De-biasing "Gold Standard" Classification for Galaxy Morphologies

Guillermo Cabrera -- Astroinformatics Laboratory, Center for Mathematical Modeling, University of Chile

Christopher Miller -- Department of Astronomy and Department of Physics, University of Michigan

Jeff Schneider -- The Robotics Institute/ School of Computer Science, Carnegie Mellon University

Cabrera, Miller, Schneider, 2014 ApJS submitted

Cabrera, Miller, Schneider, 2014 ICPR in press

De-biasing "Gold Standard" Classification for Galaxy Morphologies

Guillermo Cabrera -- Astroinformatics Laboratory, Center for Mathematical Modeling, University of Chile

Christopher Miller -- Department of Astronomy and Department of Physics, University of Michigan

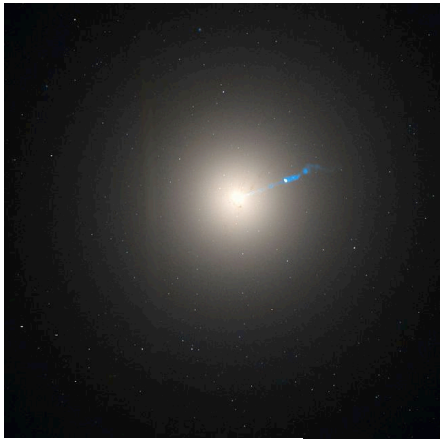
Jeff Schneider -- The Robotics Institute/ School of Computer Science, Carnegie Mellon University

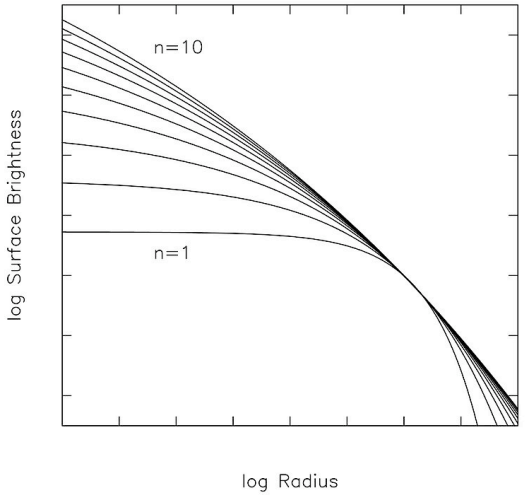
Cabrera, Miller, Schneider, 2014 ApJS submitted

Cabrera, Miller, Schneider, 2014 ICPR in press

Galaxy Morphologies from Light Profiles

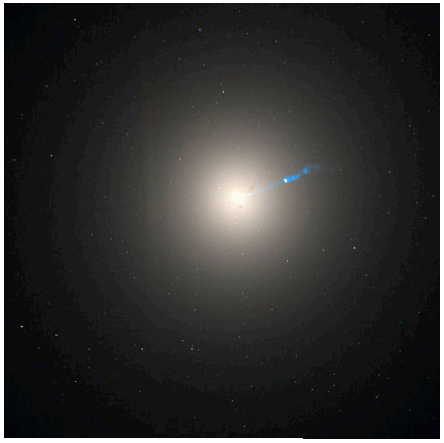
- Can we use the light profile alone to classify the morphologies of galaxies into two simple classes: spiral disks and ellipticals? Why?
  - The light profile is a common characterization (unlike 2D)
  - The shape of the profile is meaningful in terms of internal galaxy properties
  - Galaxy properties /morphologies are correlated.
  - High-level understanding of galaxy formation and evolution
- When using supervised machine learning: can we trust the "gold standard" morphology catalogs?

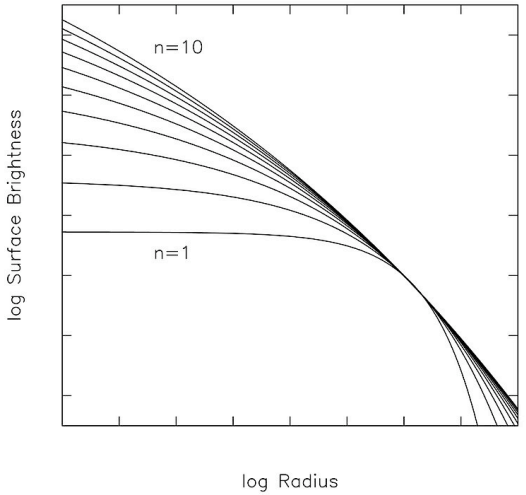




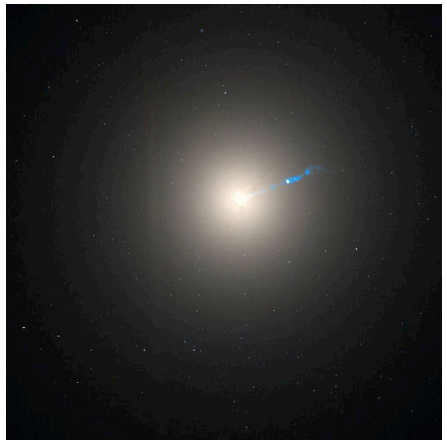
Galaxy Morphologies from Light Profiles

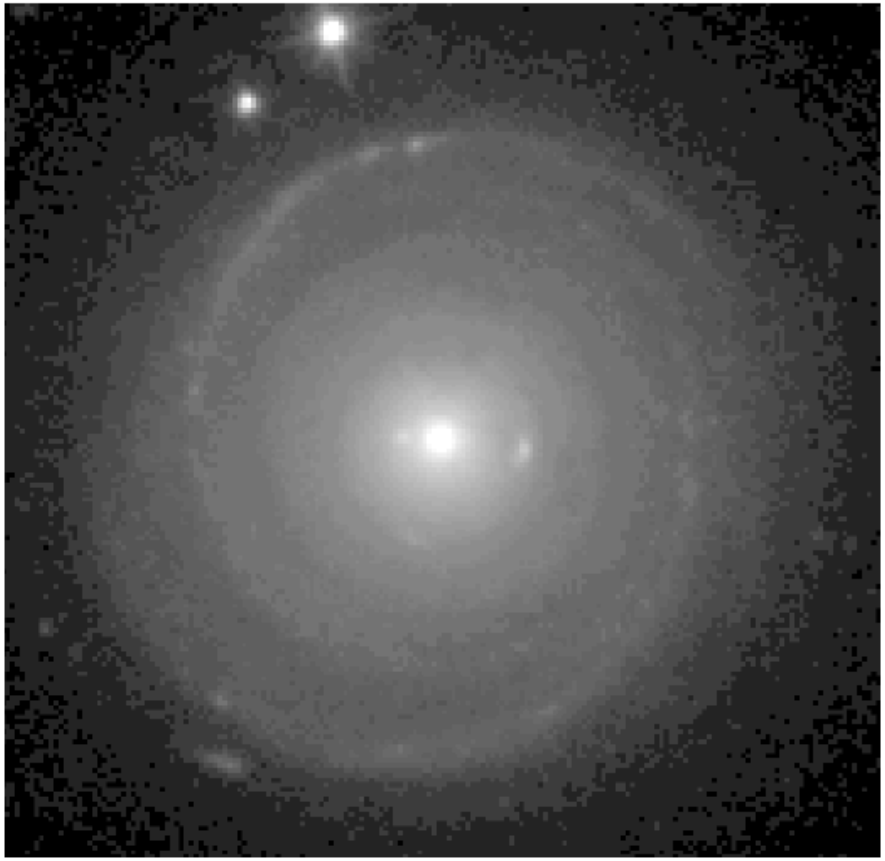
- ☑ Can we use the light profile alone to classify the morphologies of galaxies into two simple classes: spirals and ellipticals? Why?
  - The light profile is a common characterization (unlike 2D)
  - The shape of the profile is meaningful in terms of internal galaxy properties
  - Galaxy properties /morphologies are correlated.
  - High-level Understanding of Galaxy Formation and Evolution
- ☒ When using supervised machine learning: can we trust the "gold standard" morphology catalogs?

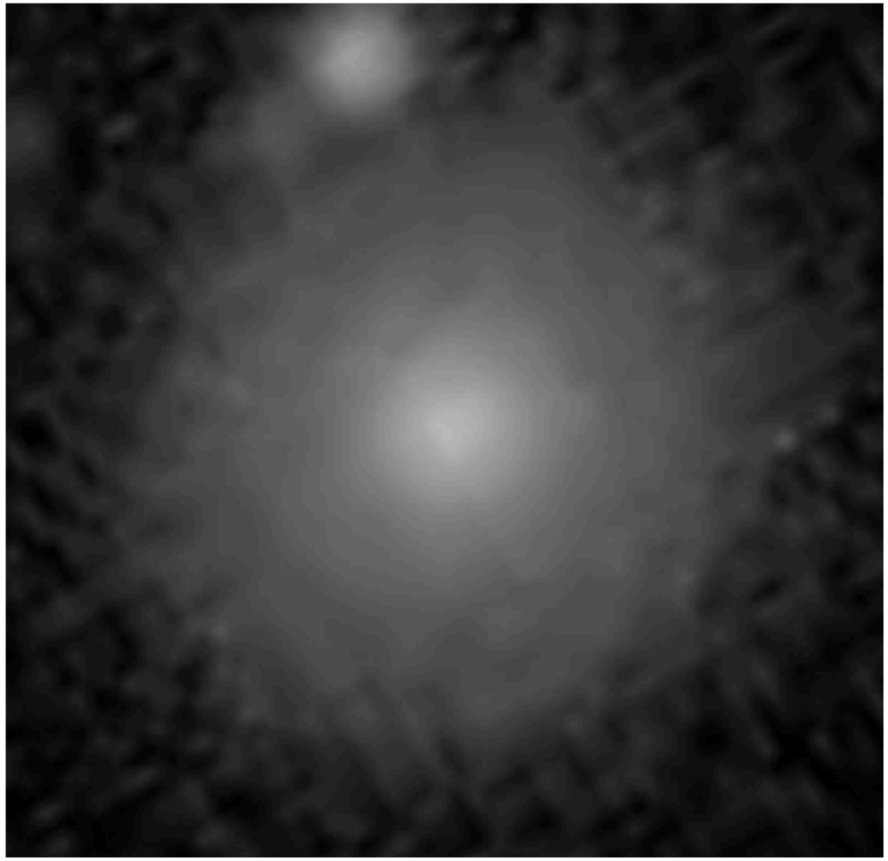




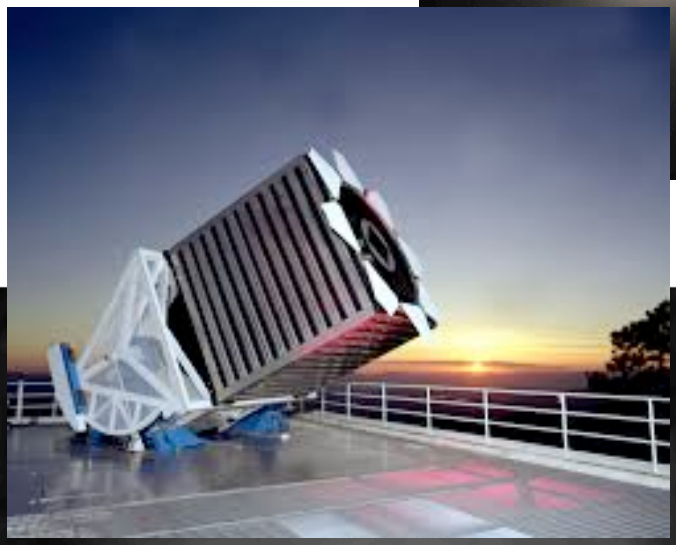
Spiral or Elliptical?

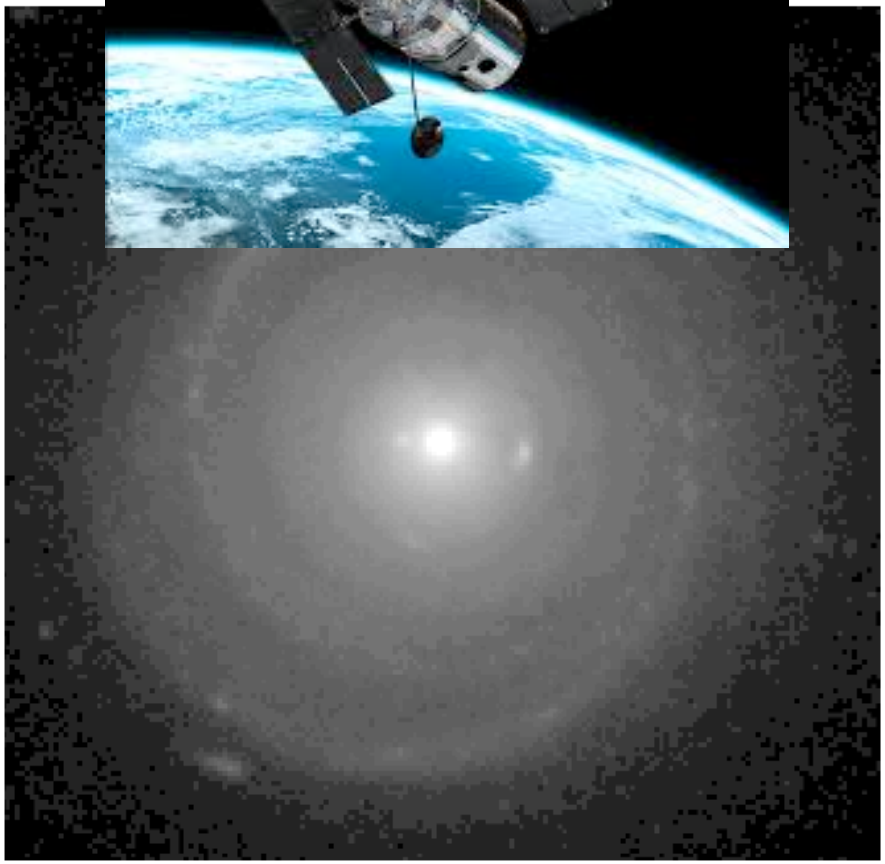


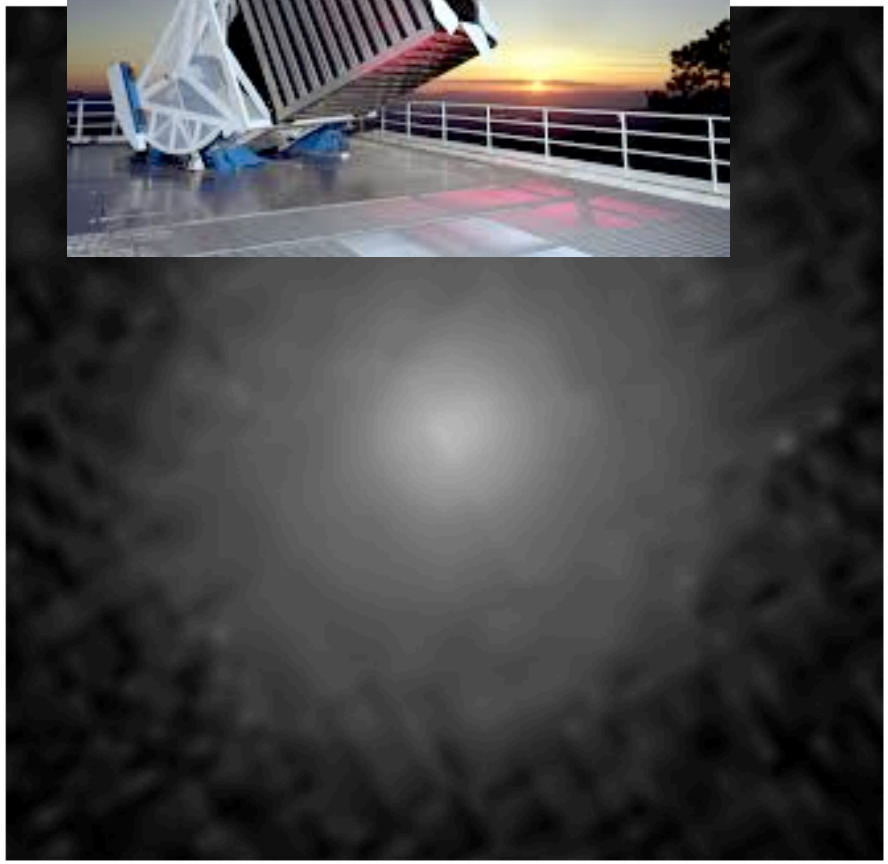




Spiral or Elliptical?





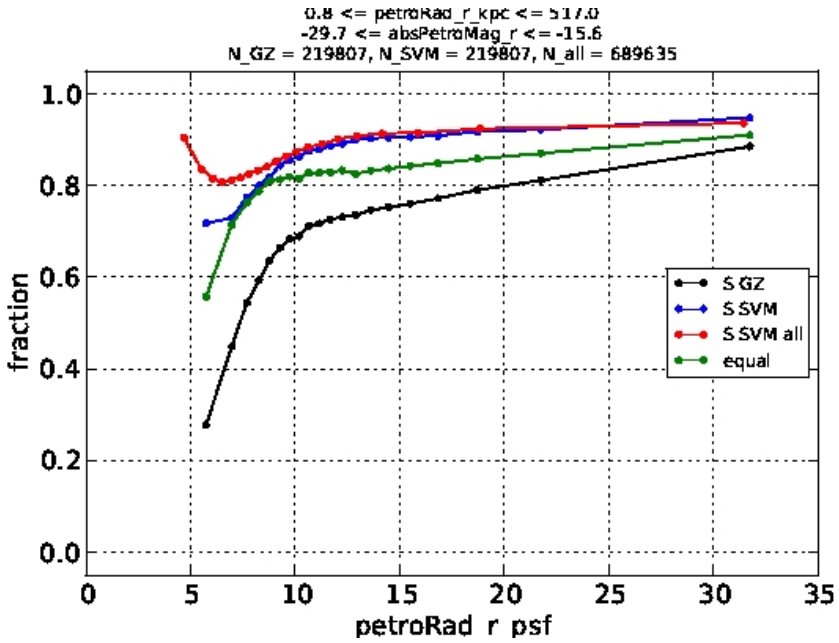


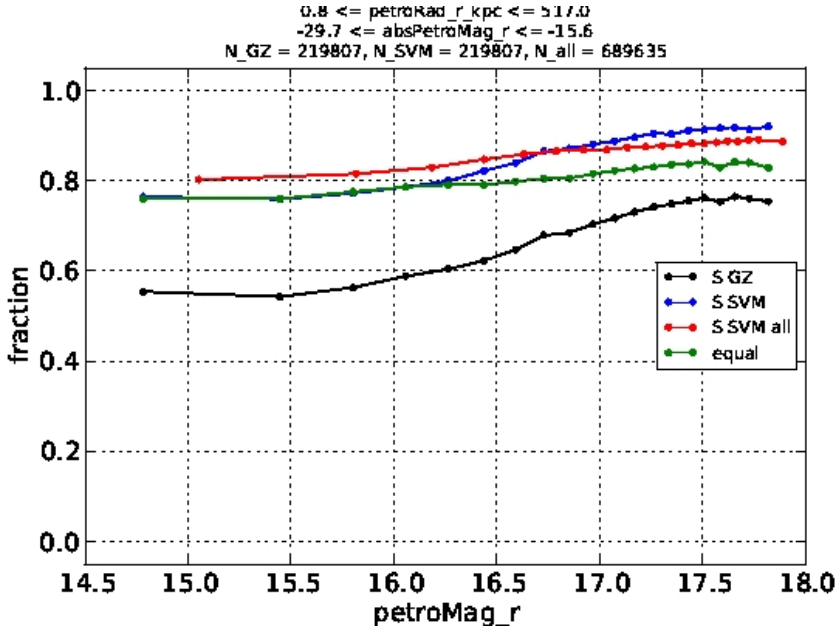
Resolution and Size

Object “Resolution”

Apparent size on the sky

Classification Bias

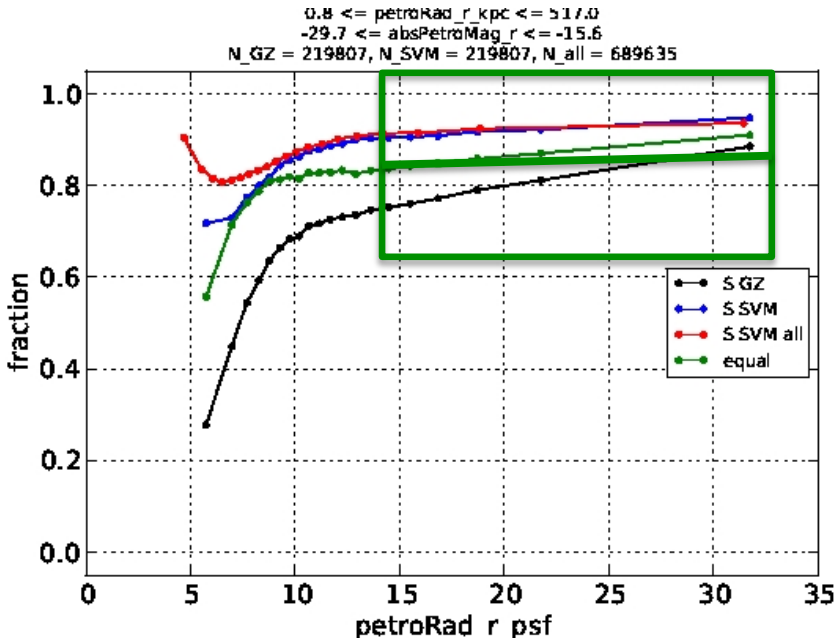


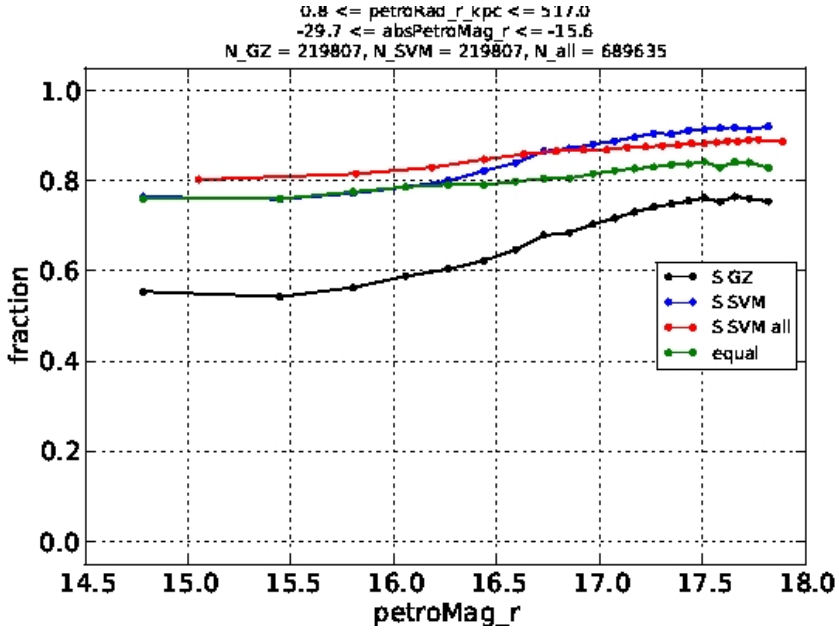


We DO NOT expect the fraction of types to fundamentally vary as a function of the observable properties.

We DO expect the fractions to depend on the intrinsic properties of the galaxies.

Classification Bias

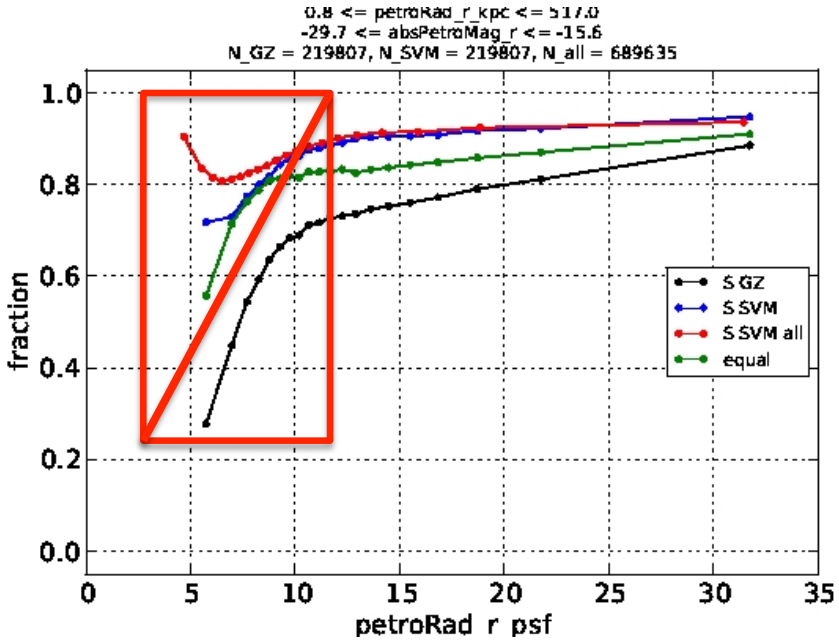


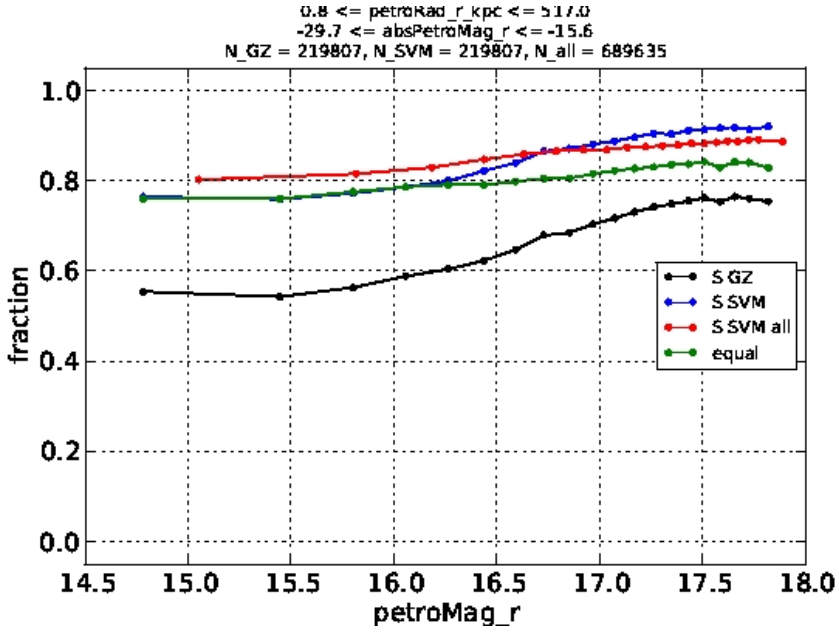


We DO NOT expect the fraction of types to fundamentally vary as a function of the observable properties.

We DO expect the fractions to depend on the intrinsic properties of the galaxies.

Classification Bias and Training Sets





We DO NOT expect the fraction of types to fundamentally vary as a function of the observable properties.

We DO expect the fractions to depend on the intrinsic properties of the galaxies.

Quantifying Classification Bias

$$\sigma^2_{j,k,q} = \frac{1}{N_{\mathcal{A}_{j,q}}} \sum_{l=1}^{N_{\mathcal{A}_{j,q}}} (r_{j,l,q,k} - r_{k,q})^2 \qquad L = \sqrt{\frac{1}{K N_{\mathcal{B}} N_{\alpha}} \sum_{j,k,q} {\sigma_{j,k,q}}^2}$$

- Define the expectation in finite bins for a given observable property A.
- Compute the ℓ^2 distance to the expectation of the observable given the truth.

- Sum the ℓ^2 distance over all observables.
- L measures the overall classification bias and should be <0> in the limit of large N.

Quantifying Classification Bias

fractions

$$\sigma^2_{j,k,q} = \frac{1}{N_{\mathcal{A}_{j,q}}} \sum_{l=1}^{N_{\mathcal{A}_{j,q}}} \left(r_{j,l,q,k} - r_{k,q} \right)^2$$

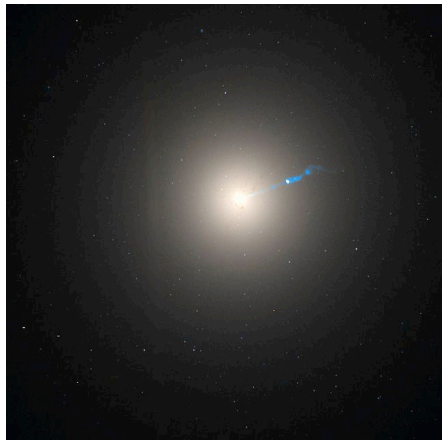
$$L = \sqrt{\frac{1}{K N_{\mathcal{B}} N_{\alpha}} \sum_{j,k,q} {\sigma_{j,k,q}}^2}$$

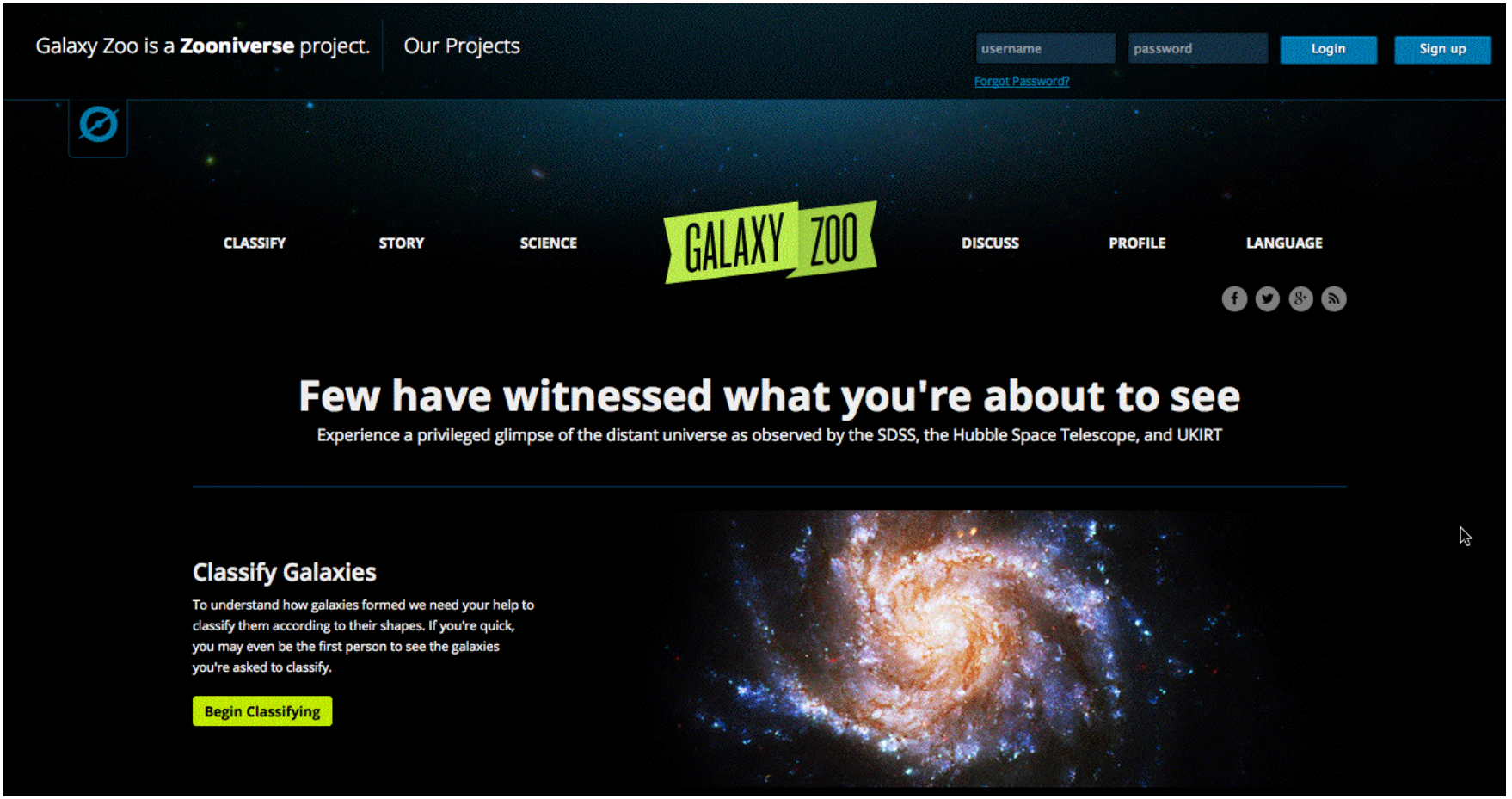
- Define the expectation in finite bins for a given observable property A.
- Compute the ℓ^2 distance to the expectation of the observable given the truth.

- Sum the ℓ^2 distance over all observables.
- L measures the overall classification bias and should be <0> in the limit of large N.

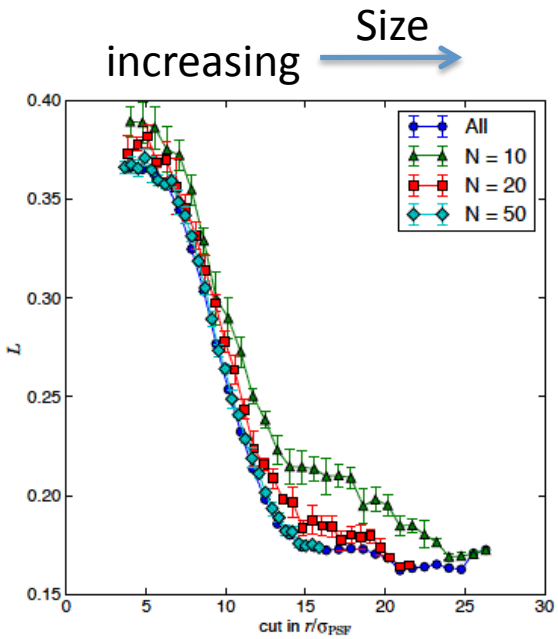
The Data: Galaxy Zoo

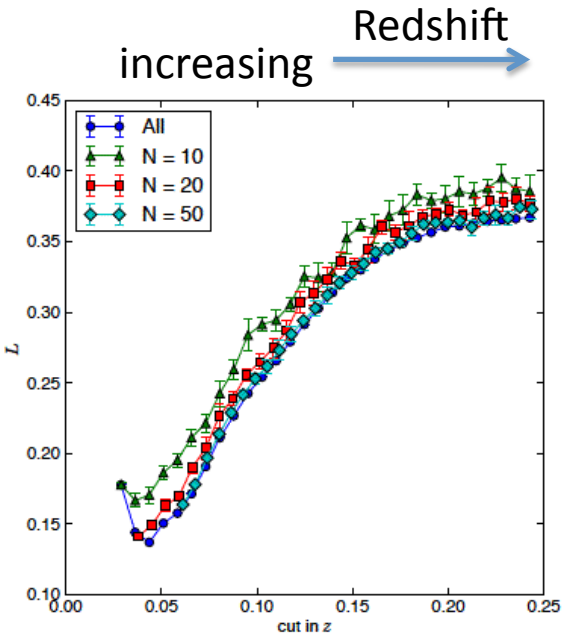
Lintott et al.

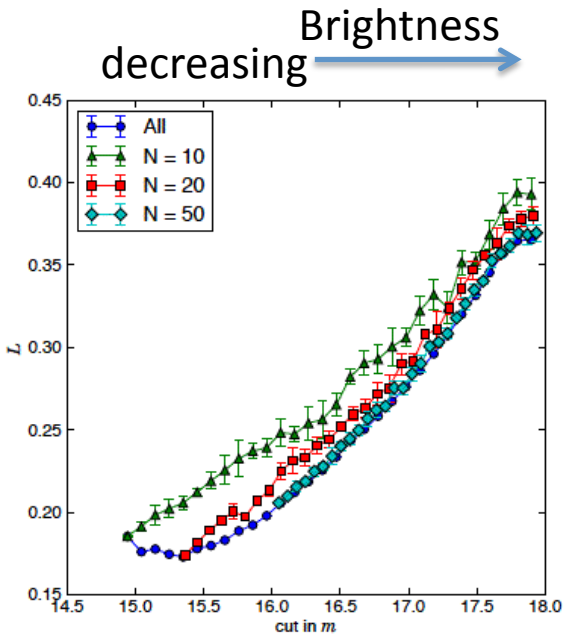




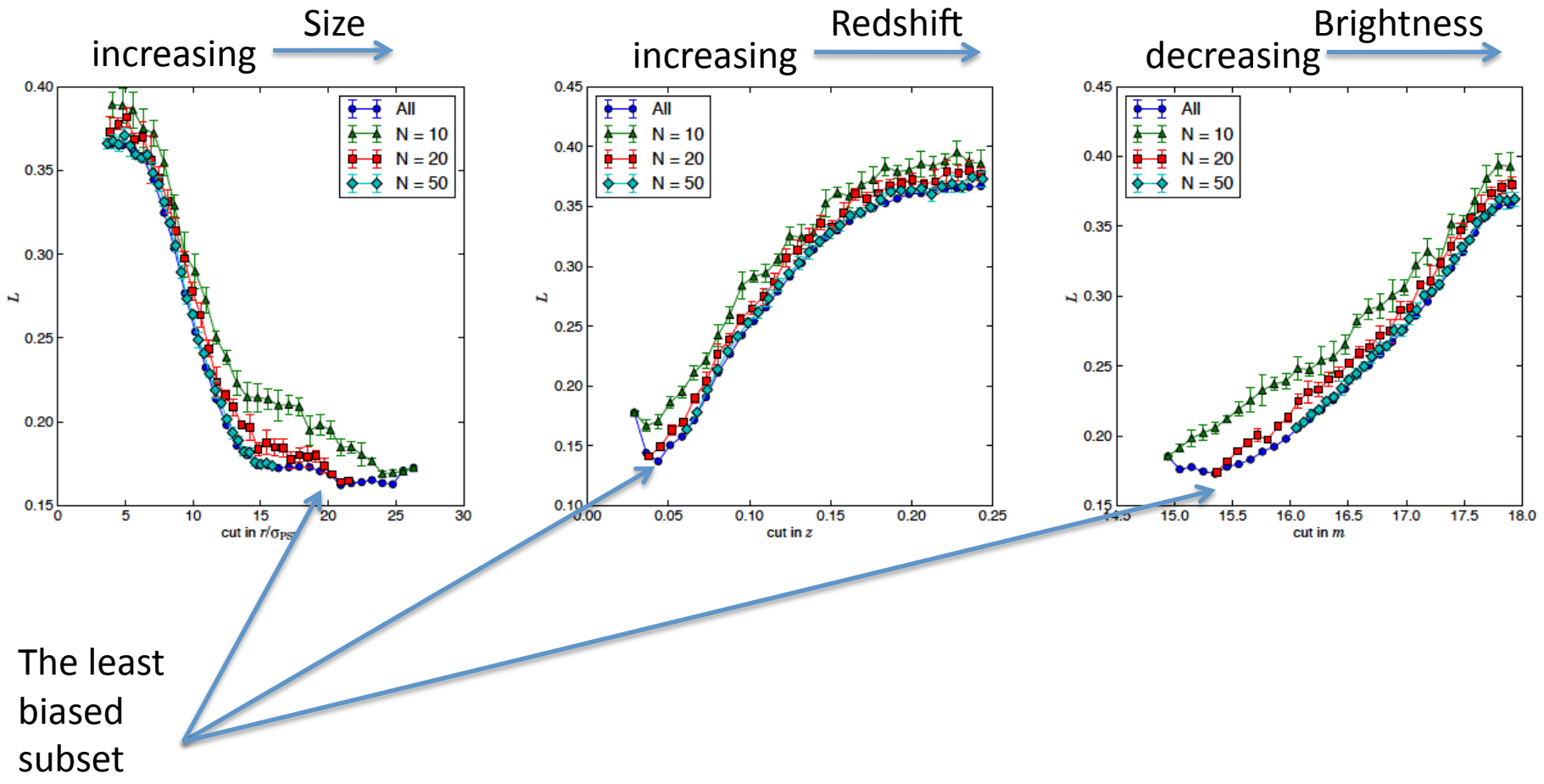
Galaxy Zoo Classification Bias







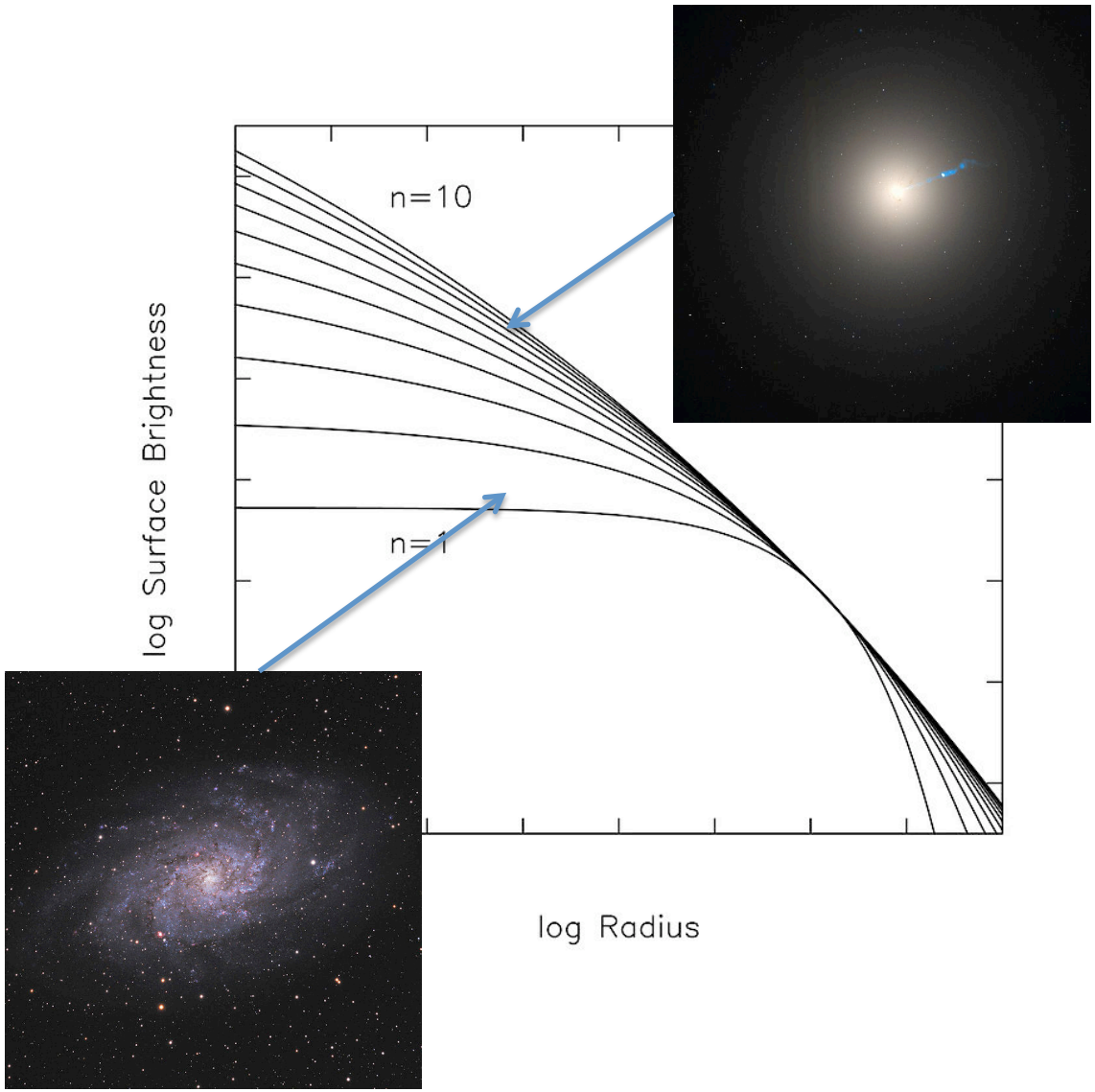
Galaxy Zoo Classification Bias



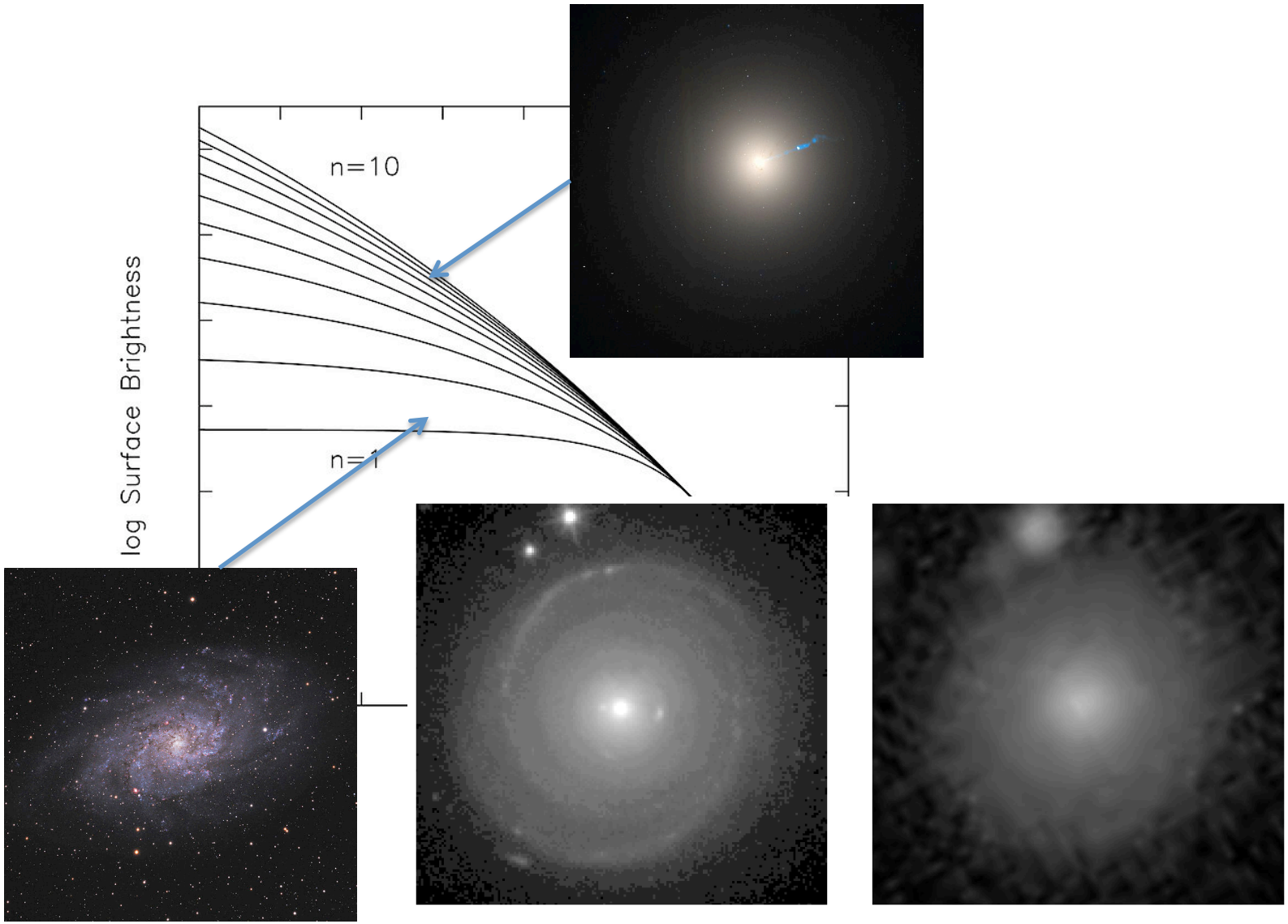
Part I: Classification Bias in the "Gold Standard Datasets"

- Astronomers have defined "Gold Standard" Morphological Datasets (not a library!)
  - Reference Catalogue of Bright Galaxies (de Vaucouleurs 1964,1995)
  - Uppsala General Catalogue of Galaxies (Nilson 1973)
  - Revised Shapley Ames Catalogue (Sandage & Tammann 1980)
  - SDSS (Fukugita et al. 2007)
  - Galaxy Zoo (Lintott et al. 2011)
- Classifiers are teams of people
  - Of one, two or three experts (Classic)
  - Of 100s of non-experts (Galaxy Zoo)
- All "Gold Standard" Morphological Catalogs are Biased

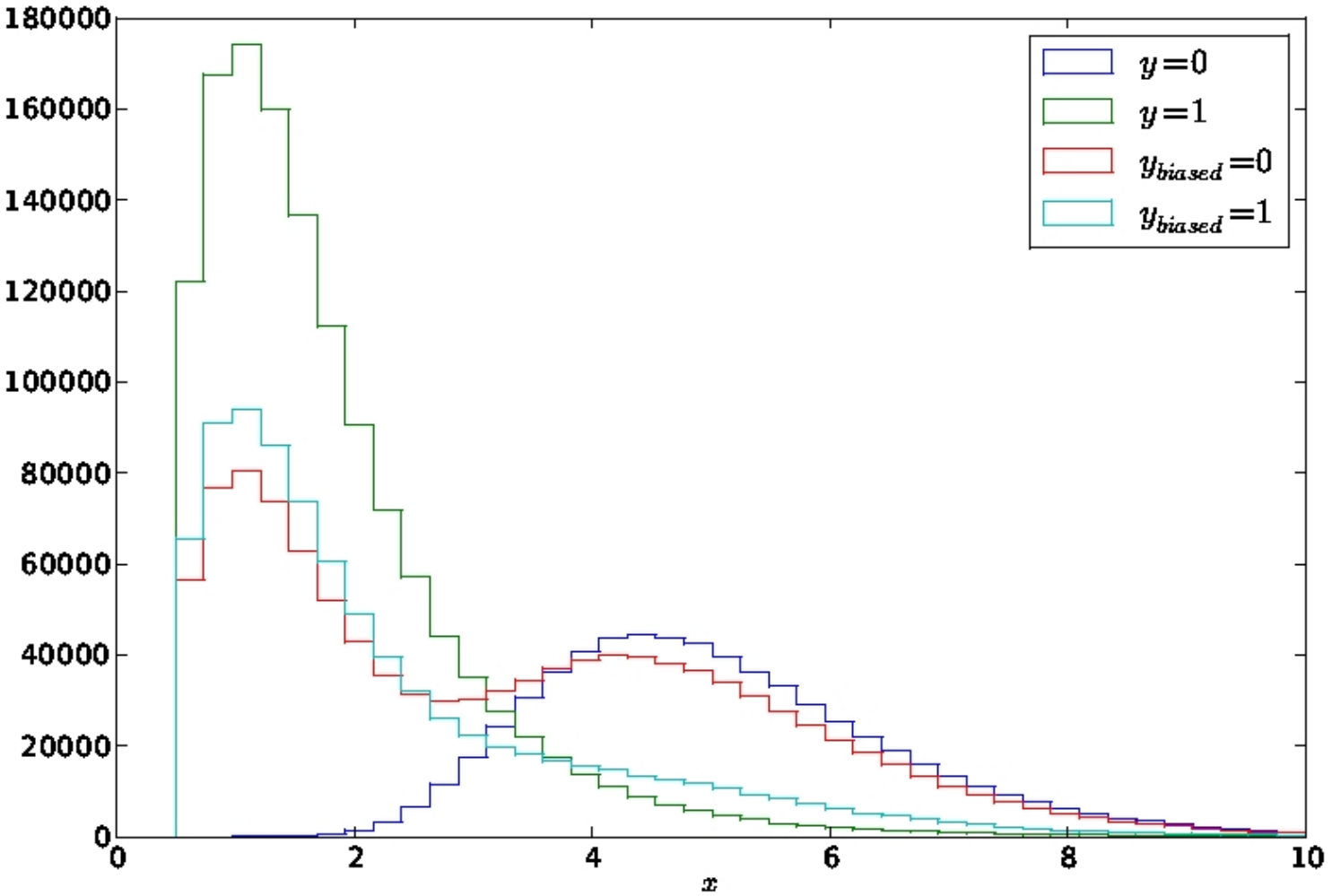
Part II: Classifying with the Sérsic Light Profile



Part II: Classifying with the Sérsic Light Profile

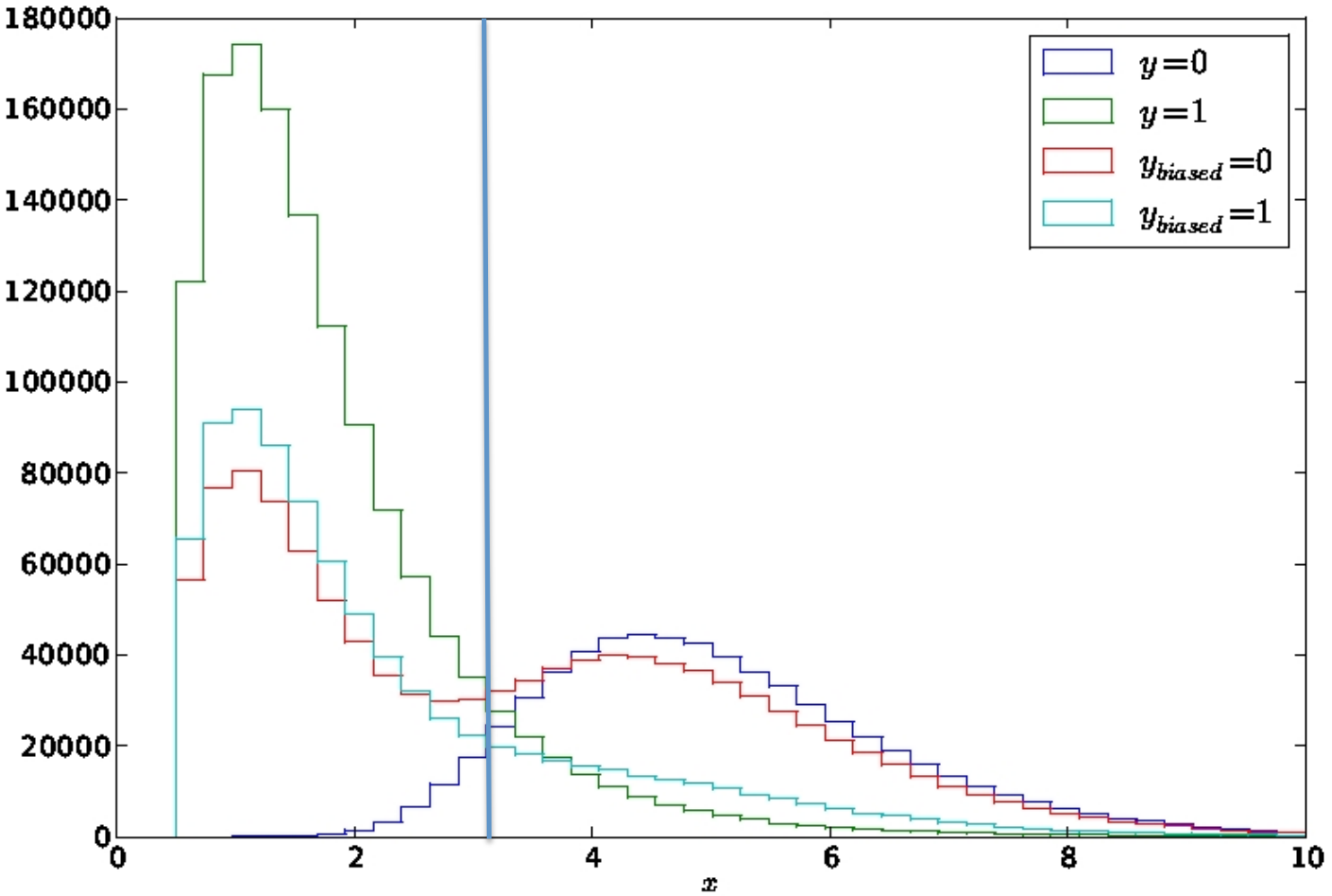


Single Parameter Classification with Bias: Simulation



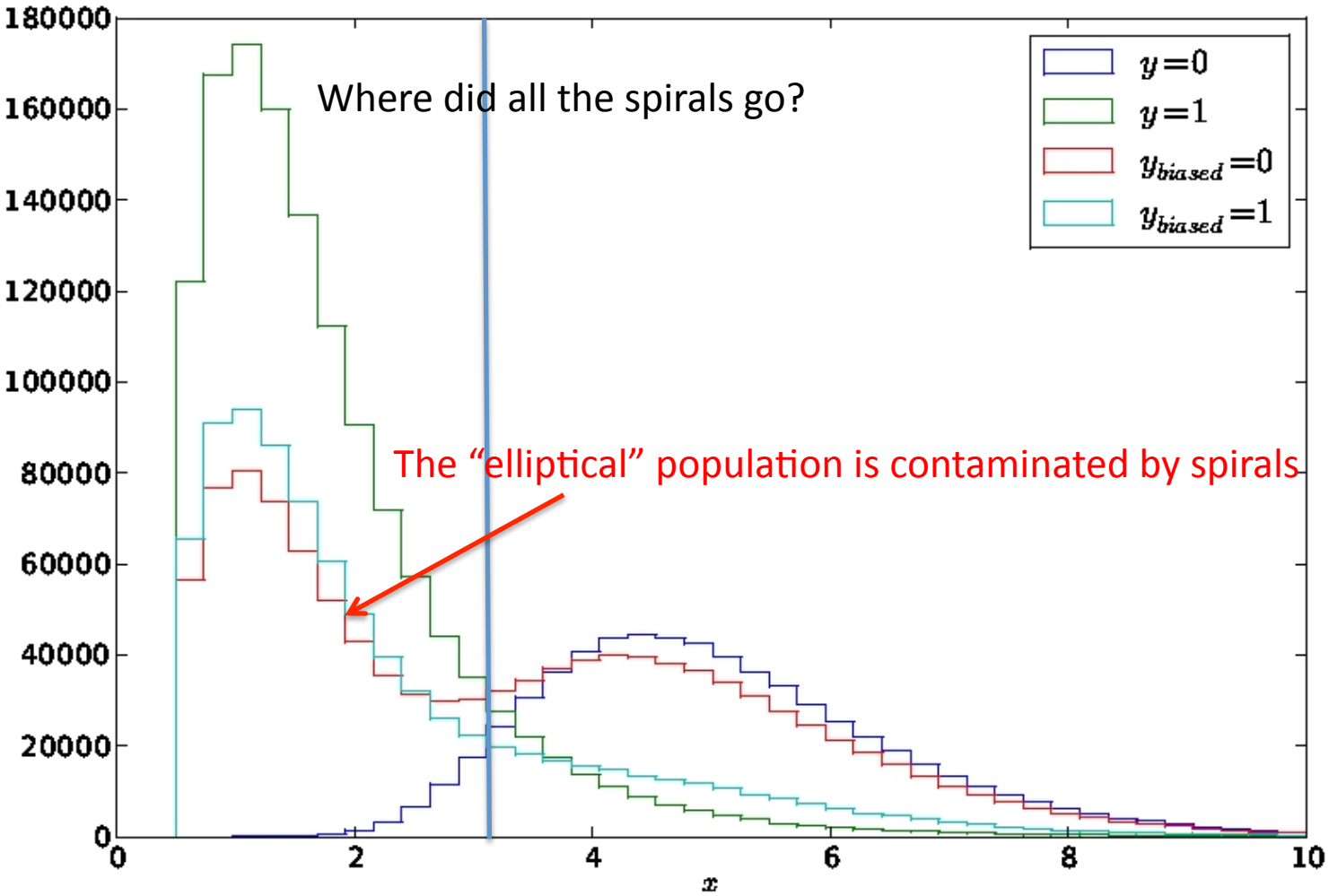
e.g.: the Sérsic index

Single Parameter Classification with Bias: Simulation



e.g.: like the Sérsic index

Single Parameter Classification with Bias: Simulation

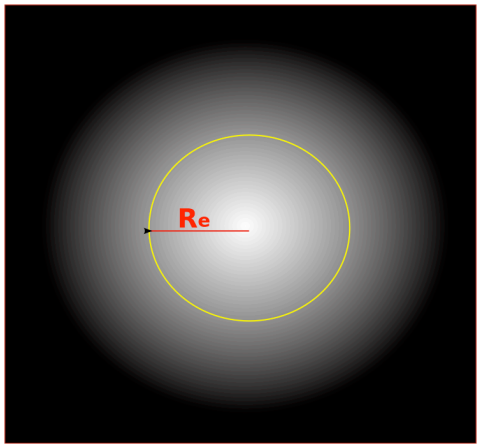


e.g.: the Sérsic index

The Sérsic Light Profile

$$I^{\mathrm{S}}(\xi) = I_0 e^{-b_n(\frac{\xi}{R_e})^{1/n}}$$

$$\xi = \sqrt{x^2 + \frac{y^2}{(1-\epsilon)^2}}$$

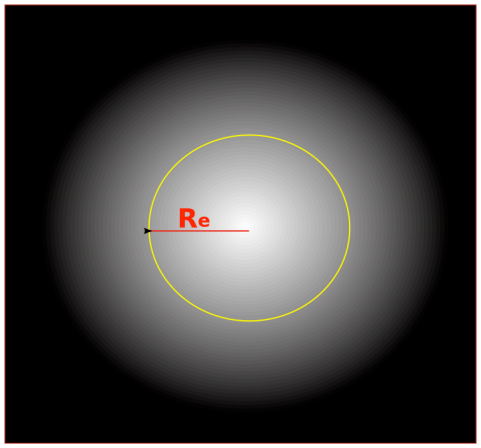


Sérsic parameters: R_e ϵ n

The Sérsic Light Profile

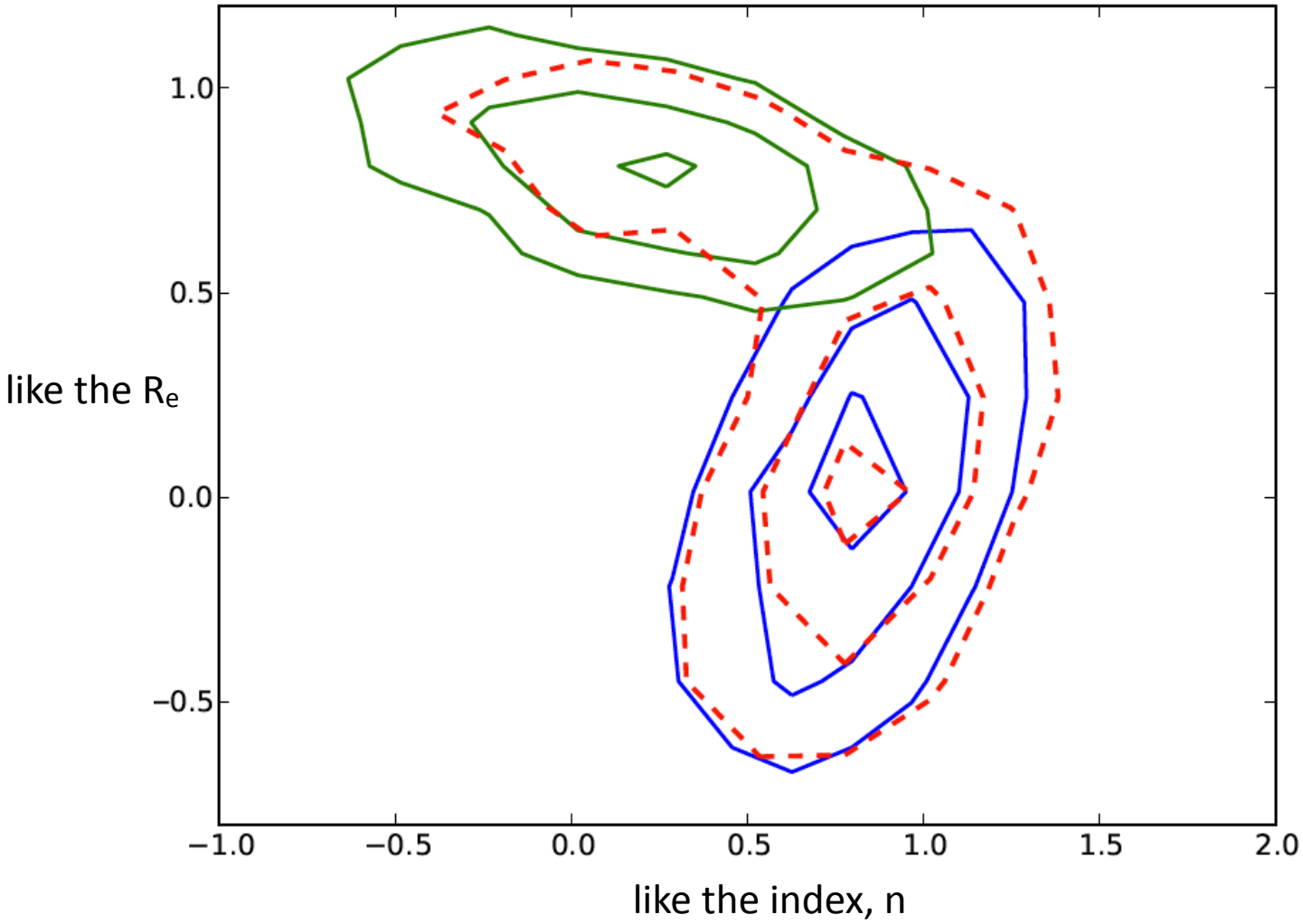
$$I^{\mathrm{S}}(\xi) = I_0 e^{-b_n\left(\frac{\xi}{R_e}\right)^{1/n}}$$

$$\xi = \sqrt{x^2 + \frac{y^2}{(1-\epsilon)^2}}$$

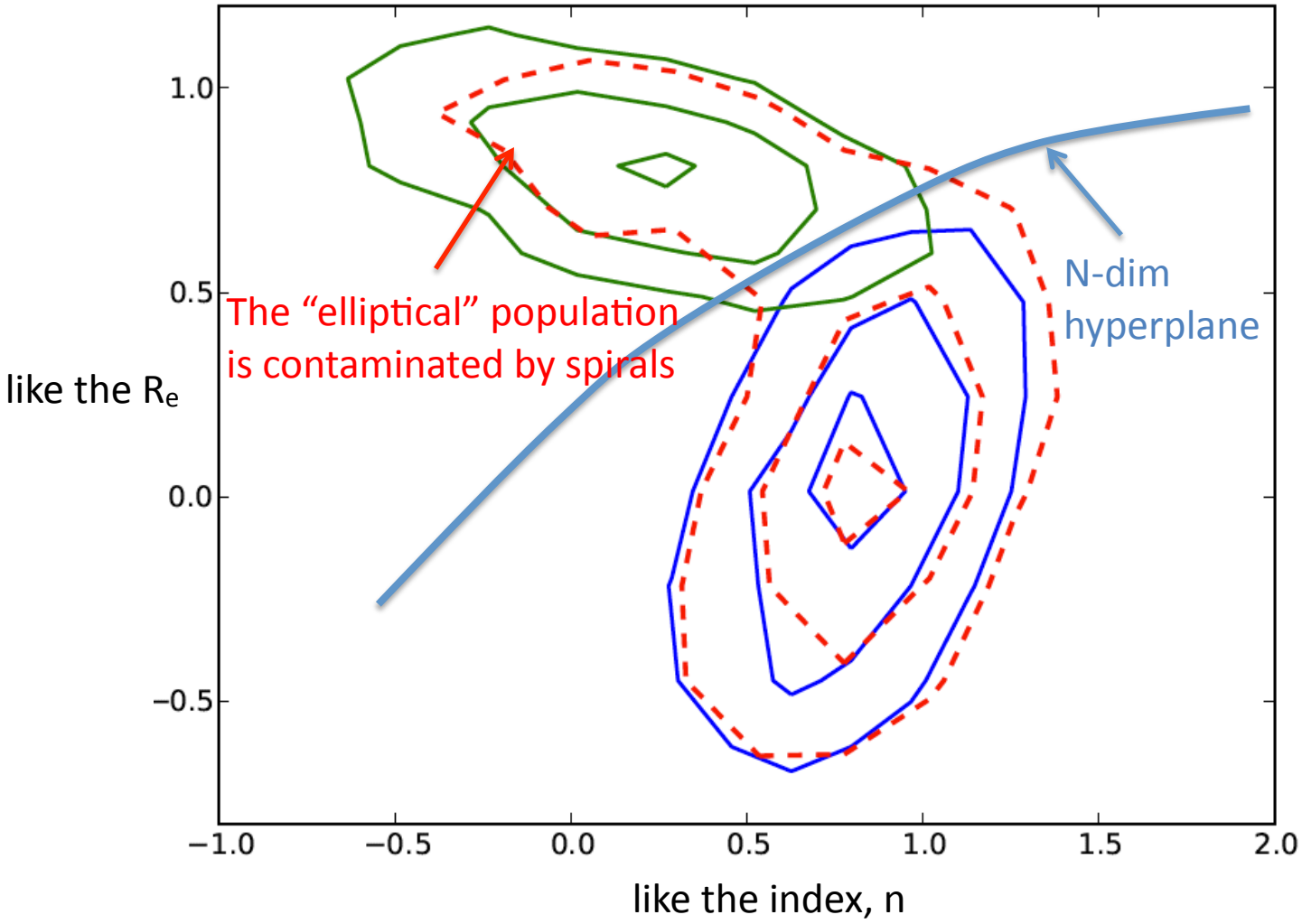


Sérsic parameters: R_e ϵ n

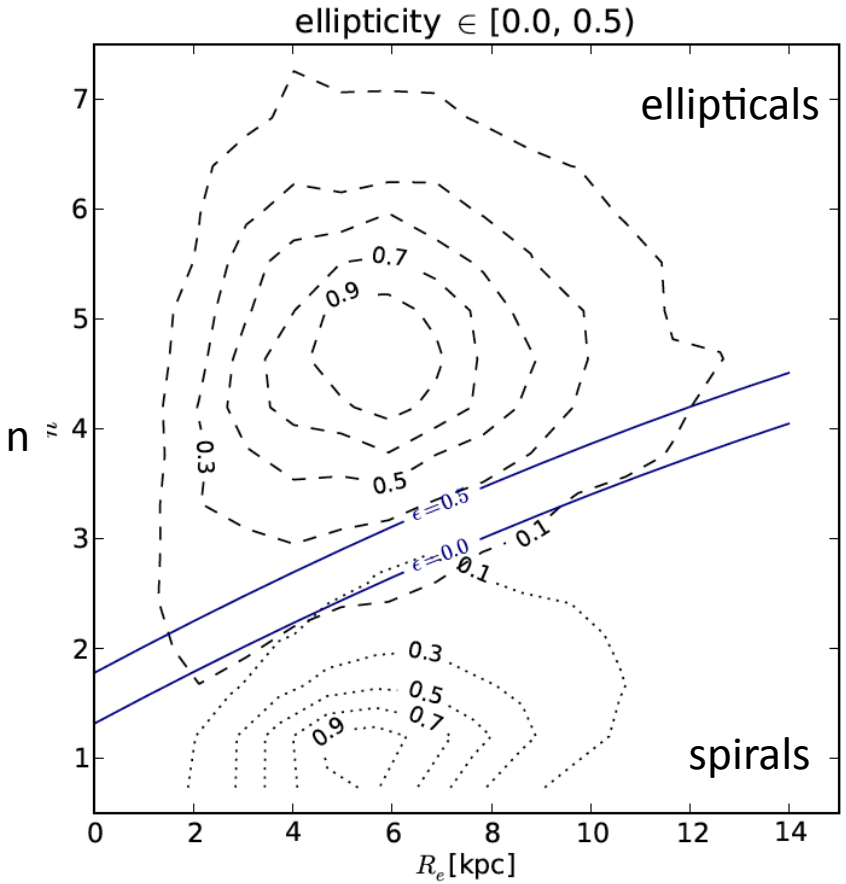
From 1D to Multi-dimensional Classification



From 1D to Multi-dimensional Classification



How do we separate spirals and ellipticals using the Sérsic parameters?



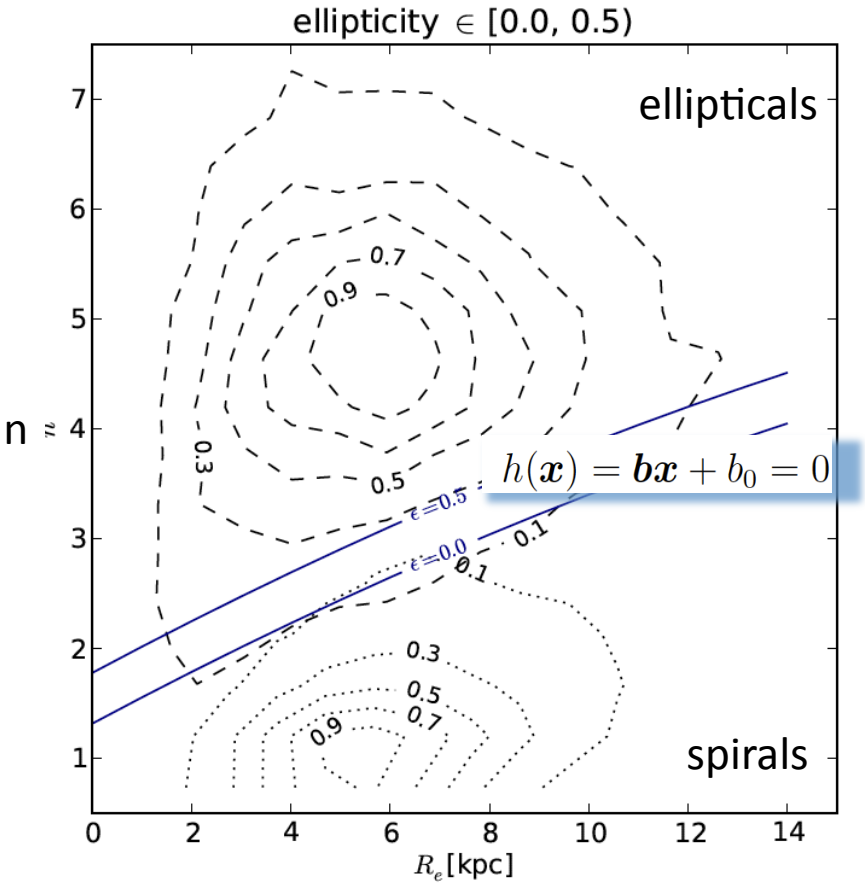
Support Vectors, b

$$h(\boldsymbol{x}) = \sum^{N_{\mathrm{SV}}} a_i \hat{y}_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b_0$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = (\gamma R_{ei} R_e + \rho)^2 + \gamma^2 (n_i n + \epsilon_i \epsilon)$$

-SVM models the training set as points in the parameter space.
-These points are mapped, so that the separate classes are “optimally” divided.
-New examples are then mapped into that same space and predicted.

How do we separate spirals and ellipticals using the Sérsic parameters?



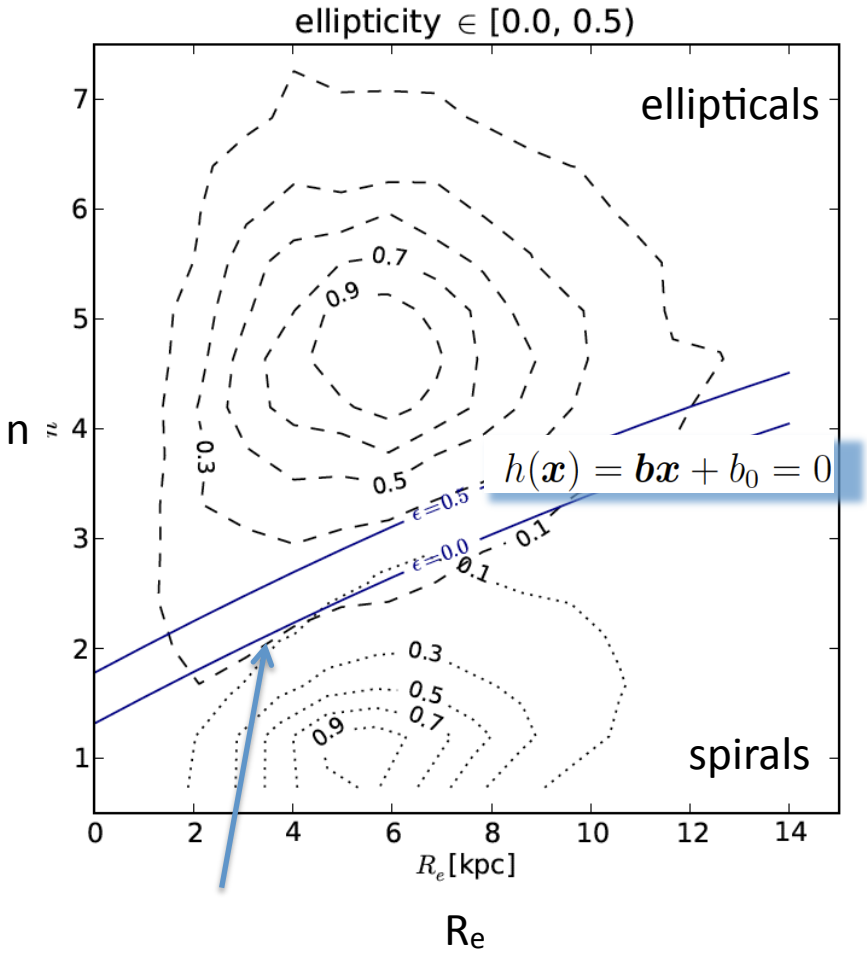
Support Vectors, b

$$h(\boldsymbol{x}) = \sum^{N_{\mathrm{SV}}} a_i \hat{y}_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b_0$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = (\gamma R_{ei} R_e + \rho)^2 + \gamma^2 (n_i n + \epsilon_i \epsilon)$$

-SVM models the training set as points in the parameter space.
-These points are mapped so that the separate classes are “optimally” divided.
-New examples are then mapped into that same space and predicted.

How do we separate spirals and ellipticals using the Sérsic parameters?



Support Vectors, b

$$h(\boldsymbol{x}) = \sum^{N_{\mathrm{SV}}} a_i \hat{y}_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b_0$$

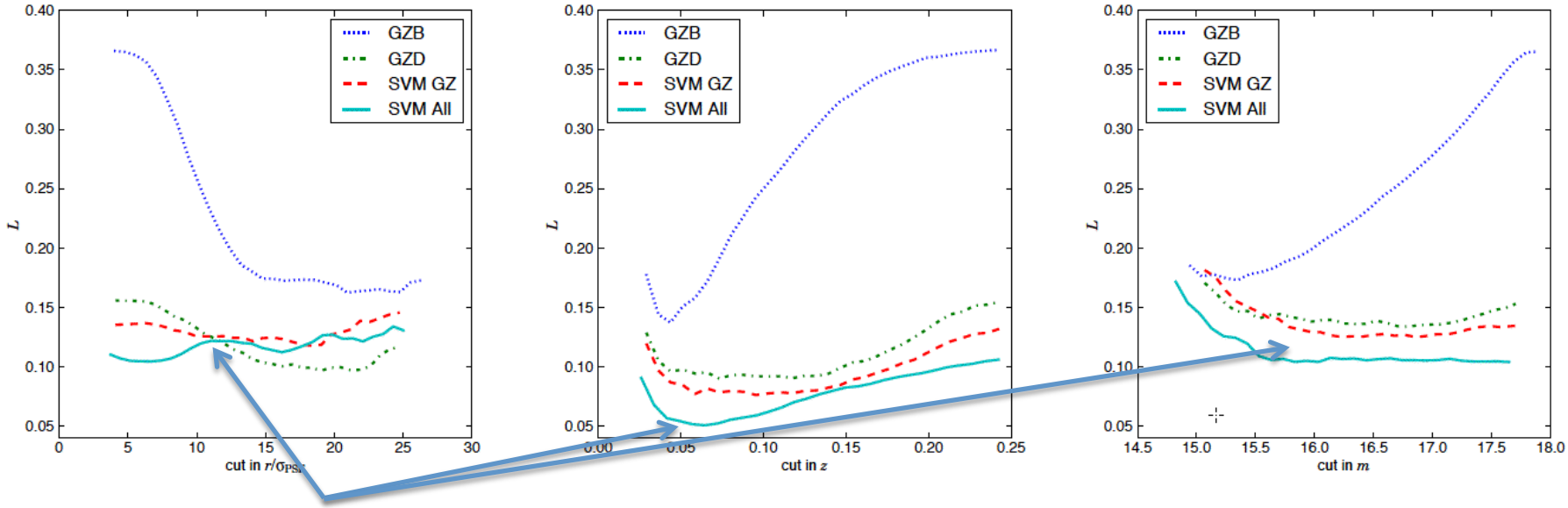
$$K(\boldsymbol{x}_i, \boldsymbol{x}) = (\gamma R_{ei} R_e + \rho)^2 + \gamma^2 (n_i n + \epsilon_i \epsilon)$$

-SVM models the training set as points in the parameter space.
-These points are mapped so that the separate classes are "optimally" divided.
-New examples are then mapped into that same space and predicted.

$$h(n, R_e, \epsilon) = AR_e^2 + Bn + CR_e + D\epsilon + E$$

The GZ SVM Classifications:
Accuracy, Precision *and* Bias

Cabrera, Miller, Schneider,ApJS 2014--submitted



Low observational bias

$f(c_i\|c_j)$	elliptical	spiral
elliptical	$92.1 \pm 0.3\%$	$4.9 \pm 0.1\%$
spiral	$7.9 \pm 0.3\%$	$95.1 \pm 0.1\%$

High accuracy/precision

A Generalized Approach:

Simultaneous De-Biasing and Classification

- Define a likelihood:

$$\mathrm{P}(\mathcal{D}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} \mathrm{P}(\hat{y}_i|\boldsymbol{x}_i, \boldsymbol{\alpha}_i, \boldsymbol{\Theta})$$

For example, in 2D Sérsic: $x = \{R_e, n\}$ $\alpha = \{m_r, \text{radius/PSF}\}$ $\Theta = \{\theta_1, \theta_2, w\}$

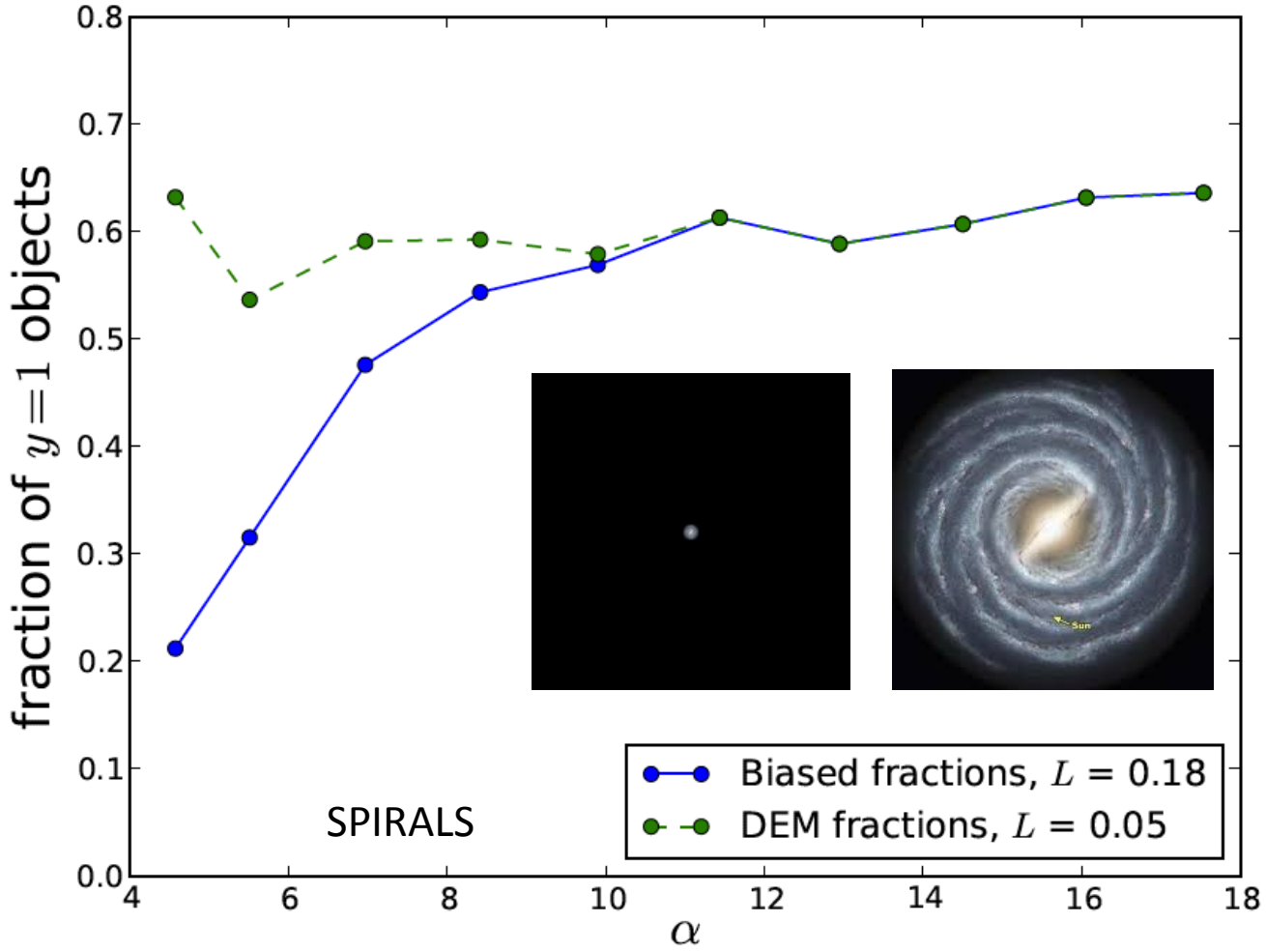
- Maximize via Expectation-Maximization

$$\boldsymbol{\Theta}_{\mathrm{ML}} = \max_{\boldsymbol{\Theta}} \{\ln \mathrm{P}(\mathcal{D}|\boldsymbol{\Theta})\}$$

Cabrera, Miller, Schneider, 2014 ICPR in press

A Simultaneous Approach via EM and Informative Priors: Lowering the Bias in Morphological Classifications

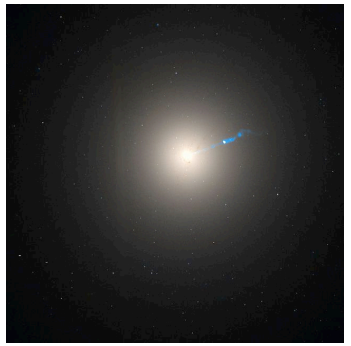
Cabrera, Miller, Schneider, 2014 ICPR in press



Radius/PSF

Cabrera, Miller, Schneider, 2014 ICPR in press

Summary



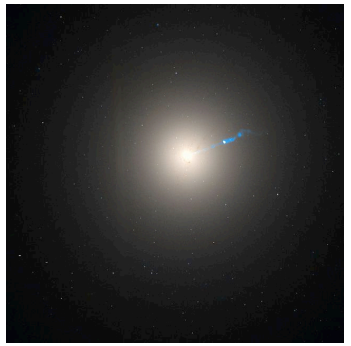
- Classification Bias (Inter-disciplinary component)
  - "gold standard" data sets (in astronomy) are rarely gold standards.
  - It's not *only* about accuracy and contamination rates (e.g., ROC curves, confusion matrices, etc). The formalism of bias needs to be included.
  - A generalized simultaneous approach works best
    - Include prior probabilistic information about bias and classifications
    - Generalizes to multi-model classifications

- **Galaxy Sérsic Light Profile (Domain science component)**
  - **Does a very good job of recovering classifications w/o the need for other correlated parameters like color**
  - **Fixes the problem of a single-parameter classifiers**
  - **They are easy to measure and come "free" with modern astronomical catalogs**

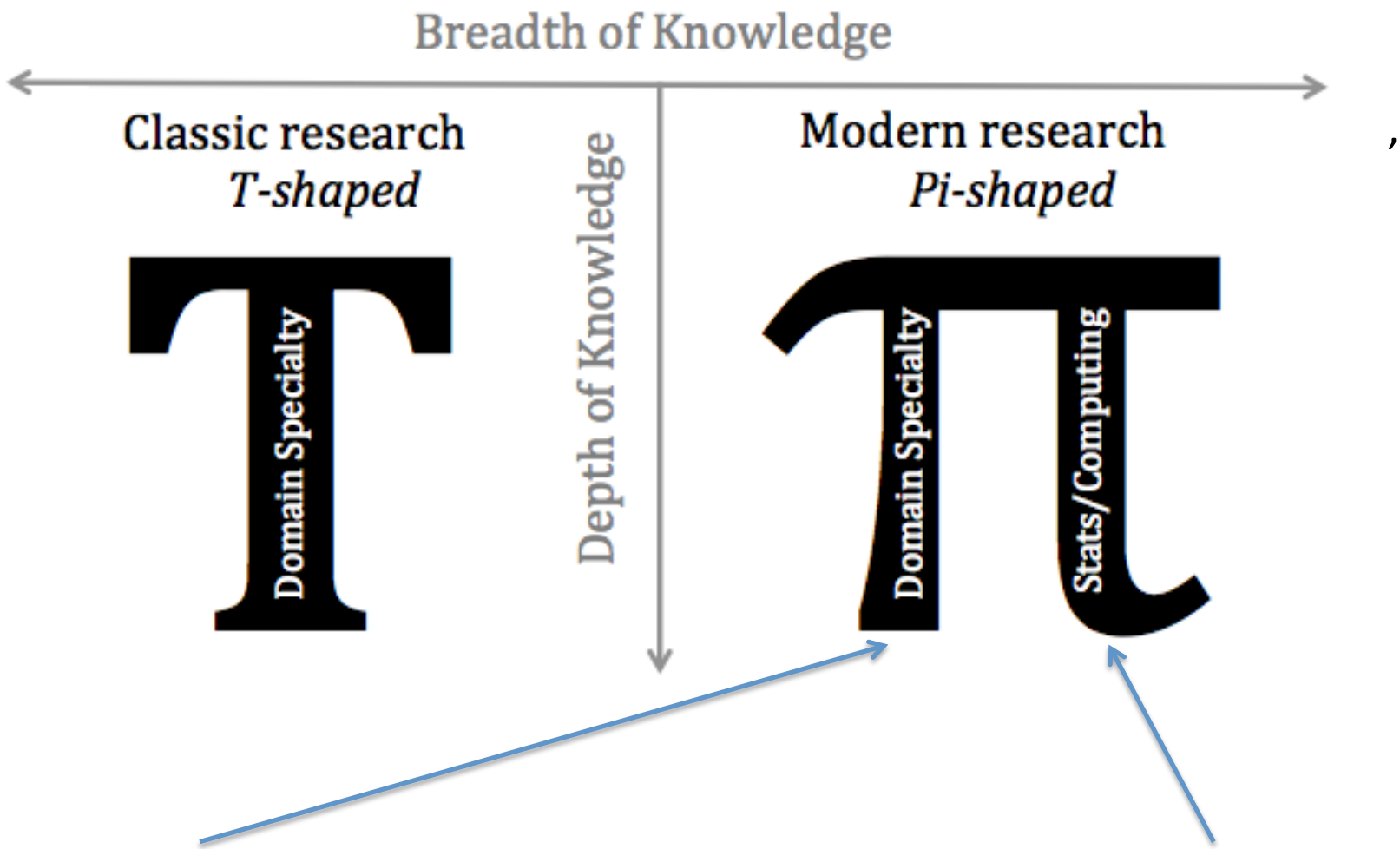
Cabrera, Miller, Schneider, 2014 ApJS submitted

Cabrera, Miller, Schneider, 2014 ICPR in press

Summary



-



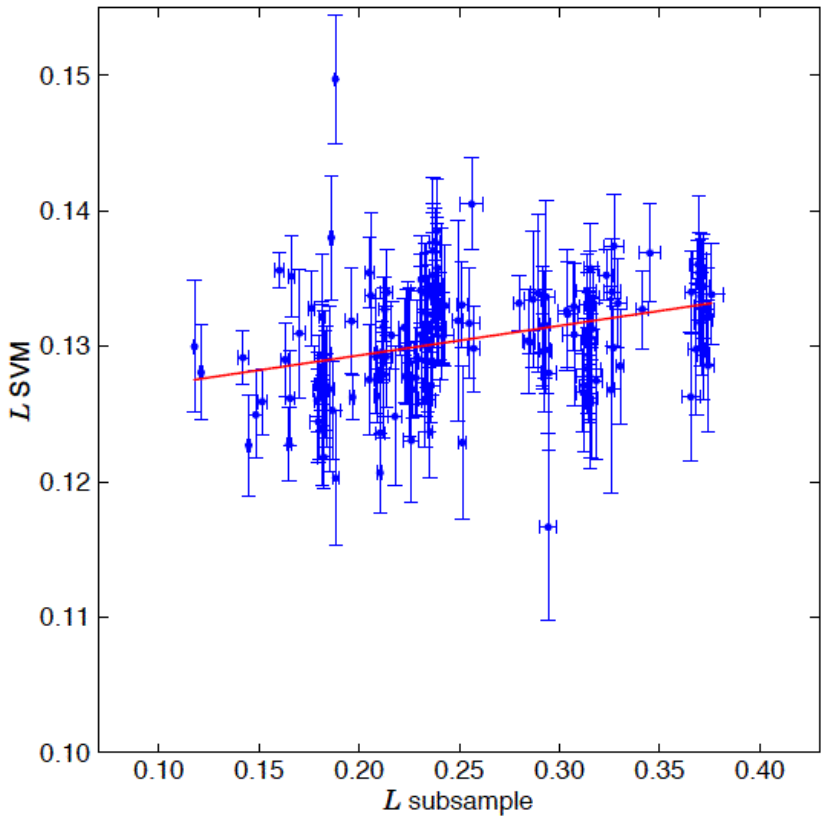
,

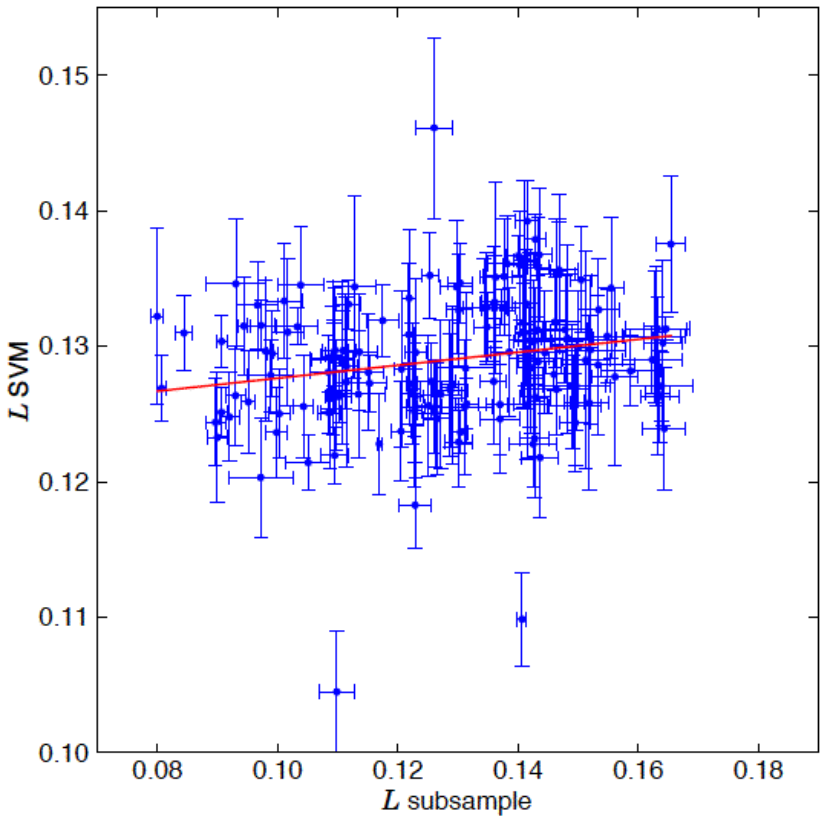
-

Cabrera, Miller, Schneider, 2014 ApJS submitted

Cabrera, Miller, Schneider, 2014 ICPR in press

Choosing the least biased subset





Simulations

	$\theta = 0.03$, changed labels = 17%			
	Ground Truth	Biased	WLR	DEM
Accuracy	92.89 ± 0.82	88.75 ± 1.13	89.78 ± 1.23	92.98 ± 0.9
AUC	0.981 ± 0.002	0.977 ± 0.005	0.978 ± 0.005	0.981 ± 0.003
L	0.073 ± 0.032	0.087 ± 0.035	0.079 ± 0.034	0.076 ± 0.034
L data	0.075 ± 0.038	0.218 ± 0.031	-	-

	$\theta = 0.05$, changed labels = 25%			
	Ground Truth	Biased	WLR	DEM
Accuracy	92.49 ± 0.77	78.27 ± 1.76	79.89 ± 1.71	92.22 ± 0.62
AUC	0.979 ± 0.003	0.974 ± 0.003	0.975 ± 0.003	0.979 ± 0.003
L	0.055 ± 0.018	0.204 ± 0.037	0.188 ± 0.034	0.056 ± 0.017
L data	0.056 ± 0.02	0.293 ± 0.027	-	-

	$\theta = 0.1$, changed labels = 35%			
	Ground Truth	Biased	WLR	DEM
Accuracy	93.44 ± 0.83	51.41 ± 2.77	50.35 ± 3.17	93.35 ± 0.9
AUC	0.983 ± 0.003	0.977 ± 0.005	0.977 ± 0.005	0.983 ± 0.003
L	0.072 ± 0.032	0.45 ± 0.041	0.461 ± 0.041	0.069 ± 0.031
L data	0.061 ± 0.023	0.343 ± 0.033	-	-

Classification Bias in the "Gold Standard Datasets"

data-set	L for 1,843 objects using $5 \times 5 \times 4$ bins
Fukugita	0.191
Huertas-Company	0.141 ± 0.011
GZ biased	0.318 ± 0.019
GZ debiased	0.156 ± 0.014