Automated unsupervised search for unusual objects in large databases

--

the LSST perspective

J. Sánchez Almeida

Instituto de Astrofísica de Canarias, Spain

Coll.:
C. Muñoz-Tuñón,
J. A. L. Aguerri,
Y. Ascasibar,
A. Morales-Luis,
D. Elmegreen,
B. Elmegreen,
A. de Vicente,
C. Allende,
...

Summary

- ❑ Unusual Objects:
  outliers in classifications of large databases
- ❑ K-means automated classification algorithm
- ❑ Example: eXtremely-Metal Poor (XMP) galaxies
- ❑ Possible uses of Kmeans for LSST data handling and analysis
- ❑ Conclusions

Unusual Objects: outliers in the classification of large data sets

Rare objects are often extremely telling from a physical point of view (the XMP example will be detailed)

Any systematic search for such objects rely on some kind of classification of a large dataset, so that rare objects stick out as outliers of the classes.

The dataset MUST me large, otherwise it would not contain unusual objects.

The classification of large datasets MUST be automated and, possibly, unsupervised.

K-means: an automated classification algotithm

If you can represent the objects of a dataset points in a high dimensional space, then k-means finds clusters of points in this space.

Pros:

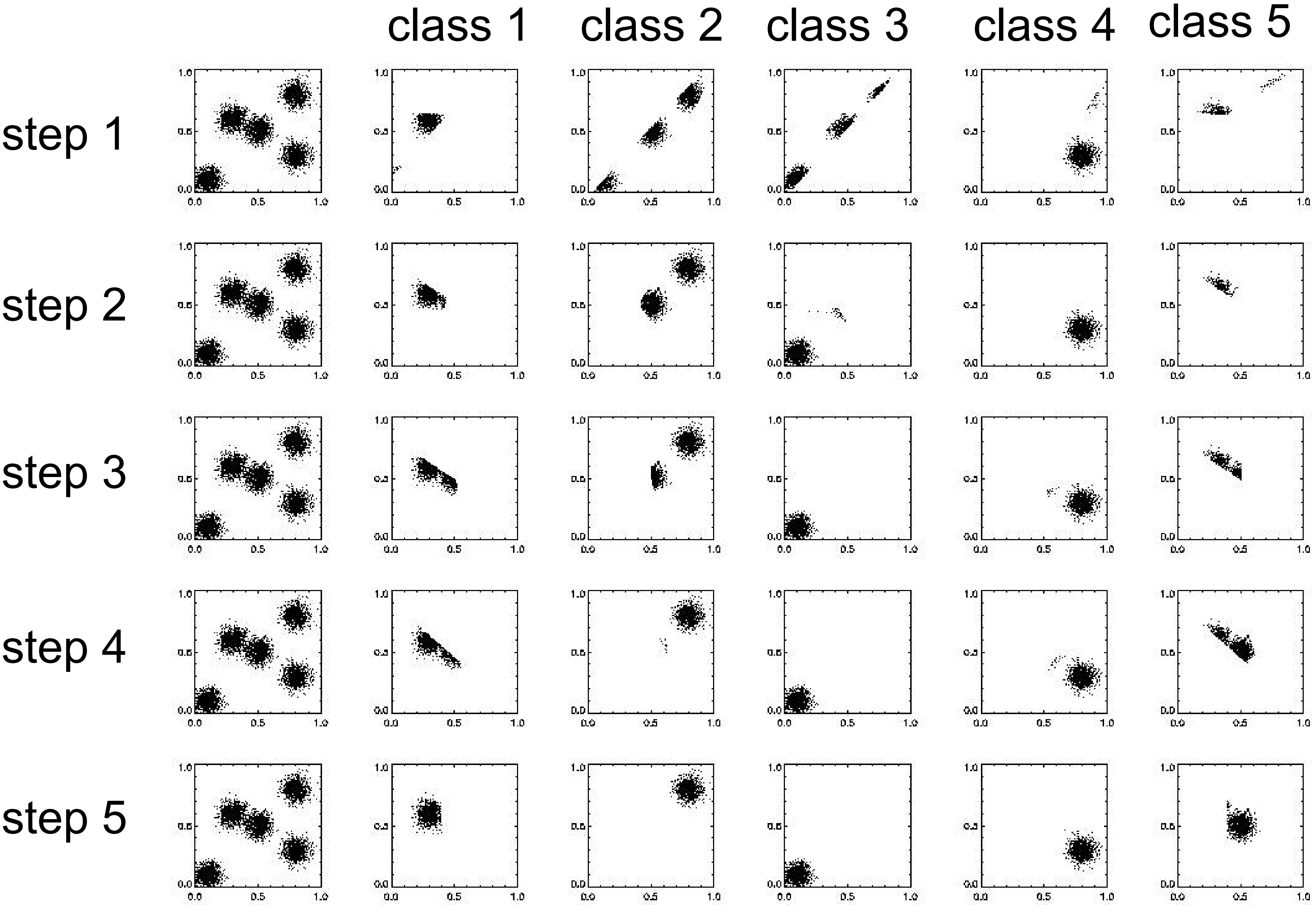
- automated, unsupervised
- robust, a workhorse able to cope with most problems
- works in thousands of dimensions with millions of points
- easy to parallelize
- decides the cluster number
- cluster centers are physical objects
- probability of good assignation … easy way to find outliers

Cons:

- the inferred clusters may be slices of real clusters
- no physical interpretation given (may be seen as a Pro)
- ...

How does k-means work?

Example in two dimensions



class 1 class 2 class 3 class 4 class 5

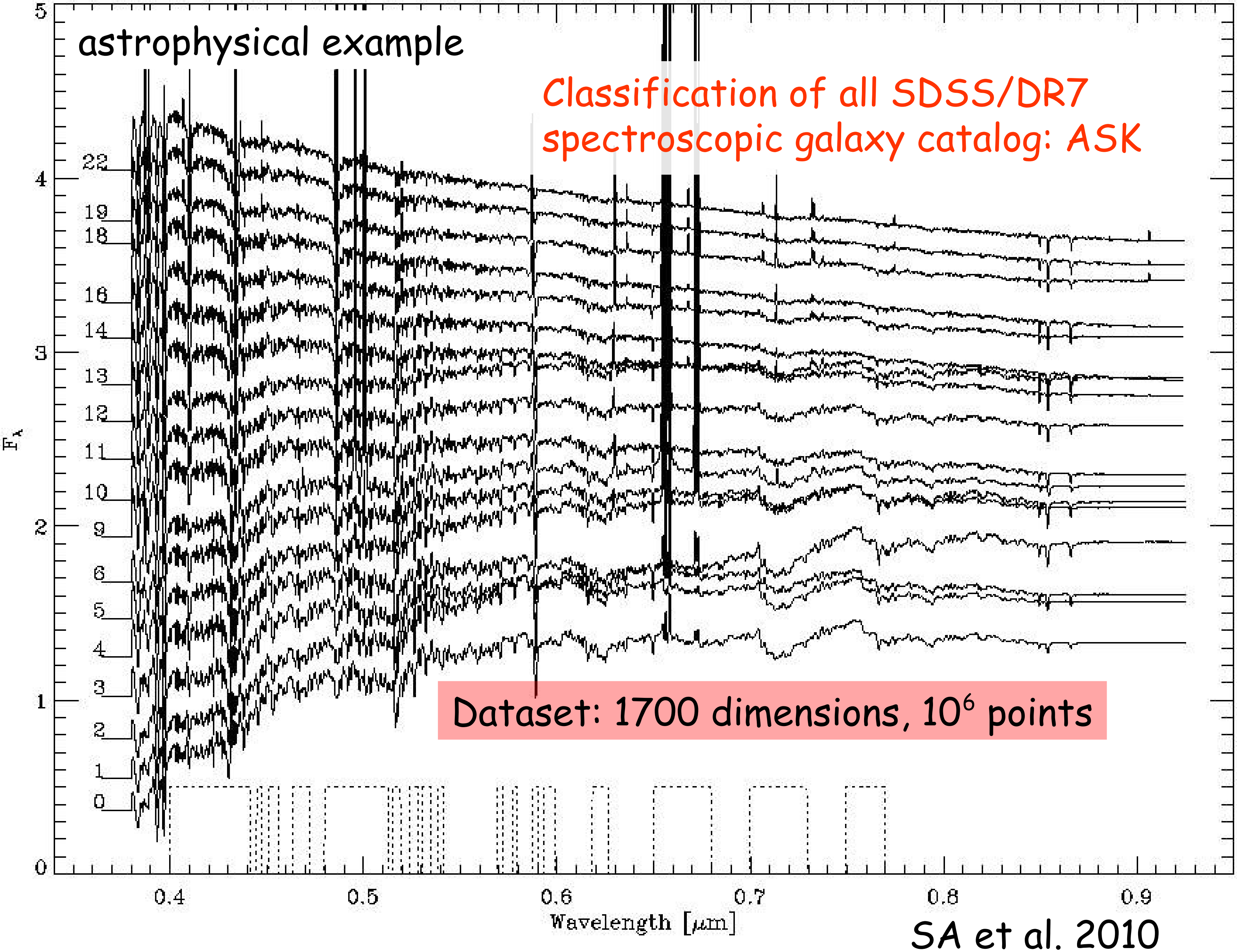
step 1

step 2

step 3

step 4

step 5



astrophysical example

Classification of all SDSS/DR7 spectroscopic galaxy catalog: ASK

Dataset: 1700 dimensions, 10^6 points

SA et al. 2010

Outliers of the ASK classification

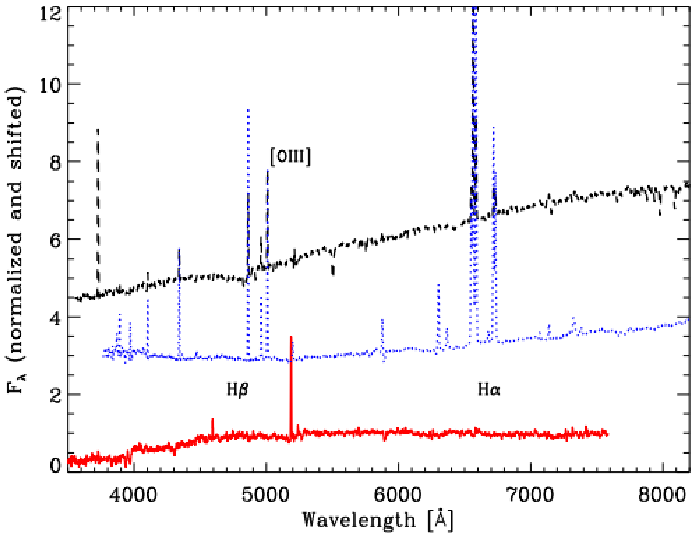


Fig. 1. Spectra of several outliers. Bottom (red solid line): red galaxy with abnormal emission line. Middle (blue dotted line): emission line galaxy with a continuum that upturns both in the blue and in the red. Top (black dashed line): extremely red object with emission lines.

- noisy and mis-reduced spectra
- QSO
- wrong redshifts
- red galax with emission lines
- blue galax with unusual continuum
- double peak QSO
-green peas and relatives

SA et al. 2013

Some uses of K-means

- Identifying problems in the automated reduction pipelines

- Pre-processing of large data sets (e.g., IFU spectra) so that similar data are interpreted the same way (SA+00,ApJ, 532, 1215)

- New spectral classification of galaxies (ASK; SA+10,ApJ,714,487) and stars (SA&AP,13,ApJ,673,50)

- S/N improving by stacking (SA+09,ApJ,698,1497)

- Targeted searches, e.g., systematic search for XMP galaxies (ML+11,ApJ,743,77)

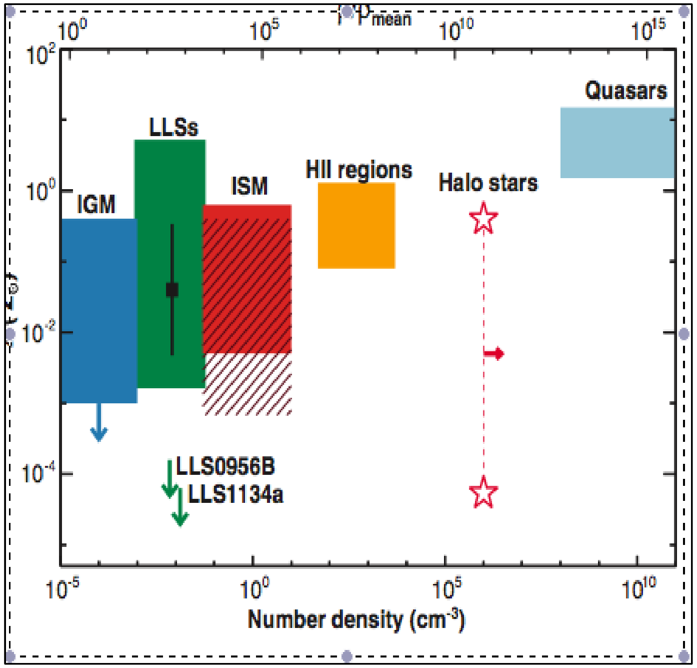
- Discovery of rare objects – both large data sets and classification needed (SA&AP,13,ApJ,673,50, SA+13,RMxAC,42,111)

- Others (e.g., morphological classification of galax … never tried)

Example: eXtremely Metal Poor galaxies

The Big-Bang just produces H and He (plus traces of Li, Be, and B).

Low metallicity targets (from the IGM to stars) are therefore primitive unevolved systems.



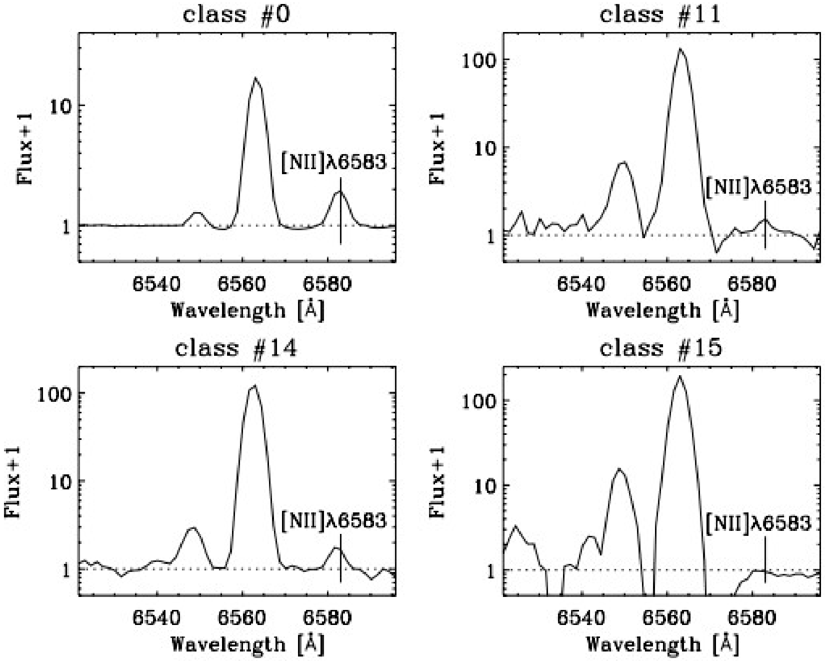
Fumagalli et al. (2011,Science)

Many such unevolved galaxies are to be expected according to the ΛCDM paradigm.

$Z/Z_o \geq 10^{-2}$ for all local galaxies (HII gas metallicity)

Galaxies with $10^{-1} \geq Z/Z_o \geq 10^{-2}$ are rare

In view of the small number of XMP and on its potential interest, we carried out a systematic search for these galaxies in the SDSS-DR7 (the largest data release available a the starting time).



Morales-Luis et al.(2011)

Using k-means, we classify all SDSS-DR7 spectra in a narrow spectral region around Ha, whose shape is extremely sensitive to metallicity (e.g., Denicolo et al. 2002, Pettini & Pagel 2004)

In the end, we find only 32 targets out of the 10^6 DSSS galaxies – 0.01% of the galax with emission lines

Most of them (24/32) turn out to be cometary!

The spectrum decides de shape !!!

Why are XMP galaxies tadpole-cometary galaxies?

We do not know for sure yet ... but the best explanation available turns out to be both surprising and of far-reaching implications in galaxy formation.

Extensive follow up work has shown:

- XMP are disks, dynamically thick (Elmegreen et al. 12; SA et al. 13)
- They are surrounded large amounts of metal poor neutral HI gas ($M_{HI}/M_{*}\approx20$; Filho et al. 13).
- The head of the comet is a **giant HII region of low metallicity compared to the rest of the disk** (SA et al. 13; 14).

✶ All this results combined suggest the **XMPs** to be **disks in early stages of assembling** with its **star-formation sustained by direct accretion of external pristine gas**.

✶ Such process, dubbed **cold-flow** accretion, is to be **expected** according to **numerical simulations** of galaxy formation (Dekel et al. 2009), but they have not been observed are in local universe (Cresci et al. 2010).

✶ Probably a **common phenomenon** in the local universe, ...

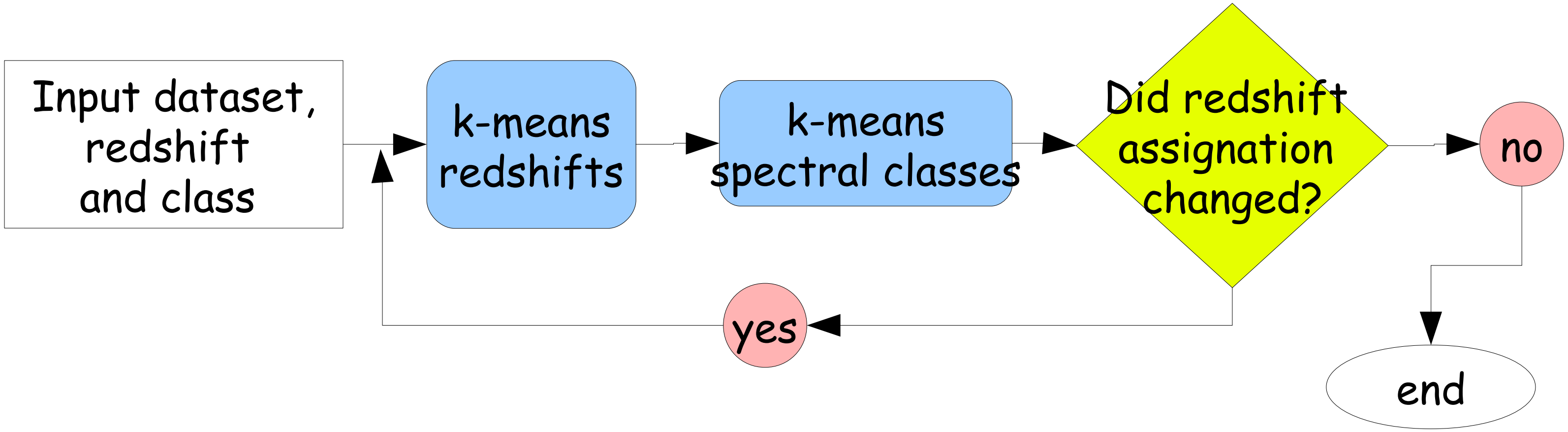
Possible uses of Kmeans for LSST data handling and analysis

Obvious Applications

- ❑ Identifying problems in the automated reduction pipelines. Many outliers are failures of the pipeline
- ❑ Spectral classification from known (photometric) redshifts and the *ugrizy* magnitudes, including AGN selection
- ❑ Discovery of unusual objects (as outliers of the spectral classification).
- ❑ Selection of class templates, for in-depth studies

New Applications

- Self-consistent spectral classification and photometric-redshift determination. Never tried but: main factor affecting the 'observed' color is redshift. Nested k-means classifications needed.


Input dataset, redshift and class
k-means redshifts
k-means spectral classes
Did redshift assignation changed?
no
yes
end

- Morphological galaxy classification based directly on the images. Never tried. Difficult.

- Spectral-temporal classification of the objets – the time represented by the first Fourier components?

Conclusions

✶ K-means is a robust workhorse algorithm able to classify large astronomical datasets without a-priory knowledge of their physical properties

✶ K-means are works well with the present databases (e.g., SDSS).

✶ K-means may be useful for the handling and data analysis of some LSST outcomes in various stages of reduction and archiving.

Kmeans:
A tool to look for the needle in the haystack