



EFIGI

Anthony Baillard

Projet EFIGI
Classification automatique de
galaxies

Journées PaRISTIC
au 21/11/2005

Institut d'Astrophysique de Paris,
UMR7095 CNRS
98bis, boulevard Arago
75014 Paris

ACI-MDA EFIGI
ACtion Incitative
Masse de Données en Astrophysique
<http://www.efigi.org>



Plan

- Présentation du projet
- Introduction à l'astronomie extragalactique
- Contexte scientifique
- Méthodologie
- Données
- Résultats
- Conclusion



EFIGI



Présentation du projet



Instituts partenaires

- 7 instituts
 - Institut d'Astrophysique de Paris (Emmanuel Bertin)
 - Laboratoire Traitement et Communication de l'Information (Henri Maître)
 - Laboratoire de Recherche de l'EPITA (Thierry Géraud)
 - Laboratoire d'Astrophysique de Marseille (Stéphane Arnouts)
 - Laboratoire d'Astrophysique de l'Observatoire Midi-Pyrénées (Pascal Fouqué)
 - Centre de Recherche Astronomique de Lyon (Georges Paturel)
 - Service d'Astrophysique du Commissariat à l'Energie Atomique (Alexandre Refregier)





Projet pluridisciplinaire

- Astronomie extragalactique
 - Acquisition de données
 - Expertise et spécification
 - Exploitation des résultats
- Traitement du signal
 - Traitement d'images
- Intelligence artificielle
 - Apprentissage automatique



Aspects concrets

- Objectif
 - Fournir un outil de classification automatique d'images quelconques de galaxies
- 1 an de travail
 - Collecte d'échantillons de données significatifs
 - Développement d'un prototype fonctionnel
 - Définition d'un système de classification
 - Développement d'une interface web de classification manuelle



EFIGI

Astronomie extragalactique

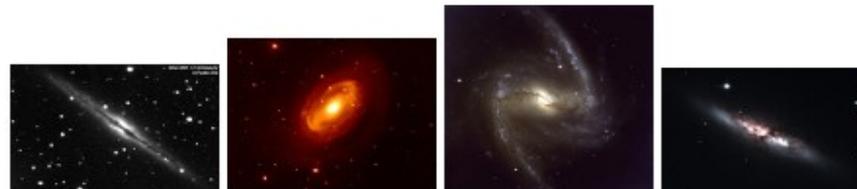


Galaxies

- Ensembles auto-gravitant en 3 dimensions
- Étoiles, gaz, poussière et matière noire (invisible)
- Forme
 - Composante sphéroïdale (bulbe, population âgée)
 - Disque (bras spiraux, population jeune)



(a) Elliptique (b) Elliptique (c) Lenticulaire (d) Lenticulaire (e) Spirale de face

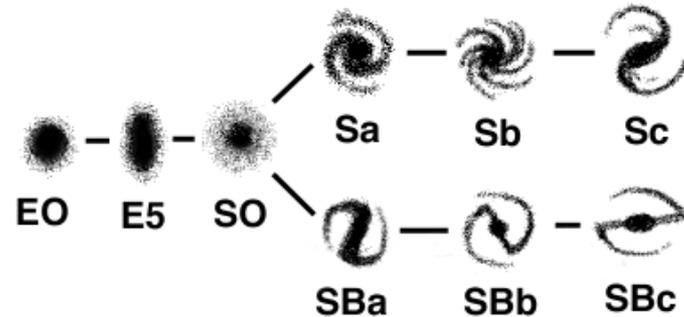


(f) Spirale par la tranche (g) Spirale (h) Spirale barree (i) Irreguliere



Classifications

- Hubble (*tuning fork*, 1936)
 - Elliptiques E0-E5
 - Lenticulaires S0
 - Spirales Sa-Sc
 - Spirales barrées Sba-SBc
 - Irrégulières
- Morgan (1958,1959)
- DDO (Van Den Bergh 1960)
- *Revised Hubble Type* (de Vaucouleurs 1959)
- *Revised morphological types* (1981)





EFIGI

**Contexte
scientifique**



Travaux existants

- Buta et Combes (1996)
 - Revue de caractéristiques morphologiques
- Abraham, Van Den Bergh et Nair (2003)
 - Coefficient de Gini
- Goderya et al. (2004)
 - Caractéristiques de forme globale
- Naim (1996)
 - Galaxies singulières (*peculiar*)
- Refregier (2003) et Bernstein et Jarvis (2002)
 - Shapelettes
- Kelly et McKay (2004)
 - Shapelettes multi-bandes



Objectifs

- Limitations des méthodes existantes
 - Dépendance aux conditions d'observation
 - Résolution, longueur d'onde, rapport signal sur bruit
- Gestion d'images de qualité variable
 - Base de projection
 - Convolution par la réponse impulsionnelle
- Classification
 - Type de Hubble
 - Détection d'attributs spécifiques
 - Bandes de poussières, anneaux, ...



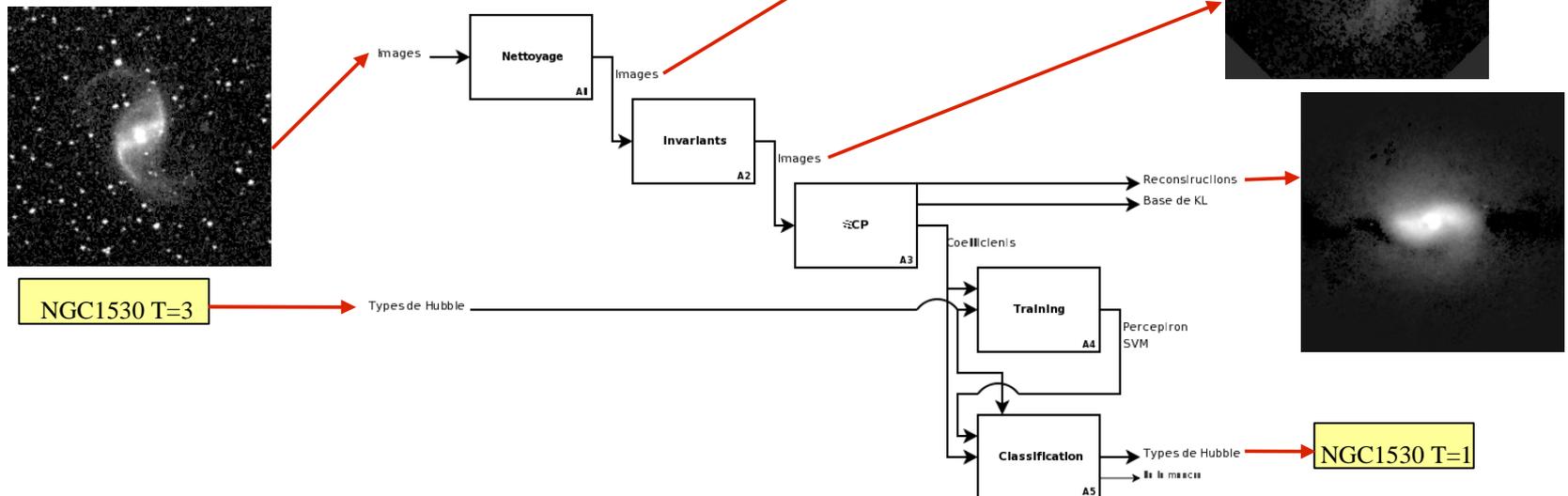
EFIGI

Méthodologie



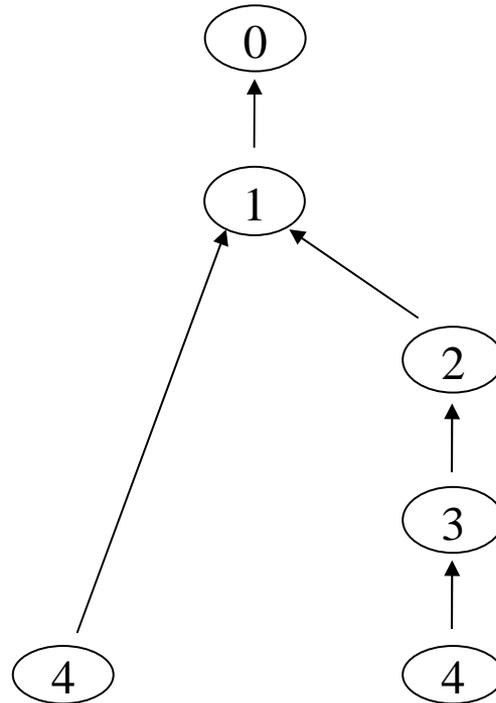
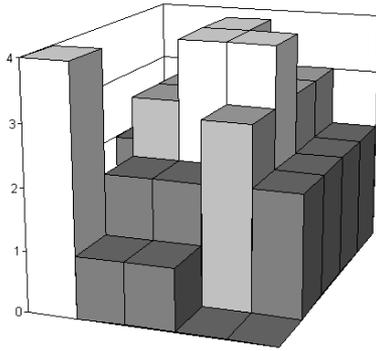
Processus global

- “Nettoyage”
- Réduction de dimensionnalité
 - Calculs d'invariants
 - Analyse en Composantes Principales (ACP)
- Apprentissage supervisé





Nettoyage (1/2)



- 1 nœud représentatif par niveau et par composante
- Information
 - Niveau, étiquette, aire, etc.
- Dépend de la relation de voisinage choisie
- Lignes de niveau



Nettoyage (2/2)

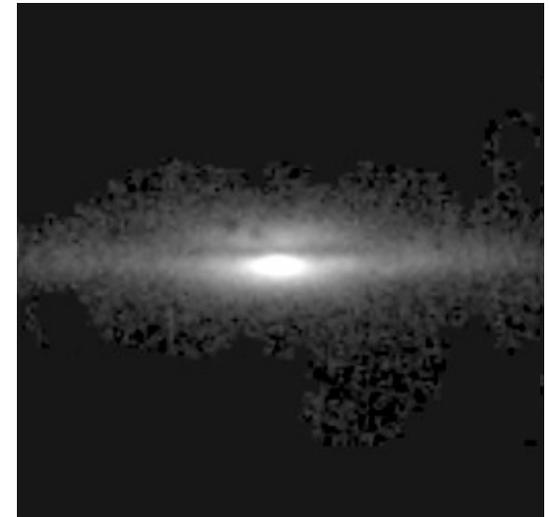
- Opérateurs successifs
 - Calcul de l'arbre des composantes C4
 - Ouverture d'aire (paramétrable)
 - White-top-hat
 - Seuillage (paramétrable)
 - Étiquetage
 - Élimination de la composante centrale
 - Dilatation C8
 - Minimum local





Calcul d'invariants

- Remise en forme des images
 - vecteurs ne dépendant que de la forme des galaxies et non de leur présentation
- Étapes
 - Soustraction du fond de ciel
 - Norme en variance
 - Centrage (barycentre itératif)
 - Rotation (moments du 2ème ordre)
 - Recalage d'échelle (moments du 2ème ordre)
 - Rognage
 - Retournements (moments du 3ème ordre)
 - Retournements (asymétrie rotationnelle)





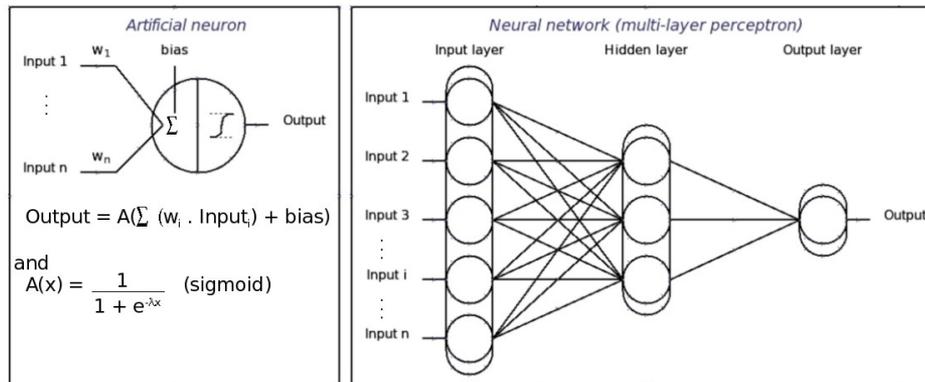
ACP

- Transformée de Karhunen-Loève
- Maximisation de la variance de l'échantillon
 - $\mathbf{u} = \mathbf{W}\mathbf{x}$
 - \mathbf{u} vecteur projeté de dimension \mathbf{m}
 - \mathbf{W} matrice de passage (Base de Karhunen-Loève)
 - \mathbf{x} vecteur original de dimension \mathbf{d}
 - Individus : \mathbf{d} images
 - Variables : \mathbf{s} pixels ($\mathbf{w} \times \mathbf{h}$)
 - Matrice \mathbf{x} : $\mathbf{d} \times \mathbf{s}$
- Espace dual (données ou attributs)
- Vecteurs de \mathbf{u} décorrélés



Apprentissage supervisé 1/2

- Perceptron multi-couches

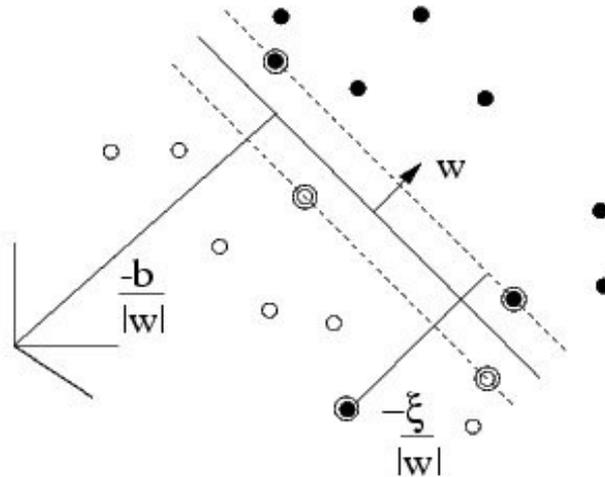


- Classifieur "Précoce/Avancé"
 - 2x2x2
- Classifieur type T de Hubble
 - $30 \times N_H \times 1$
 - N_H généralement compris entre 15 et 70



Apprentissage supervisé 2/2

- SVM (Support Vector Machine)
 - Calcul d'un hyperplan de séparation linéaire entre 2 classes
 - Notion de marge
 - Utilisation de points significatifs (Vecteurs de support)
 - Positifs
 - Négatifs
 - Projection dans un espace de dimension possiblement infinie
 - Fonction noyau à base radiale (RBF) généralement gaussienne





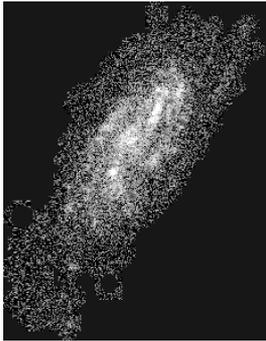
EFIGI

Données

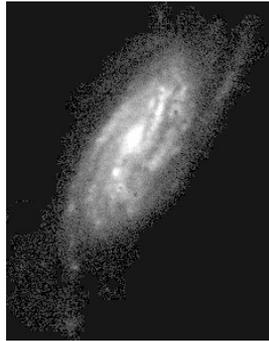


Images

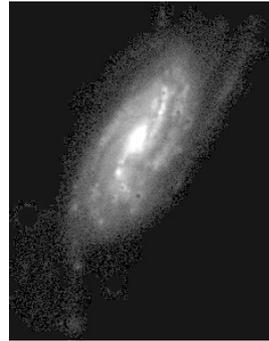
- SDSS (*Sloan Digital Sky Survey*)
 - Homogène, récent, de qualité et multi longueur d'ondes
 - *SDSS deblender*
 - Utilisation des trois bandes centrales



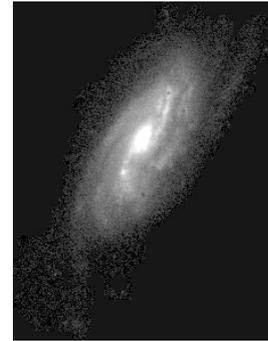
u
(ultraviolet)



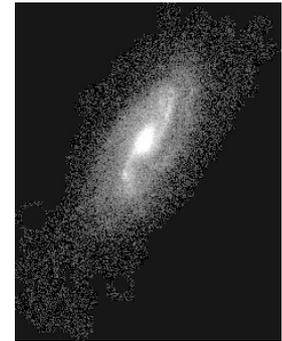
g
(bleu-vert)



r
(orange)



i
(très proche infrarouge)



z
(proche infrarouge)



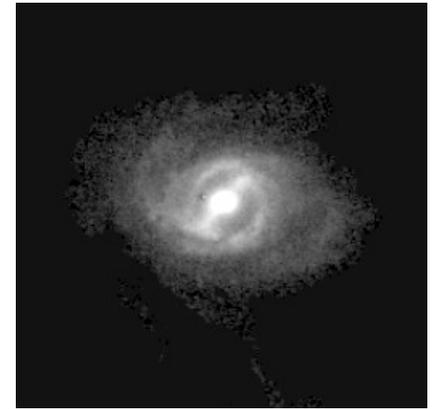
Classification

- PGC (Principal Galaxies Catalog)
 - Objets limités en diamètre
 - *Revised morphological types*
 - Utilisation du type de Hubble
- Revised Morphological Types
 - NASA Goddard Space Flight Center
 - Préparé par Theresa A. Nagy et Robert S. Hill (Avril 1981)
 - Révisé par Suzan E. Gessner and Lee E. Brotzman (Décembre 1990)
 - Fusionne les classifications existantes
- Jeu de données actuel
 - 774 galaxies
 - Utilisation de la bande g



Revised Morphological Types

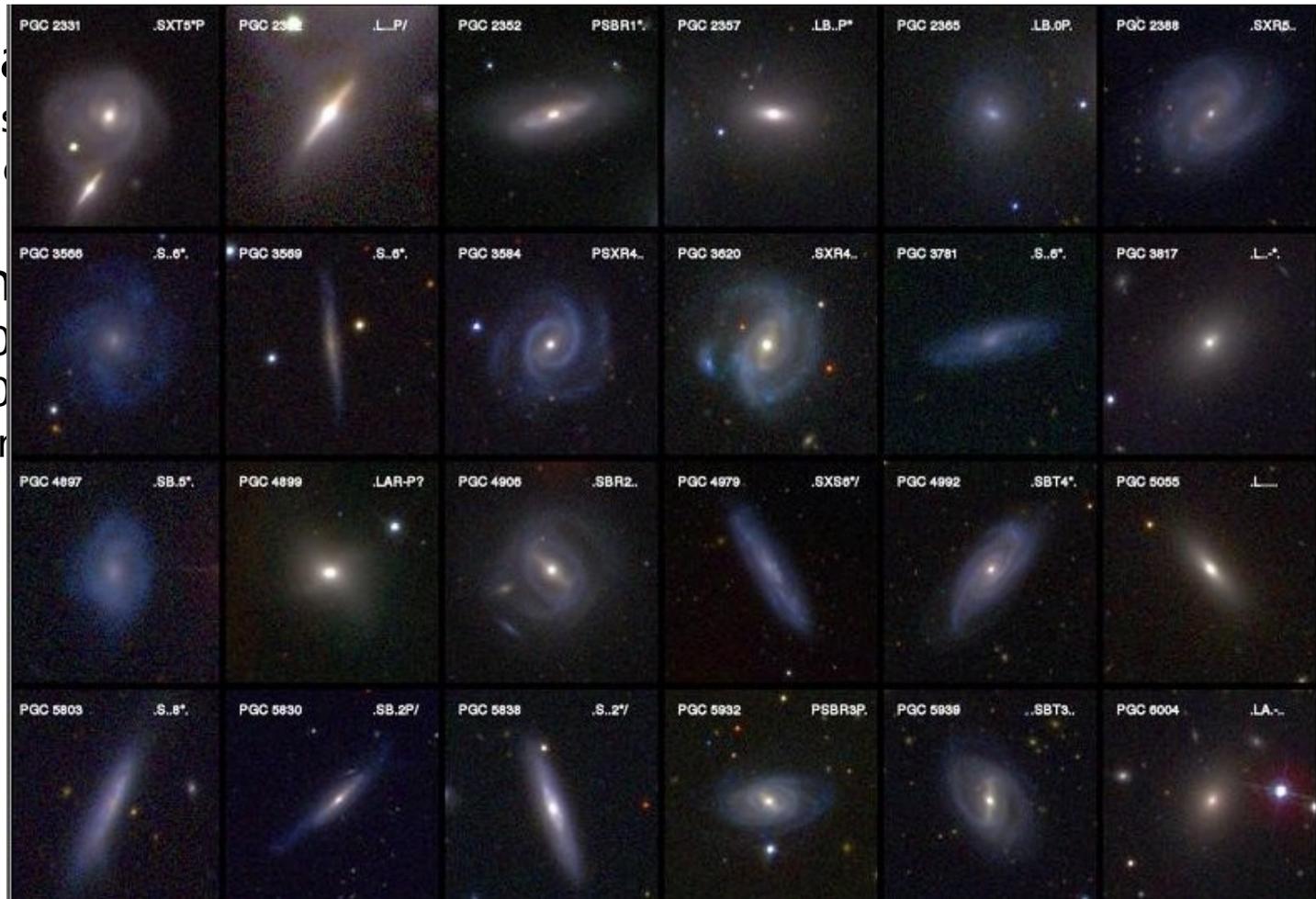
- 5 caractères principaux
 - Classe, Famille, Variété, Stade
 - Spirale, Barrée, Anneau interne, $b = .SBR3$
- 1 caractère supplémentaire en fin de chaîne
 - Incertain *
 - Douteux ?
- 1 caractère supplémentaire en fin de chaîne
 - *Peculiarities P*
 - *Spindle S*
- Options pour le 1er caractère
 - Compact C, Anneau R ou pseudo-anneau P





Collecte de données

- Grande quantité de données
 - Millions de galaxies
 - Seuls les objets les plus intéressants sont sélectionnés
- Echantillon représentatif
 - 11 000 galaxies
 - 55 000 galaxies
 - 2 To de données



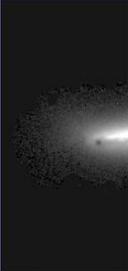


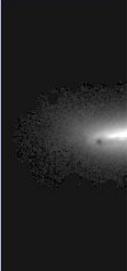
Classification manuelle

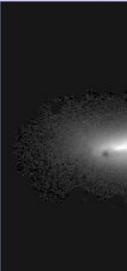
You are currently logged in as **Anthony Baillard**
[Main page](#)[Instructions](#)[How to classify](#)

Data set (Click on the dropdown menu)

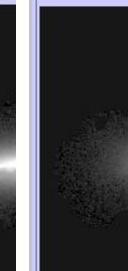
You are currently logged in as **Anthony Baillard**
[Main page](#)[Instructions](#)[How to classify](#)

PGC: 

PGC: 

PGC: 

PGC: 

PGC38212_g 

Summary

Result

- Type: Face-on spiral
- Stage: Scd
- Cross section: AB(r)
- Arms thickness: 2/4
- Arms flocculence: 2/4
- Arms curvature: 2/4
- Arms regularity: 2/4
- Dust band: 2/4
- Multiplicity: 1
- Perturbation: 2/4
- Contamination: 2/4

Contamination



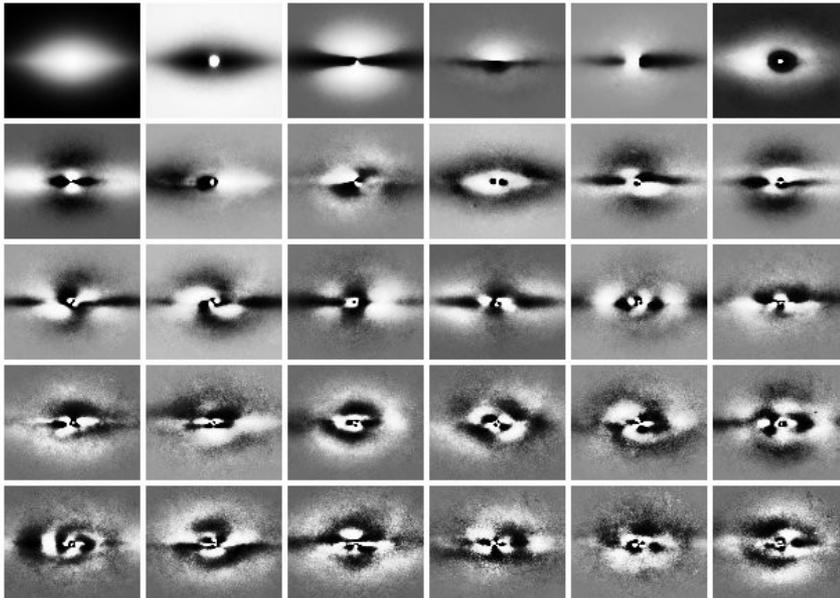


EFIGI

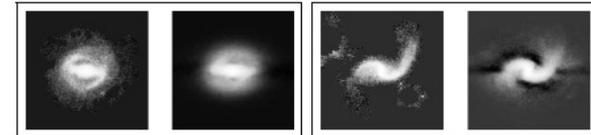
Résultats



Bases de projection

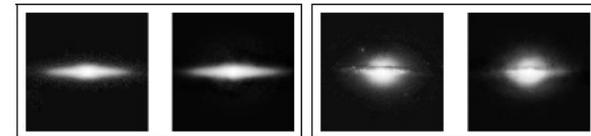


30 premières composantes d'une base de KL.



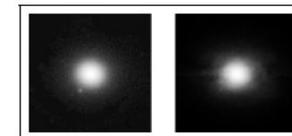
(a) Anneau

(b) Spirale de face



(c) Spirale par la tranche

(d) Bande de poussière



(e) Elliptique

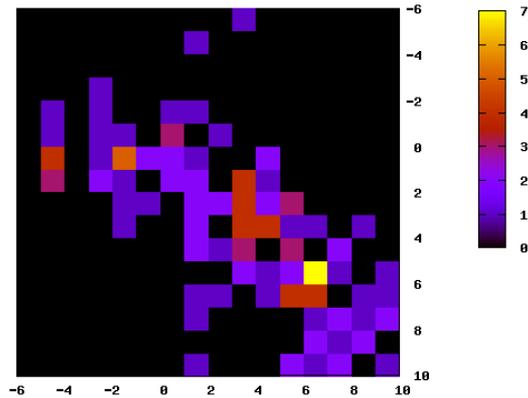
Reconstruction d'images à partir de 30 composantes de KL.



Performances de classification

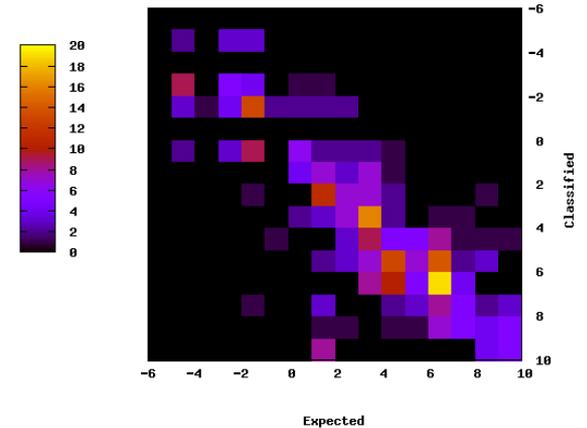
MLP

68% à deux types de Hubble près



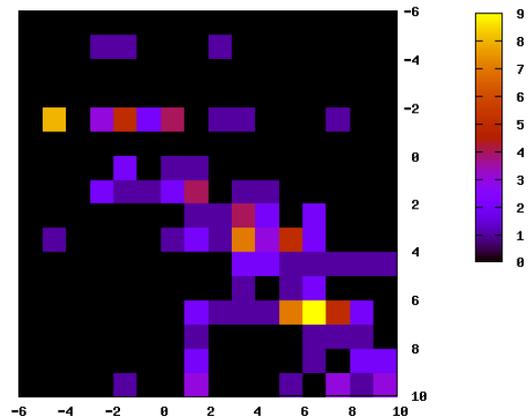
Astronome

Hubble type - Classified against Expected



SVM

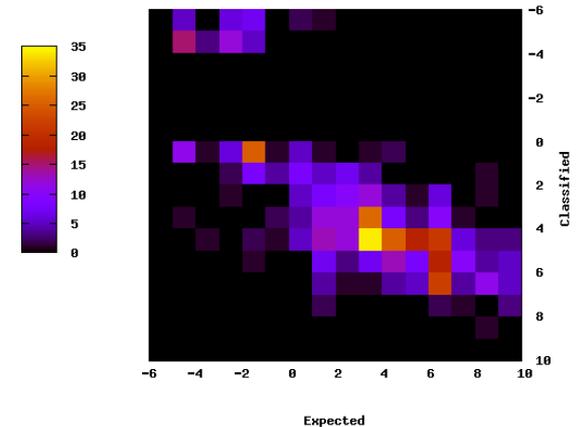
75% à deux types de Hubble près



Astronome

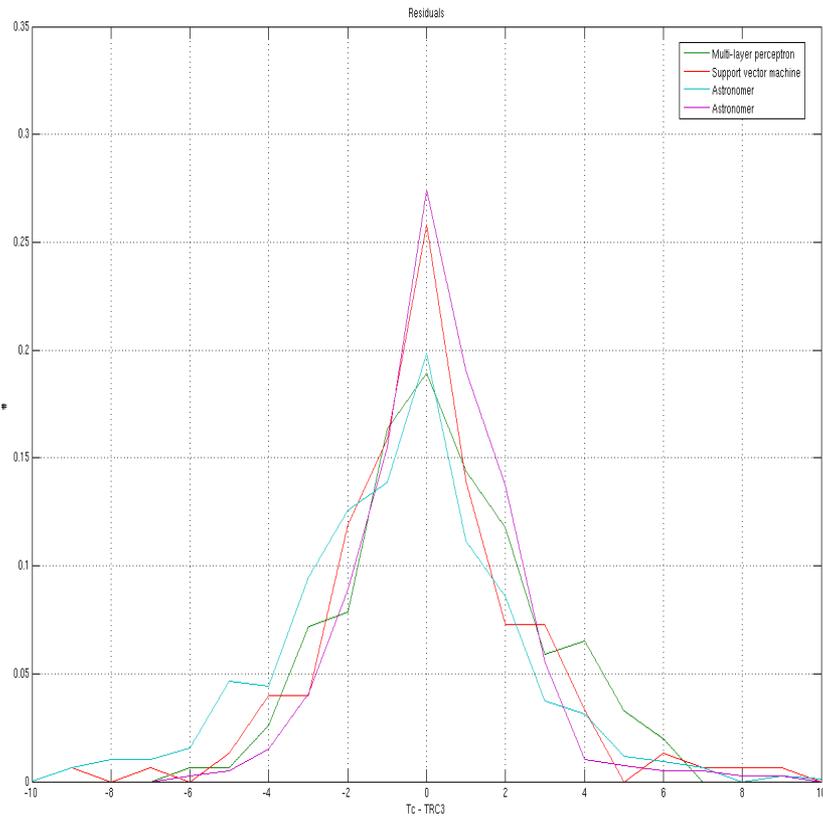
Mise en évidence d'un biais

Hubble type - Classified against Expected





Comparaisons



Courbes de résidus
Type T donné – Type T du RC3

- σ_1 et σ_2 (Odewahn 2004)
- Régression linéaire
 - Pente
 - Déviation standard σ_1

$$\sigma_2 = \frac{\sum (T_c - T_{RC3})^2}{(N-1)}$$

Classifieur		σ_1	σ_2
Astronome (violet)	0,869	0,57	2,03
Astronome (cyan)	0,646	1,49	3,18
MLP	0,810	0,83	2,97
SVM	0,833	0,70	2,66
Odewahn 2004	0,946	1,83	1,54



EFIGI

Conclusion



Critique

- + Classificateur opérationnel
- + Reconstructions efficaces
 - 30 composantes de KL contre 160 pour les shapelettes
- Performances de classification insuffisantes
 - Spirales
- Limitations de la classification des experts
- Type *T* de Hubble
- Galaxies bien résolues



Perspectives

- Nouveau jeu de données
 - 10 000 images du SDSS en 5 bandes
 - Classification de référence erronée
- Nouveaux invariants
 - Départ des bras spiraux
- Nouvelles bases de projection
 - Analyse en Composantes Indépendantes
 - Base analytique
- Nouveaux systèmes de classification
 - k -moyennes
 - Sélection de caractéristiques sur la base de KL (Marie Lienou)



Contacts et remerciements

- <http://www.efigi.org>
- Coordinateur Emmanuel Bertin bertin@iap.fr
- Ingénieur EFIGI baillard@iap.fr
- Merci à
 - Emmanuel Bertin, Marine Campedel, Reda Dehak, Pascal Fouqué, Thierry Géraud, Jean-François Le Borgne, Marie Lienou, Henri Maître, Yannick Mellier, Roser Pelló
 - L'équipe TERAPIX à l'IAP
 - Les collaborateurs d'EFIGI