Intelligence artificielle et apprentissage automatique en astronomie

Alain Brémond

Importance en astronomie

- Avec le seul mot-clef : « neural network » et de janvier 2021 à août 2022 : 1 719 publications.

Introduction

- Nécessité en astronomie :
  - Complexité des données astronomiques
  - Taille des données à traiter

• Complexité des données astronomiques

- **Initiales** : longueurs d’ondes étendues
  - Fréquentes : IR/visuel, radio
  - Plus rares : UV, gamma etc,
- **Secondaires** : catalogues de propriétés mesurées (GAIA),
- Variations dans le temps.
- Simulations numériques.

Taille des données à traiter

- Le giga octet (1 milliard d’octets) est largement dépassé,
- Les 100 tera octets (100 X 10^{12} octets) ont été dépassés il y 20 ans.
- On va atteindre les exa octets (10^{18} octets) !

Pour traiter ces masses de données...

- Une première étape : le **data mining**.
- Il permet :
  - De traiter le volume
  - Et la diversité des sources d’information (ex. la NSA).

Exemples en astronomie :

- Le SDSS (Sloan Digitized Sky Survey) :
  - découverte d’étoiles à grandes vitesses propres,
  - découverte d’une galaxie naine satellite de notre Galaxie (cachée) par les étoiles.
- Et bien d’autres découvertes.

Les grandes méthodes de l'IA

- Apprentissage supervisé, **apprentissage automatique**
- Apprentissage non supervisé, **intelligence artificielle**, apprentissage profond

Apprentissage supervisé

- Un humain a classé les données sur un échantillon de taille réduite, en fonction de variables qui ont été mesurées. Le logiciel sera ensuite capable de faire le même travail sur un très grand fichier.
  - Exemples : Iris, étoiles ou galaxies ?

Apprentissage non supervisé

- On fournit au logiciel une base de données et celui-ci devra classer ces données sans l'aide d'un humain.
  - Exemple : classement de photos, découverte d'objets anormaux dans une base de données astronomiques, classement des types d'objets dans le Hubble Deep Field South....

Quelques exemples de logiciels

- Les plus simples, méthodes statistiques :
  Régressions linéaire multiple, régression logistique...
- Les plus utilisés en astronomie :
  - Réseaux de neurones et leurs dérivés
  - Arbres de décision et dérivés
- Et beaucoup d’autres.

Explications

- Réseaux de neurones
- Arbres de décision

Aujourd’hui : une grande diversité d’études et de méthodes

- Plan
  - Les données astronomiques
  - Les objectifs de l’IA
  - Les techniques
  - Où en est -on ?

Les données astronomiques sont vite volumineuses

- 1- Les images
- 2- La spectroscopie
- 3- La photométrie
- 4- Les données temporelles

Les images

- En N et B : grille de pixels avec les coordonnées ad et dec et un flux (0-256)

- En couleur : un cube de données avec 4 tables superposées : couches RVB et BN.

La spectroscopie

- C’est une image en pixels (table de valeurs)
- Ou un cube de données : une image par longueur d’onde.

La photométrie

- Mesures précises d’un flux après calibration
- Utilisation de filtres

Les séries temporelles

- Variations d’intensité d’une source :
  - Exemples : étoiles variables, détection d’une planète extrasolaire,
  - En radioastronomie : pulsars,
  - Ondes gravitationnelles,
  - Simulations.....

Les buts

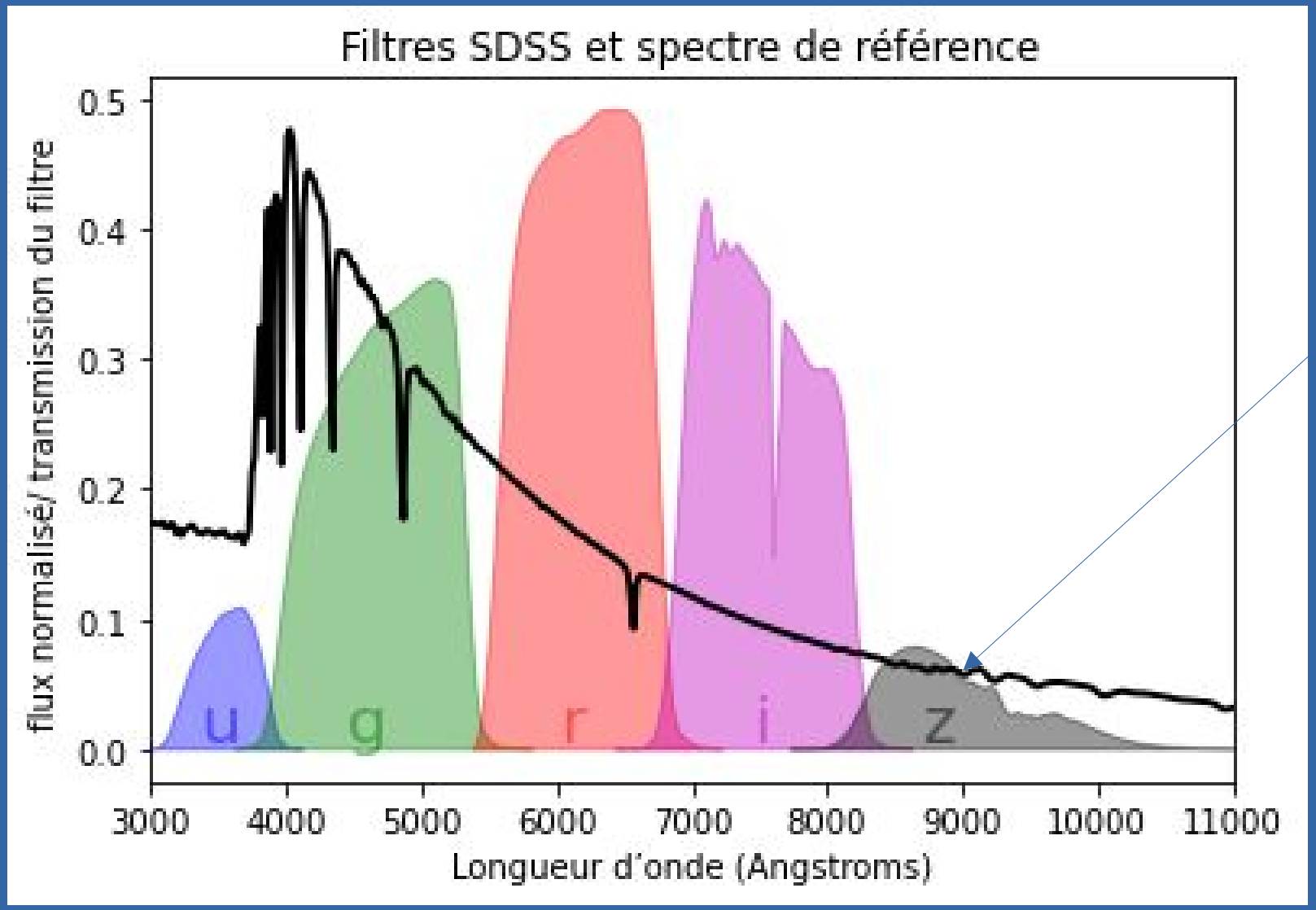
- Classification
- Régression
- Regroupement
- Prévision
- Génération de données manquantes
- Découverte
- Nouvelles connaissances

Classification

- Dans une exploration photographique du ciel : identifier des catégories d’objets.
  - - A partir des images
  - - Ou à partir de valeurs mesurées sur ces objets.

Exemple personnel (simple) de classification

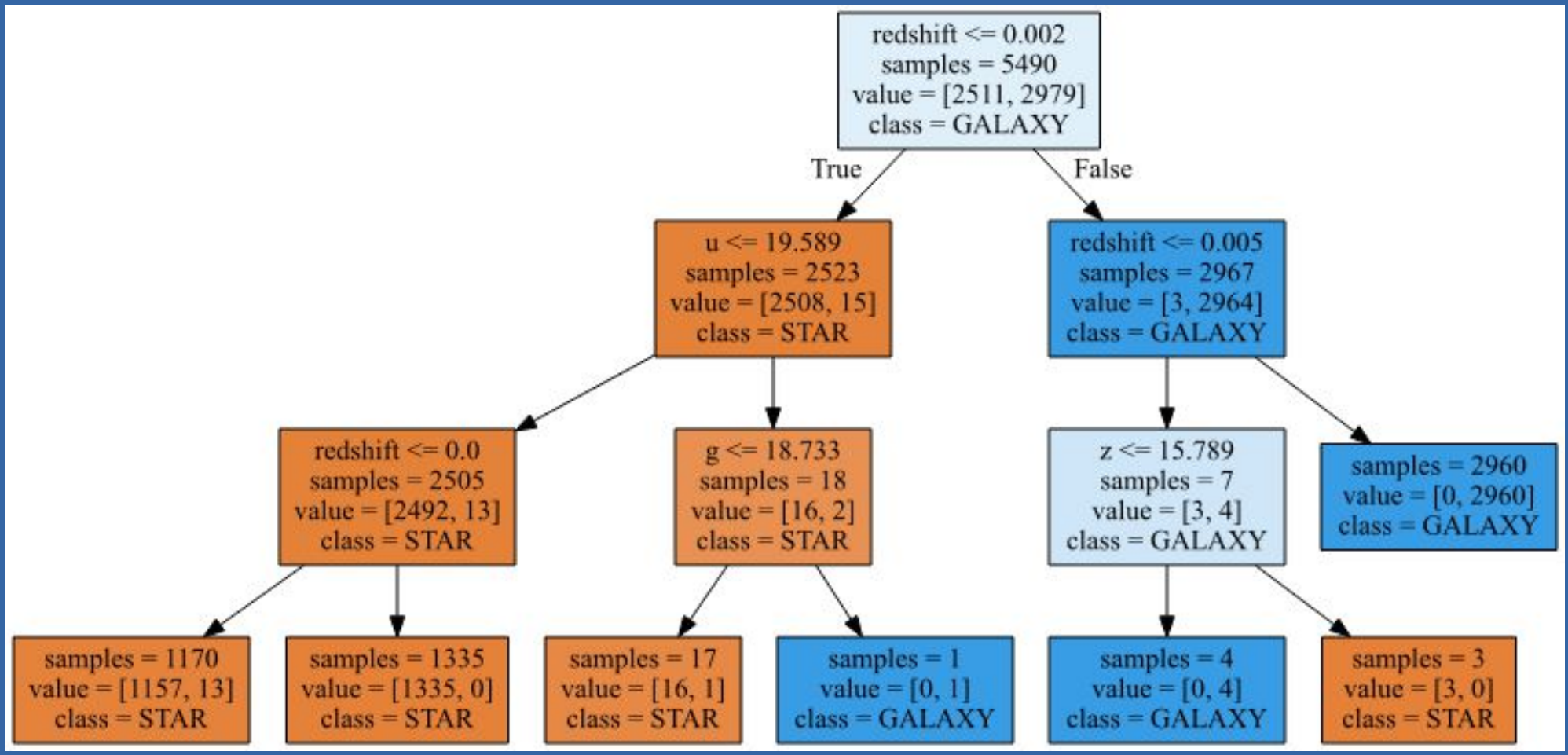
- Une base de donnée avec :
  - 4998 galaxies
  - 4152 étoiles
  - 850 quasars

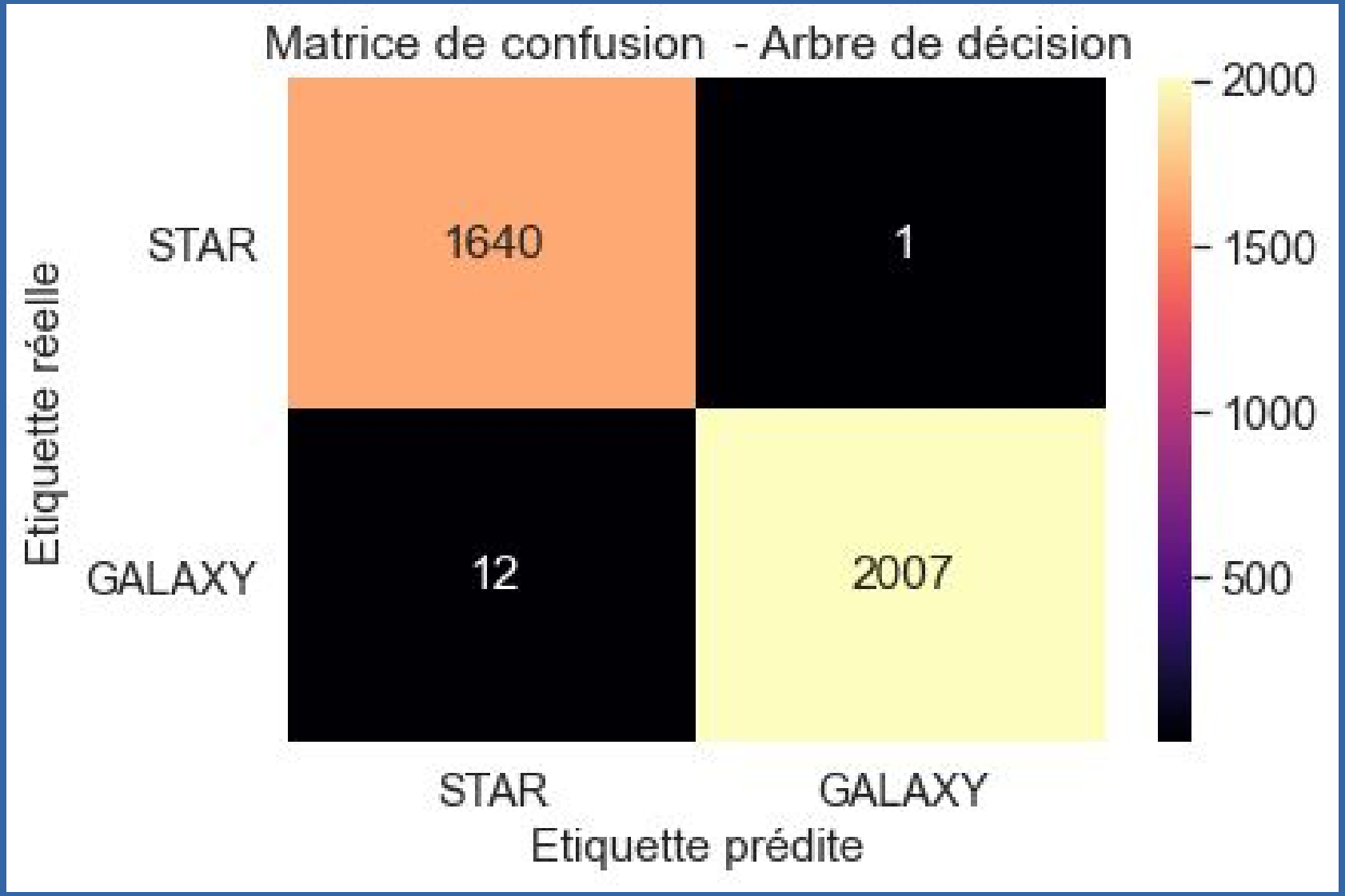


En noir : spectre de Vega

- Apprentissage **supervisé** pour distinguer les étoiles des galaxies
- Les variables :
  - Redshift
  - Magnitudes du système u g r i z
- Un échantillon tiré au sort pour l'apprentissage (6862), le reste pour tester le modèle (2288).
- La méthode : les arbres de décision

- Résultats :
  - Sur le fichier d’apprentissage : 99.74 % de bien classées
  - Sur le fichier de test : 99.64 %
  - C’est cette performance qu’on s’attend à trouver lorsqu’on voudra classer de **nouveaux objets**.


redshift <= 0.002
samples = 5490
value = [2511, 2979]
class = GALAXY
True
False
u <= 19.589
samples = 2523
value = [2508, 15]
class = STAR
redshift <= 0.005
samples = 2967
value = [3, 2964]
class = GALAXY
redshift <= 0.0
samples = 2505
value = [2492, 13]
class = STAR
g <= 18.733
samples = 18
value = [16, 2]
class = STAR
z <= 15.789
samples = 7
value = [3, 4]
class = GALAXY
samples = 2960
value = [0, 2960]
class = GALAXY
samples = 1170
value = [1157, 13]
class = STAR
samples = 1335
value = [1335, 0]
class = STAR
samples = 17
value = [16, 1]
class = STAR
samples = 1
value = [0, 1]
class = GALAXY
samples = 4
value = [0, 4]
class = GALAXY
samples = 3
value = [3, 0]
class = STAR


Matrice de confusion - Arbre de décision
STAR
GALAXY
1640
1
12
2007
Etiquette réelle
STAR
GALAXY
Etiquette prédite
2000
1500
1000
500

Régression

- Ici, on ne veut plus définir des catégories mais une valeur numérique continue : par exemple une température en fonction de paramètres mesurées au télescope.

Regroupement (clustering)

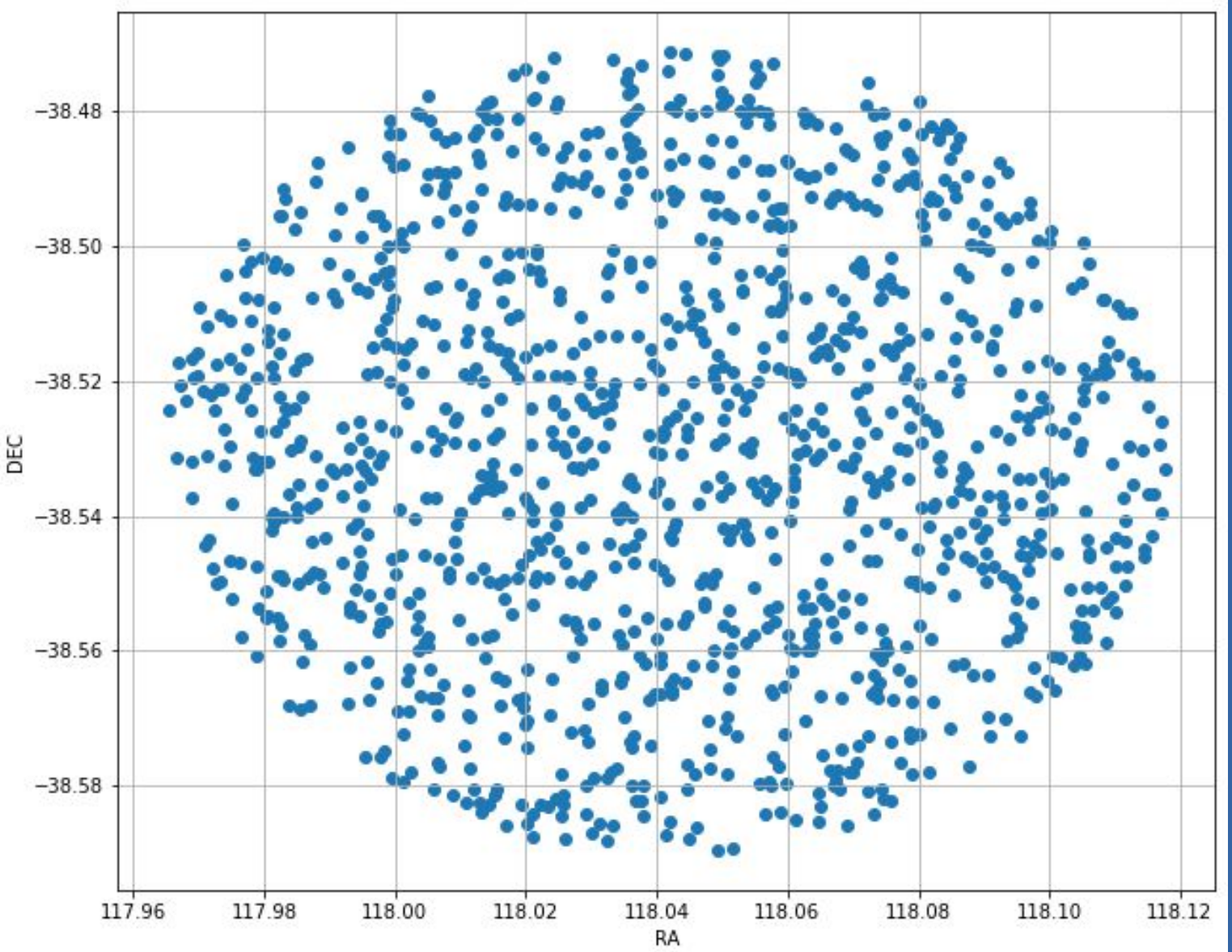
- Une étoile appartient-elle à une amas ?
- 2e exemple personnel avec Gaia DR2 : rechercher quelles étoiles dans un champ donné peuvent appartenir à un amas.

Exemple personnel

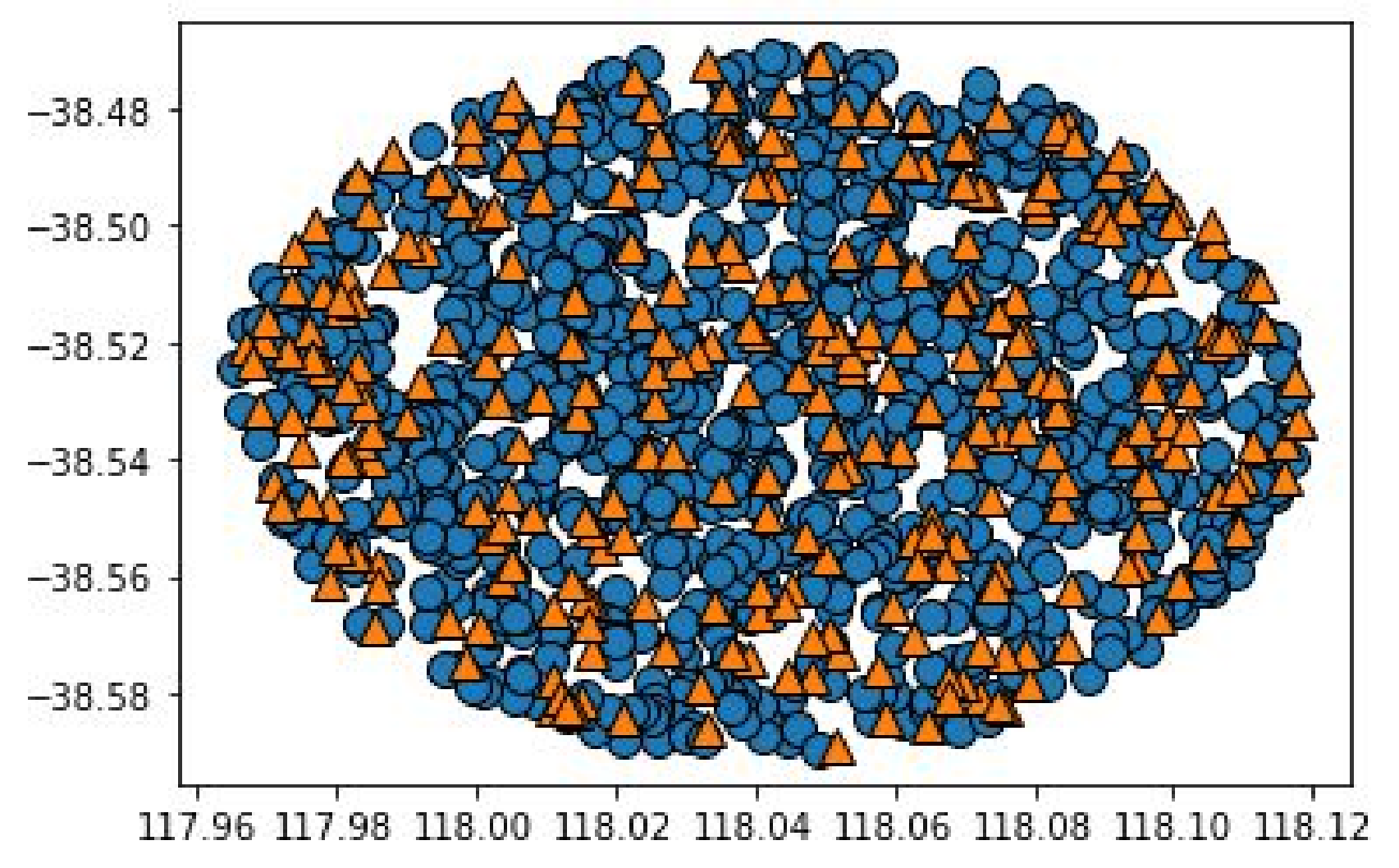
- GAIA DR2
- Sélection autour de NGC 2477, amas ouvert
- Critères :
  - Ad : 118.04, dec : -38.53 cône de 0.06°, phot mag g <=20

- Paramètre stellaires utilisés :
  - Ra, dec, parallaxe, mouvements propres, indice bp-rp

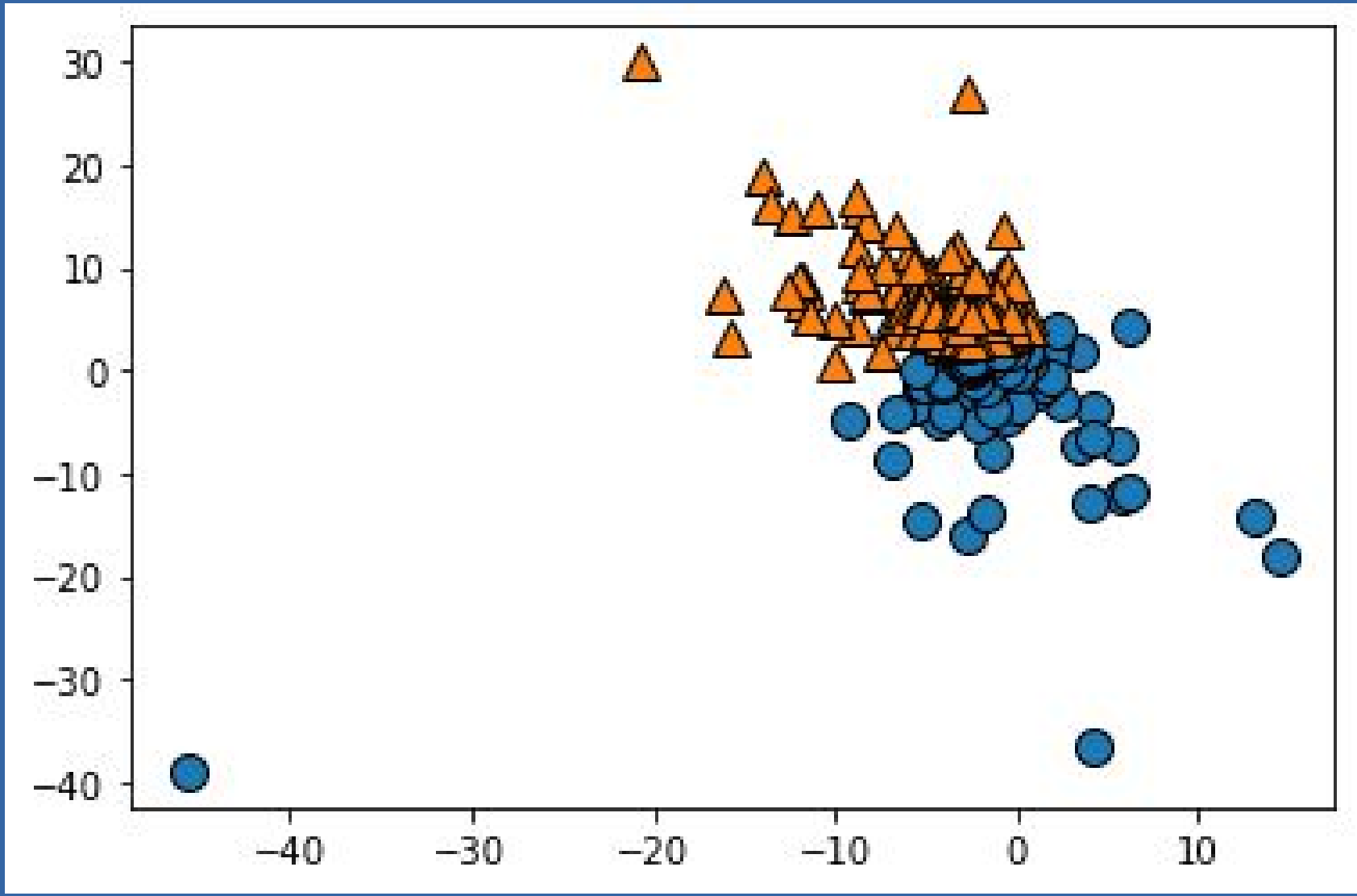
- On obtient 1 238 étoiles :

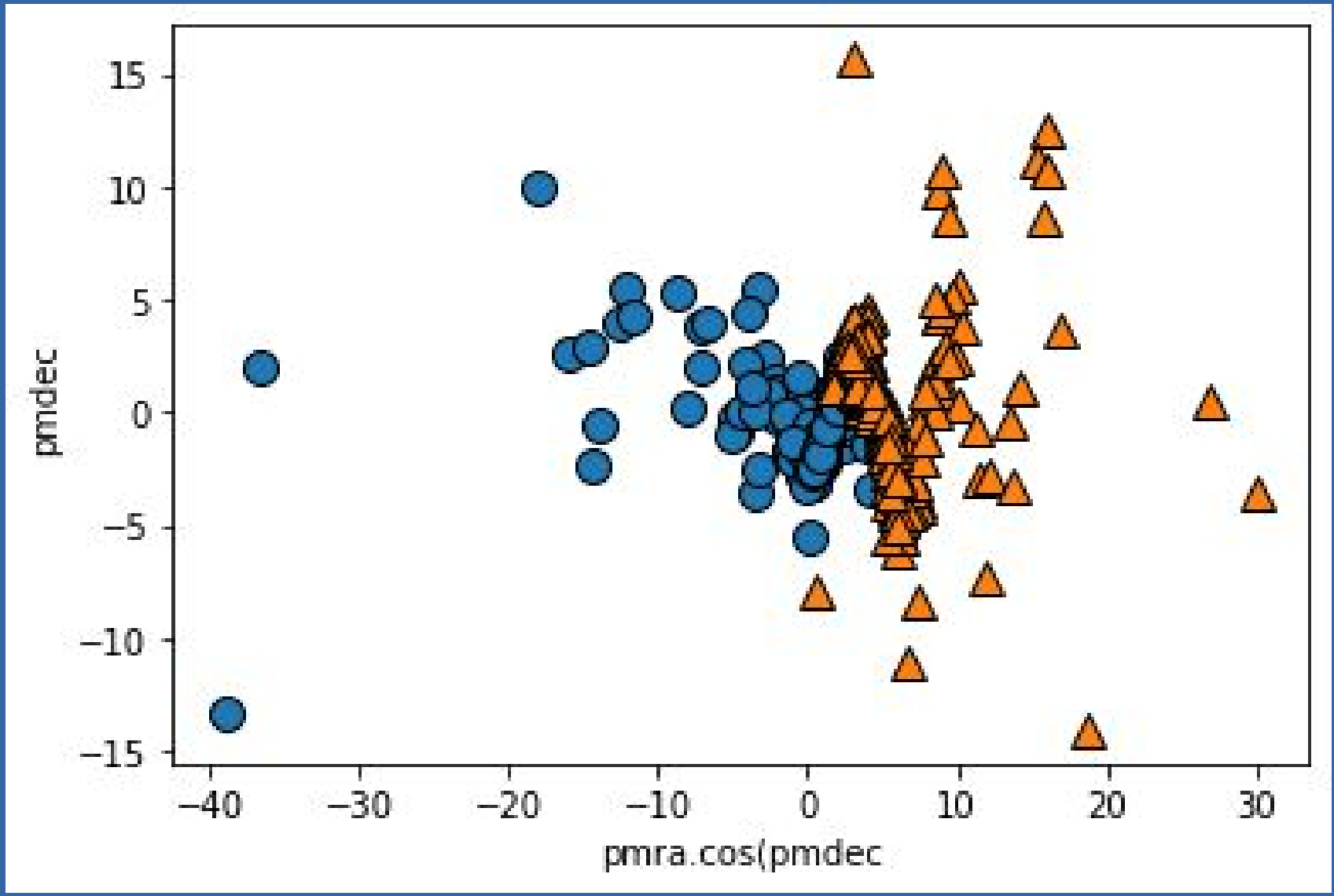

−38.48
−38.50
−38.52
−38.54
−38.56
−38.58
DEC
117.96
117.98
118.00
118.02
118.04
118.06
118.08
118.10
118.12
RA

Application de la méthode de recherche d’amas (agglomerative clustering).

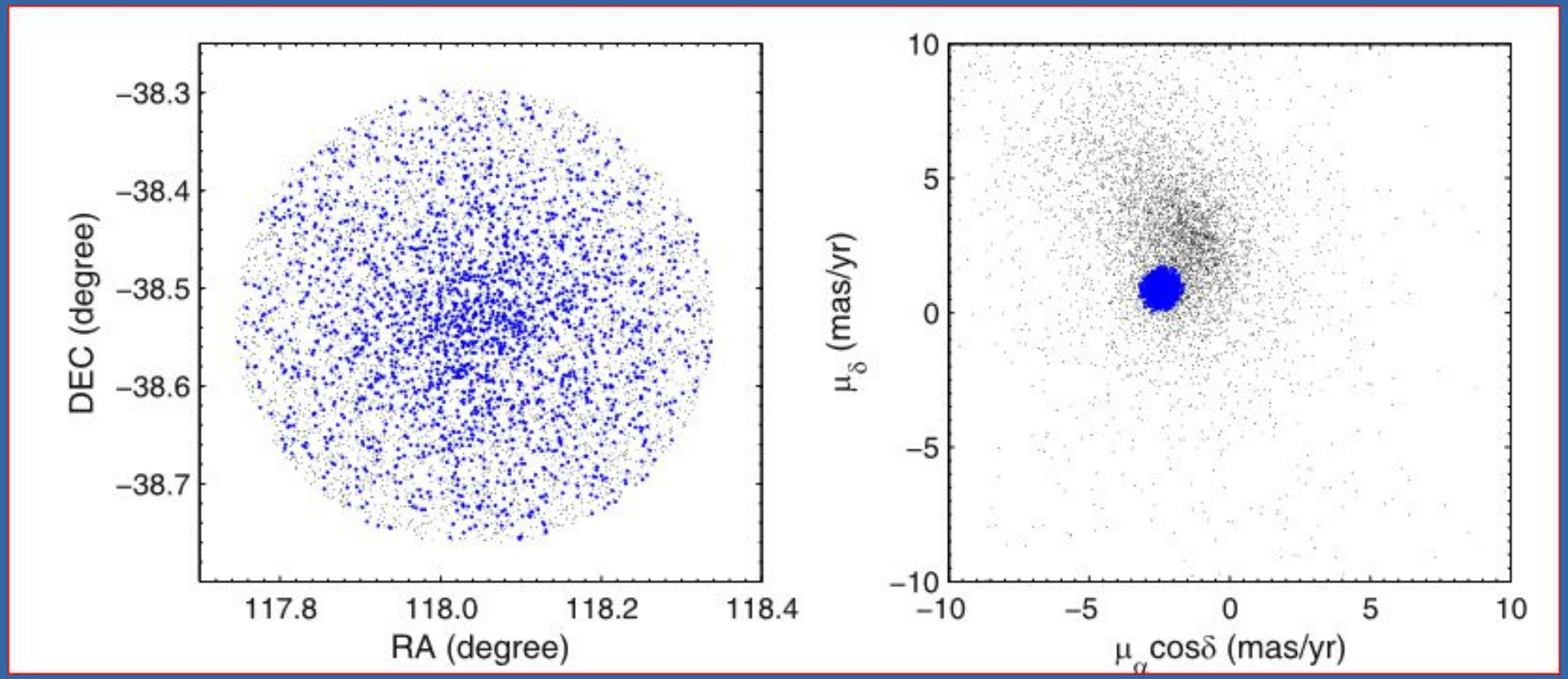

−38.48
−38.50
−38.52
−38.54
−38.56
−38.58
117.96
117.98
118.00
118.02
118.04
118.06
118.08
118.10
118.12

- On affiche les mouvements propres des étoiles (l’amas est en bleu):


30
20
10
0
−10
−20
−30
−40
−40
−30
−20
−10
0
10


15
10
5
0
−5
−10
−15
pmdec
−40
−30
−20
−10
0
10
20
30
pmra.cos(pmdec

- Avec beaucoup plus d’étoiles, un champ plus large, une autre méthode de clustering et d’autres paramètres (**et un autre ordinateur !**) :


−38.3
−38.4
−38.5
−38.6
−38.7
DEC (degree)
117.8
118.0
118.2
118.4
RA (degree)
10
5
0
−5
−10
μ_δ (mas/yr)
−10
−5
0
5
10
$\mu_\alpha \cos\delta$ (mas/yr)

Prévision

- A partir d’observations anciennes : prévoir l’évolution future : par exemple surveillance du Soleil pour prédire des éruptions dangereuses.

Reconstructions

- Dans une ensemble d’objets, toutes les mesures n’ont pas pu être réalisées ou un signal peut être localement bruité : on peut parfois reconstruire les données manquantes.

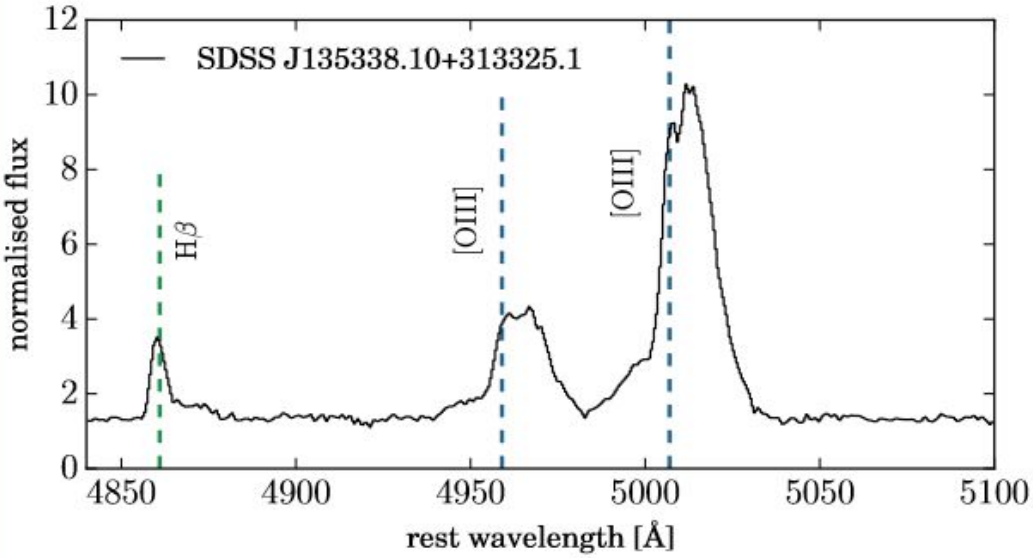
Découvertes

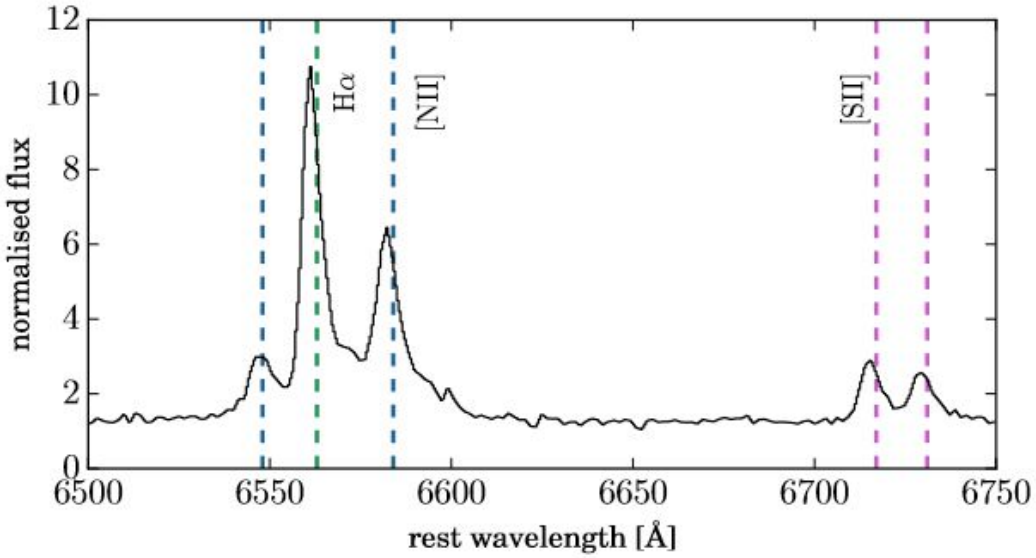
- Nouveaux objets : par exemple identification d’une galaxie naine satellite de la Galaxie parmi les milliards d’étoiles étudiées par Gaia.
- Ou découvertes d’anomalies ou de valeurs aberrantes. Des galaxies « aberrantes » dans le relevé SDSS.

Etude des objets « étranges » du SDSS

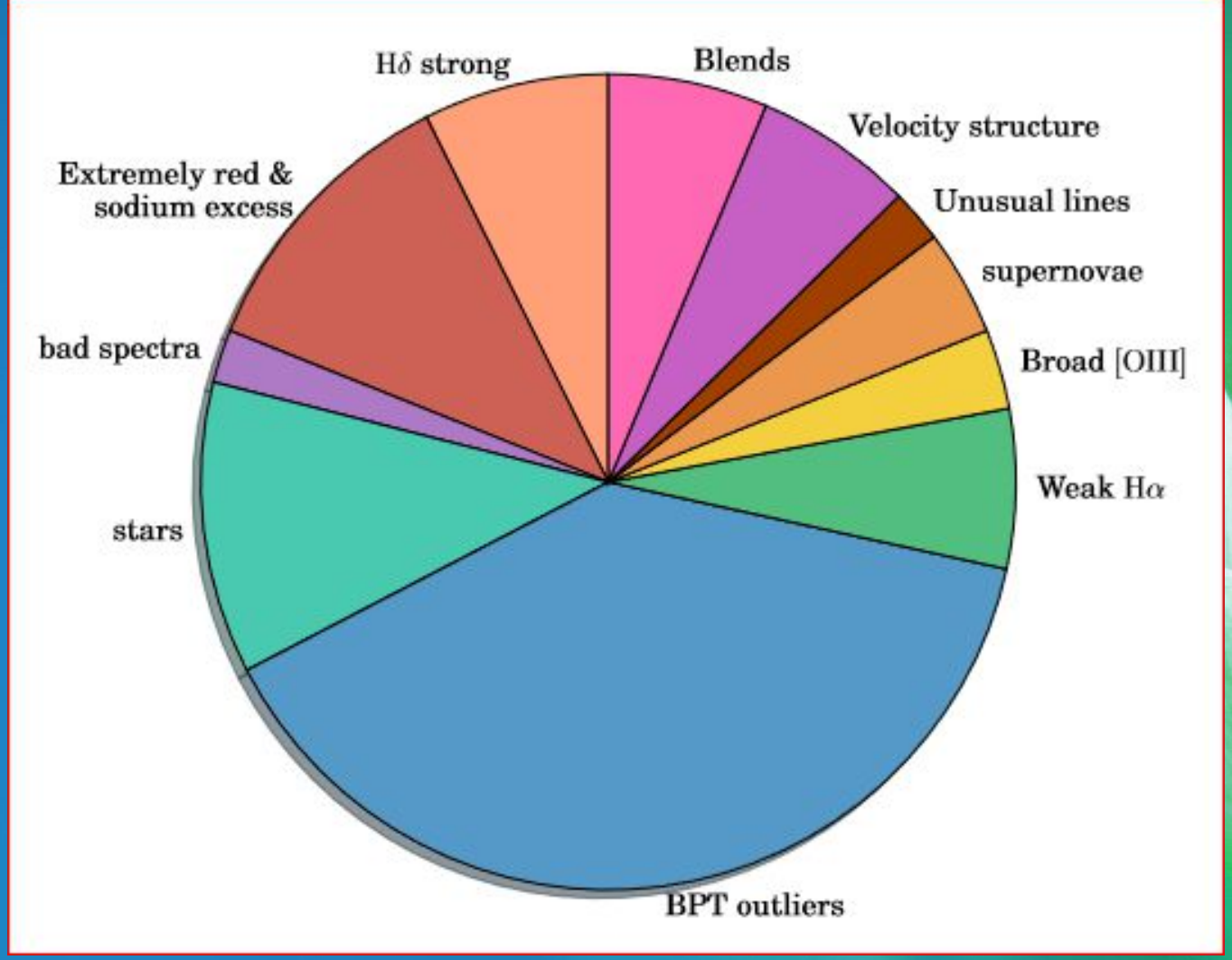
- Matériel : les spectres de toutes les galaxies du SDSS : 2 355 926 galaxies !
- Les données :
  - - les lignes : les objets
  - - Les colonnes les 15 700 longueurs d’onde.

- Utilisation de la technique des **forêts aléatoires** (RF) en IA non supervisée.
- Construction de données synthétiques à partir des distributions marginales (même taille)
- Puis application d’une IA supervisée (réelle vs synthétique).
- Cela va faire apparaître les spectres étranges.





Deux raies d’émission OIII et à D une raie décalée vers le bleu non identifiée



BPT Baldwin, Phillips, Terlevich diagramme ([OIII]/Hβ et [NII]/Hα)

- Intérêt : Une fois découvertes des catégories d’objets de type inconnus,
- On peut les extraire,
- Puis les étudier en détail.
- Cela peut permettre de faire de nouvelles découvertes sur la physique des galaxies.

Nouvelles connaissances

- Autre exemple : le problème de l’évolution des galaxies.
  - L’observation de toutes les galaxies laissent facilement imaginer une évolution en leur sein (cf Edwin Hubble).
  - Impossible à étudier individuellement car l’intervalle de temps serait trop long, même pour plusieurs vies humaines.

- Une équipe s’est intéressée à ce problème avec l’aide de l’IA, en particulier des réseaux neuronaux non supervisés
- combinés avec les simulations
- et les observations de larges bases de donnés (CANDELS du HST).

- **Etape 1** : Construire un échantillon de galaxies simulées formé à partir de données issues de la physique des galaxies.
- On fait évoluer chaque galaxie simulée (il y en a 28), de z=4 à z=1 en 19 images 3D.
- Des images 2D sont générées à partir de ces chaque simulation 3D.
- Chaque images 2D est « photographiée » sous 19 angles de vues différentes.
- Au total ~10 108 images.

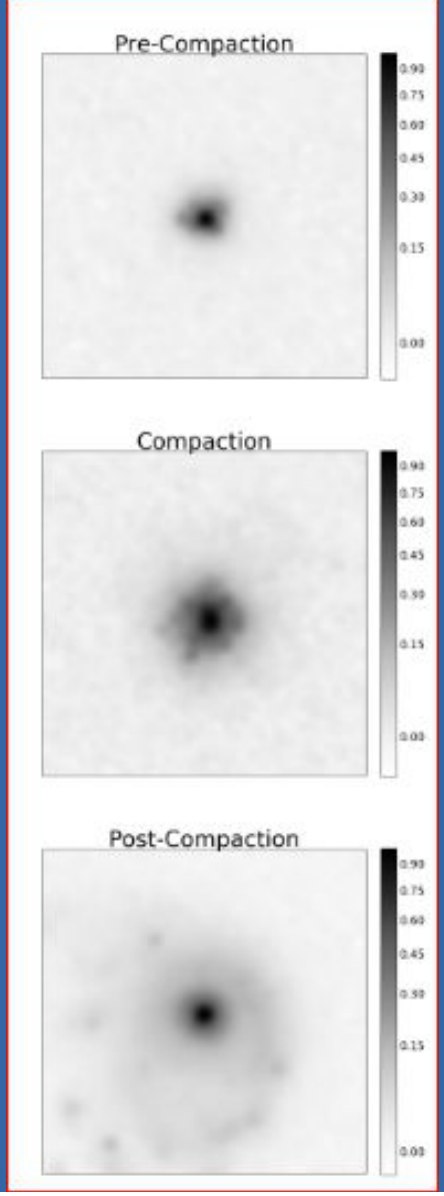
- **Etape 2** : sélection d’images de la base CANDELS du télescope Hubble de mêmes caractéristiques physiques que celles rentrées dans le modèle.

- **Etape 3** : Utilisation d’un réseau de neurones sur les données simulées pour identifier dans ces galaxies de haut redshift le phénomène dit « blue nuggets » : BN
- Détermination par le RN de trois phases : pré BN, BN et post BN. C’est un apprentissage **non supervisé**.
- Au total 10 modèles sont produits.

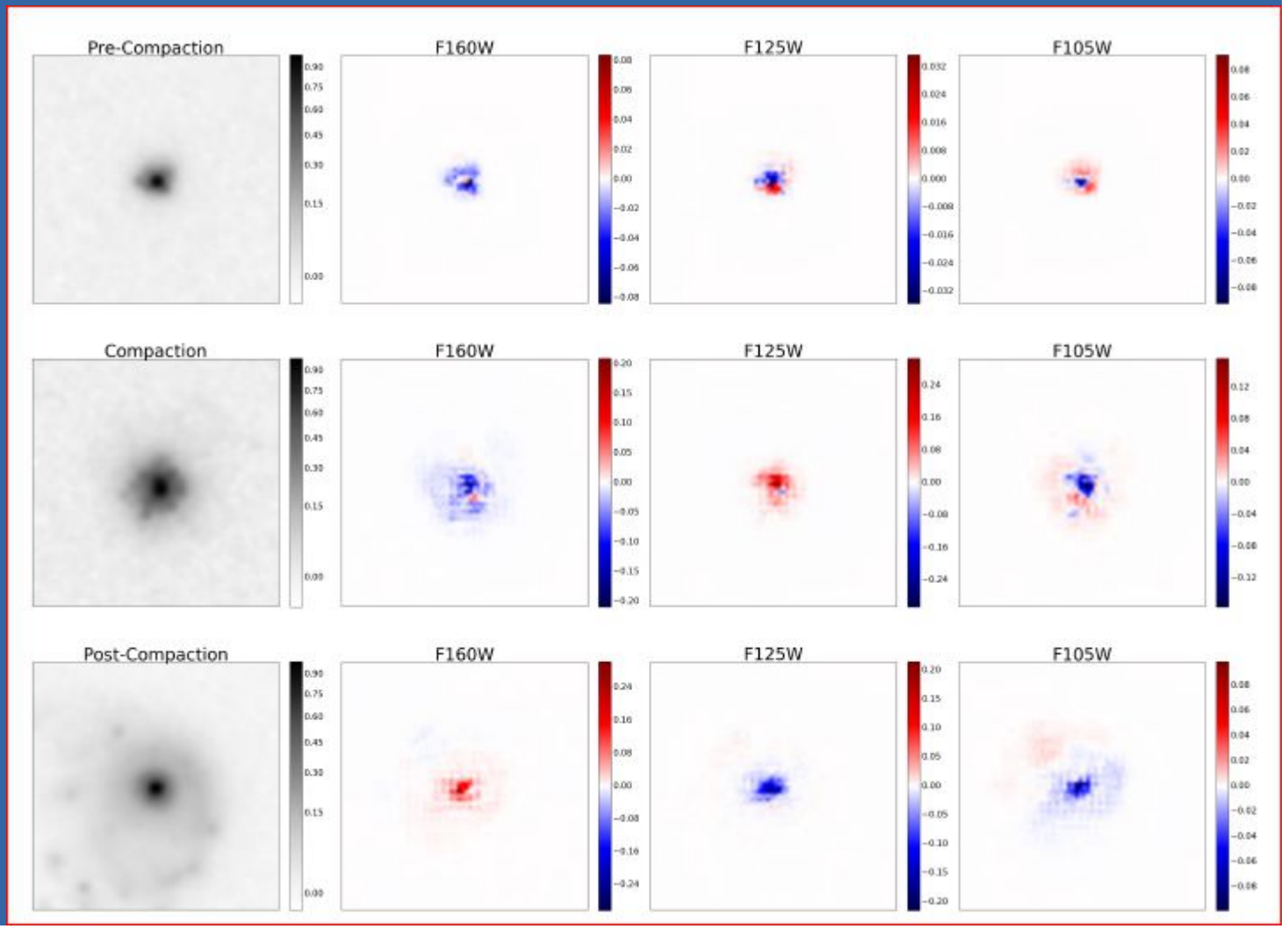
- La phase BN est une phase transitoire compacte de formation d’étoiles dans la région centrale de galaxies à haut z.

- Etape 4 : identification des phases BN dans les observations du HST/CANDELS (observations en IR)

- Il est ainsi possible de classer les galaxies à haut z en trois phases évolutives :
- Pré-compaction (pré BN)
- Compaction (BN)
- Post compaction (post BN)

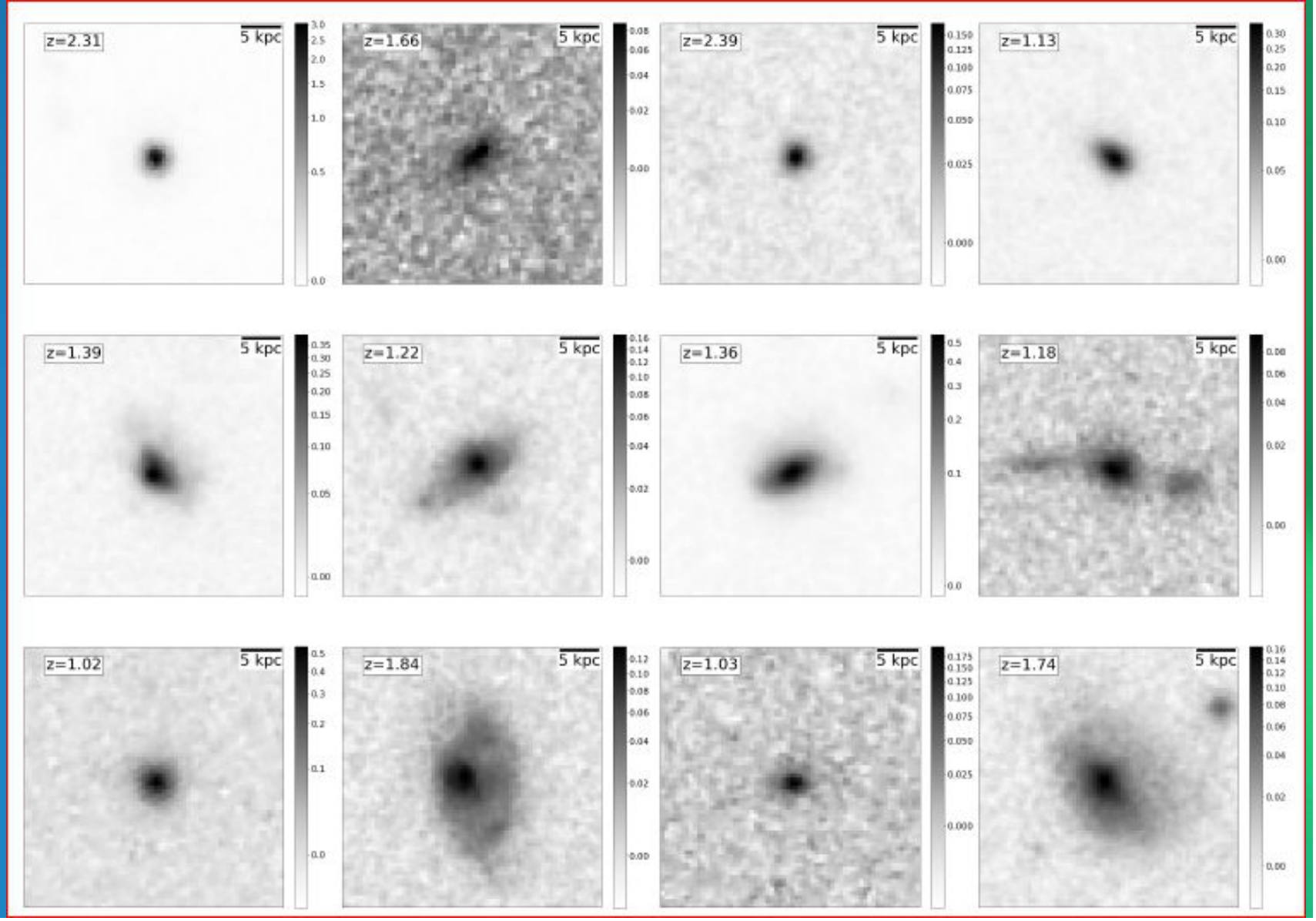


En réalité les images de galaxies de CANDELS sont données pour les trois filtres



- L’application du modèle permet de trouver des galaxies réelles de mêmes caractéristiques que celles simulées (98%).

Galaxies de CANDELS



- Il devient possible :
  - De déterminer des phases évolutives temporelles au sein de galaxies lointaines
  - Puis de mesurer certains caractéristiques de ces galaxies

Conclusions

- L’IA est devenu indispensable à beaucoup d’études en astronomie et cela dans de très nombreux domaines qui vont de l’étude du Soleil à celui des objets les plus anciens de l’Univers, en passant par les planètes extra-solaires.