Classification algorithms for Gaia

Kester Smith, Coryn Bailer-Jones, Carola Tiede, Christian Elting

MPIA, Heidelberg

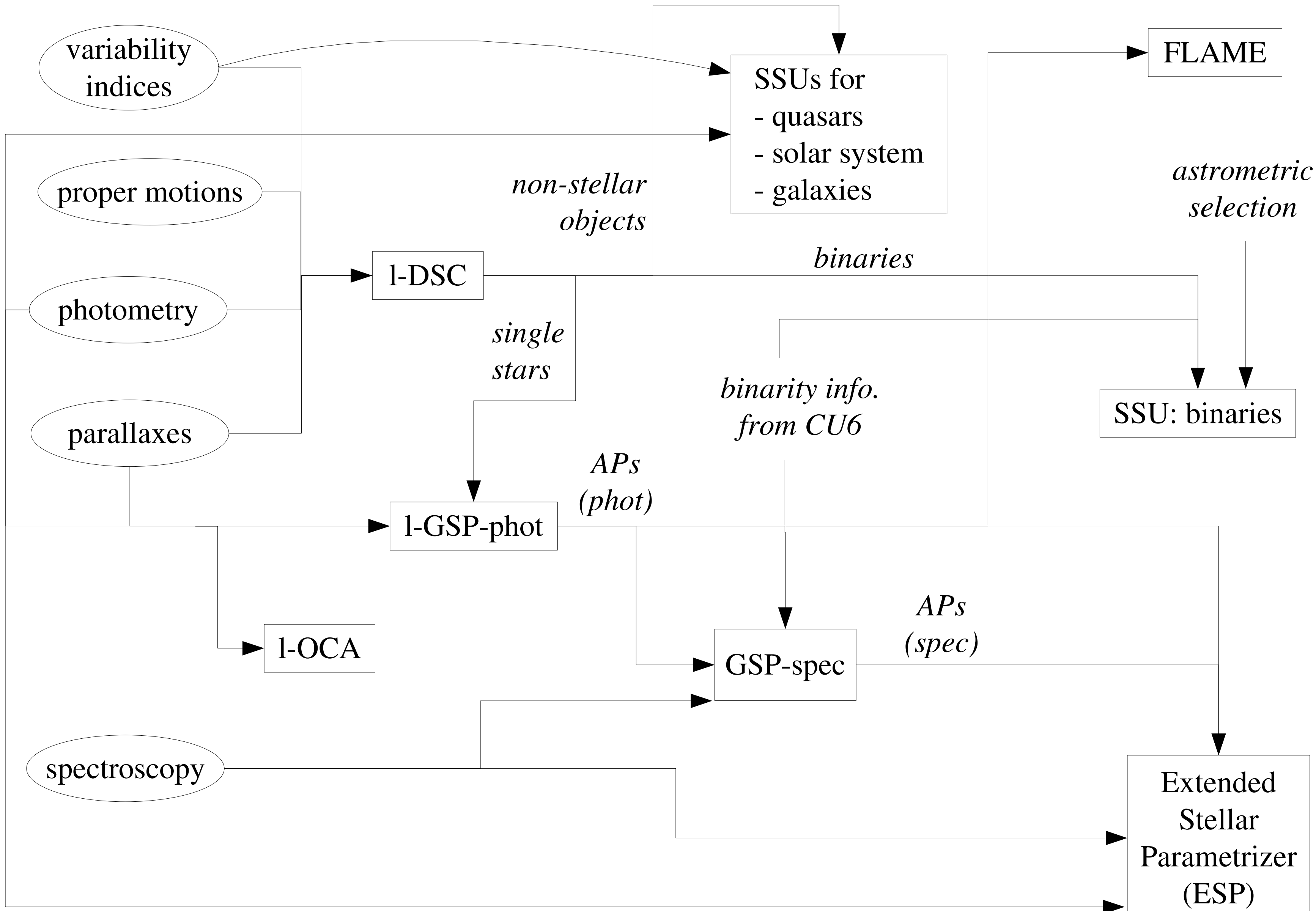
Overview

- Introduction: role of CU8 in Gaia
- Support vector classifiers
- Application to Gaia data
- A note about unbalanced datasets

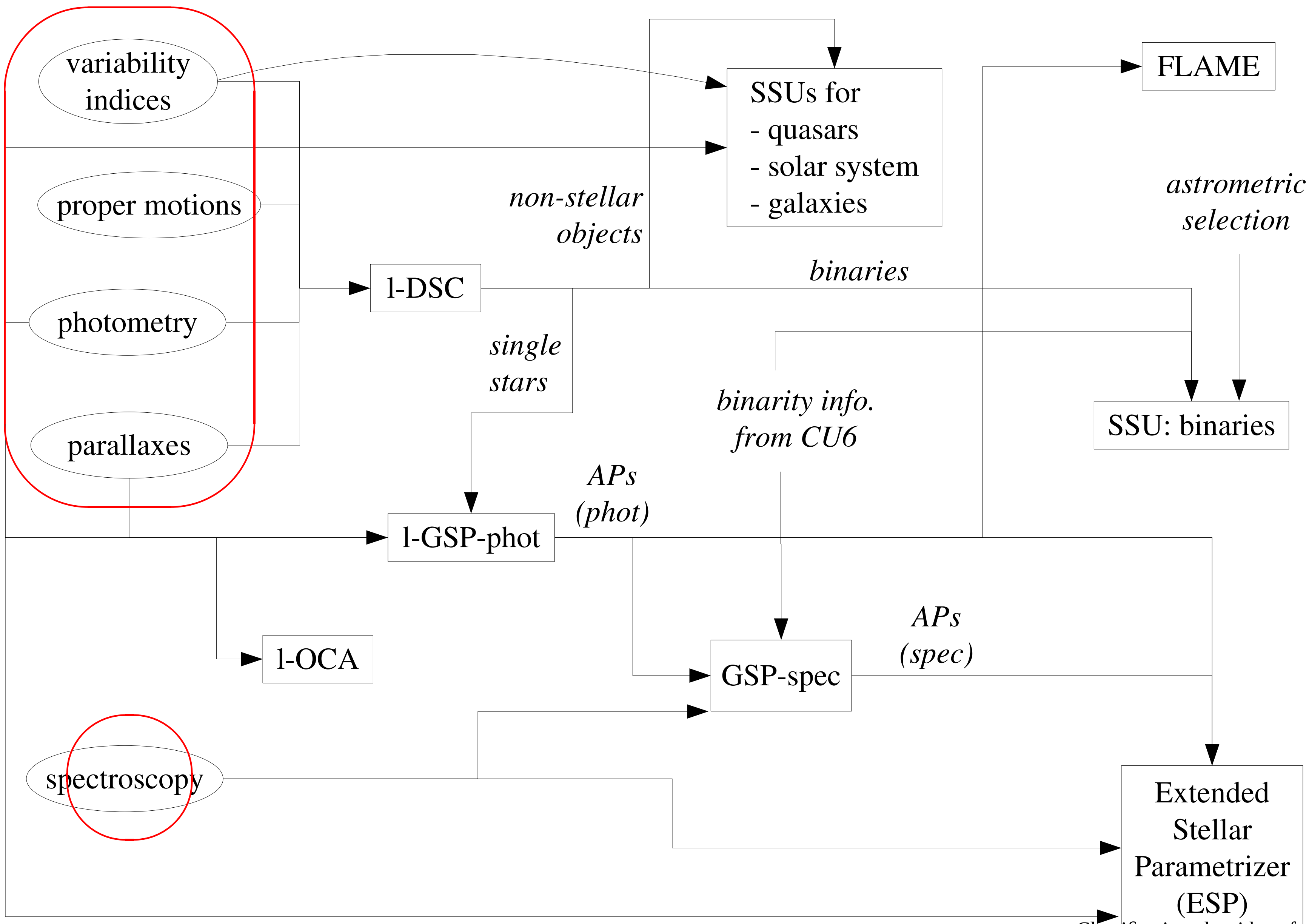
CU8 goals

CU8 will classify all Gaia sources into broad astrophysical classes, and find the main astrophysical parameters for them. It will also search for unusual sources (novelty detection).

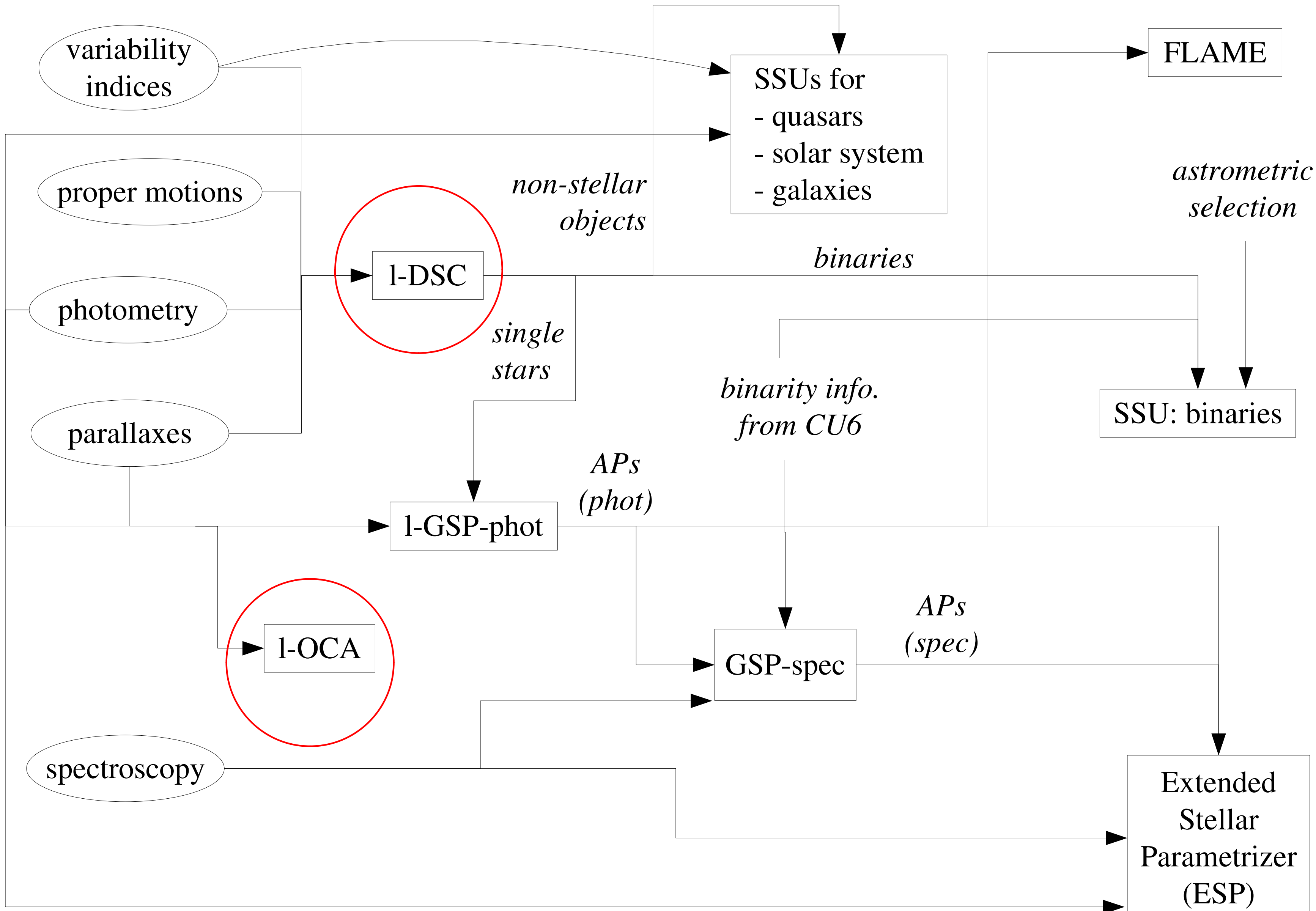
CU8 work flow



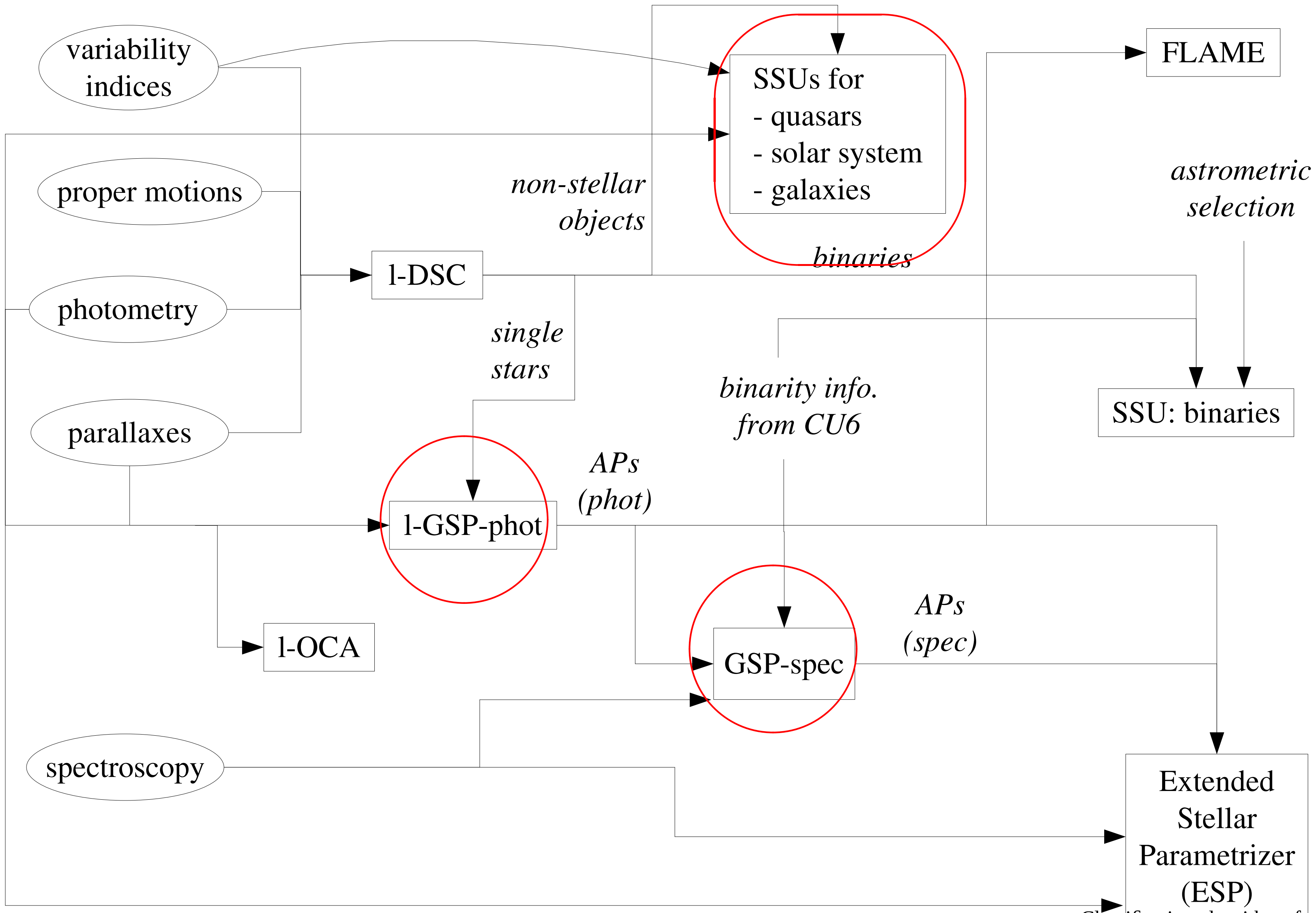
CU8 work flow



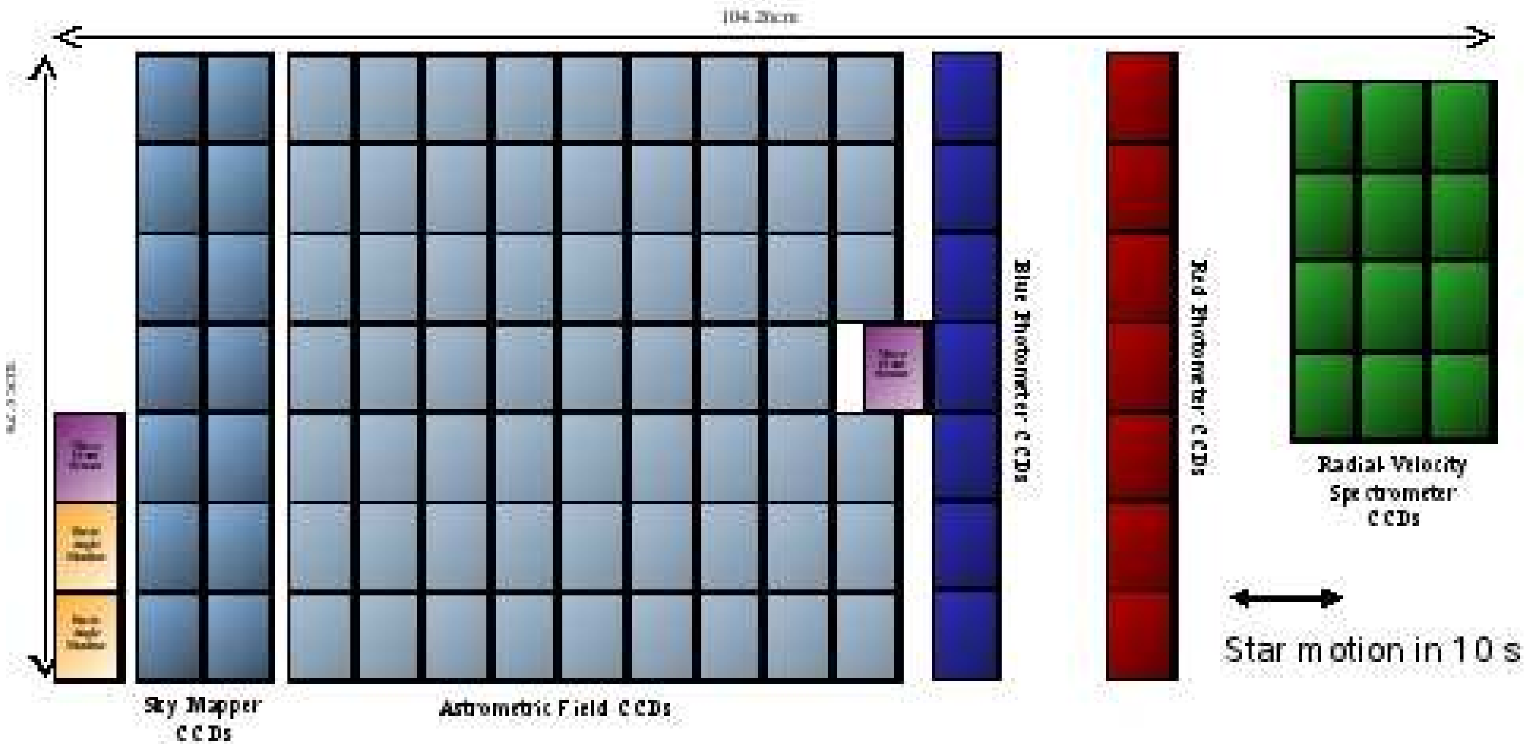
CU8 work flow



CU8 work flow



The focal plane



Total field:

- active area: 0.75 deg^2
- CCDs: 14 + 62 + 14 + 12
- 4500 x 1966 pixels (TDI)
- pixel size = 10 µm x 30 µm = 59 mas x 177 mas

Sky mapper:

- detects all objects to 20 mag
- rejects cosmic-ray events
- FoV discrimination

Astrometry:

- total detection noise: 6 e$^-$

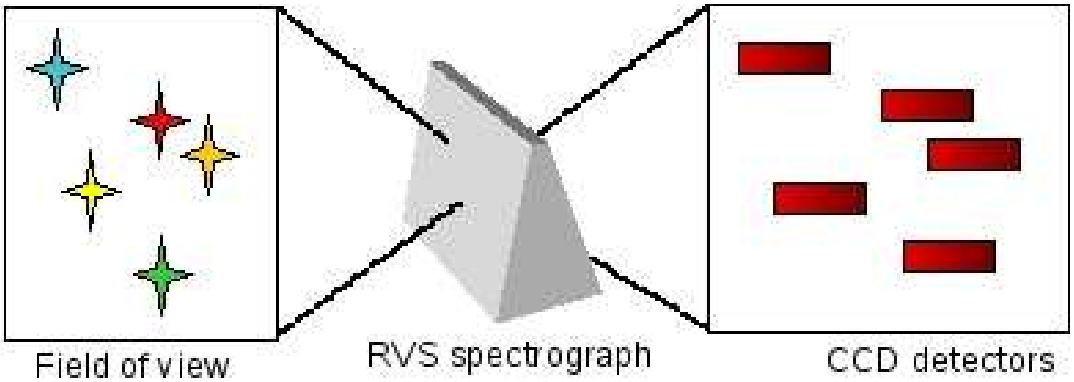
Photometry:

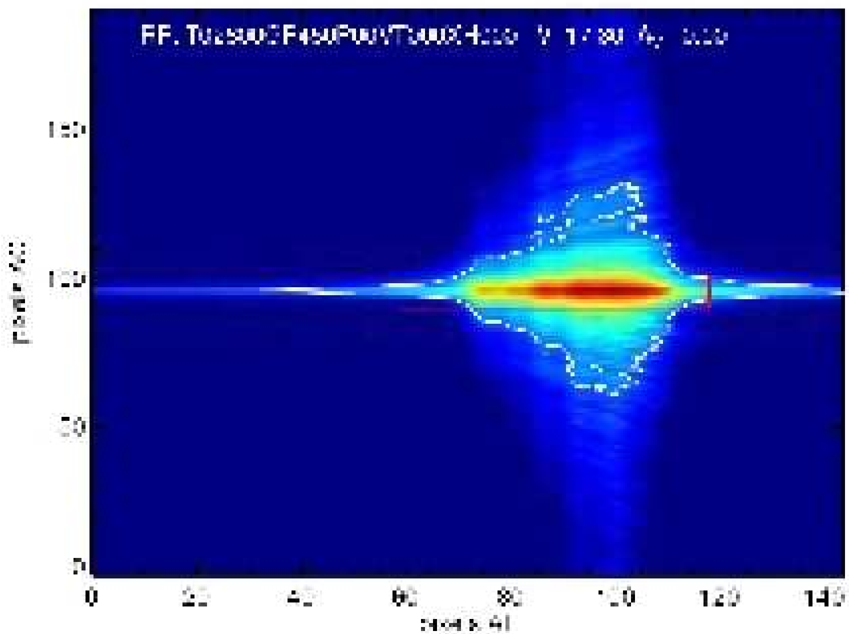
- two-channel photometer
- blue and red CCDs

Spectroscopy:

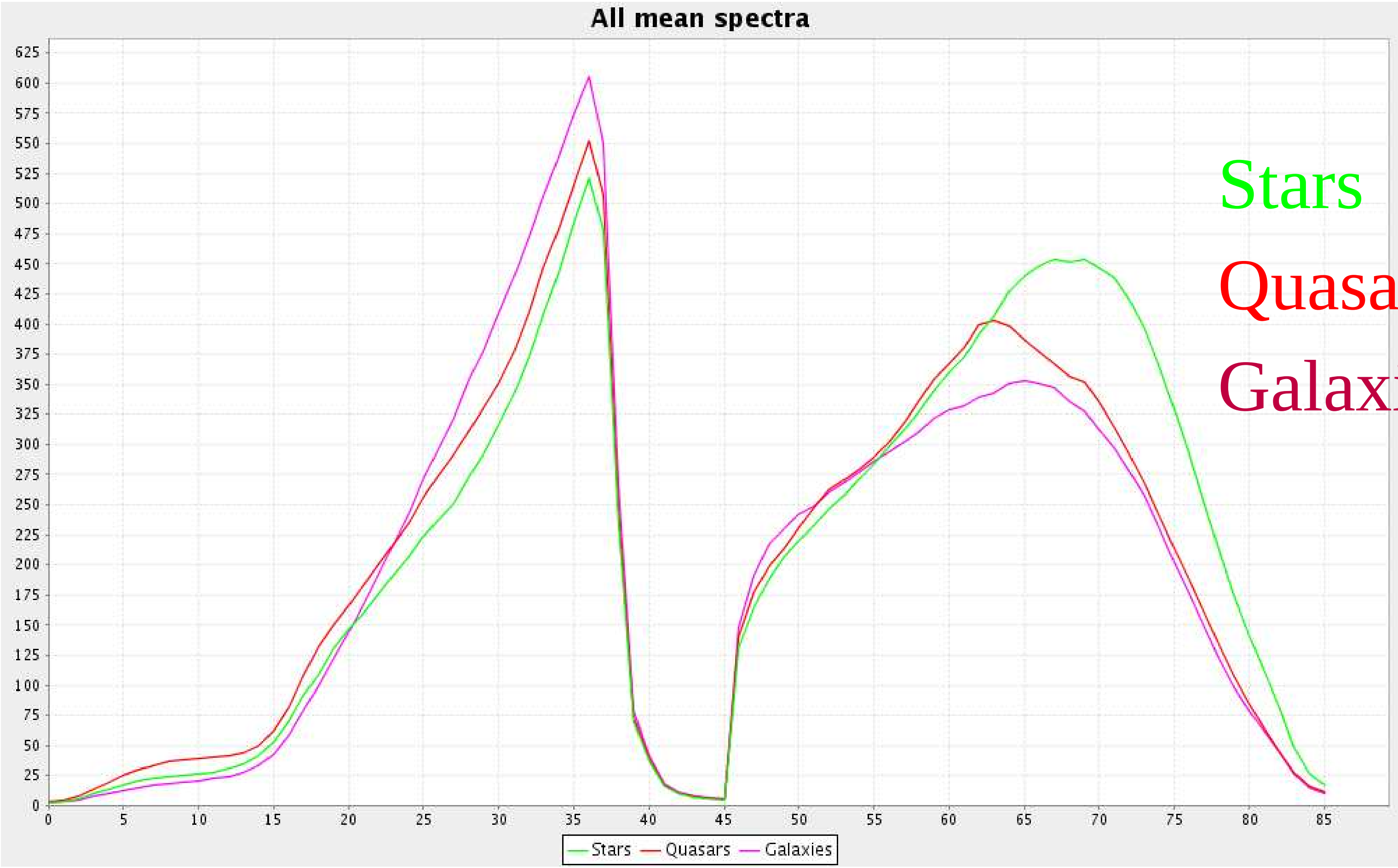
- high-resolution spectra
- red CCDs

RP/BP spectra





Photometry data

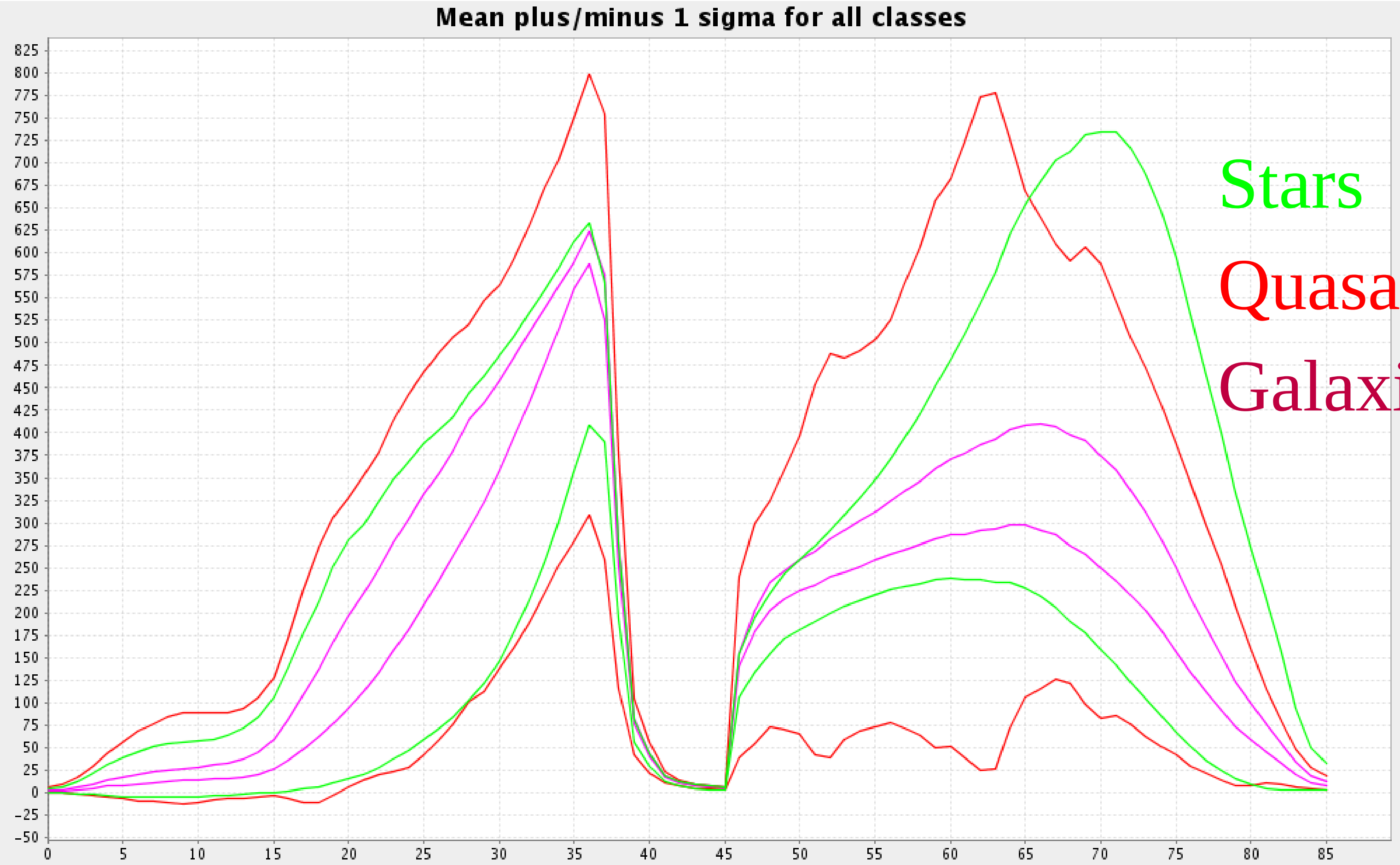


Stars

Quasars

Galaxies

Photometry data

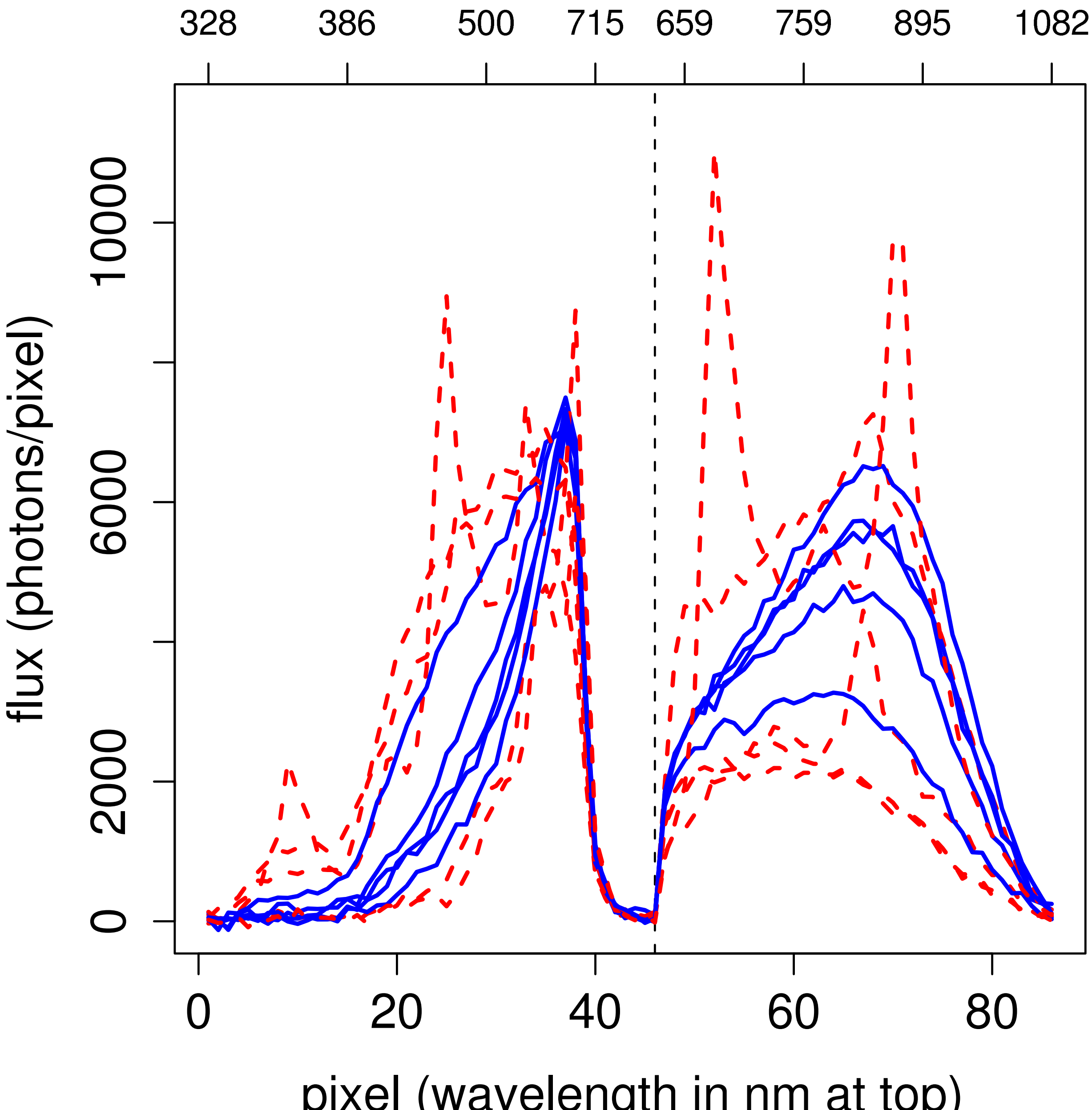


Stars

Quasars

Galaxies

Example spectra



Sample

- around 1 billion stars
- a million unresolved galaxies
- around half a million quasars
- numerous solar system objects

Main classification algorithms

Several algorithms considered.

Most successful are Neural network and Support Vector Machine (Svm)

Main classification algorithm is the Svm.

Svm concept

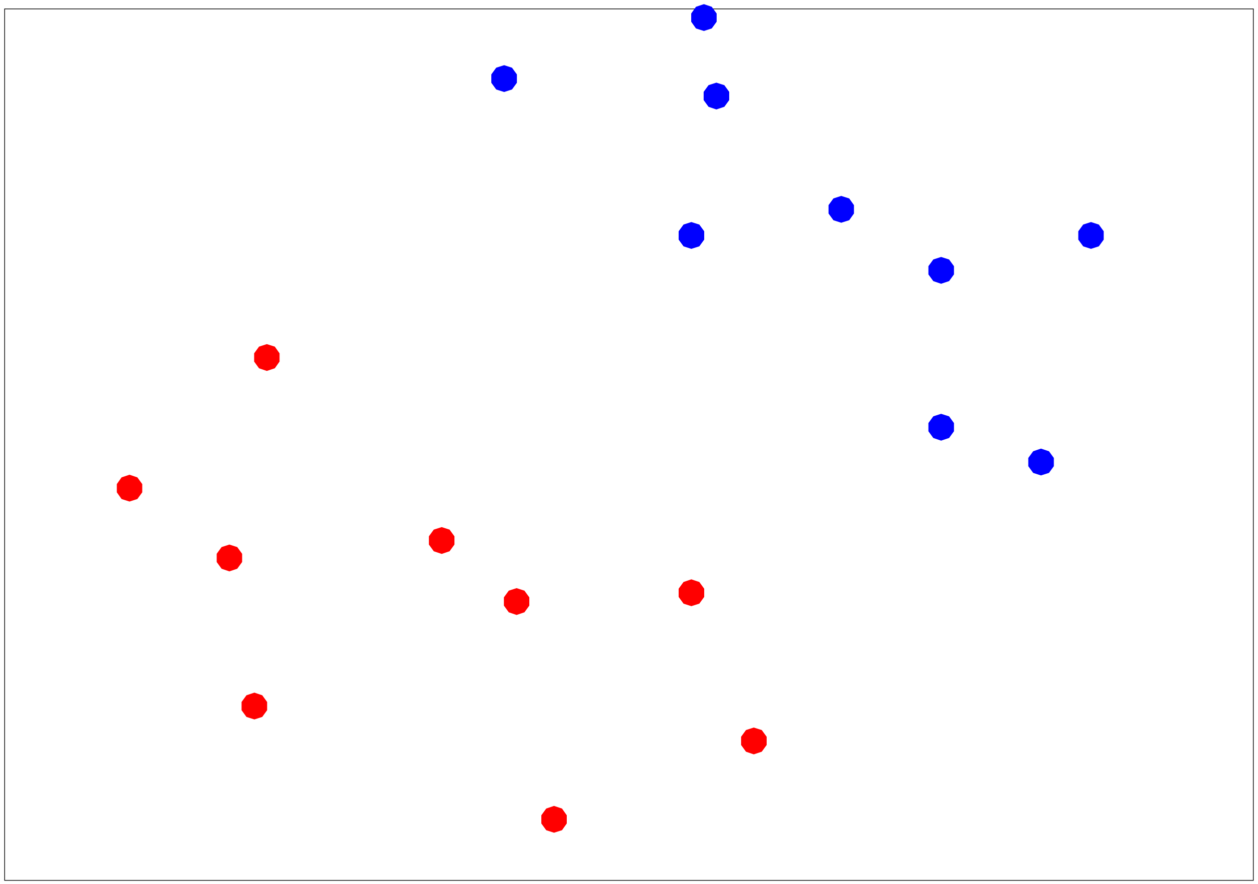
The following slides are loosely based on;
'Support Vector Machines: Hype or Hallelujah?'
Bennett and Campbell 2000

See also;
Cortes & Vapnik, 1995, 'Machine Learning' Vol 20 no. 3
or Wikipedia

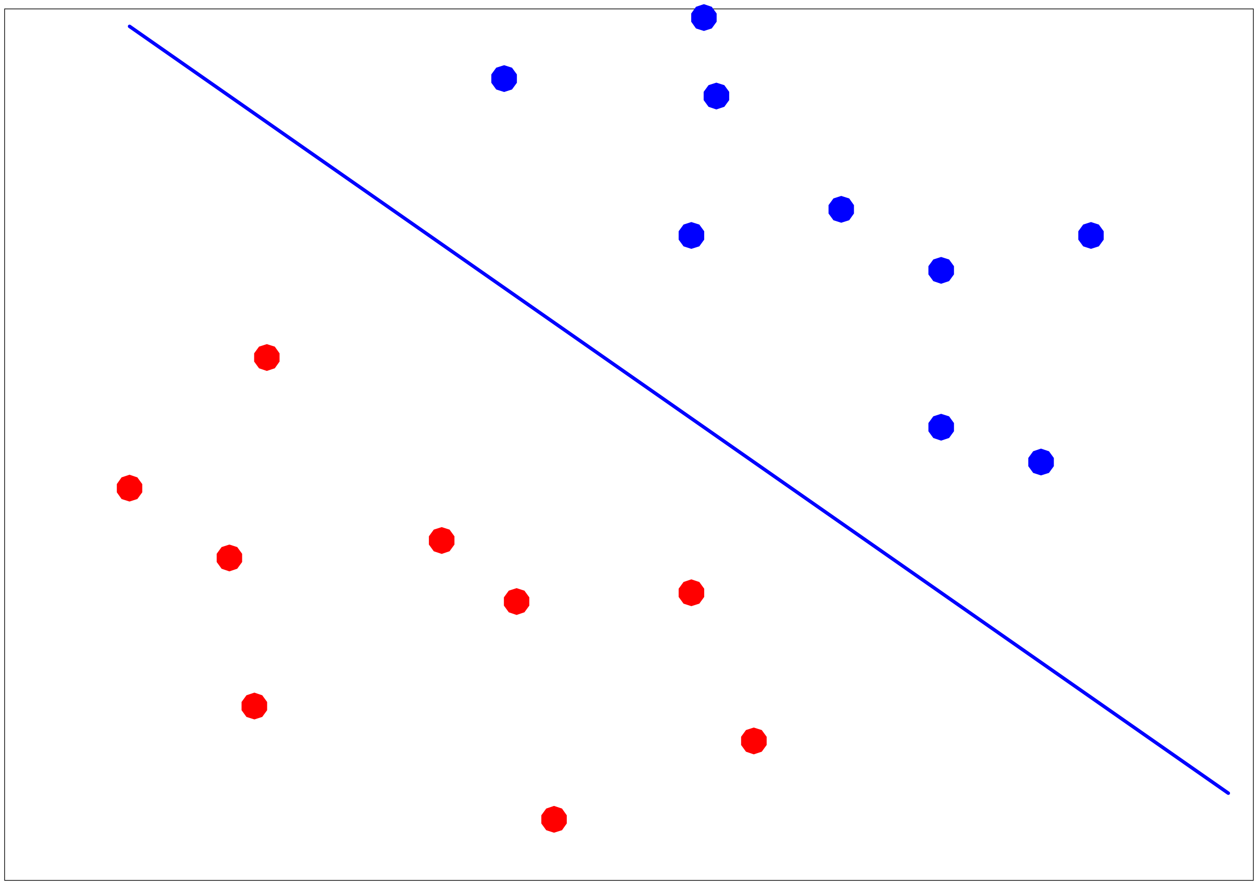
Svm implementation is available in R (library e1071)
and in weka

We use libSvm (Chang & Lin)

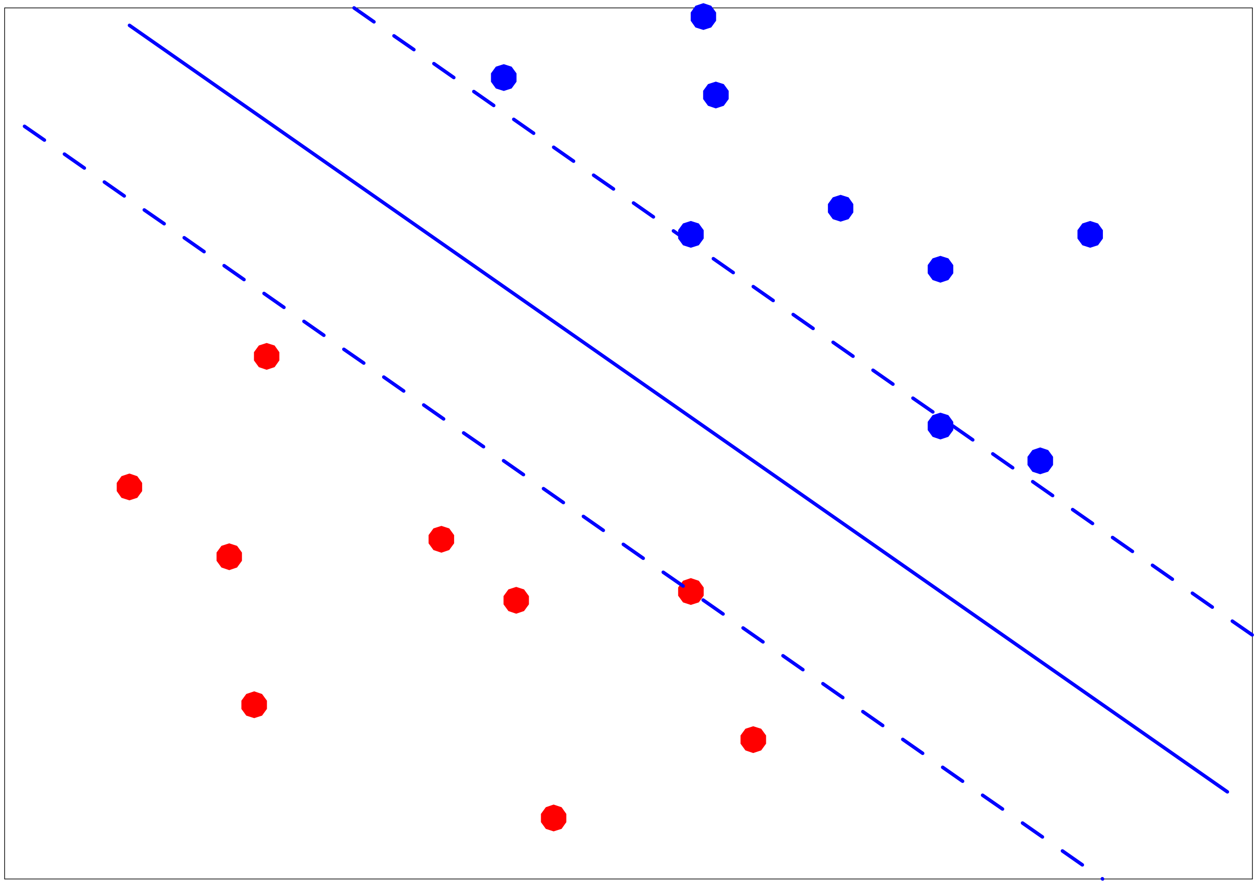
Svm concept: maximum margin



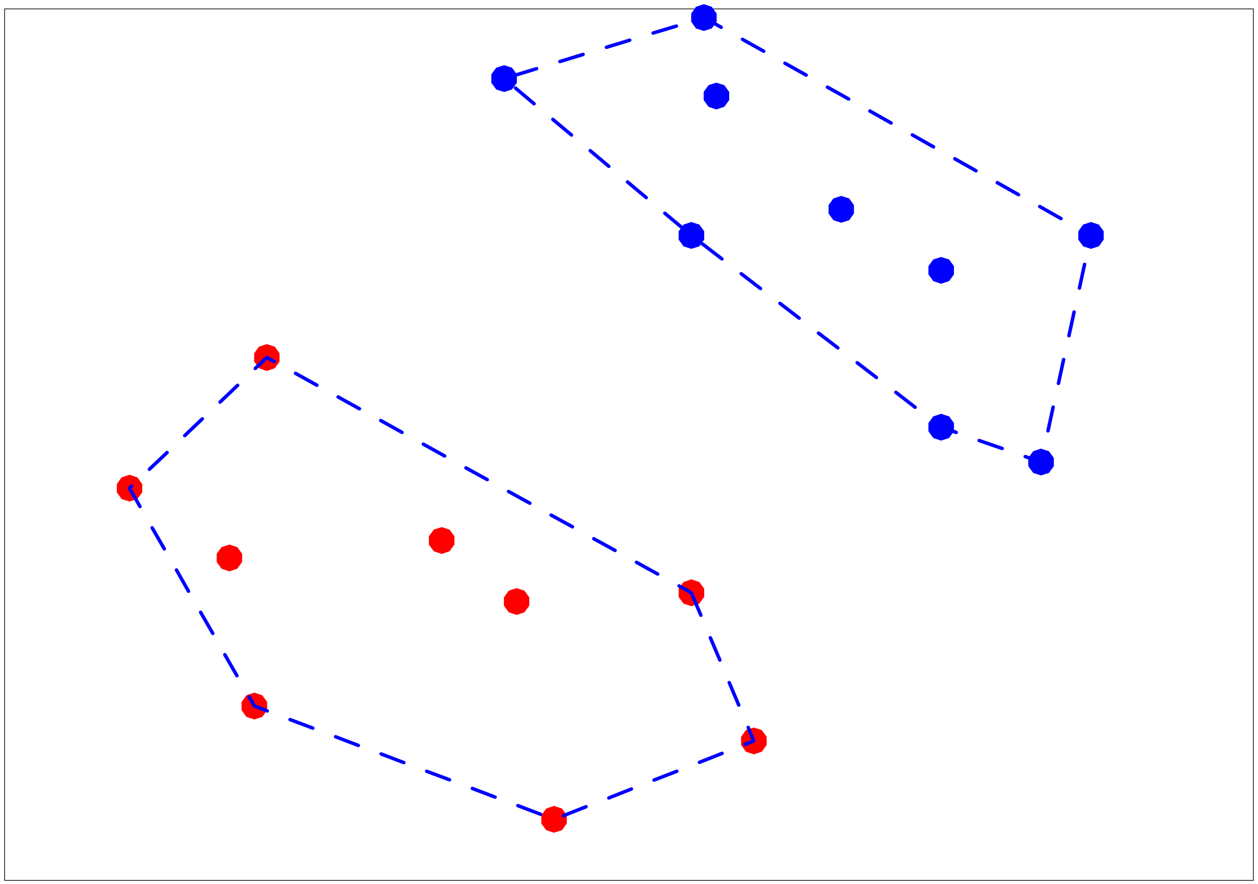
Svm concept: maximum margin



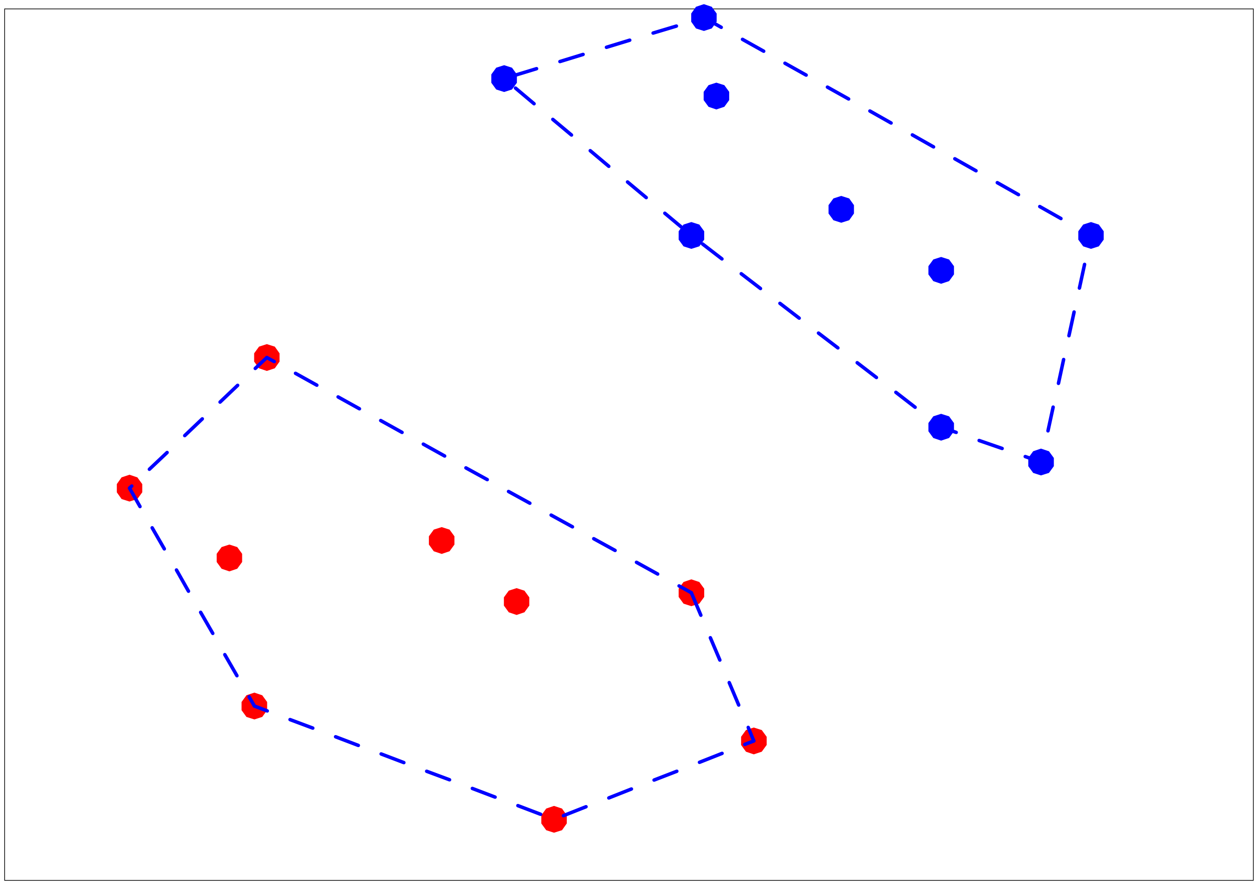
Svm concept: maximum margin



Svm concept: Convex hull



Svm concept: Convex hull



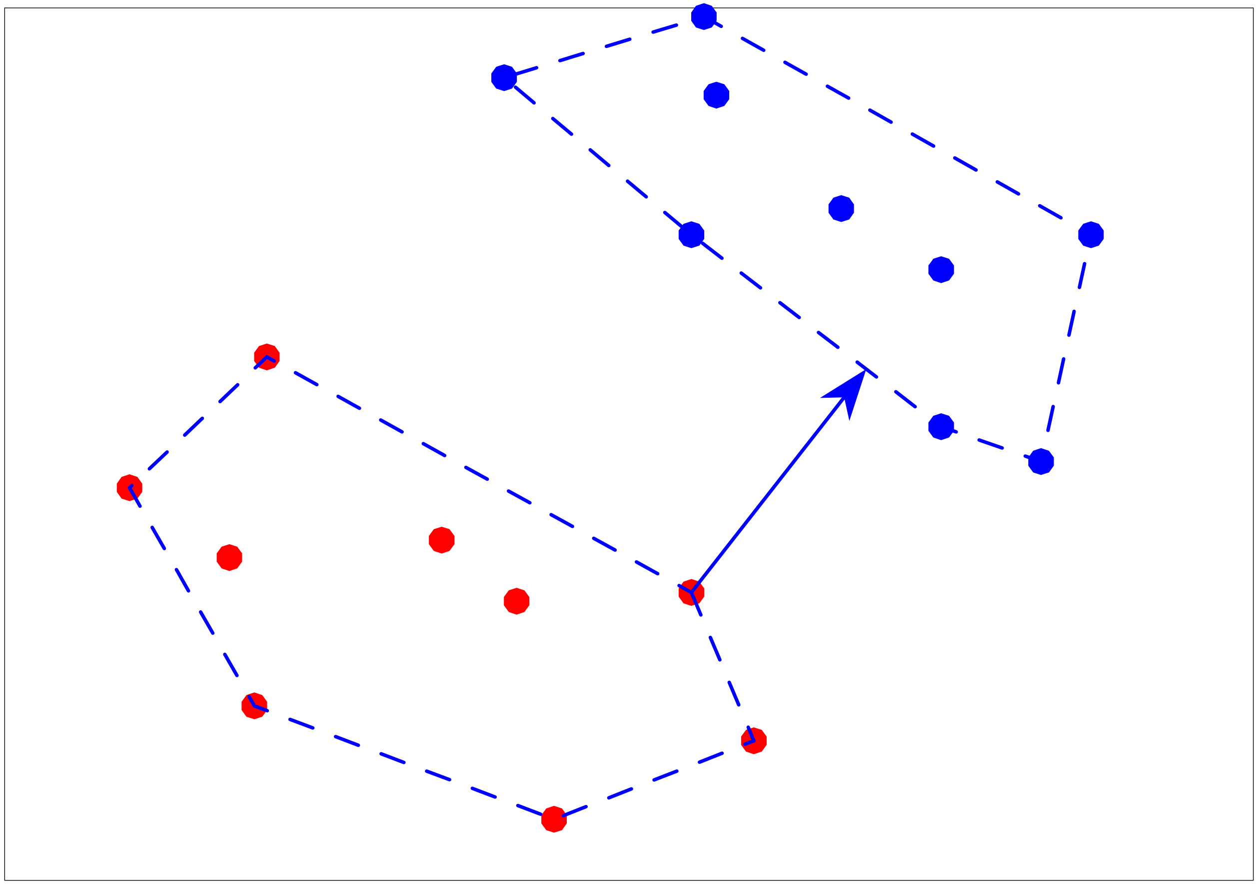
$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



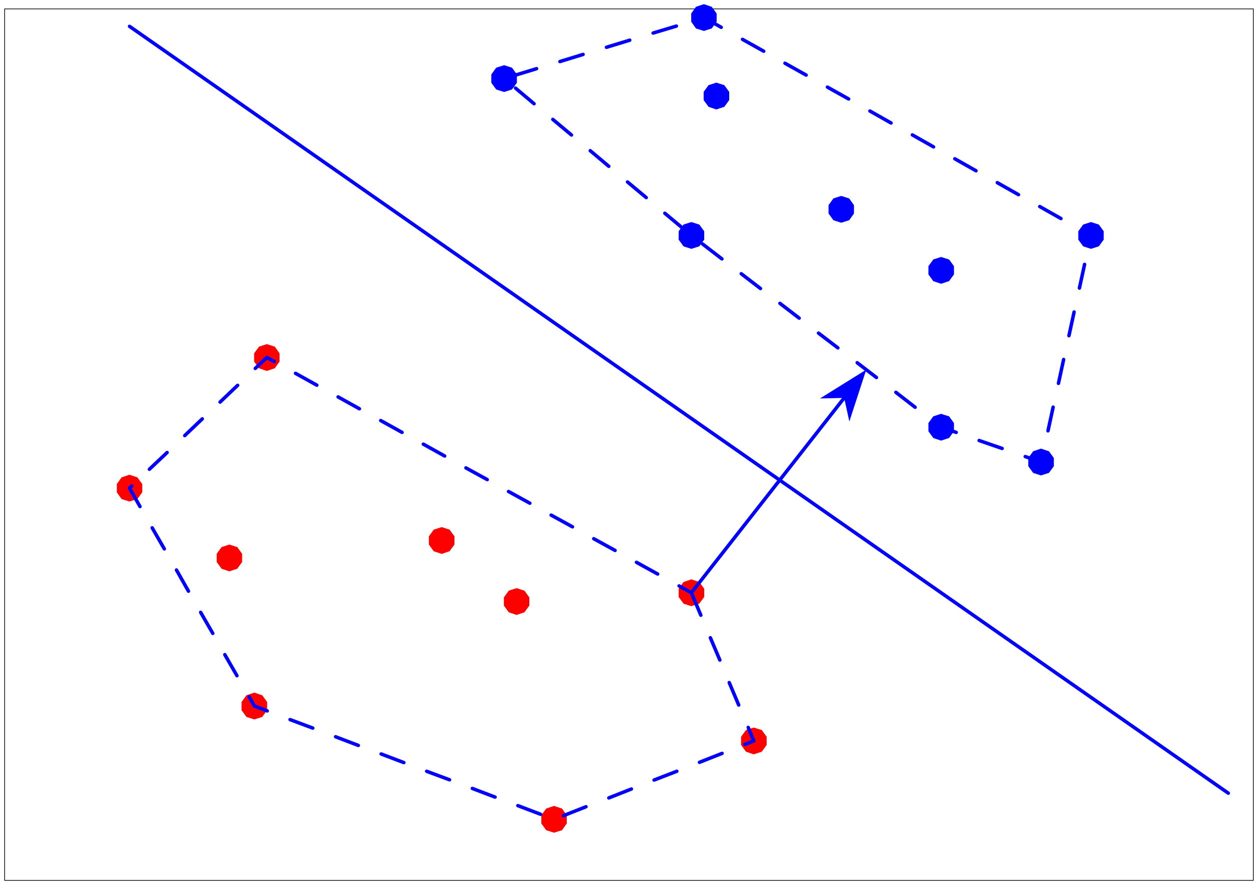
$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



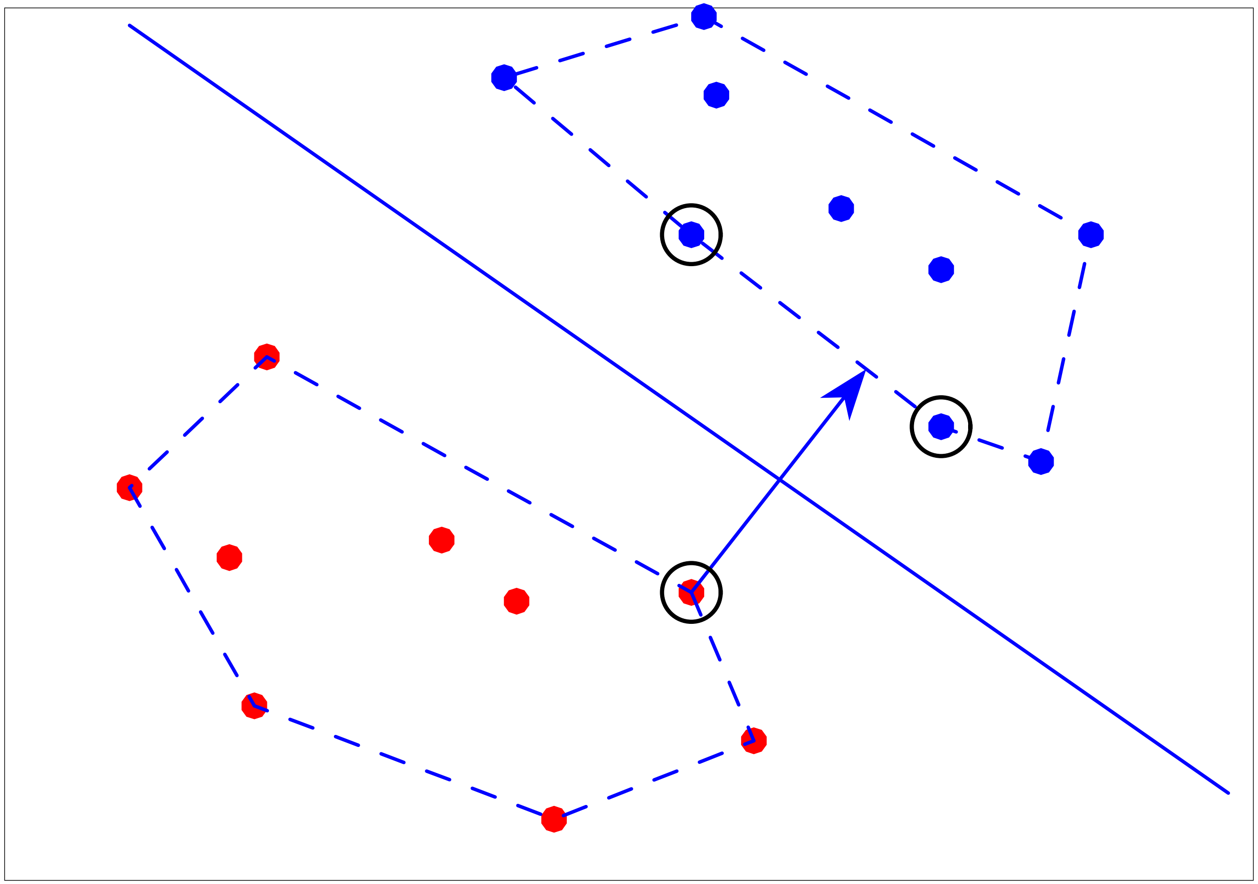
$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



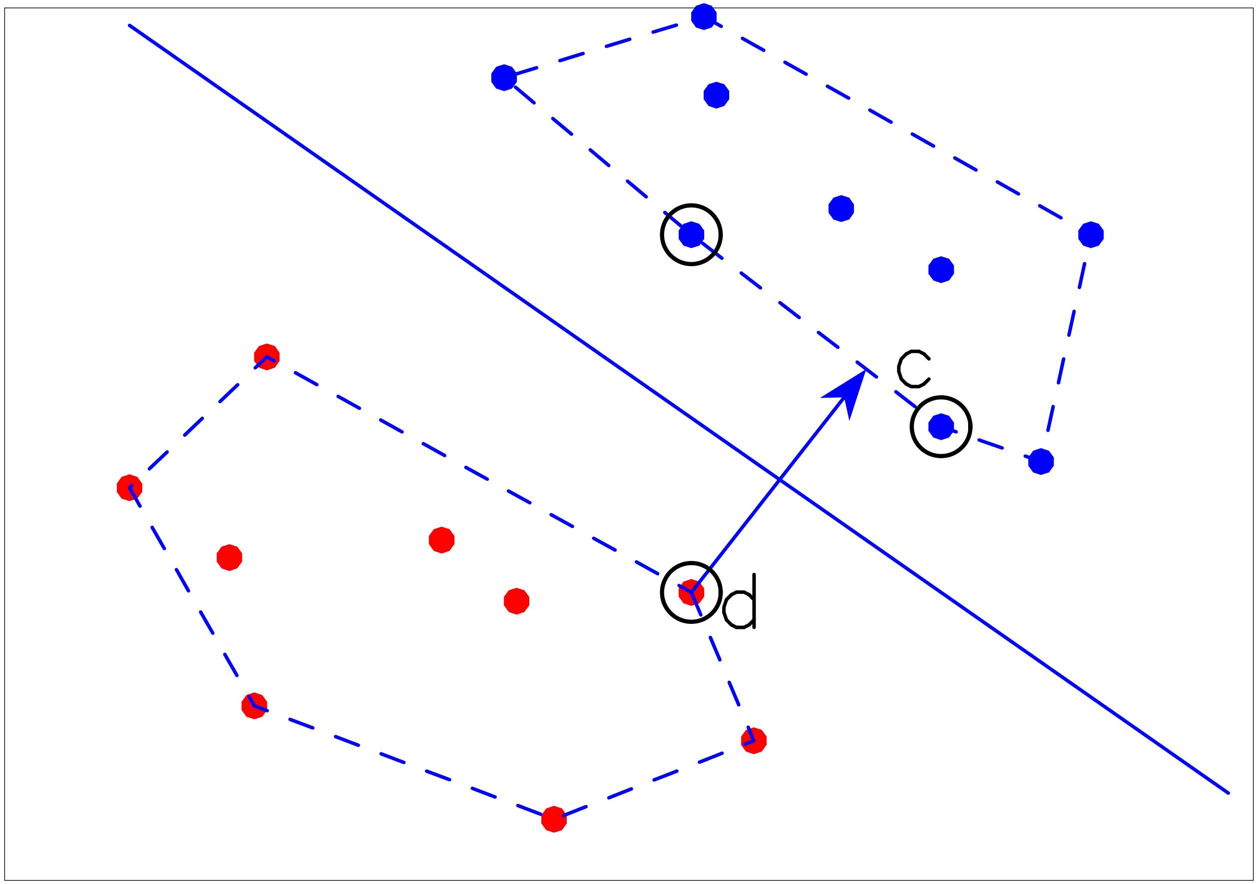
$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i \mathrm{x}_i$

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



$$\min \frac{1}{2}\|c-d\|^2$$

$$\text{c,d}=\sum_i \alpha_i \text{x}_i$$

$$\text{s.t. } \sum \alpha_i = 1$$

$$0 \leq \alpha_i \leq 1$$

Svm's

minimize

$$\min \frac{1}{2}\|c-d\|^2; \quad c,d = \textstyle\sum_i \alpha_i \mathbf{x}_i$$

$$\min \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i . \mathbf{x}_j$$

where $y_i = \pm 1$ depending on class
subject to constraint

$$\sum_{iclass1} \alpha_i = 1; \quad \textstyle\sum_i y_i \alpha_i = 0$$

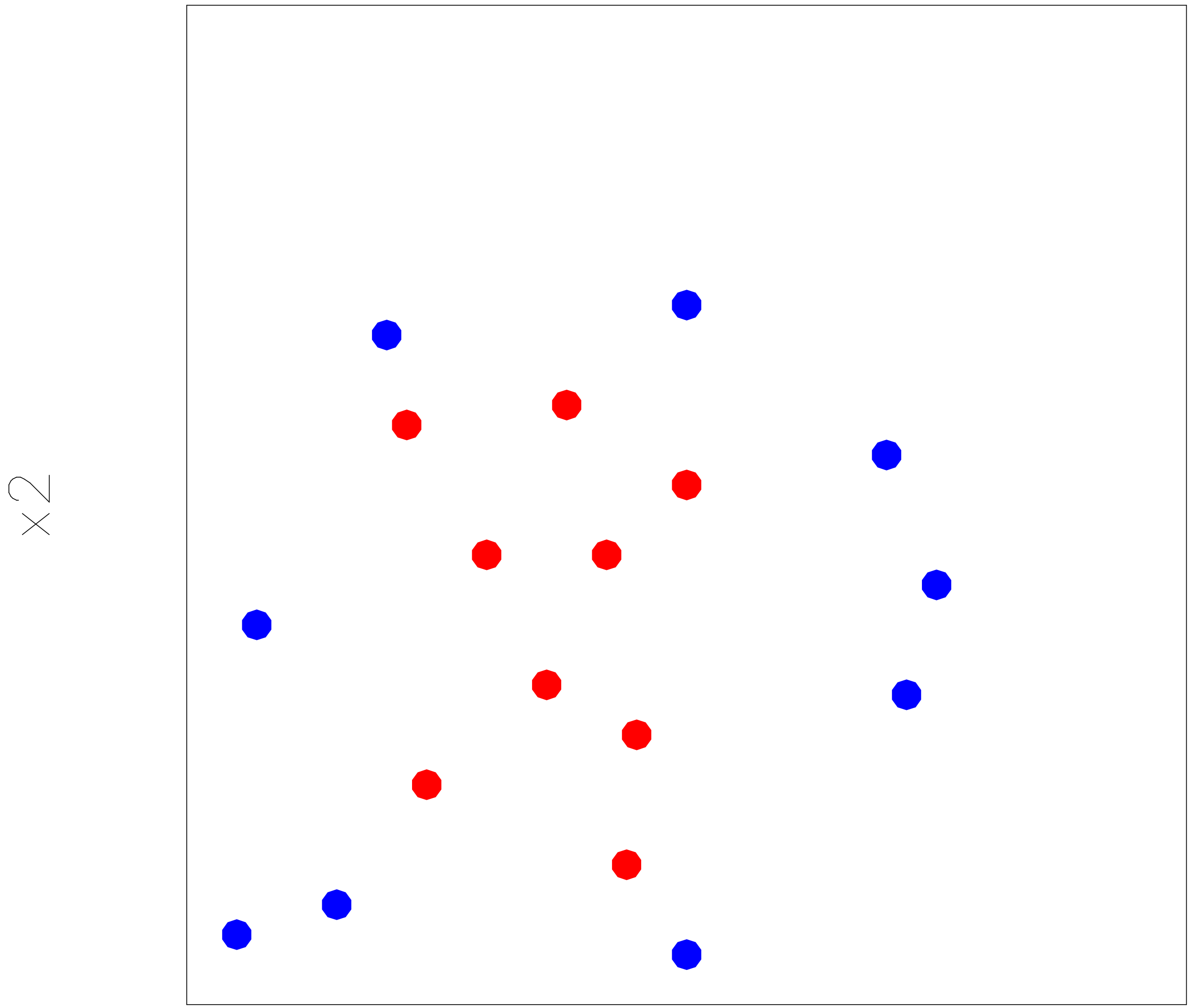
Applying the classifier

For a new point s, the decision boundary is;

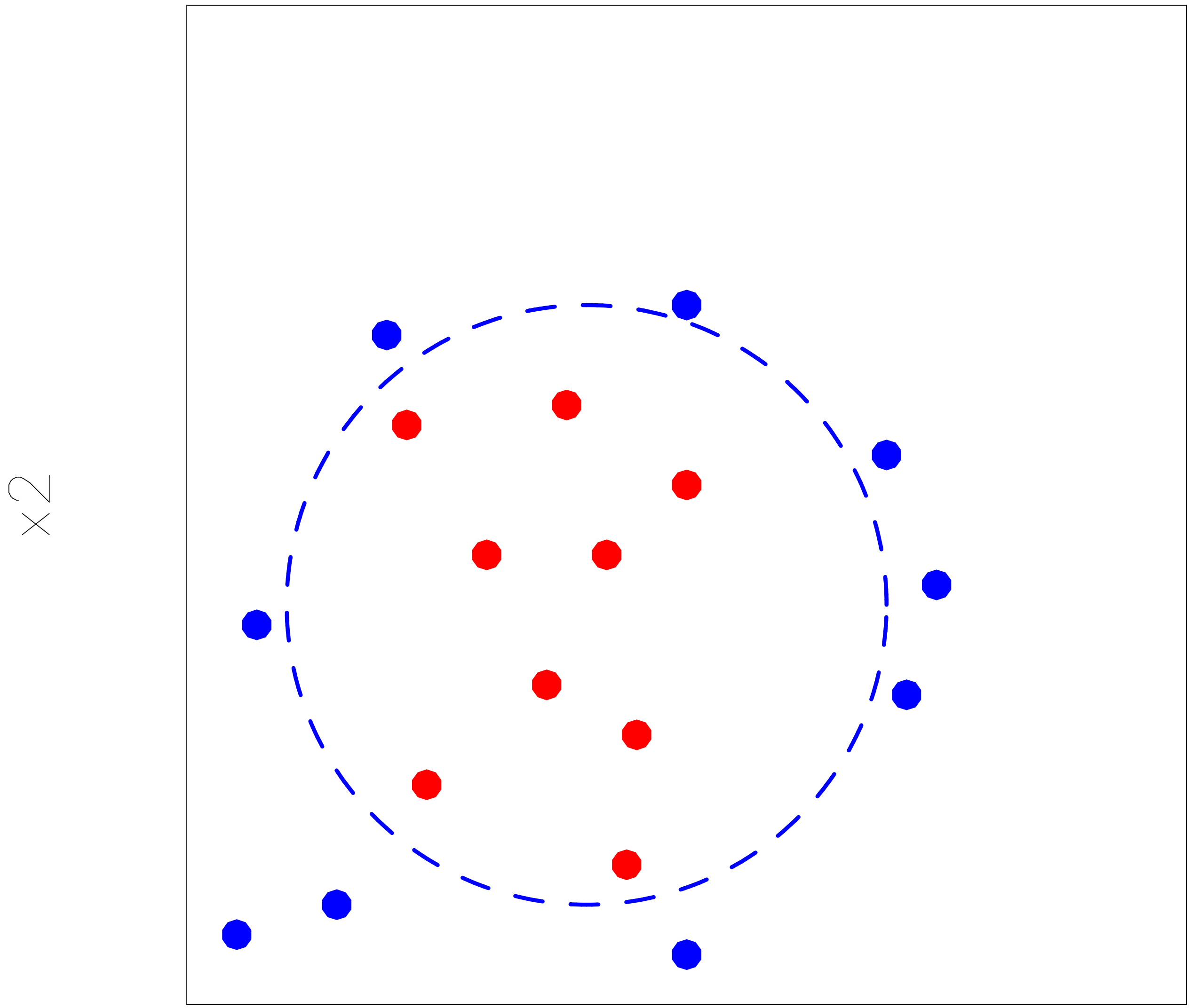
$$f = \sum_i \alpha_i y_i \mathbf{x}_i.\mathbf{s} + b,$$

$$f > 0. \quad |\mathbf{s} \rightarrow class1$$
$$f < 0. \quad |\mathbf{s} \rightarrow class2.$$

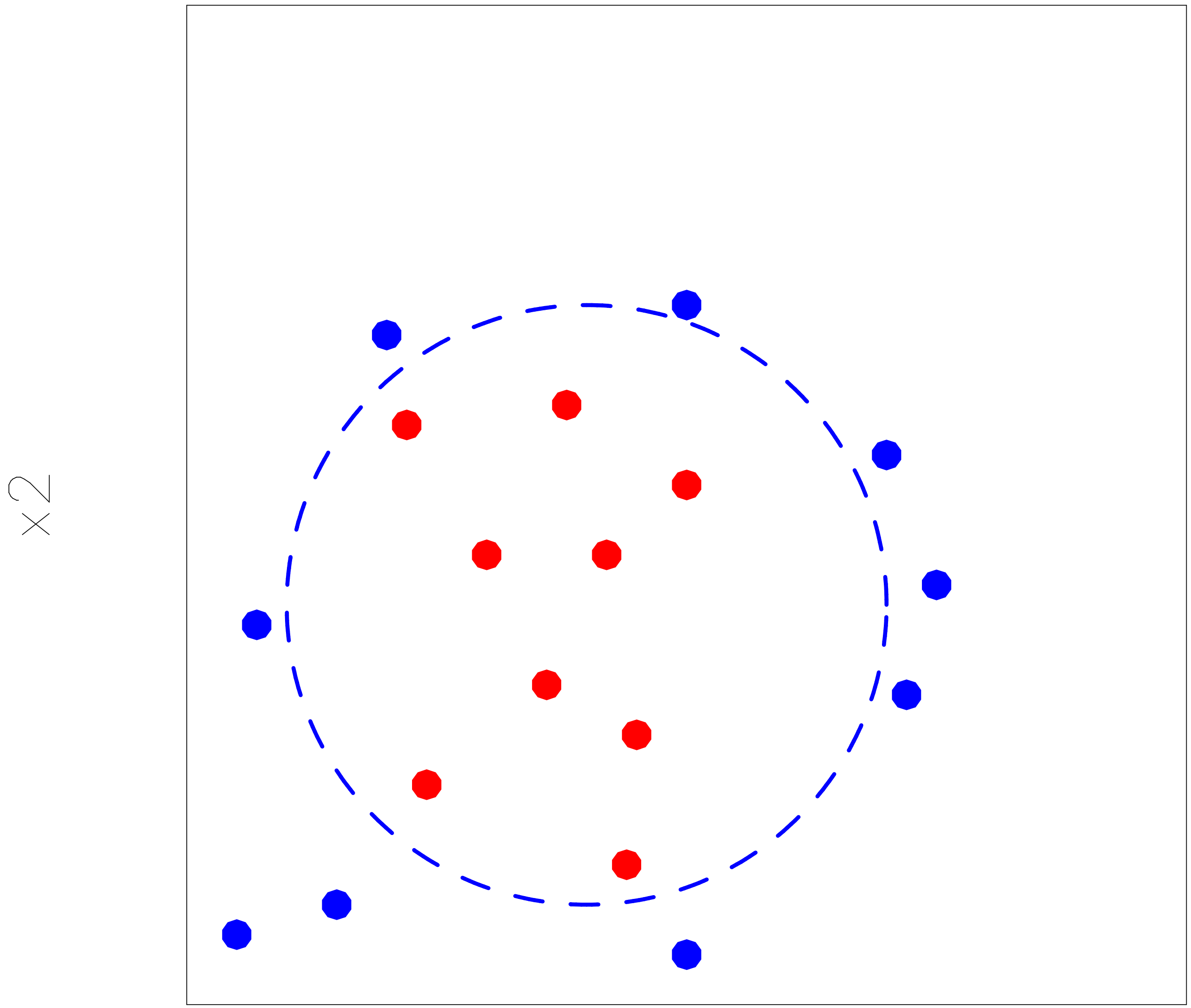
Linearly inseparable data



Linearly inseparable data

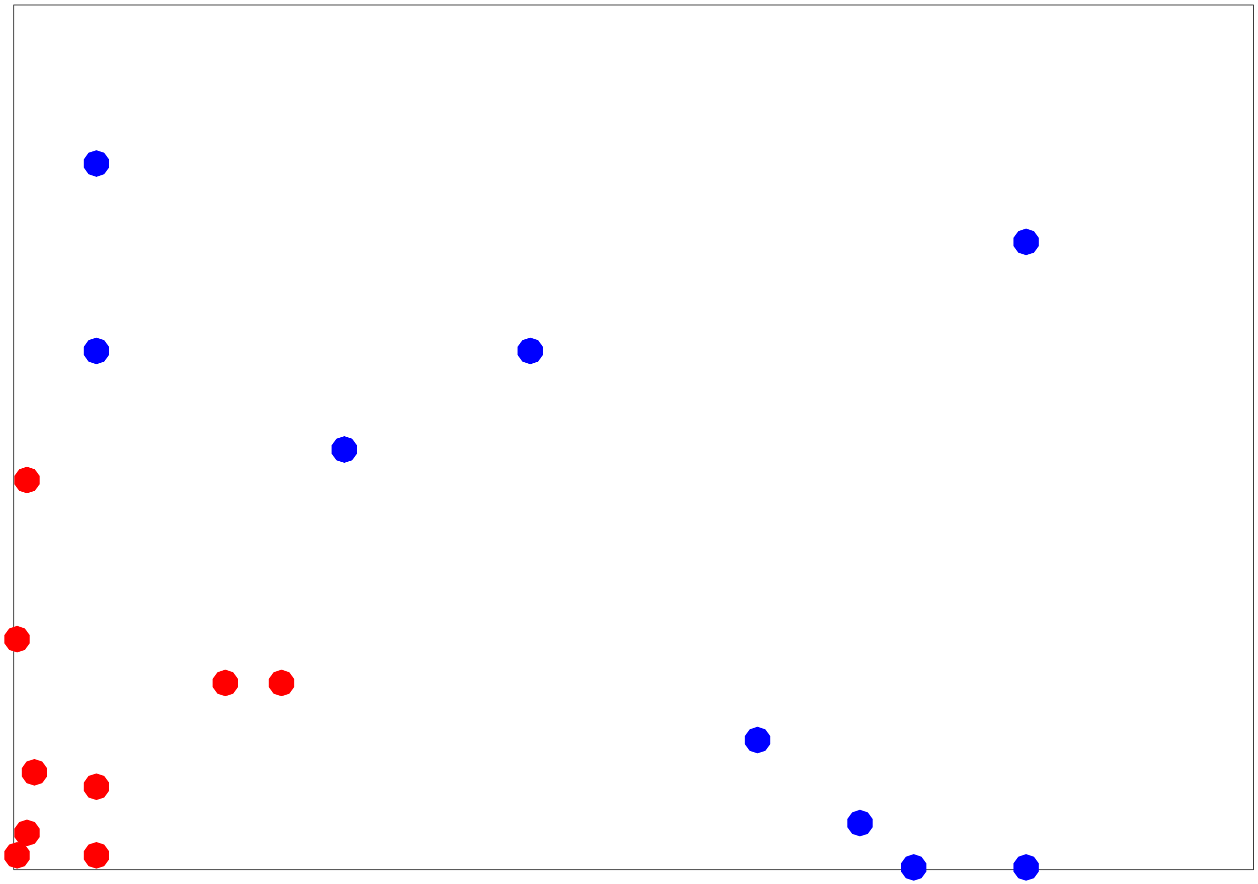


Linearly inseparable data



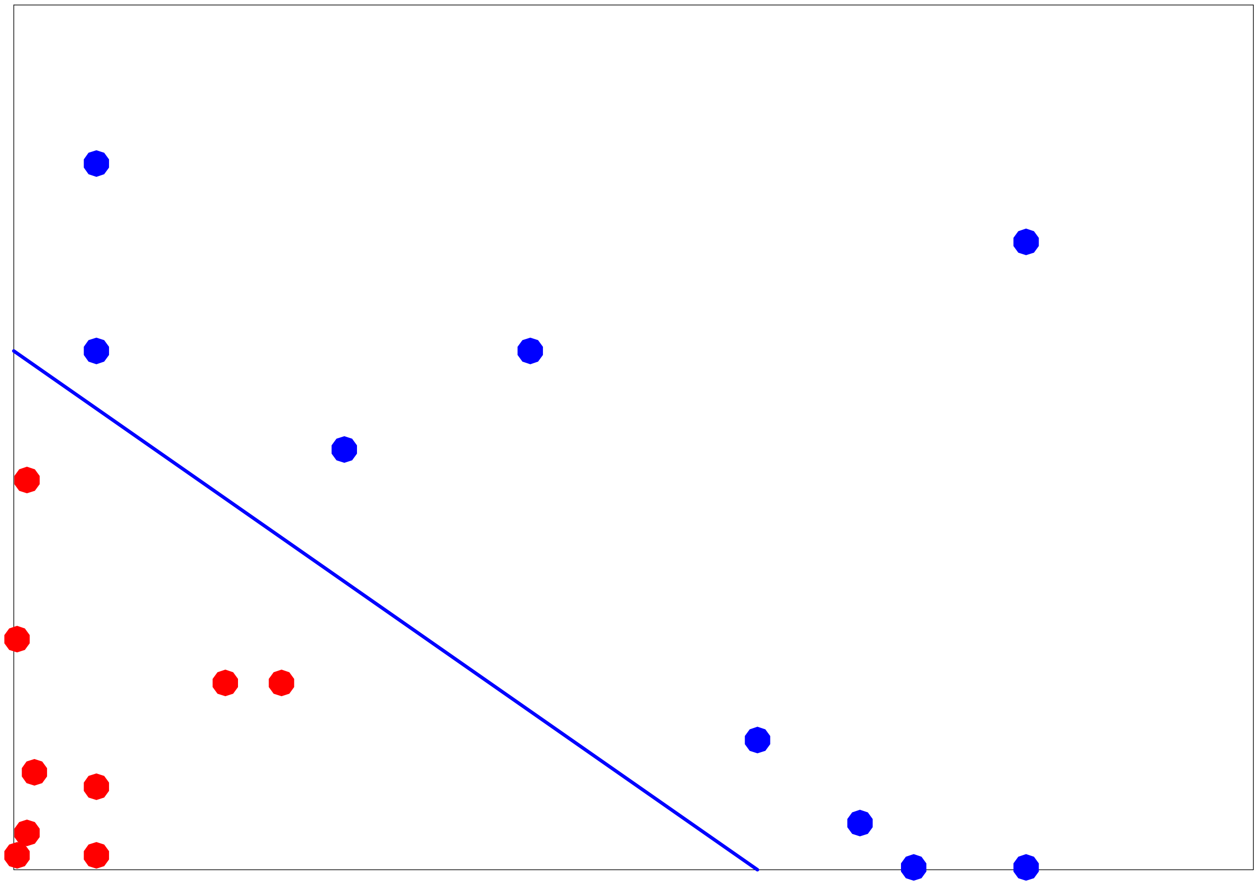
$$u, v \longrightarrow u^2, v^2$$

Linearly inseparable data



$$u, v \longrightarrow u^2, v^2$$

Linearly inseparable data



Kernel trick

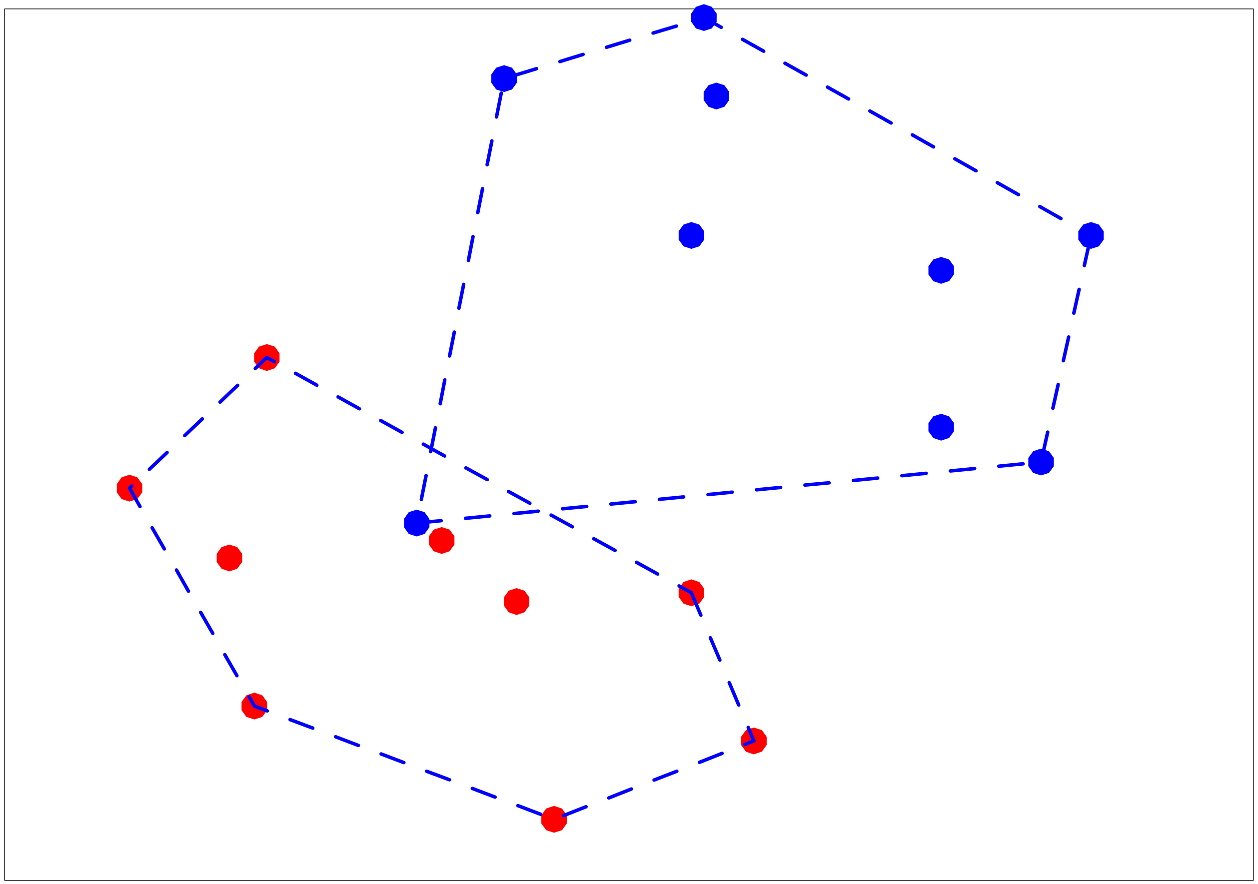
$$f = \sum_i \alpha_i y_i \mathbf{x}_i.\mathbf{s} + b,$$

replace dot products with Kernel function.
Most commonly radial basis function

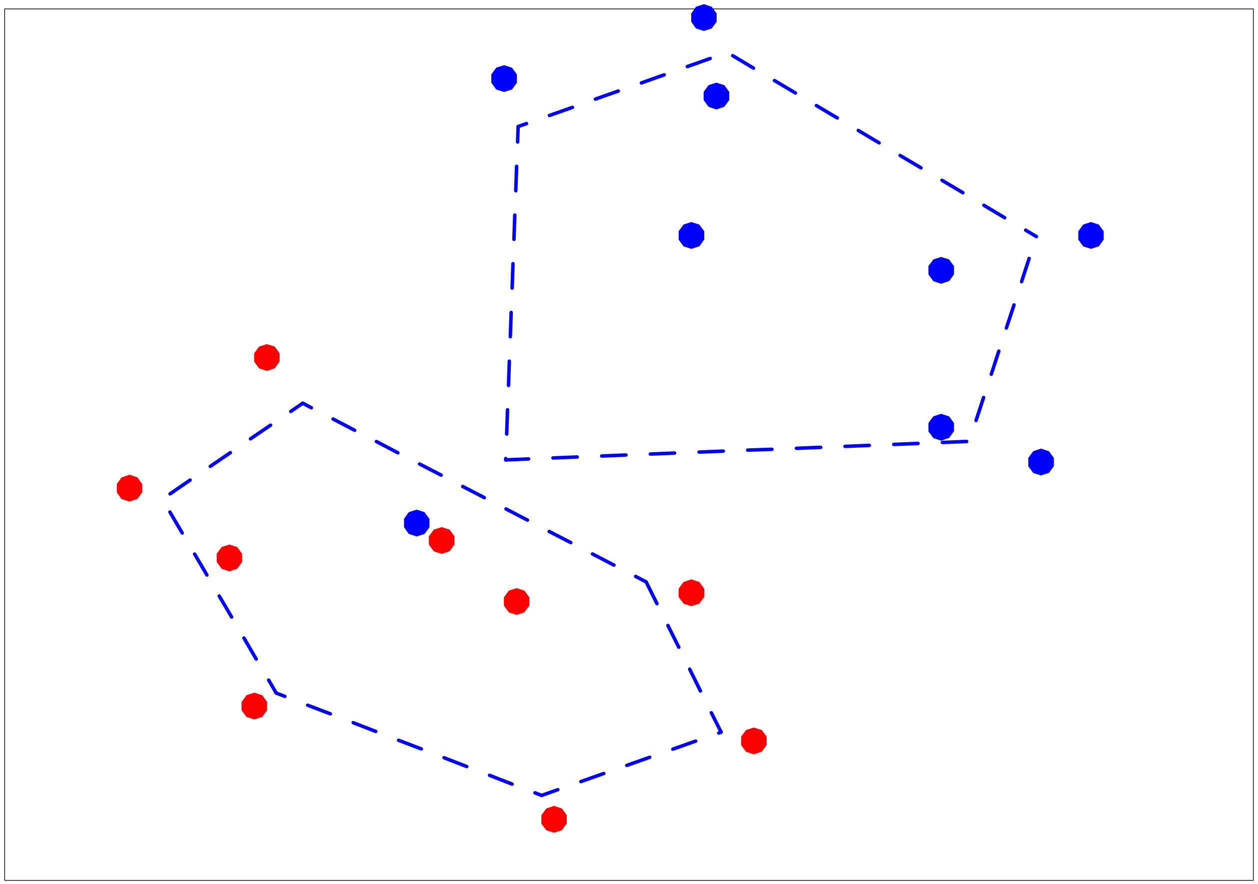
$$\exp\left(-\gamma \|x_i - s\|^2\right)$$

need to tune for γ.

Svm: Non separable cases

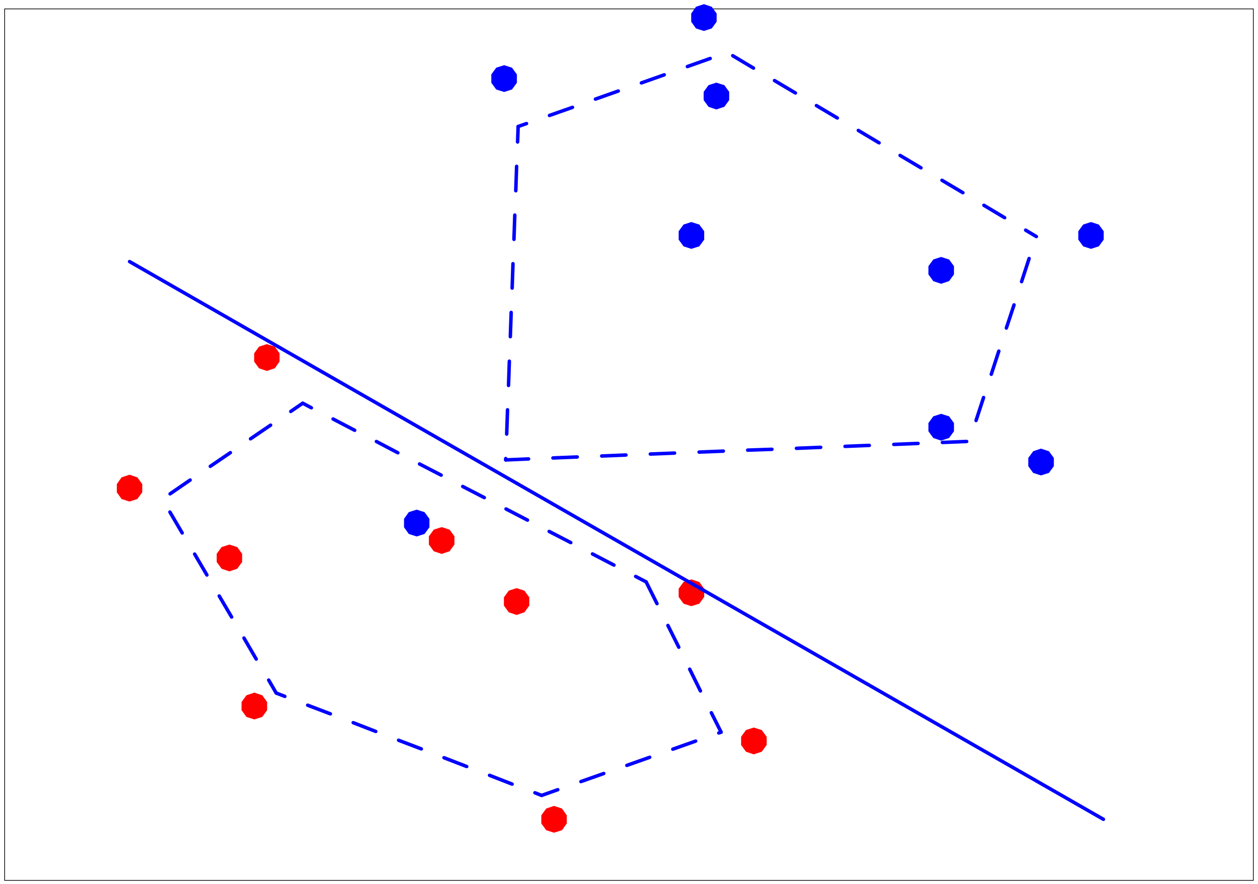


Svm: Non separable cases



$$0 \leq \alpha_i \leq D$$

Svm: Non separable cases



$$0 \leq \alpha_i \leq D$$

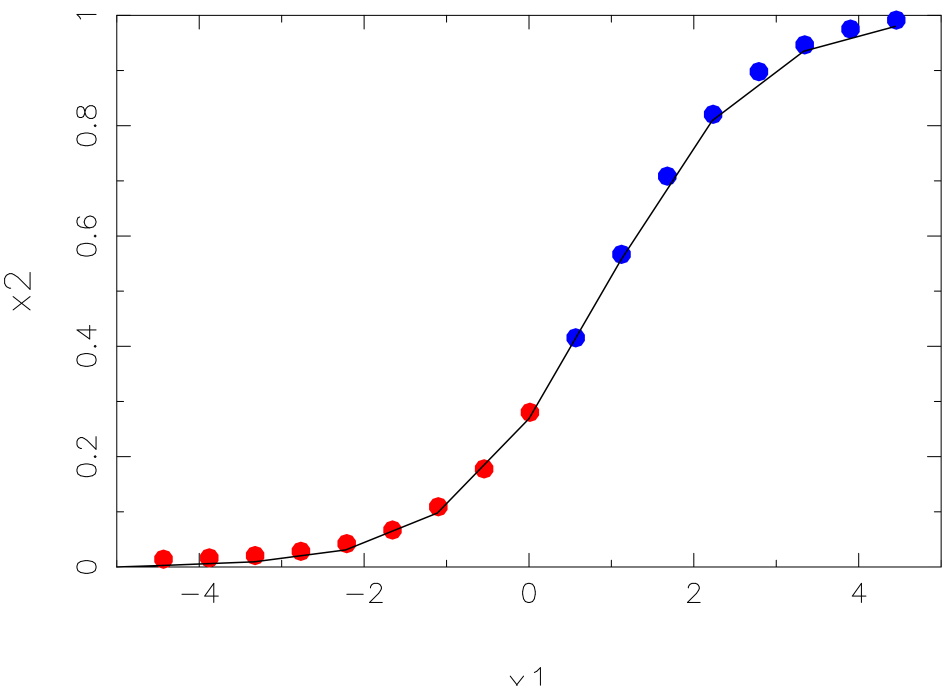
Svm

Svm's are used for

- object classification (DSC)
- stellar parameterization in Gsp-Phot
- outlier detection

Probabilities

$$P(class|f) = \frac{1}{1 + \exp\left(Af + B\right)}$$



Probabilities: pairwise coupling

- classify pairwise (star vs. binary, star vs. quasar etc.)
- Can combine all pairwise probabilities to unique set of $P(C_m)$

(Wu et al. 2004,Journal of Machine learning research)

Object classification

Currently works with 7 distinct classes. We consider 4.

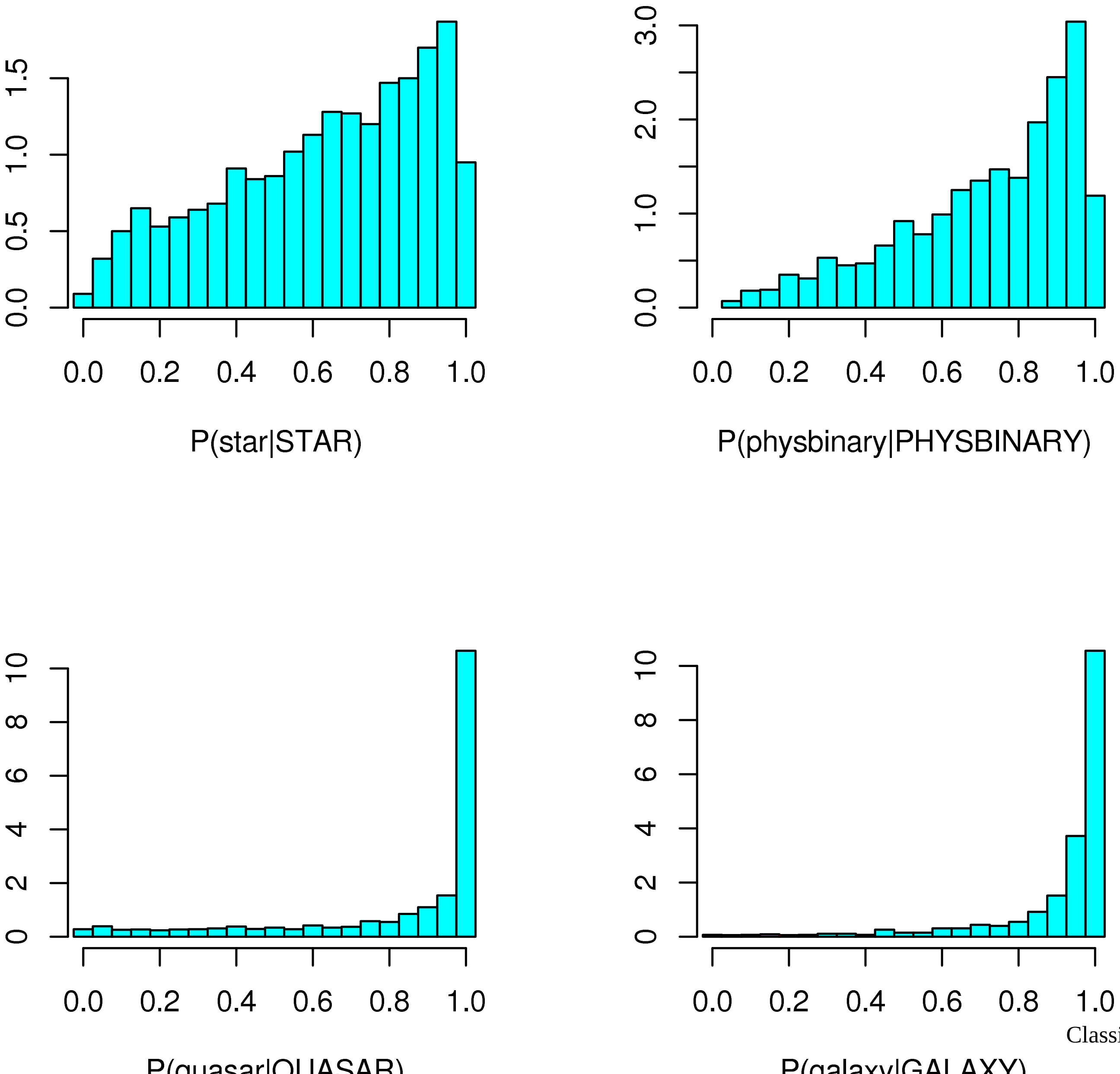
G=18.5				
Stars	81.90	15.70	1.90	0.50
Binaries	10.15	89.80	0.00	0.05
Quasars	5.45	0.55	91.60	2.40
Galaxies	1.55	0.05	0.10	98.30

Object classification

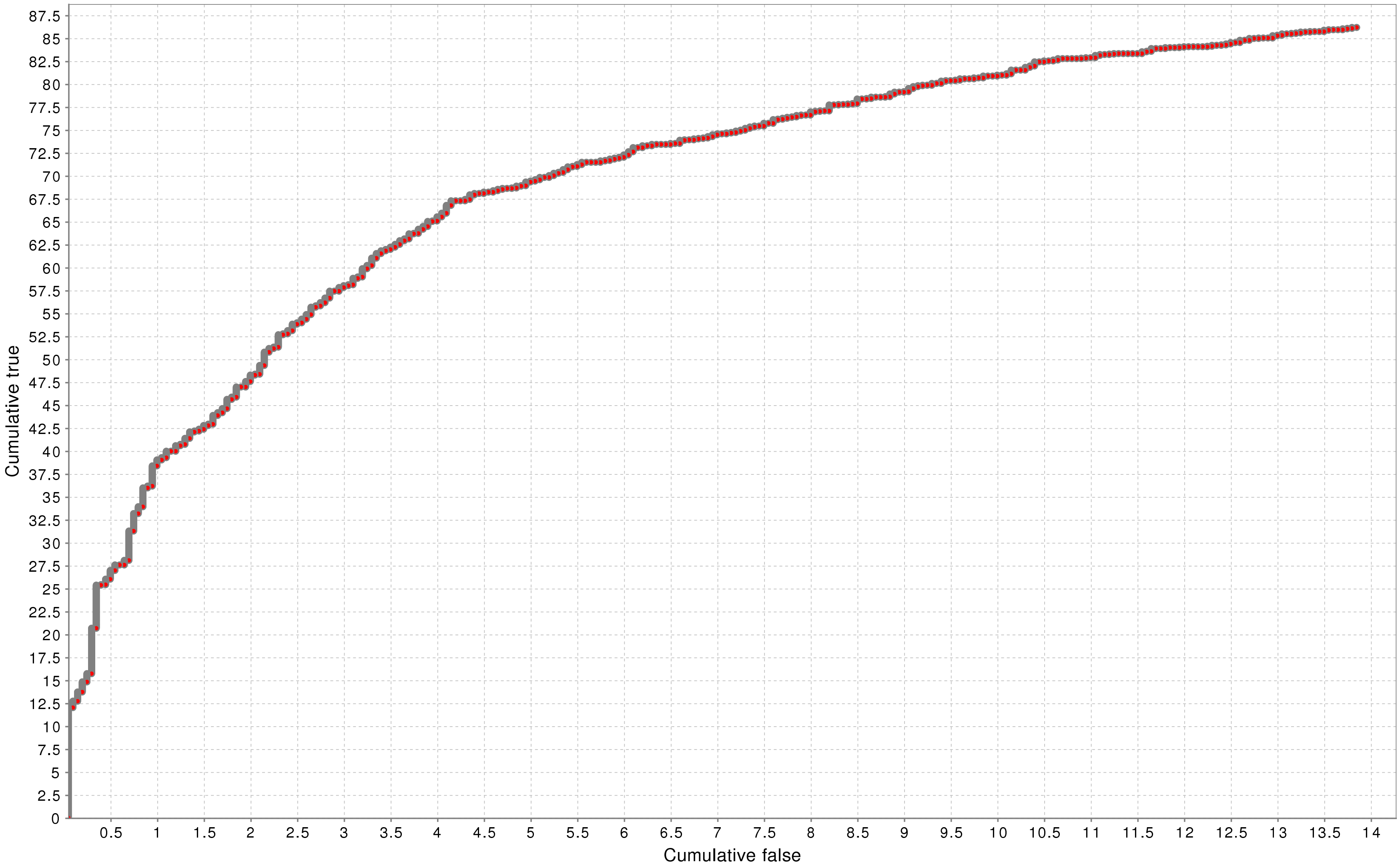
Currently works with 7 distinct classes. We consider 4.

G=18.5				
Stars	81.90	15.70	1.90	0.50
Binaries	10.15	89.80	0.00	0.05
Quasars	5.45	0.55	91.60	2.40
Galaxies	1.55	0.05	0.10	98.30

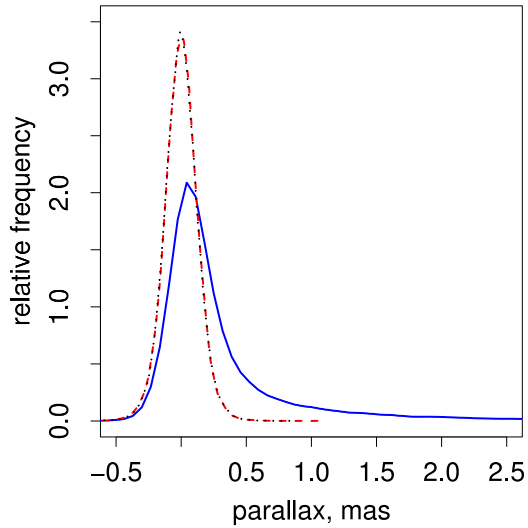
Classification performance

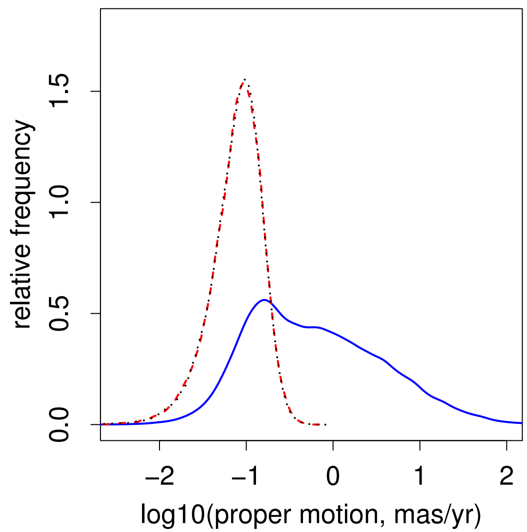


Classification performance



Astrometric data





Incorporating astrometry

Can simply append to BP/RP spectrum.
With probability, can also classify separately and combine

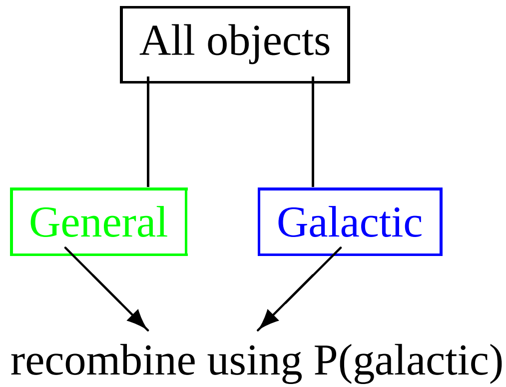
separate classifiers

Branch then classify

	Stars	Binaries	Quasars	Galaxies
Stars	90.82	9.18	0.00	0.00
Binaries	8.87	91.13	0.00	0.00
Quasars	2.04	0.90	95.77	1.28
Galaxies	0.00	0.00	0.62	99.38

SVM including parallax

Stars	93.52	6.03	0.45	0.00
Binaries	6.38	93.62	0.00	0.00
Quasars	0.76	0.14	98.91	0.19
Galaxies	0.00	0.00	0.41	99.59



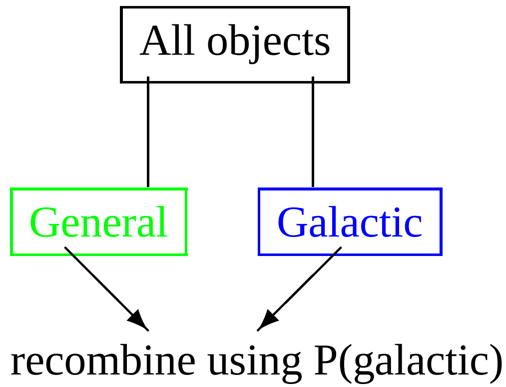
separate classifiers

Branch then classify

	Stars	Binaries	Quasars	Galaxies
Stars	90.82	9.18	0.00	0.00
Binaries	8.87	91.13	0.00	0.00
Quasars	2.04	0.90	95.77	1.28
Galaxies	0.00	0.00	0.62	99.38

SVM including parallax

	Stars	Binaries	Quasars	Galaxies
Stars	93.52	6.03	0.45	0.00
Binaries	6.38	93.62	0.00	0.00
Quasars	0.76	0.14	98.91	0.19
Galaxies	0.00	0.00	0.41	99.59



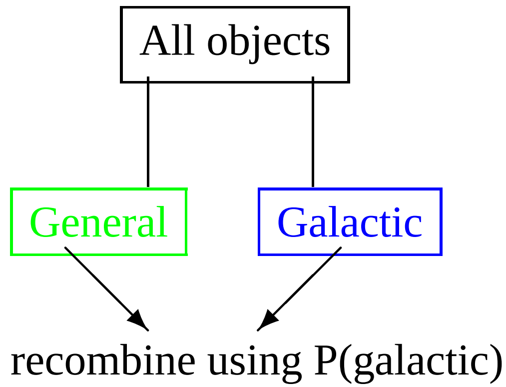
separate classifiers

Branch then classify

	Stars	Binaries	Quasars	Galaxies
Stars	90.82	9.18	0.00	0.00
Binaries	8.87	91.13	0.00	0.00
Quasars	2.04	0.90	95.77	1.28
Galaxies	0.00	0.00	0.62	99.38

SVM including parallax

Stars	93.52	6.03	0.45	0.00
Binaries	6.38	93.62	0.00	0.00
Quasars	0.76	0.14	98.91	0.19
Galaxies	0.00	0.00	0.41	99.59



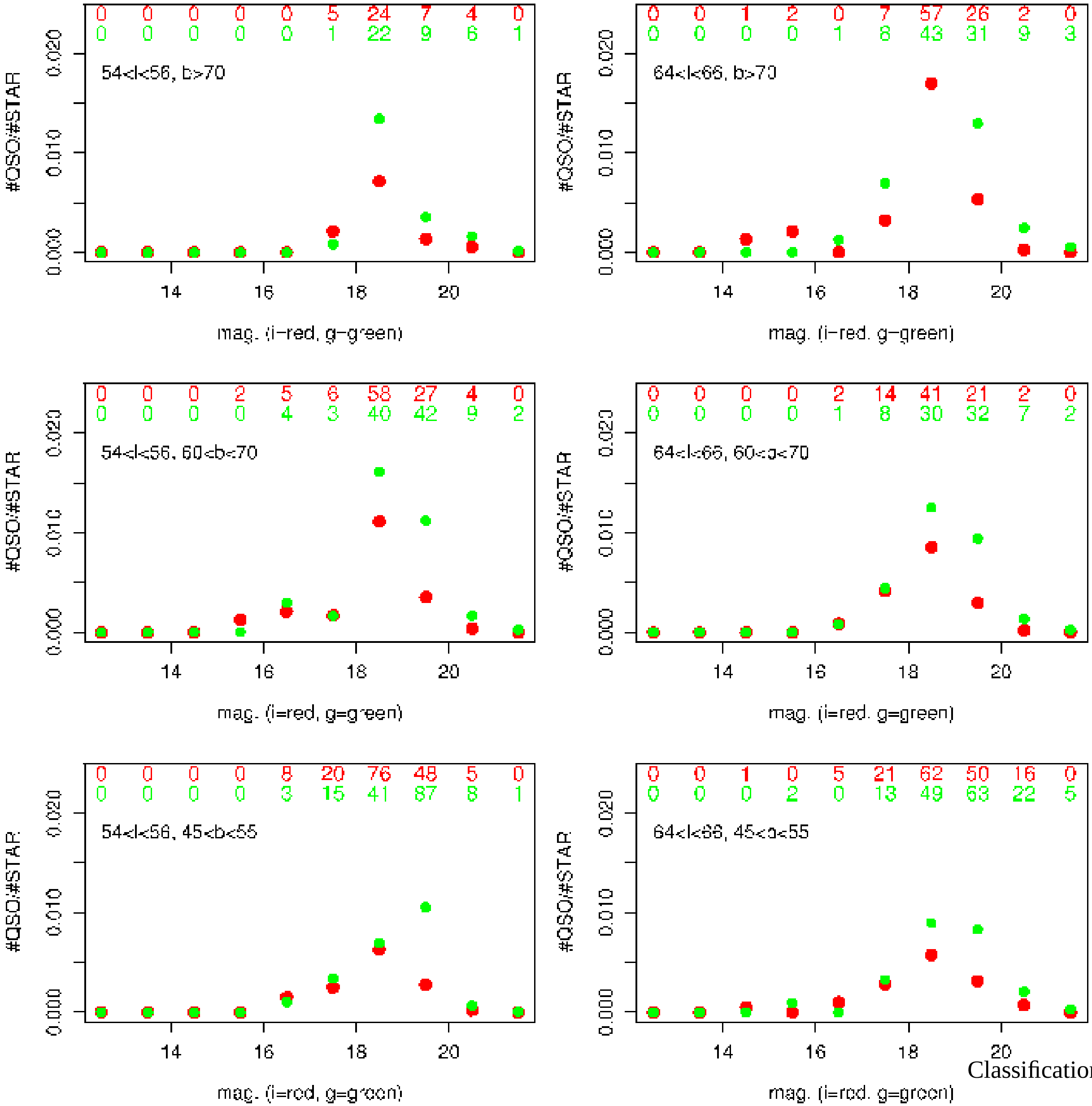
QSO sample

so far, all classifications performed on flat class distributions, i.e. $N_{stars} = N_{\mathrm{quasars}} = N_{\mathrm{galaxies}}$ In reality,

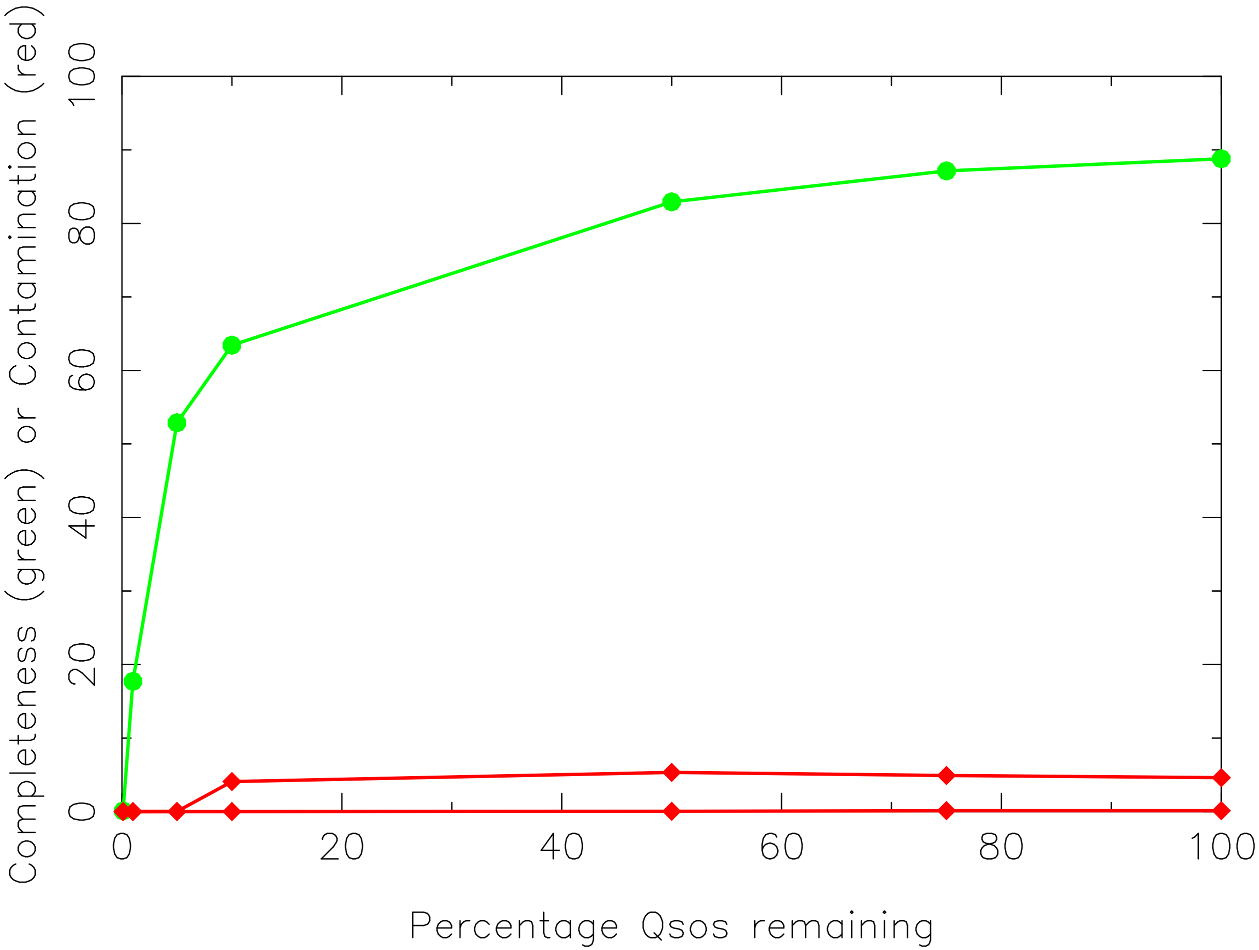
stars far outnumber quasars. Further, it is required that

Dsc provides clean QSO sample for the reference frame.

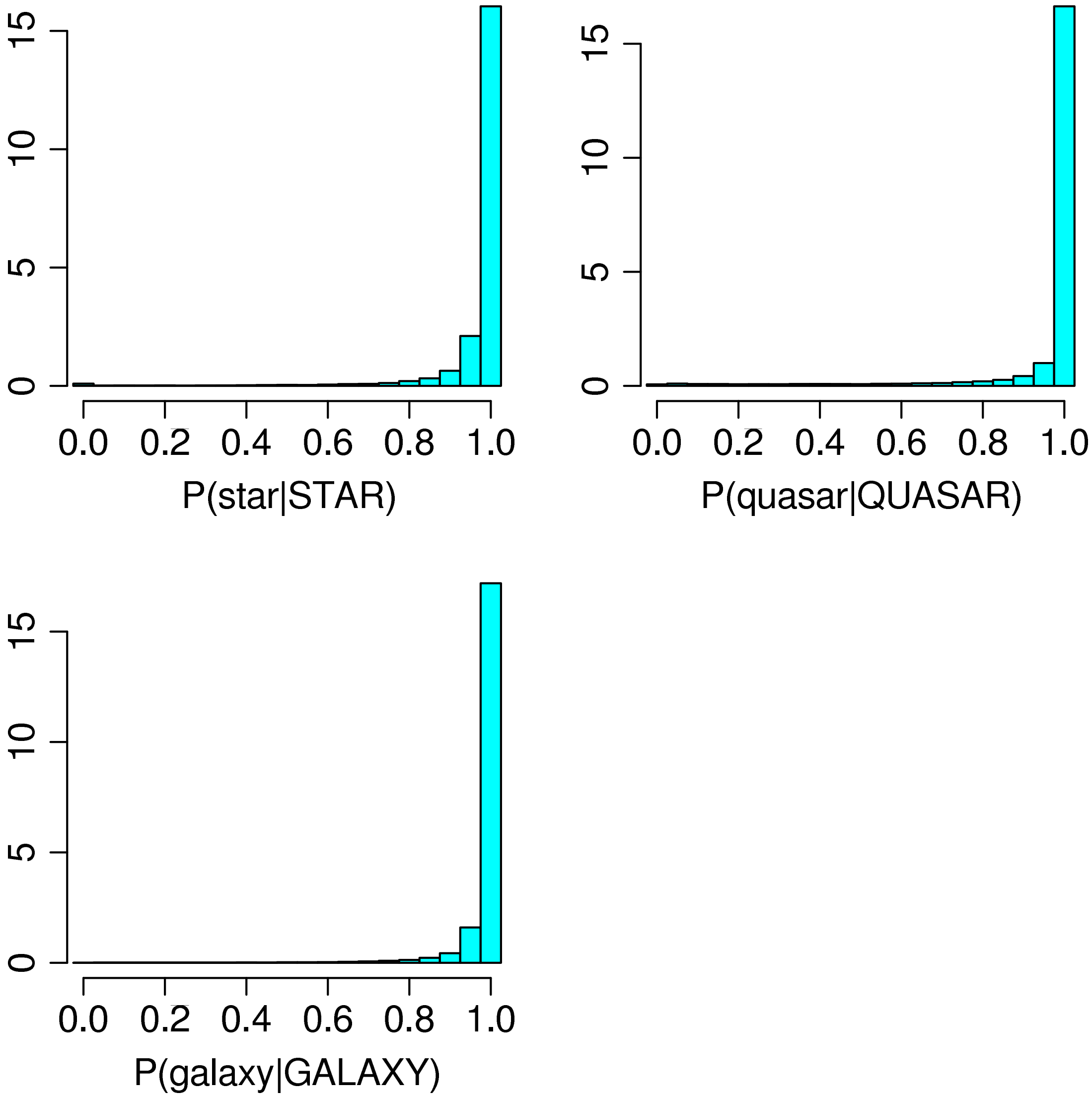
QSO distribution



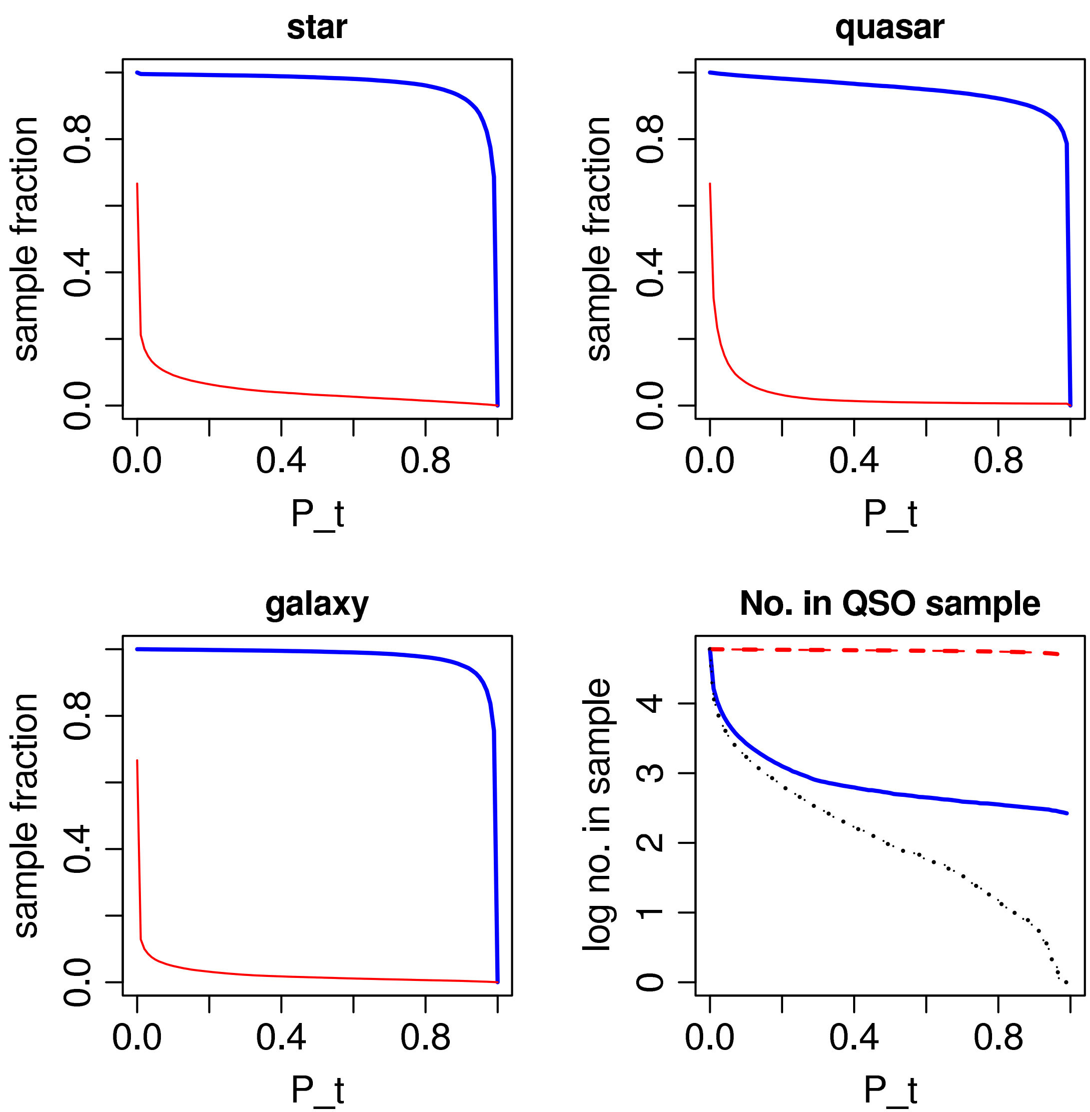
QSO training fraction



Normal classification



Normal classification



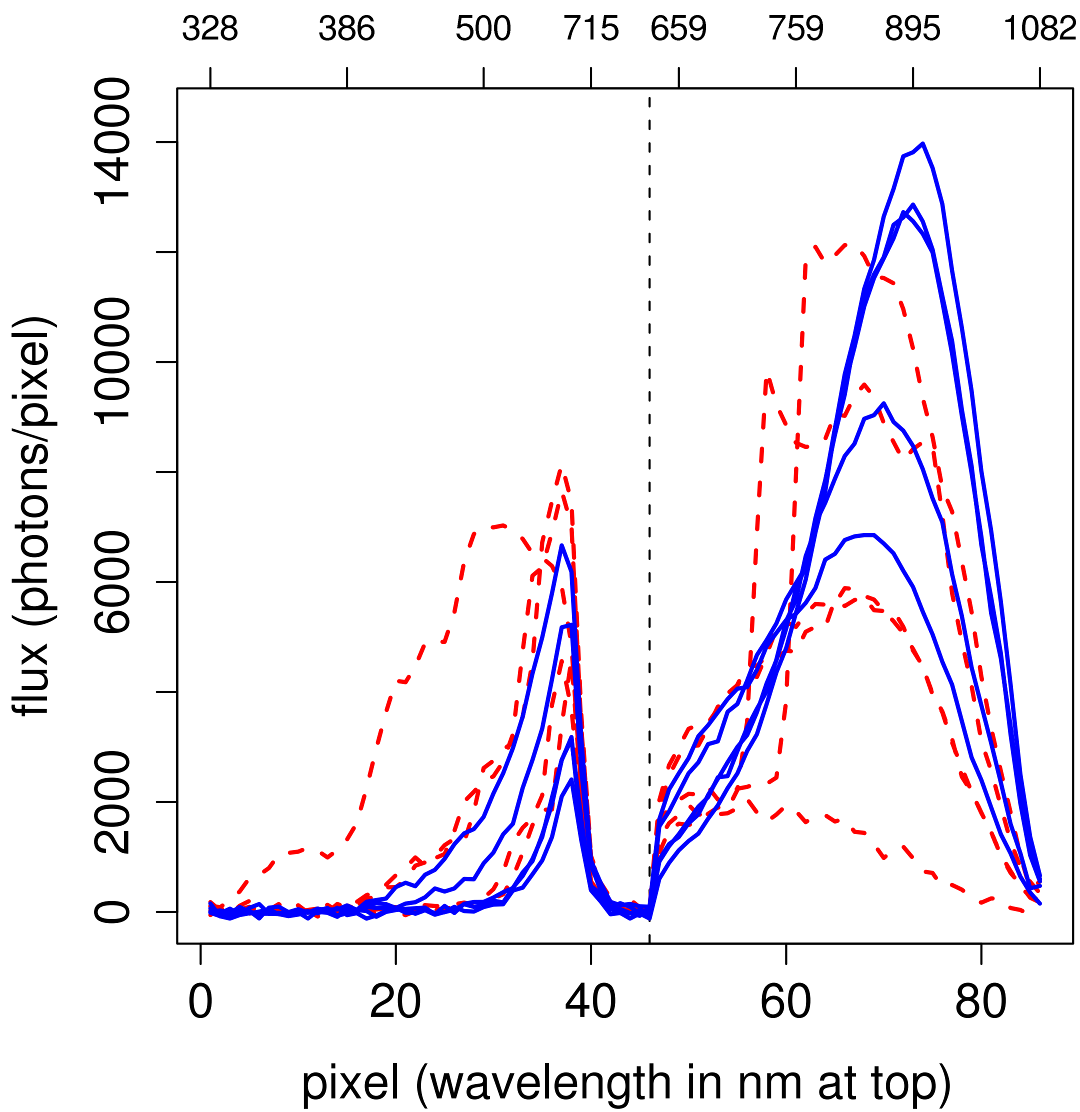
But, for rare objects...

If have large probability from classifier that object is of type A

but class A is rare

it is more likely a misclassification than that the object truly is class A

Low EW QSOs



328
386
500
715
659
759
895
1082
flux (photons/pixel)
0
2000
6000
10000
14000
0
20
40
60
80
pixel (wavelength in nm at top)

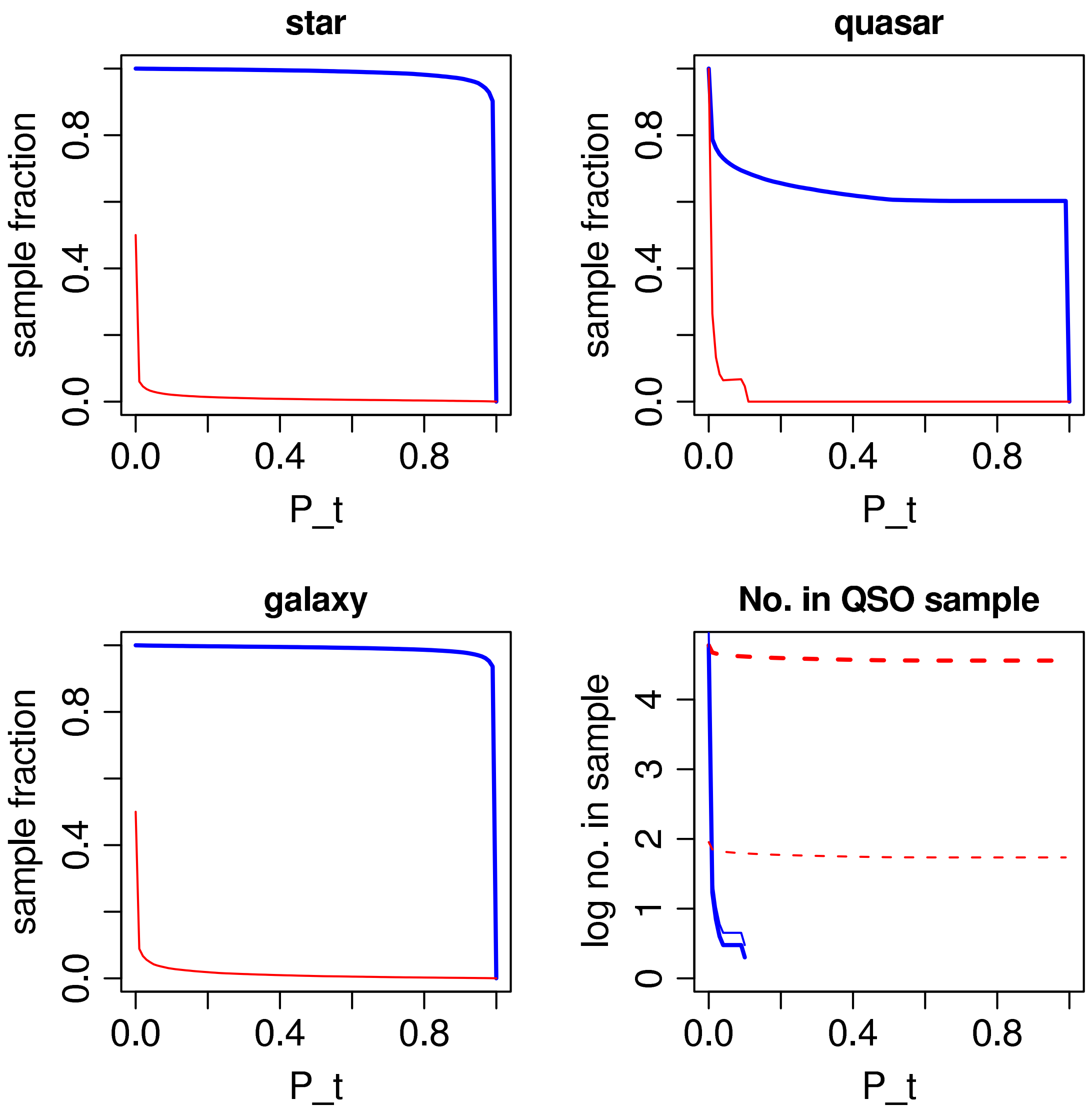
Strategy

correct for the imbalance in the test set to obtain better estimates of quasar purity in real cases.

Eliminate continuum quasars from the training set.

obtain better estimates of the quasar output sample purity

Modified probabilities



Summary

- Svm is a good candidate for the main Dsc classifier. Is robust, easy to use, needs only two parameters to be tuned
- Can estimate probabilities based on decision value, but this is not ideal.
- Can use knowledge of the quasar prior in different parts of the sky to correct the ’nominal’ model and estimate thresholds for pure quasar samples