# Galaxy Classification without Feature Extraction

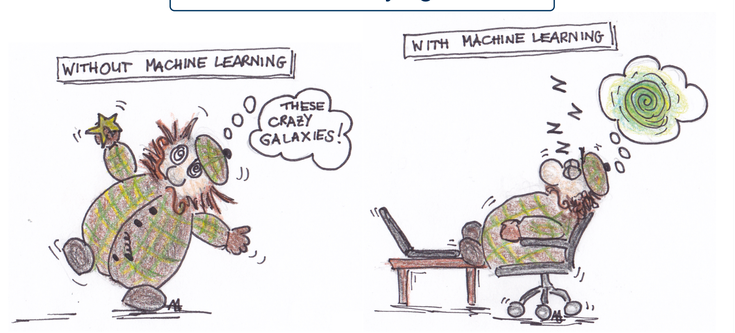

**Kai Lars Polsterer**[1] polsterer@astro.ruhr-uni-bochum.de

**Fabian Gieseke**[2] f.gieseke@uni-oldenburg.de

**Oliver Kramer**[2] oliver.kramer@uni-oldenburg.de

[1]Astronomisches Institut, Department of Physics and Astronomy, Ruhr-University of Bochum, Germany [2]Computational Intelligence, Computer Science Department, University of Oldenburg, Germany


## The Task: Classifying Galaxies





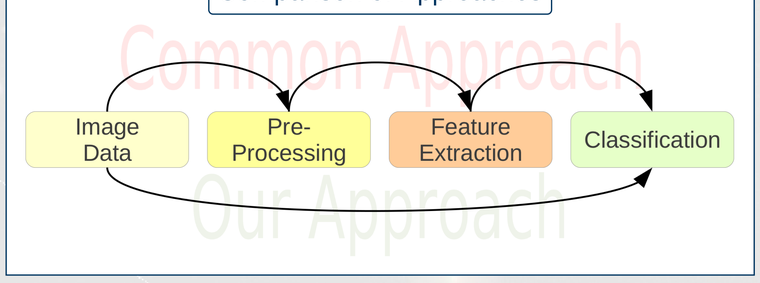
## Abstract

*The automatic classification of galaxies according to the different Hubble types is a widely studied problem in the field of astronomy. The complexity of this task led to projects like Galaxy Zoo which try to obtain labeled data based on visual inspection by humans. Many automatic classification frameworks are based on artificial neural networks (ANN) in combination with a feature extraction step in the pre-processing phase. These approaches rely on labeled catalogs for training the models. The small size of the typically used training sets, however, limits the generalization performance of the resulting models. In this work, we present a straightforward application of support vector machines (SVM) for this type of classification tasks. The conducted experiments indicate that using a sufficient number of labeled objects provided by the EFIGI catalog leads to high-quality models. In contrast to standard approaches no additional feature extraction is required.*


## Determine the Hubble Type of a Galaxy

An automated classification of galaxies is typically realized via a multi-stage approach. In the first step the image is pre-processed e.g. contrast enhancing or edge finding kernel-filters are applied. In a next step a small number of features is extracted from the image. Finally the generated features are used as input for classifiers e.g. ANNs or decision trees. [1, 2] are good examples for common approaches.

### Comparison of Approaches




The classification approach we present uses the raw image data without any feature pre-processing / extraction.

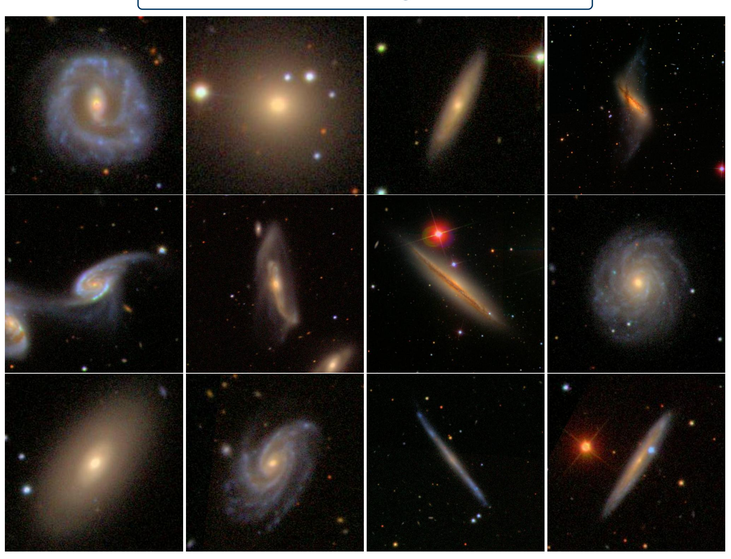
## The Classifiers: Support Vector Machines

The experiments rely on SVMs. Roughly speaking, the aim of a SVM is to find a hyperplane in a feature space which maximizes the "margin" between classes such that only a few training patterns lie within this margin [3]. The latter task can be formulated as an quadratic optimization problem, where the first term corresponds to maximizing the margin and the second term to the loss caused by patterns lying within the margin:

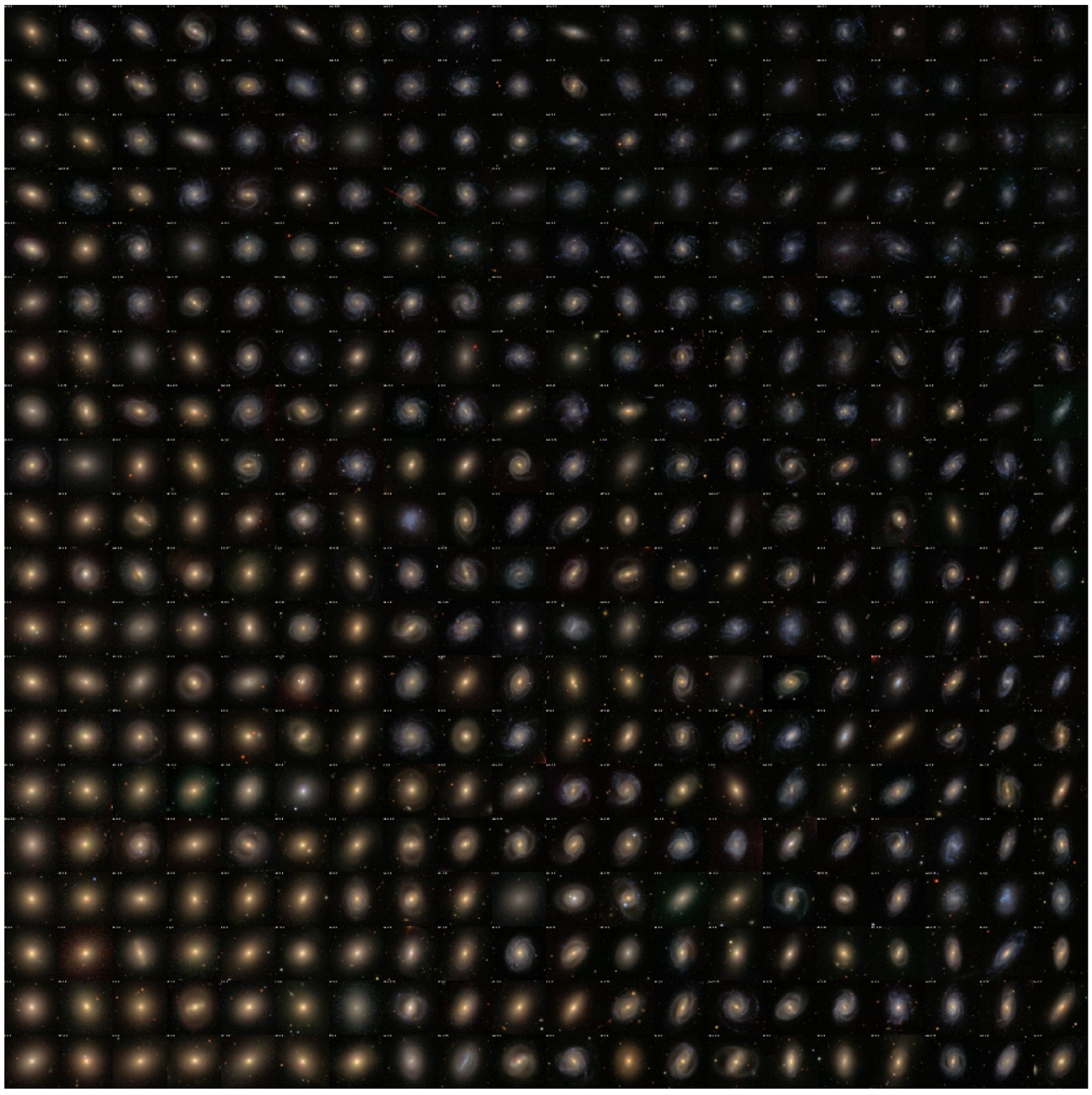
$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H}_0,\, \xi\in\mathbb{R}^n,\, b\in\mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\langle\mathbf{w},\Phi(\mathbf{x}_i)\rangle + b) \geq 1-\xi_i, \\ & \text{and } \xi_i \geq 0, \end{aligned} \tag{1}$$

where $C > 0$ is a user-defined parameter. The function $\Phi : \mathbb{R}^d \to \mathcal{H}_0$ is induced by a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $k(\mathbf{x}_i, \mathbf{x}_j) = \langle\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle$.

### The Data: SDSS Images of Galaxies



## Self-Organizing Map of the EFIGI Galaxies



A kernel function can be seen as a "similarity measure" for input patterns. The goal of the learning process is to find the optimal prediction function $f(\mathbf{x}) = \langle\mathbf{w}, \Phi(\mathbf{x})\rangle + b$. A common choice for the kernel function is the linear kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle\mathbf{x}_i, \mathbf{x}_j\rangle \tag{2}$$

or the RBF kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \tag{3}$$

with $\sigma$ as additional parameter.

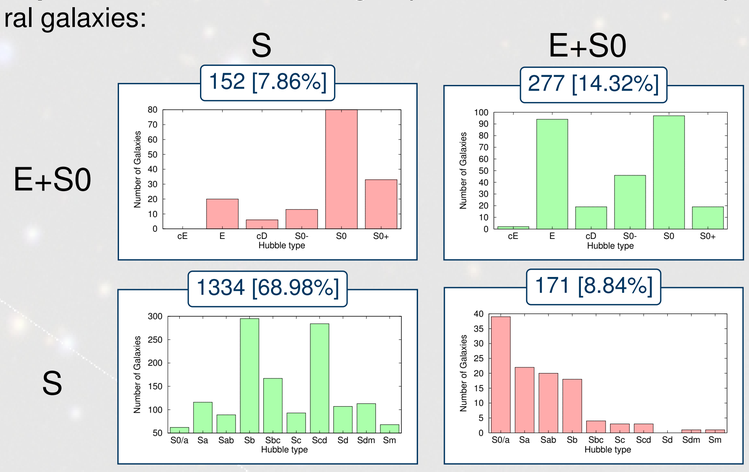
## Image Data and Labels

The presented experiments are based on image data taken from the Sloan Digital Sky Survey (SDSS) [4]. In [5] the EFIGI catalog of 4458 nearby galaxies is presented. The Hubble type and morphological features of these galaxies have been determined by a group of human experts. This catalog was used to extract the required labels for the experiments. In a first step the image data for each galaxy was retrieved as JPEG files. The resolution was adjusted to fit the whole galaxy and a $40 \times 40$ pixels$^2$ stamp was created.

## Classification Experiments and Results

Based on the available data we conducted several morphological classification experiments:

**Experiment 1** Discriminating elliptical and lenticular from spiral galaxies:




**Experiment 2** Discriminating elliptical from spiral galaxies:

A classification accuracy of 88% is reached.

**Experiment 3** Detecting a bar in a galaxy:

| | Bar | No Bar |
|---|---|---|
| No Bar | 501 [38.87%] | 37 [2.87%] |
| Bar | 733 [56.87%] | 18 [1.39%] |

## Self-Organizing Maps

By using an ANN a discrete and low-dimensional map is created which represents the input objects in their high-dimensional feature space. The resulting map is called a self-organizing map (SOM). It reflects the similarity of the input objects in high dimensions. With all EFIGI galaxies such a SOM was trained. This map clearly separates spiral from elliptical galaxies.

**Conclusions** Both, the classification experiments and the SOM show that simple classification tasks can be solved with raw features. Complex tasks like detecting a bar require other features. If a SOM shows a clear separation in a feature space concerning a certain classification task, this task could be solved with the raw features.

## References


[1] Jorge de la Calleja and Olac Fuentes. *Machine learning and image analysis for morphological galaxy classification*, MNRAS, 2004.

[2] D. B. Wijesinghe et al. *Morphological classification of galaxies and its relation to physical properties*, MNRAS, 2009.

[3] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning*, Springer, 2010.

[4] Sloan Digital Sky Survey. http://www.sdss.org, October 2011.

[5] A. Baillard et al. *The EFIGI catalogue of 4458 nearby galaxies with detailed morphology*, A&A, 2011.



**Acknowledgements** This work is based on data of the SDSS project [4]. We thank Anna Amelung for the cartoon.


Kai Lars Polsterer

Fabian Gieseke

http://www.astro.rub.de/polsterer/ADASS2011b.pdf


**ADASS 2011**, Paris, France, November 2011.