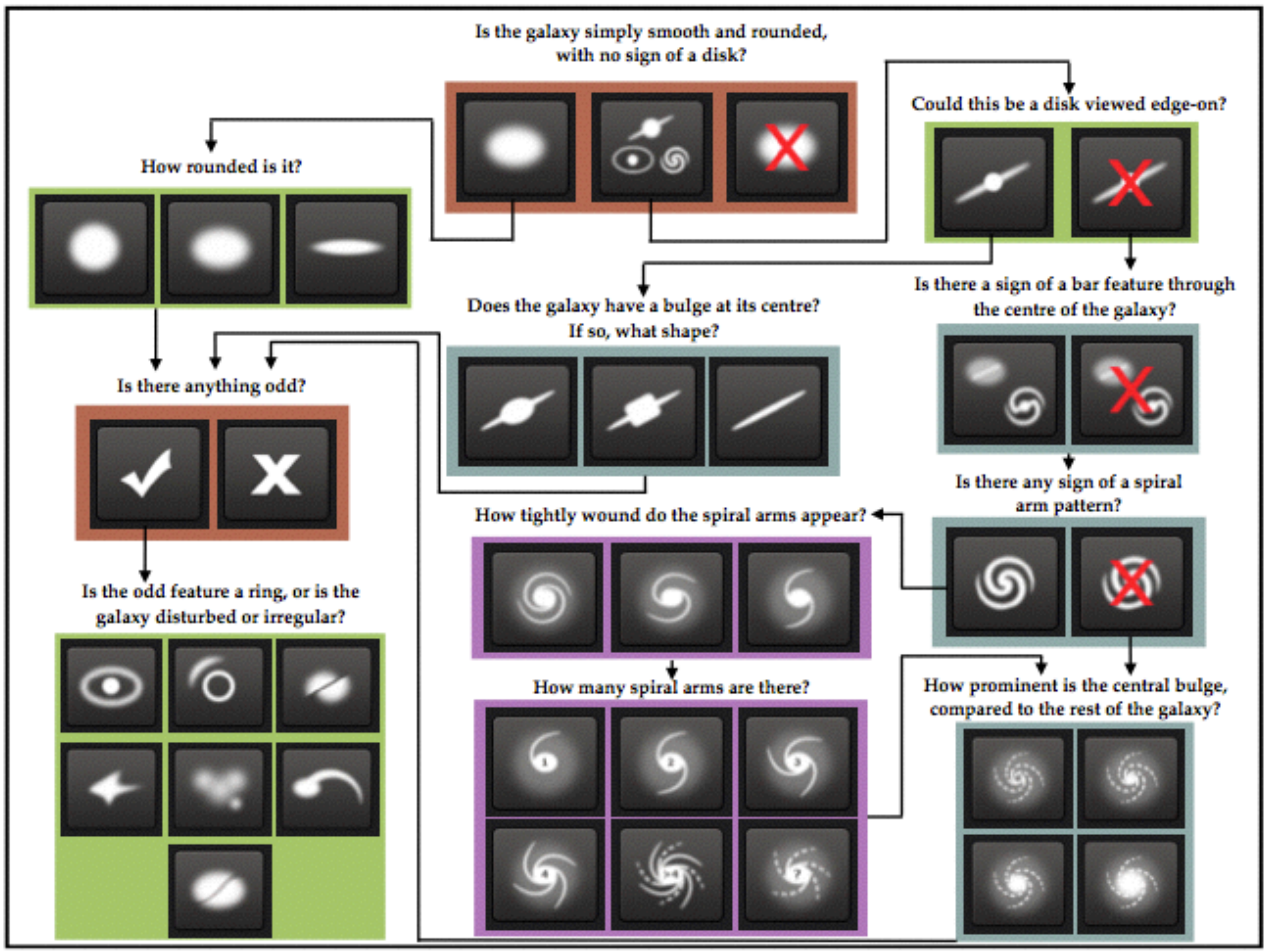


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching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

37 probabilities!

Galaxy Zoo - The Galaxy Challenge

Finished

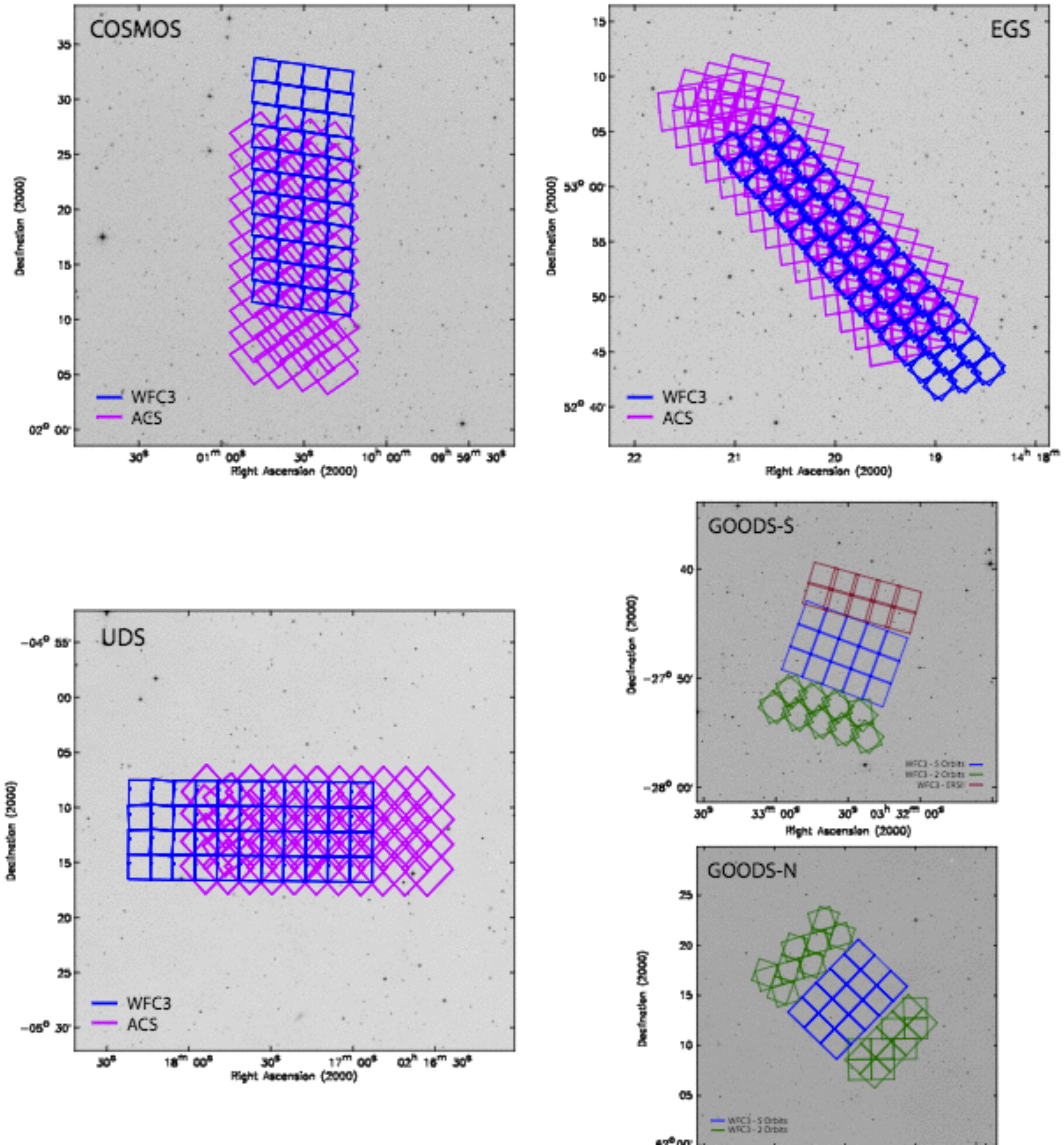
Friday, December 20, 2013 | $16,000 • 329 teams | Friday, April 4, 2014

Competition Details » Get the Data » Make a submission

Classify the morphologies of distant galaxies in our Universe

Understanding how and why we are here is one of the fundamental questions for the human race. Part of the answer to this question lies in the origins of galaxies, such as our own Milky Way. Yet questions remain about how the Milky Way (or any of the other ~100 billion galaxies in our Universe) was formed and has evolved. Galaxies come in all shapes, sizes and colors: from beautiful spirals to huge ellipticals. Understanding the distribution, location and types of galaxies as a function of shape, size, and color are critical pieces for solving this puzzle.

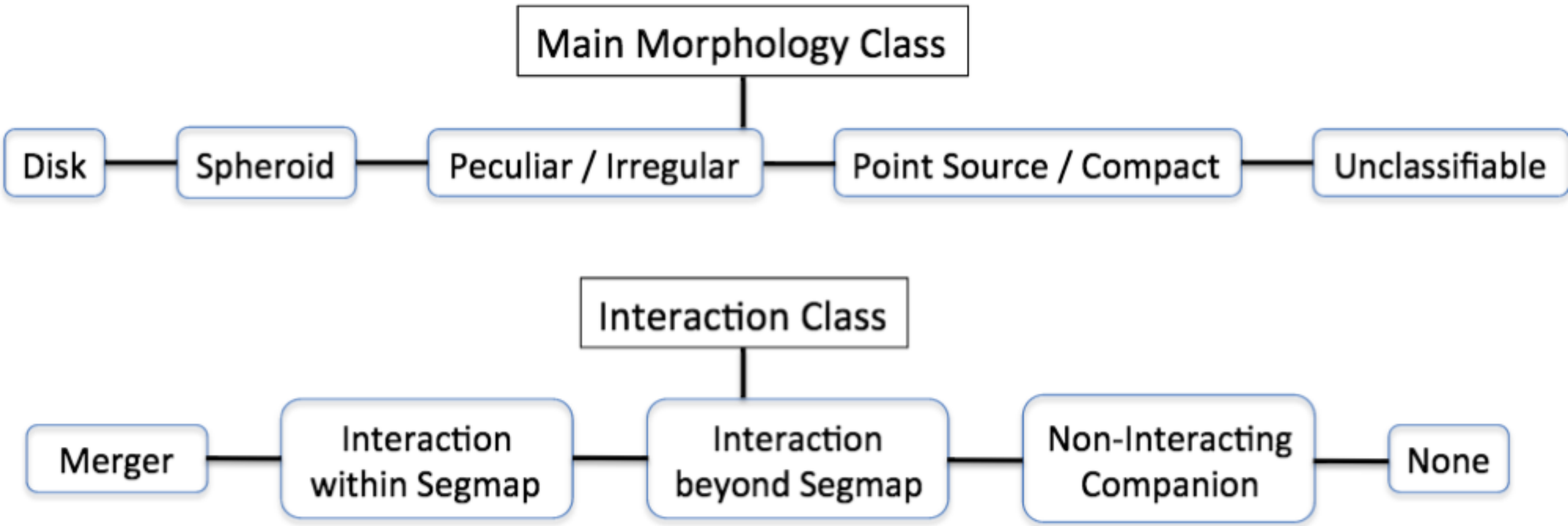
CANDELS survey



HST NIR survey of 4 cosmological fields

Rest-frame optical high-resolution moro morphologies at 1<z<3

CANDELS as a test case for EUCLID



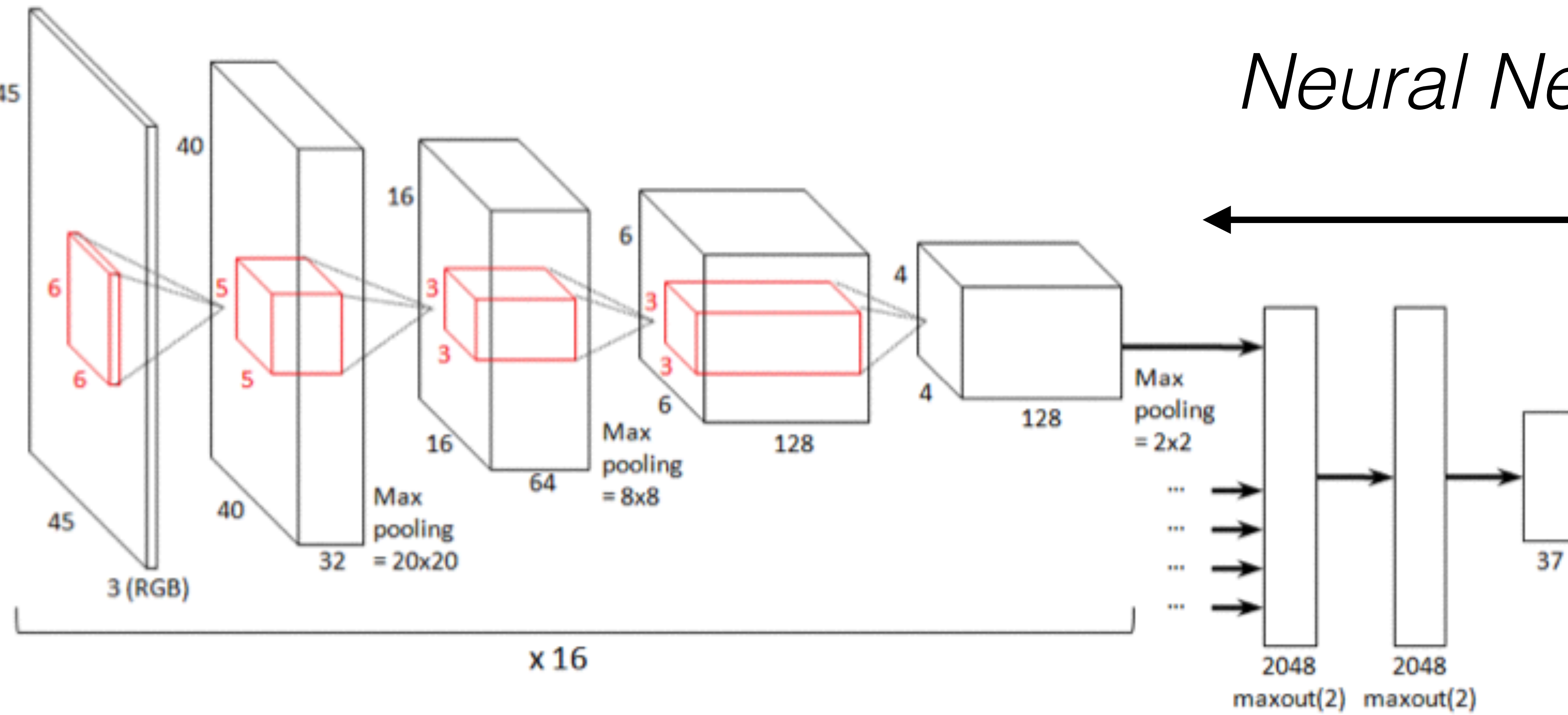
Kartaltepe+14

— Classification of galaxies in GOODS-S with **H<24.5**
- — Each galaxy is classified by 3-5 experts
  - — Fractions for ~8000 galaxy in GDS
- — Classification done in F160 (+F125,F105)

CONVNET for SDSS (60.000 galaxies)

Feature learning

Neural Network

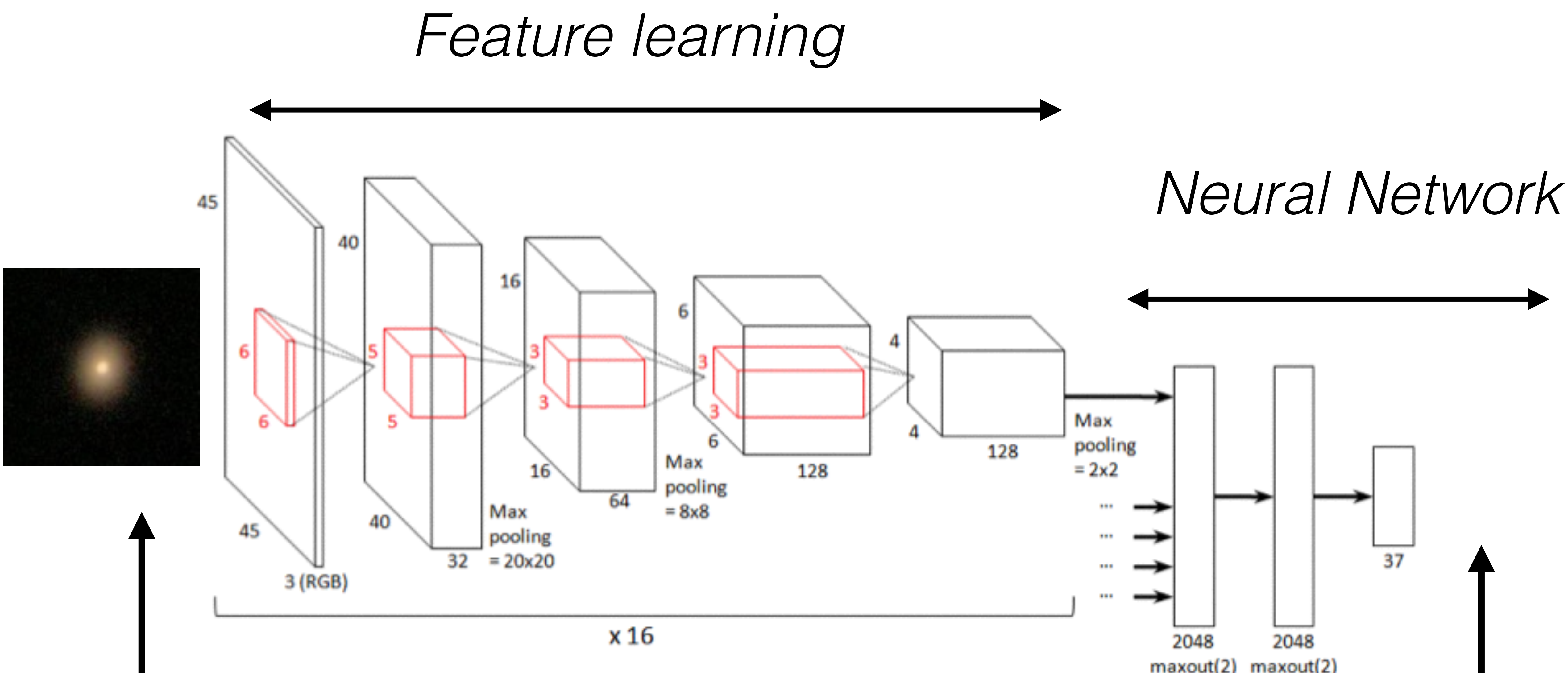


Dielman+14

***INPUT**: RGB JPEG SDSS snapshots*

***OUTPUT**: 37 GZOO probs.*

CONVNET for SDSS (60.000 galaxies)

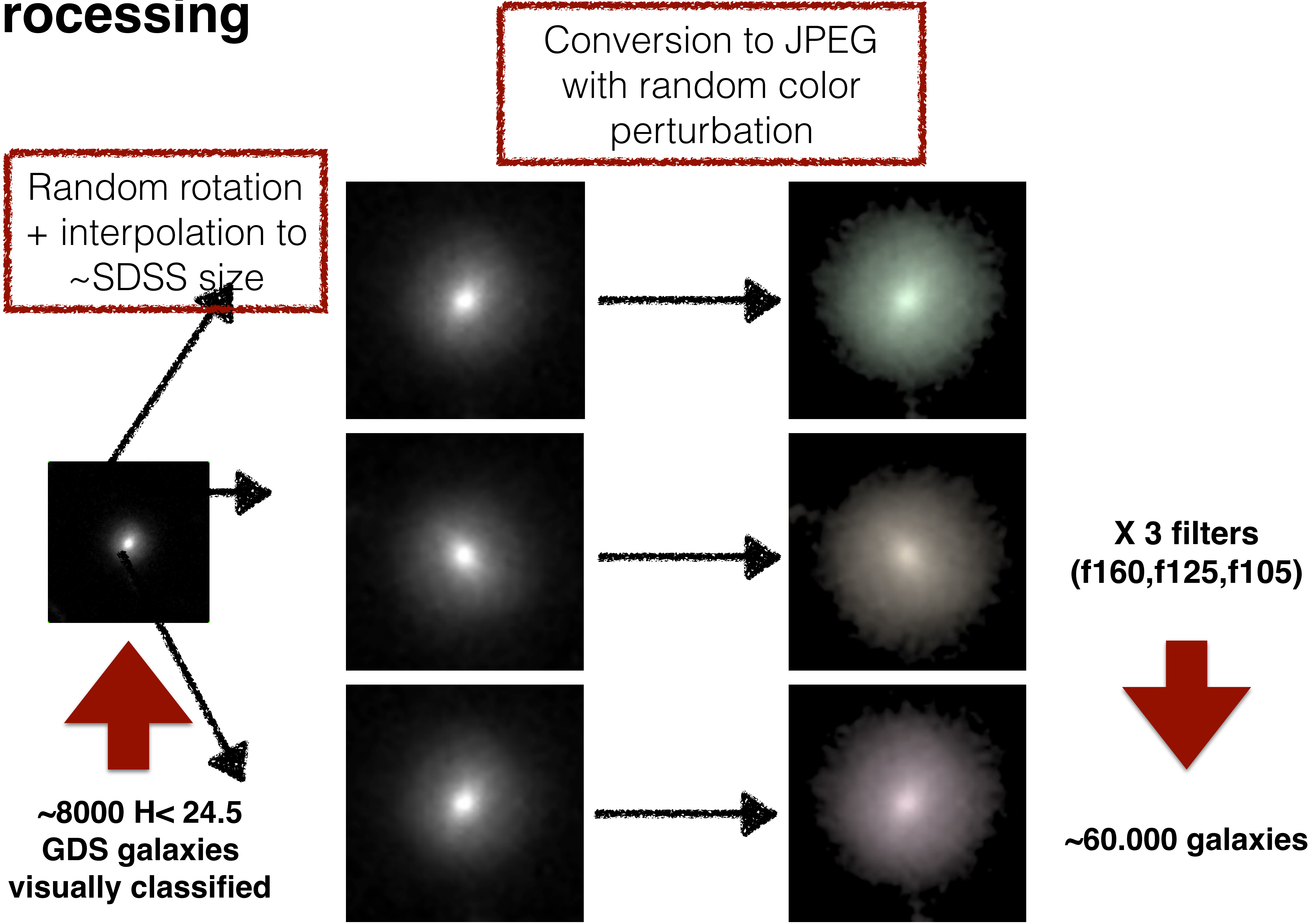


Dielman+14

INPUT: *RGB JPEG SDSS snapshots*

OUTPUT: *37 GZOO probs.*

rocessing

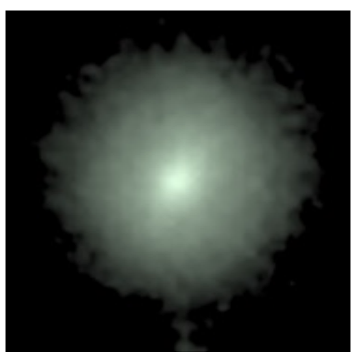

Conversion to JPEG with random color perturbation
Random rotation + interpolation to ~SDSS size
X 3 filters (f160,f125,f105)
~8000 H< 24.5 GDS galaxies visually classified
~60.000 galaxies

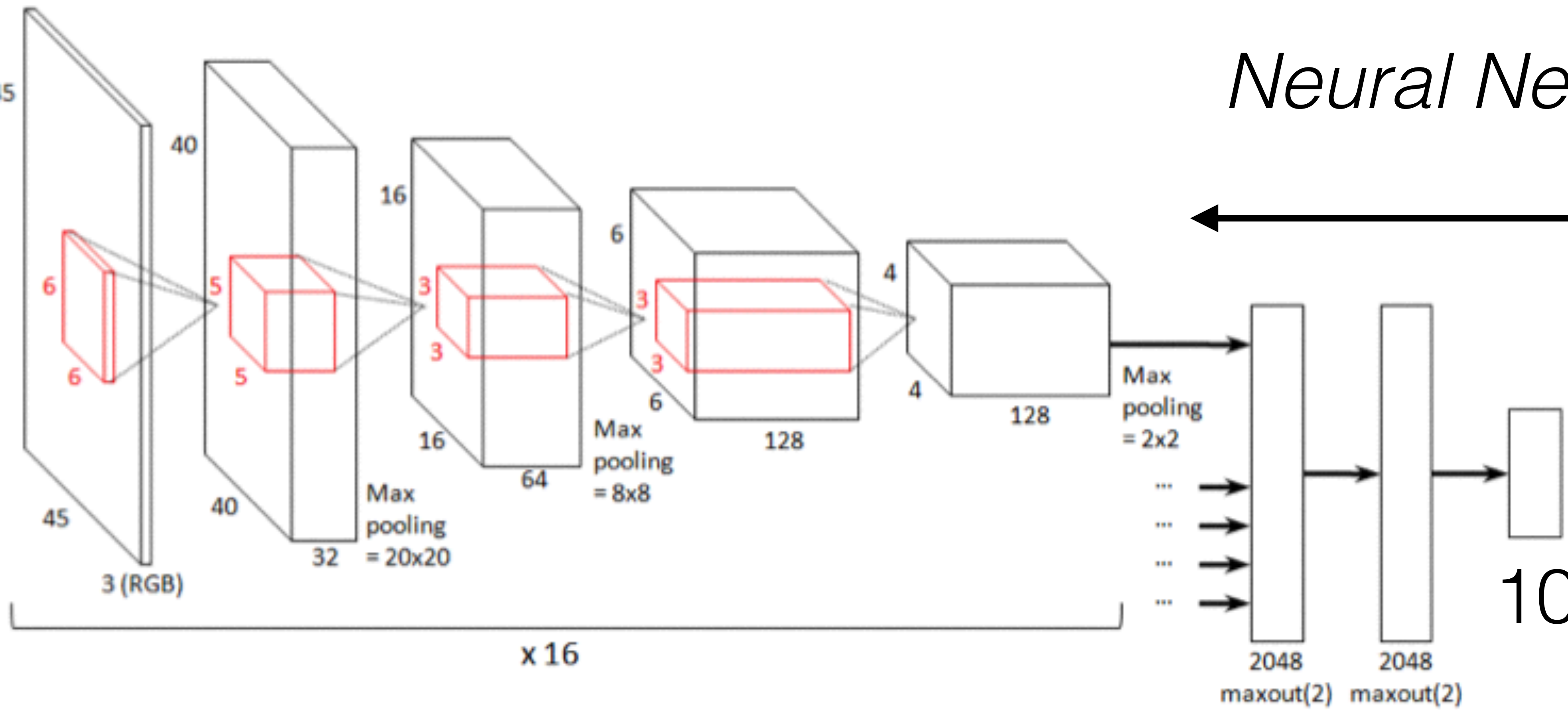
CONVNET for CANDELS

- **TRAIN**: ~50.000 redundant galaxies in GDS (~10 days)
- **CLASSIFY**: GDN, COSMOS, UDS, GDS (~8h/field)

Feature learning

Neural Network





***INPUT**: RGB JPEG GDS snapshots*

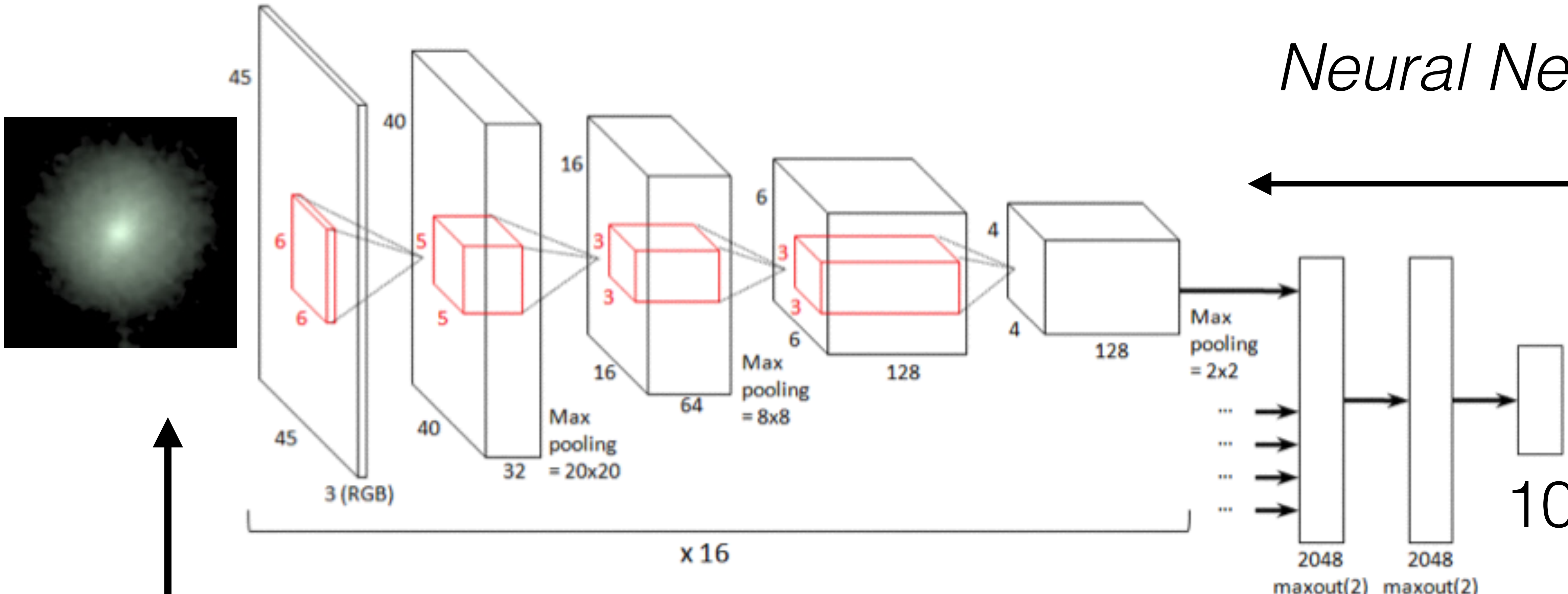
***OUTPUT**: 10 probs.*

CONVNET for CANDELS

- **TRAIN**: ~50.000 redundant galaxies in GDS (~10 days)
- **CLASSIFY**: GDN, COSMOS, UDS, GDS (~8h/field)

Feature learning

Neural Network

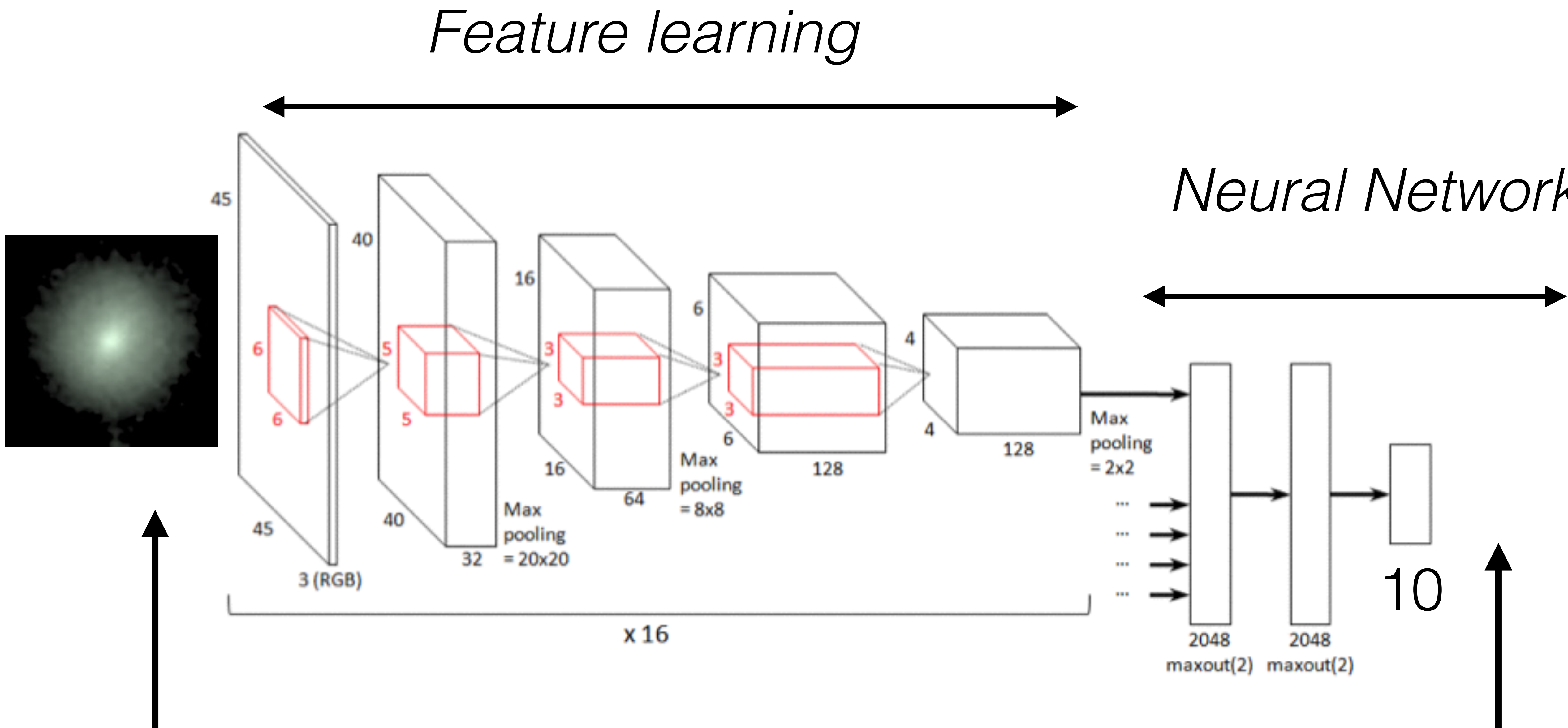


***INPUT**: RGB JPEG GDS snapshots*

***OUTPUT**: 10 probs.*

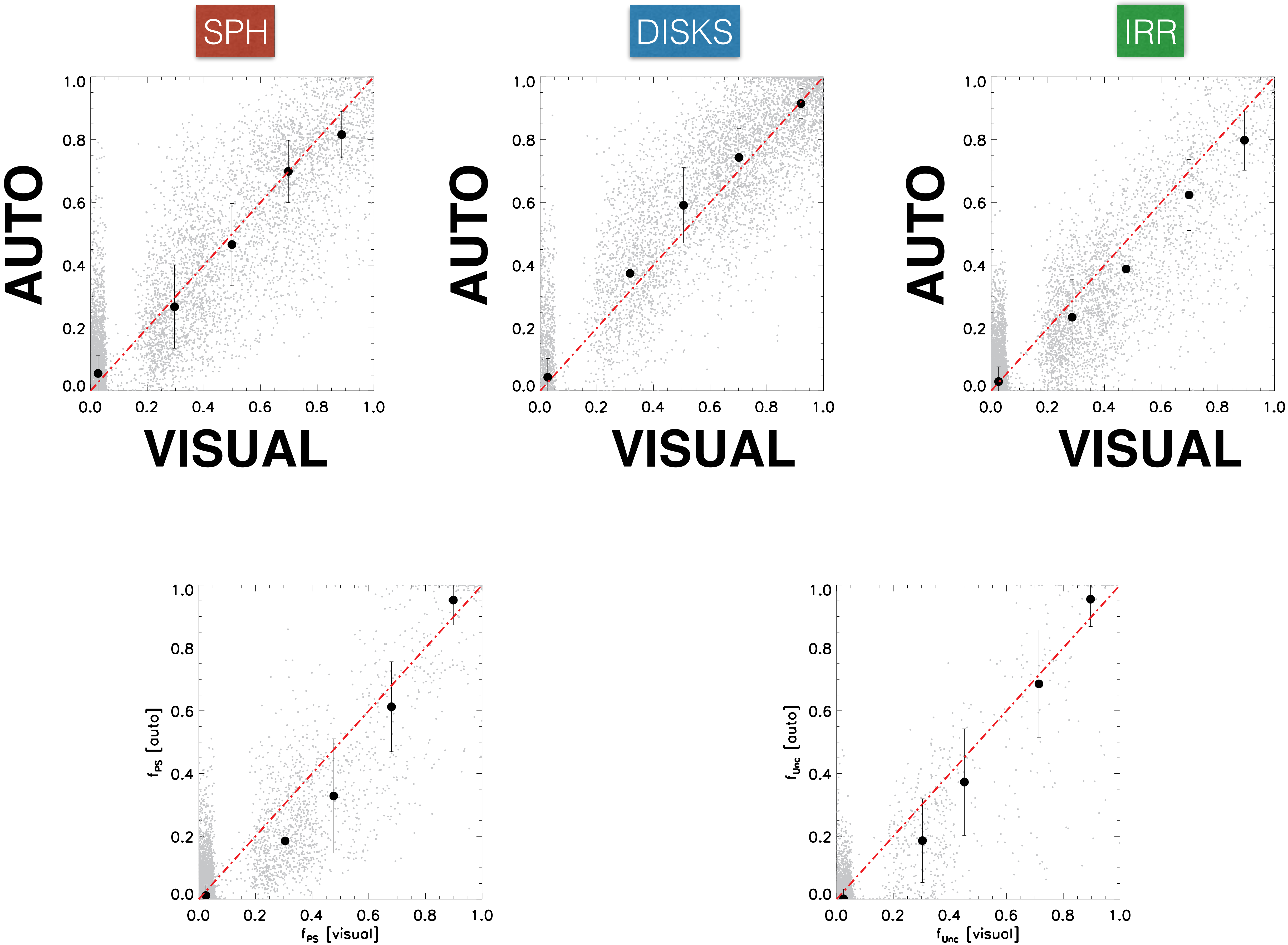
CONVNET for CANDELS

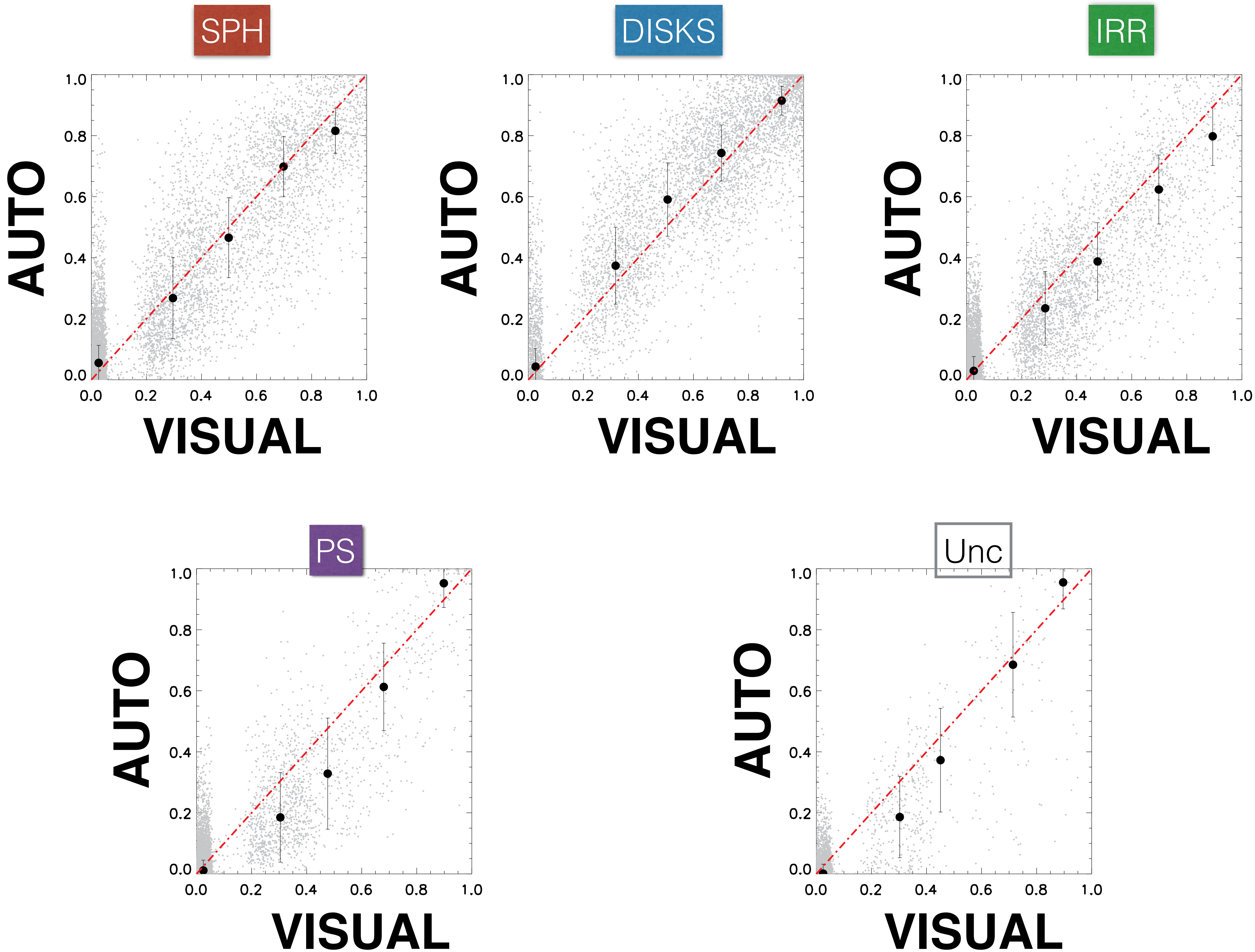
- **TRAIN**: ~50.000 redundant galaxies in GDS (~10 days)
- **CLASSIFY**: GDN, COSMOS, UDS, GDS (~8h/field)



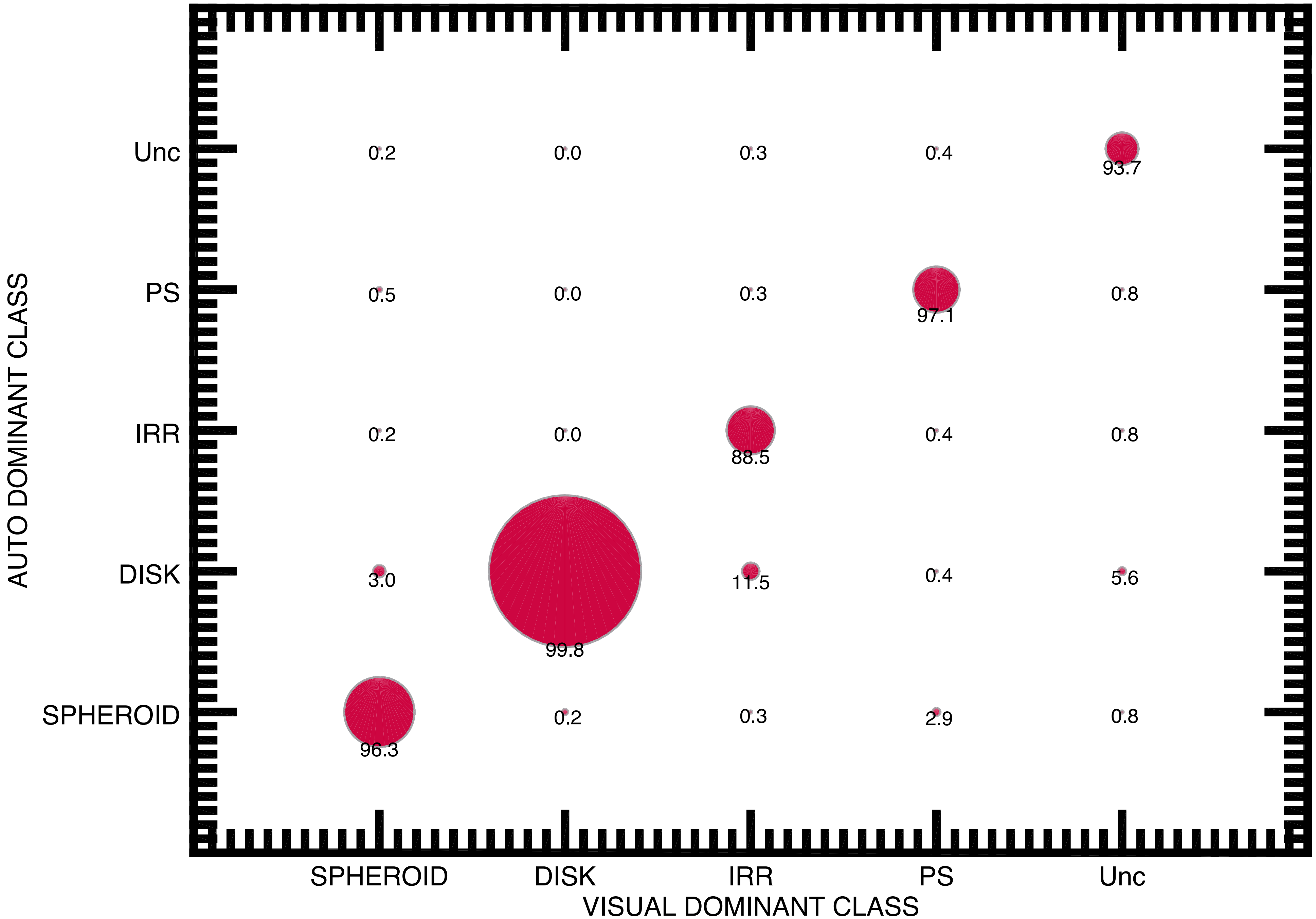
***INPUT**: RGB JPEG GDS snapshots*

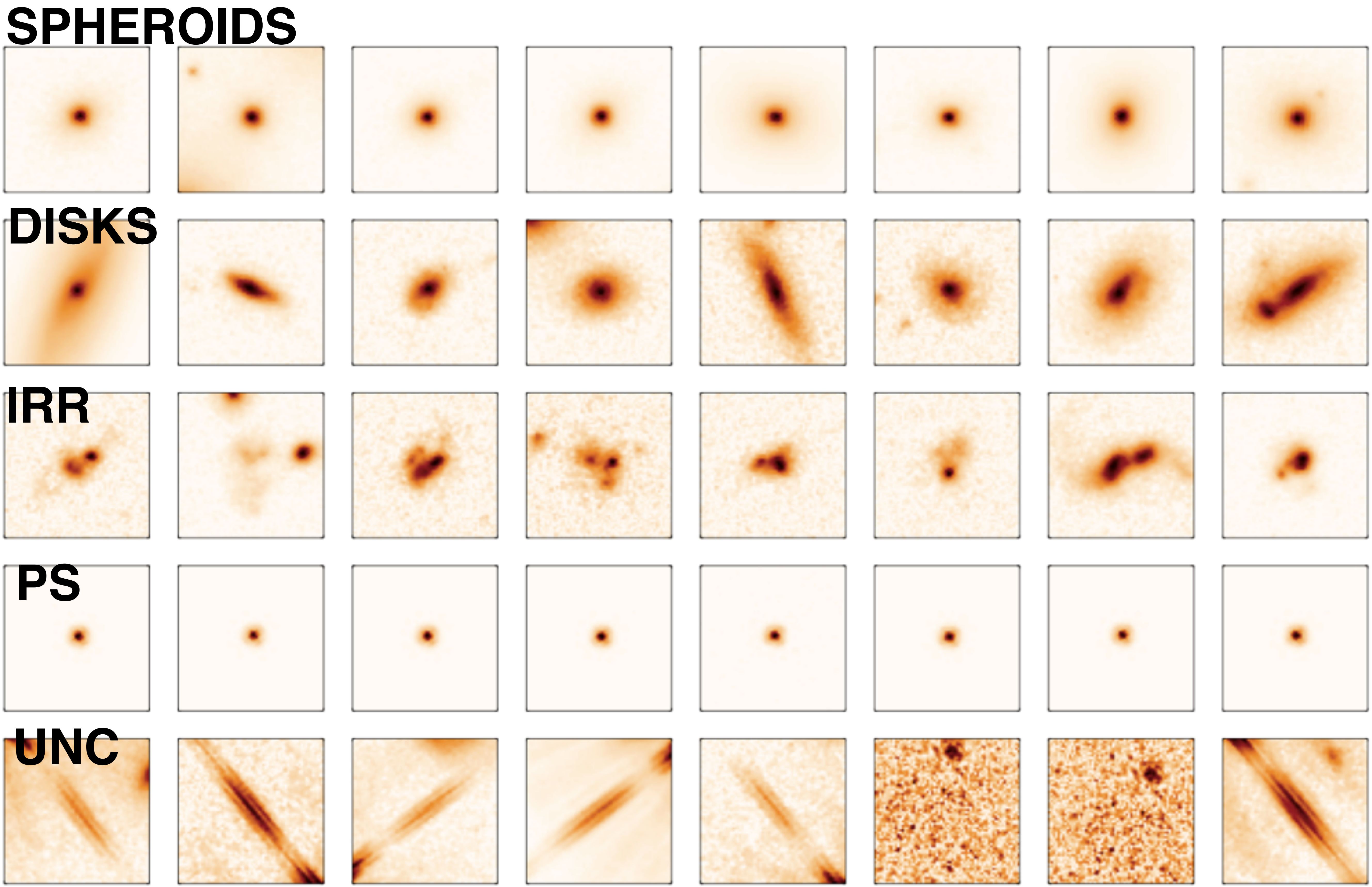
***OUTPUT**: 10 probs.*


SPH
DISKS
IRR
AUTO
VISUAL
0.0
0.2
0.4
0.6
0.8
1.0
f_{PS} [auto]
f_{PS} [visual]
f_{Unc} [auto]
f_{Unc} [visual]


SPH
DISKS
IRR
PS
Unc
AUTO
VISUAL
0.0
0.2
0.4
0.6
0.8
1.0

DOMINANT CLASS

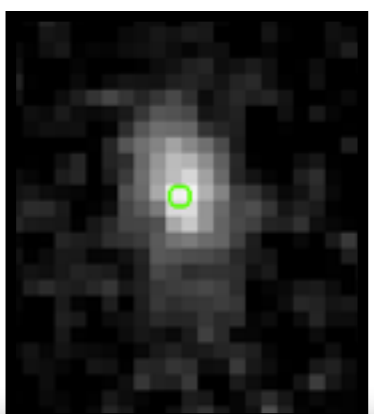



SPHEROIDS
DISKS
IRR
PS
UNC

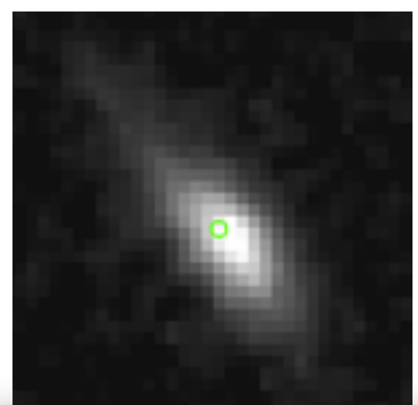
Catastrophic “errors”

fsph_v > 0.7 and fsph_a < 0.3 or fsph_v < 0.3 and fsph_a < 0.7

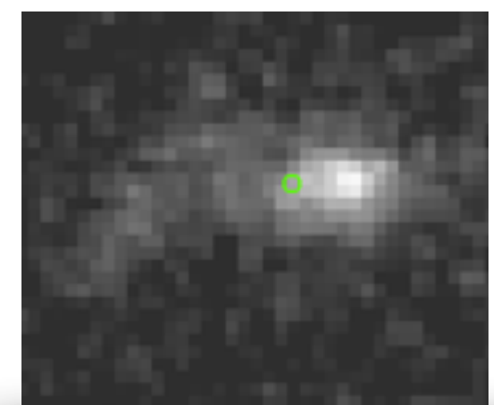
~15/8000=0.2%



fsph=0.82 / 0.25
fdisk = 0.5 / 0.76
firr = 0.0 / 0.22



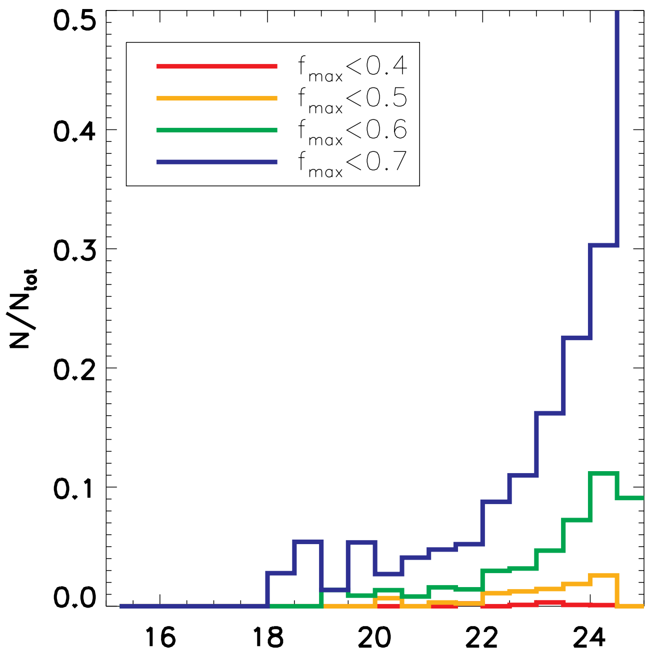
fsph=0.8 / 0.25
fdisk = 0.75 / 0.95
firr = 0.0 / 0.0



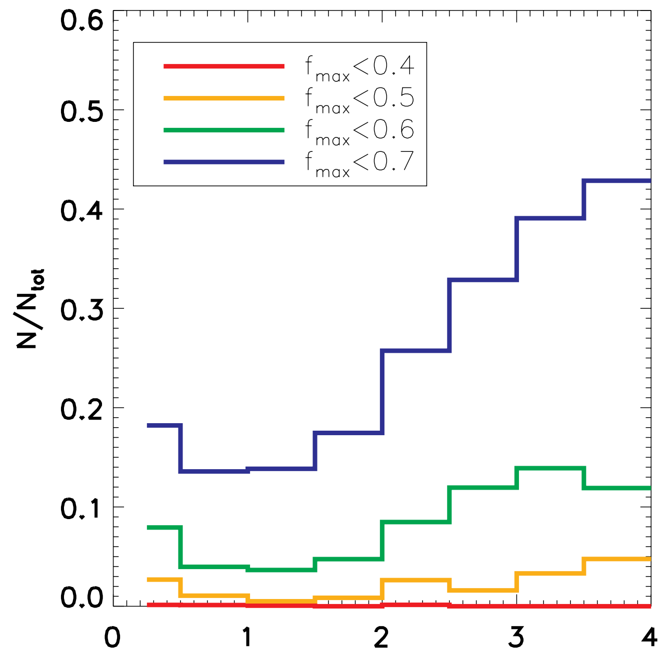
fsph=0.76 / 0.11
fdisk = 0.6 / 0.66
firr = 0.39 / 0.53

VISUAL / AUTO

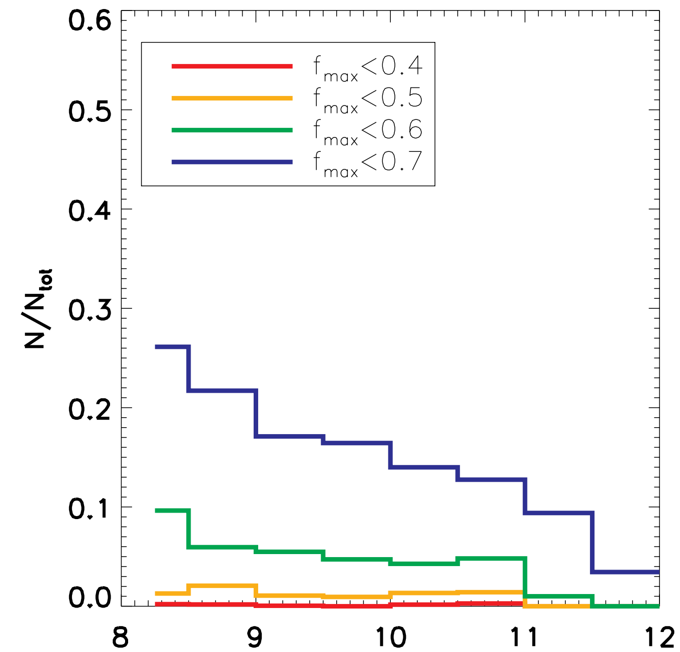
“Uncertain” objects



H mag

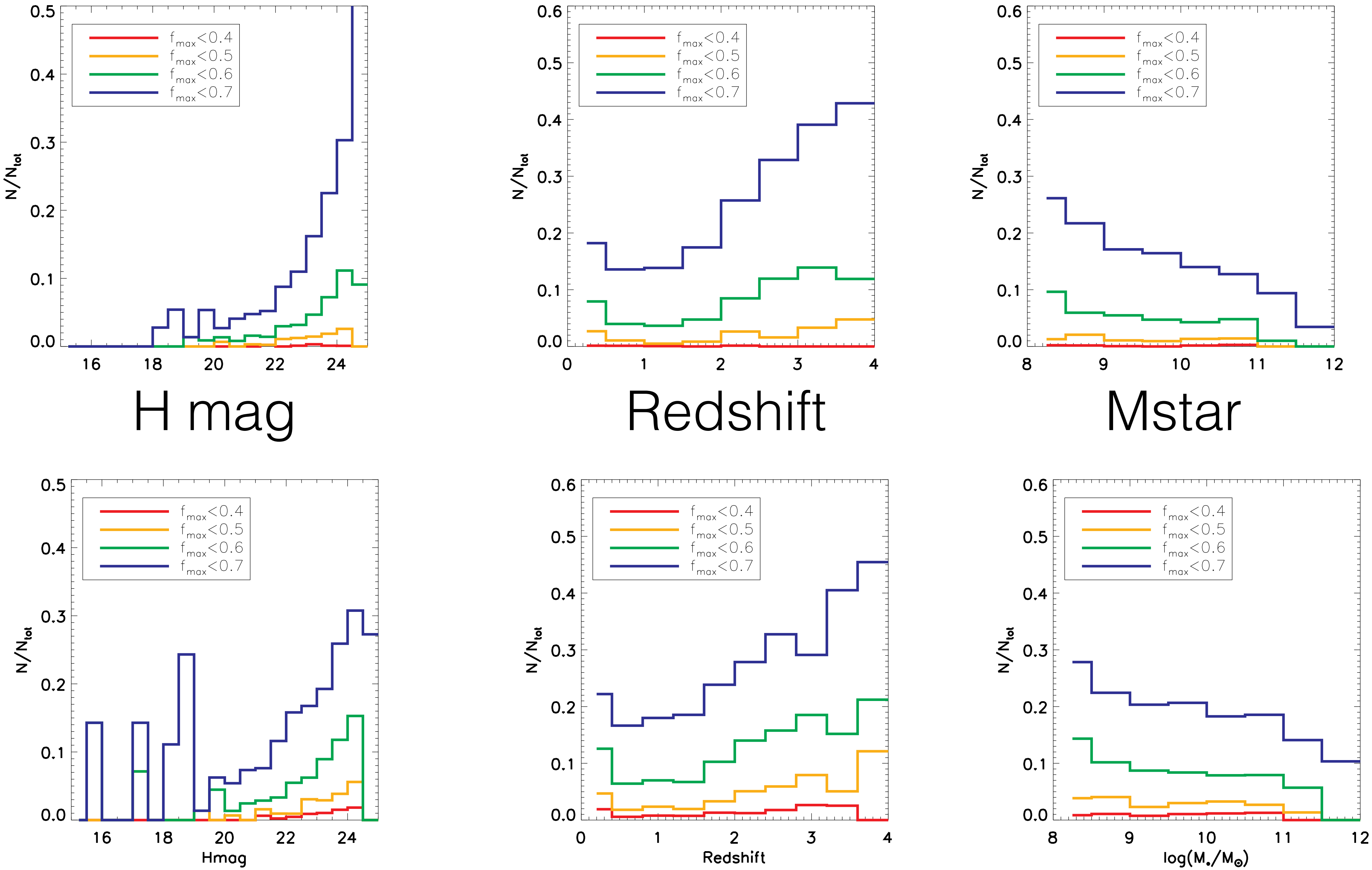


Redshift



Mstar

“Uncertain” objects

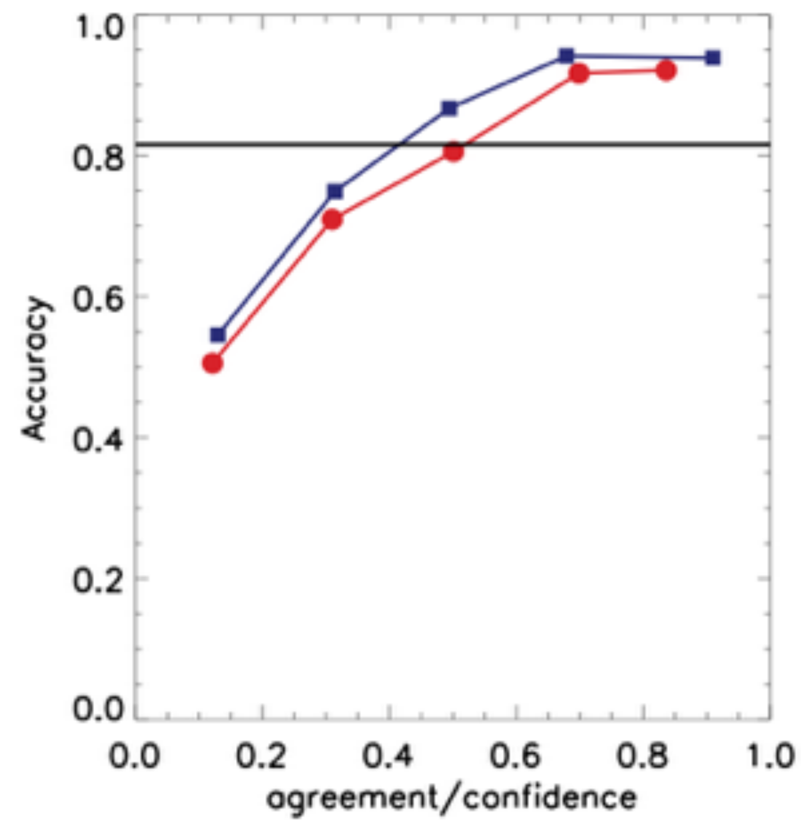


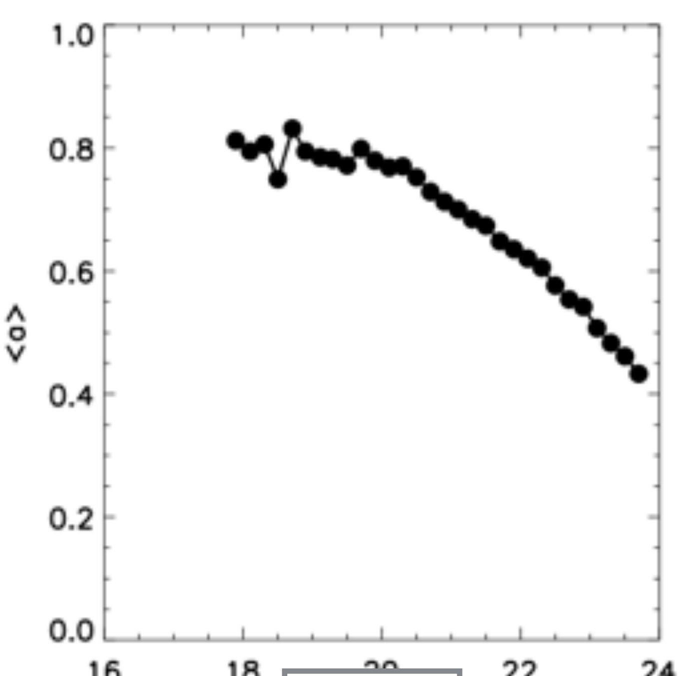
H mag

Redshift

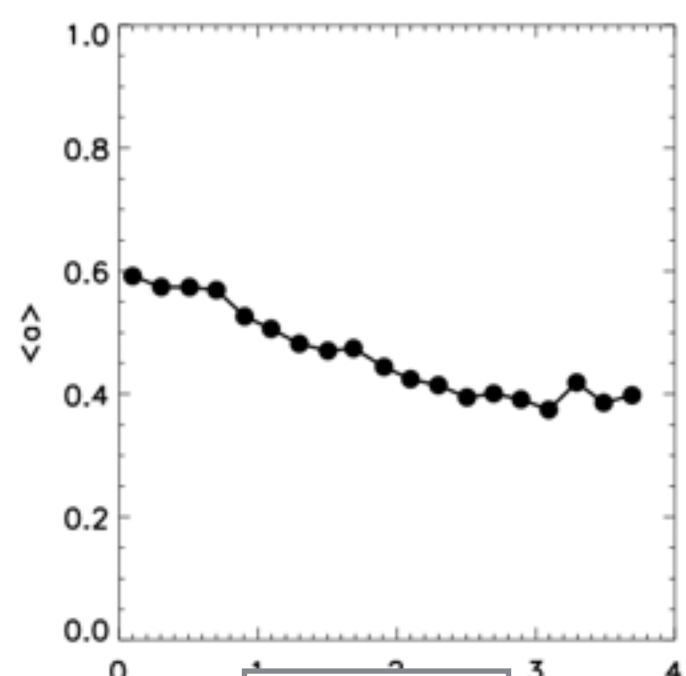
Mstar

Entropy - agreement between classifiers

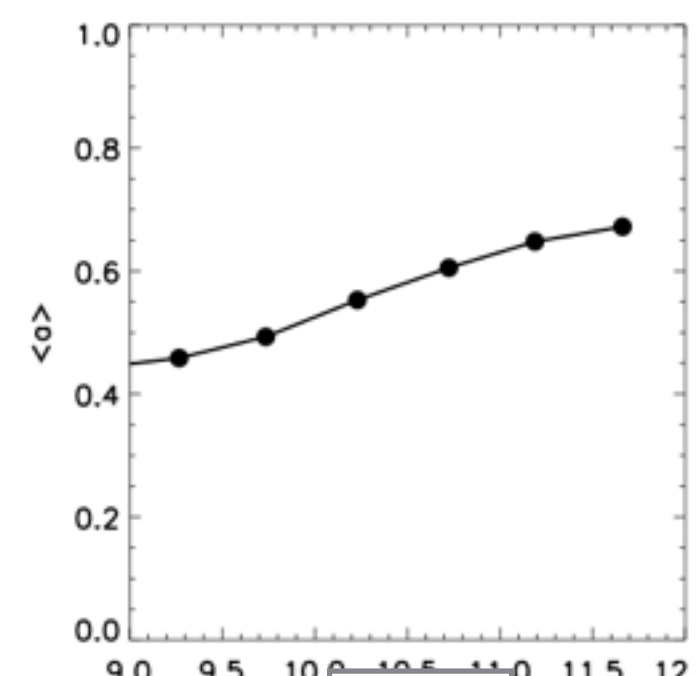




MAG



Redshift



Mass

Future work

- Extend to other, more detailed, morphological features
- Generalize
  - Main limitation, training set size!
  - numerical simulations as input (Illustris, Horizon etc…)
  - Self-learning??