Automated classification of Gaia sources

Kester Smith, Gaia AP team

Overview

- Introduction to Gaia
- Introduction to Machine Learning
- Application of ML to Gaia data analysis
- Detailed view of the Discrete Source Classifier

Overview

- Introduction to Gaia
- Introduction to Machine Learning
- Application of ML to Gaia data analysis
- Detailed view of the Discrete Source Classifier

Overview

- Introduction to Gaia
- Introduction to Machine Learning
- Application of ML to Gaia data analysis
- Detailed view of the Discrete Source Classifier

Overview

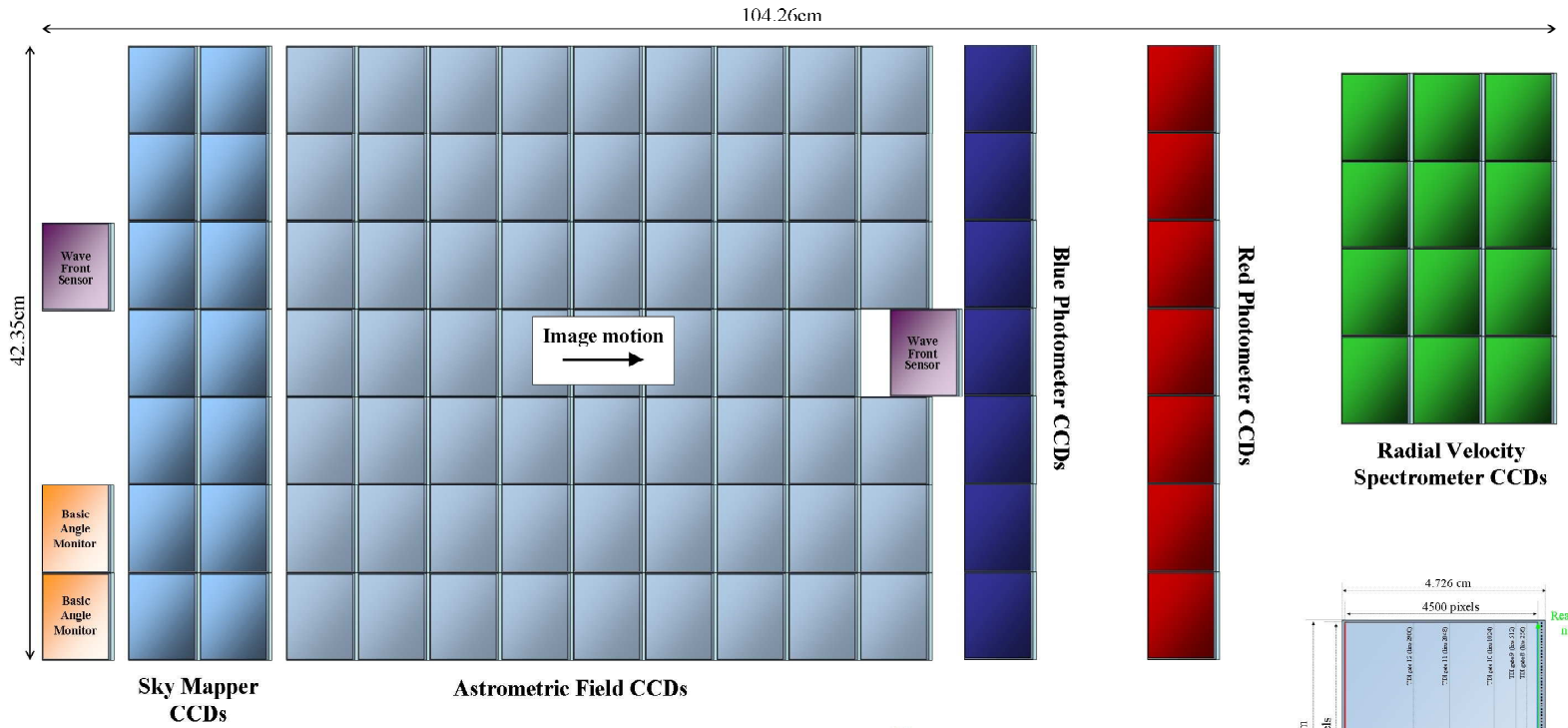
- Introduction to Gaia
- Introduction to Machine Learning
- Application of ML to Gaia data analysis
- Detailed view of the Discrete Source Classifier

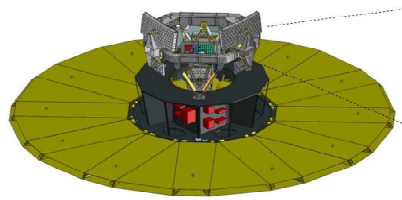
Mission characteristics.

- The satellite rotates to scan the whole sky
- 70-90 visits per field
- Limiting magnitude G=20
- Only point sources will be observed.
- $\sim$ 1 billion sources expected

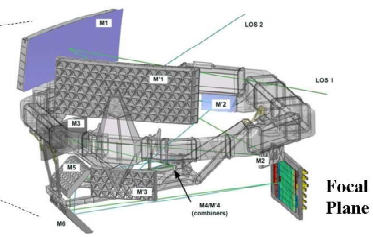
Gaia Focal Plane

106 CCDs ≈ 938 million pixels ≈ 2800 cm²

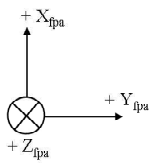


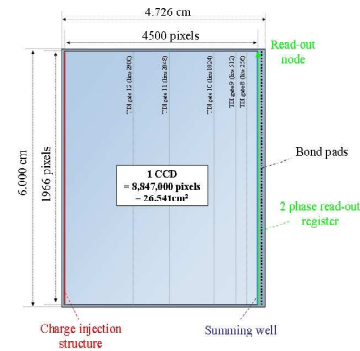


Gaia (MLI removed)



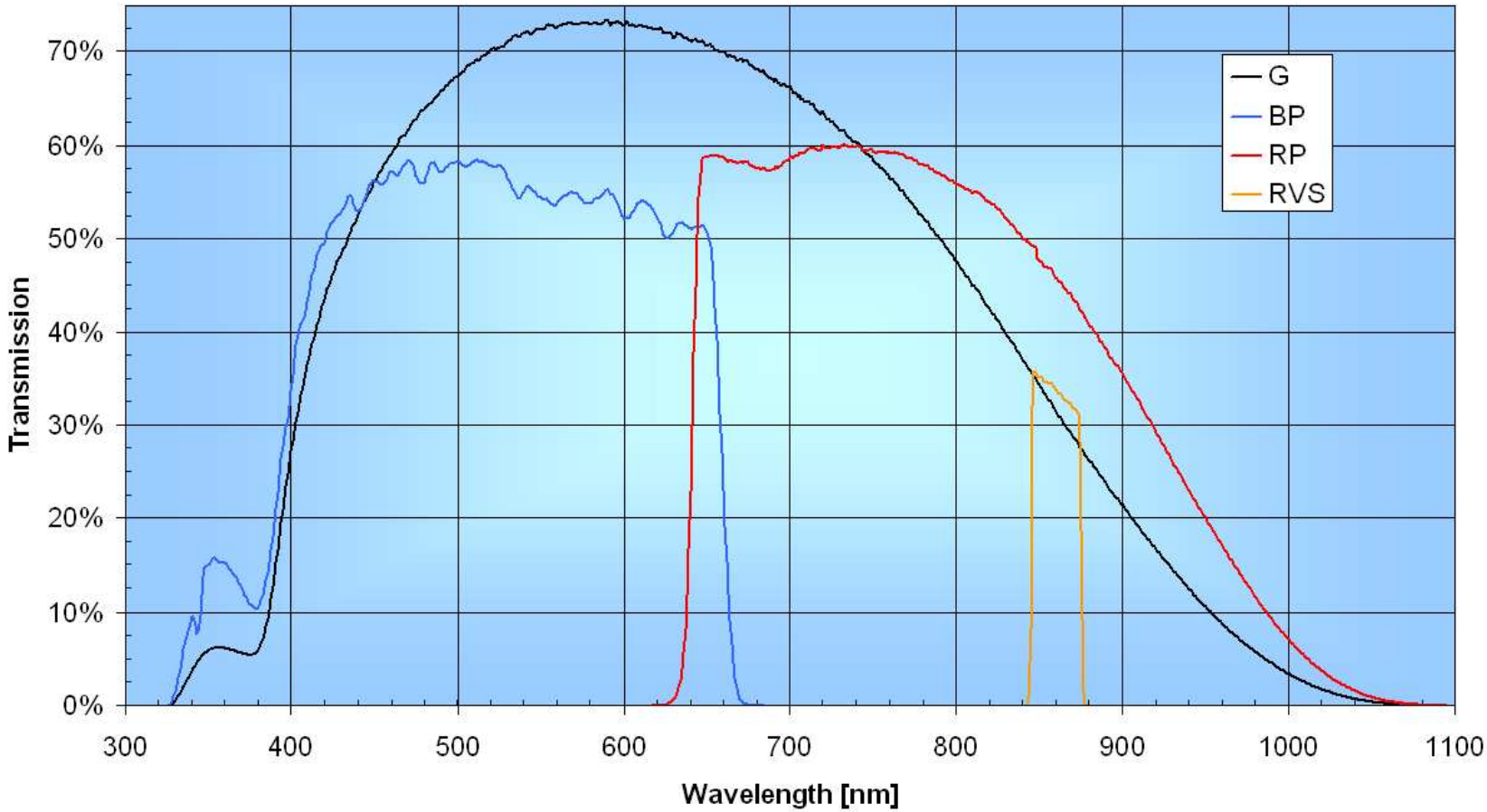
Payload Module (telescopes)

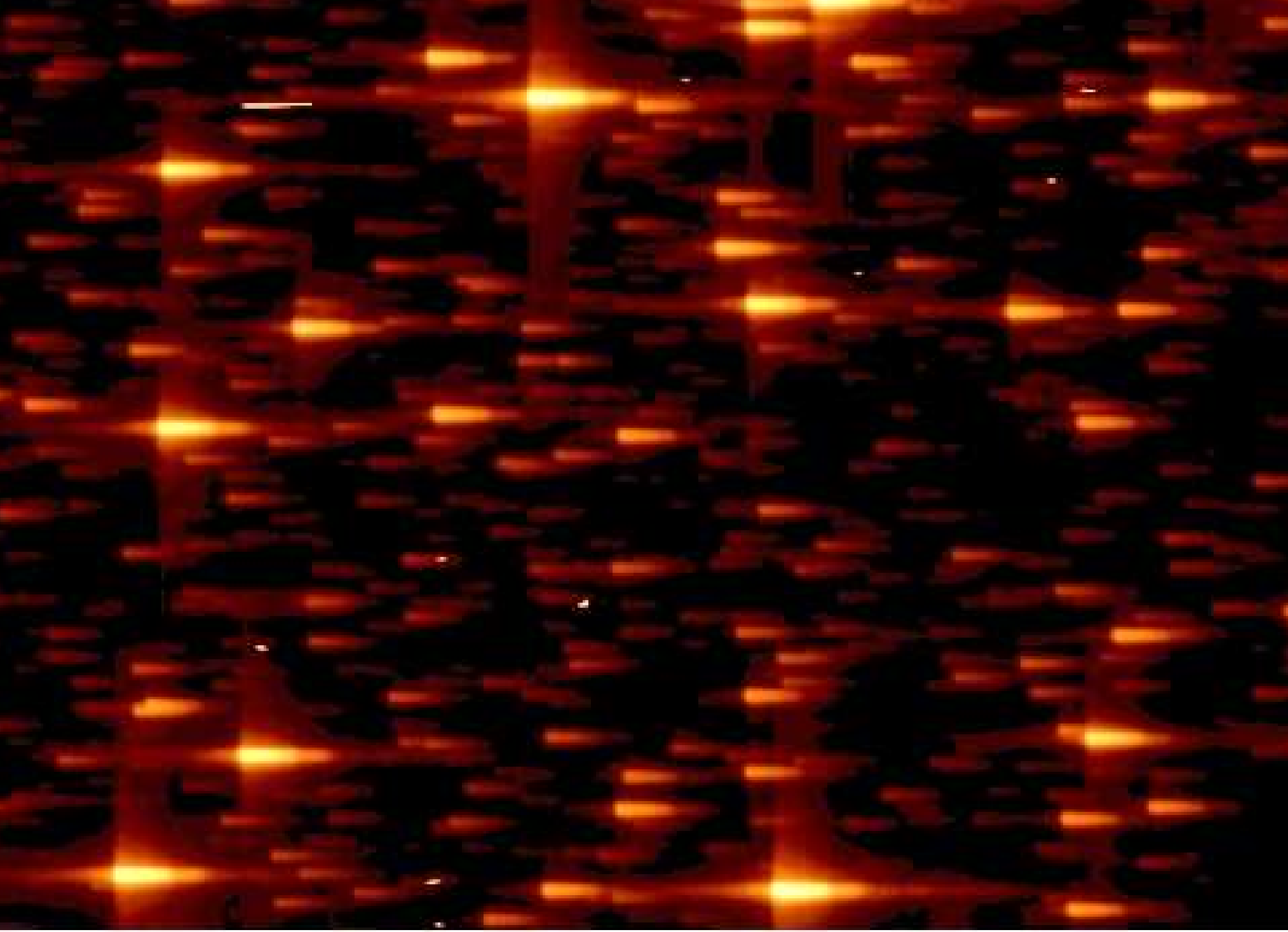


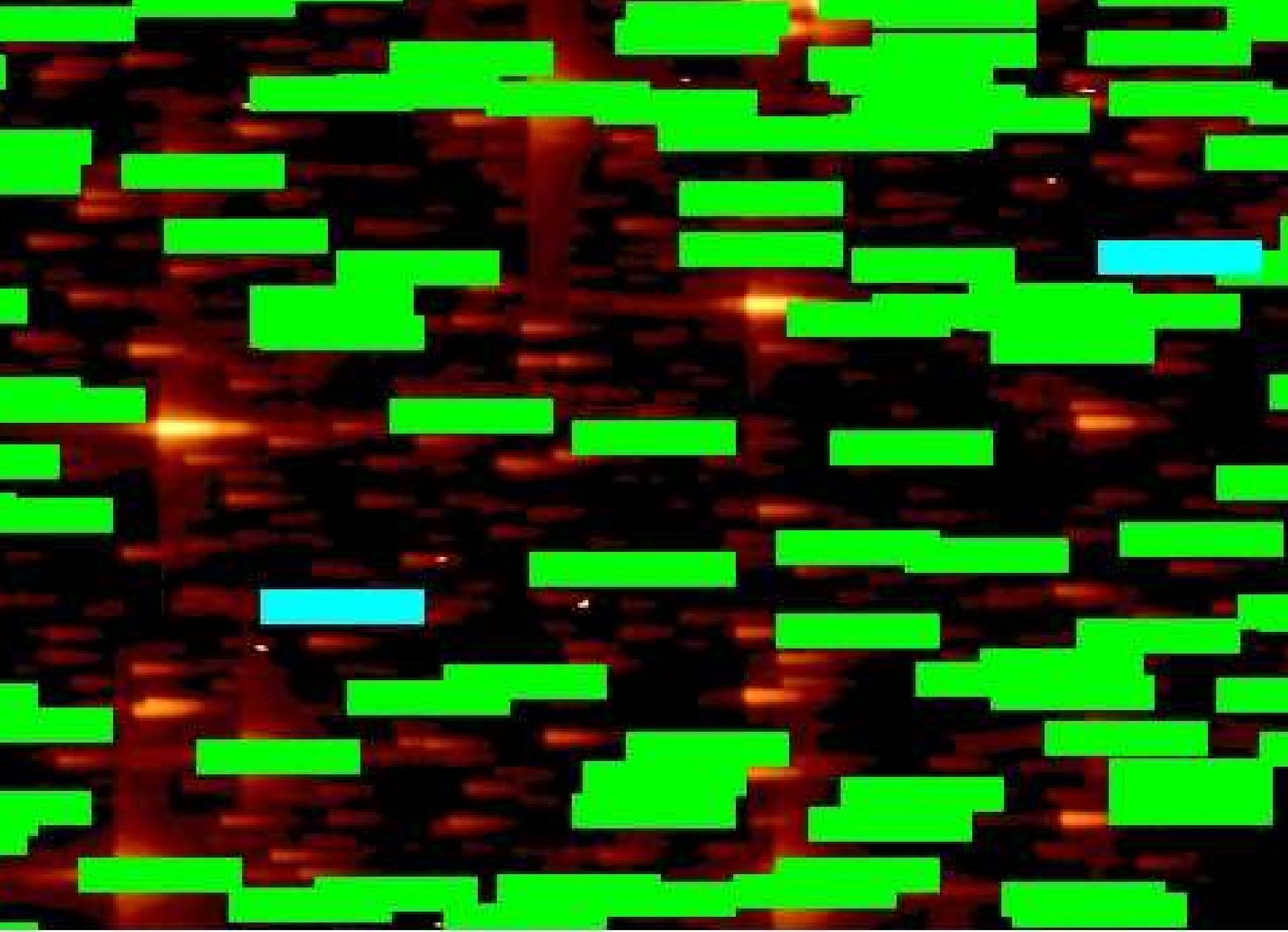


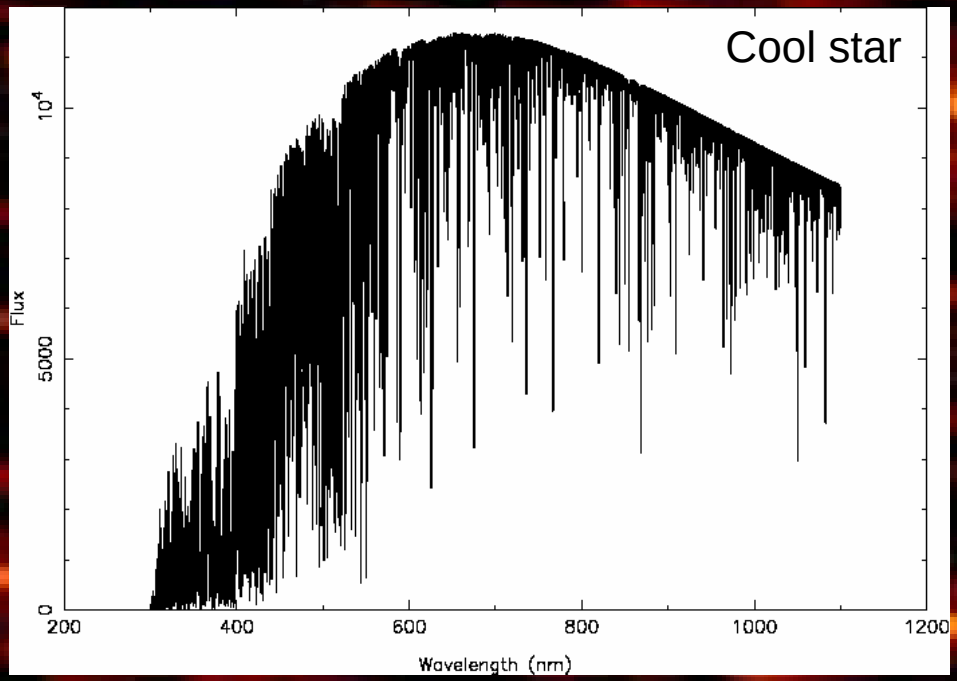
esa A.D. Short – July, 2009

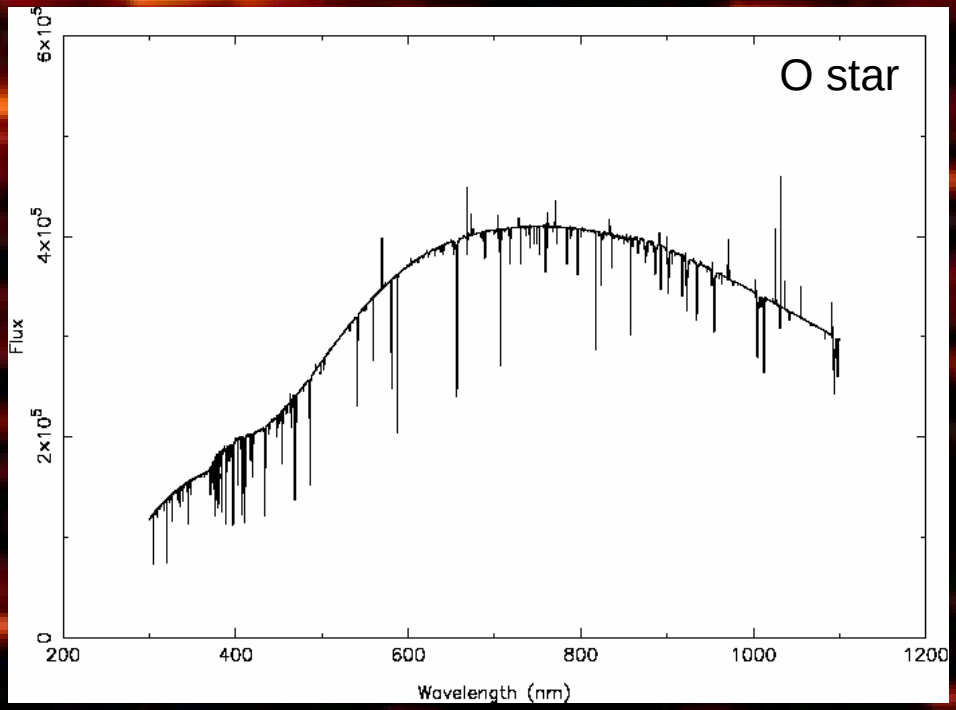
Photometric response

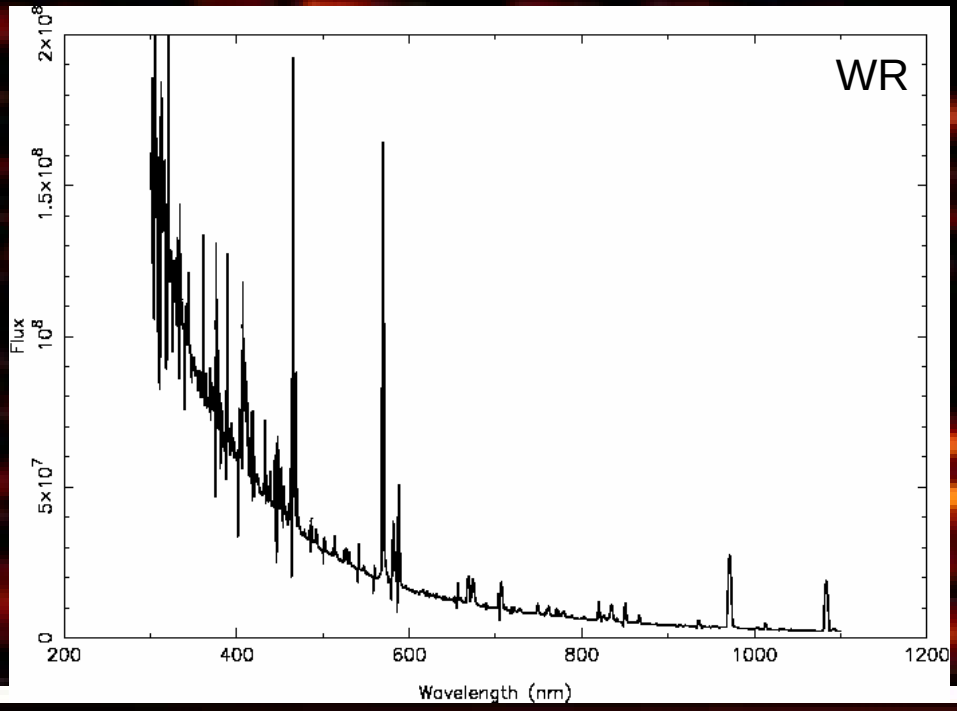


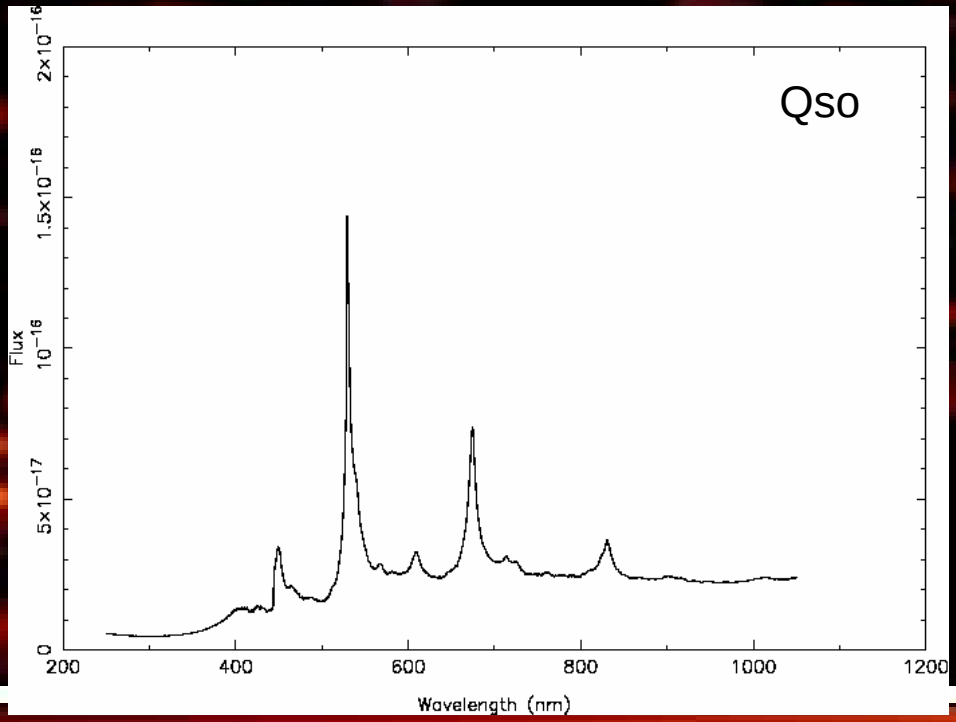


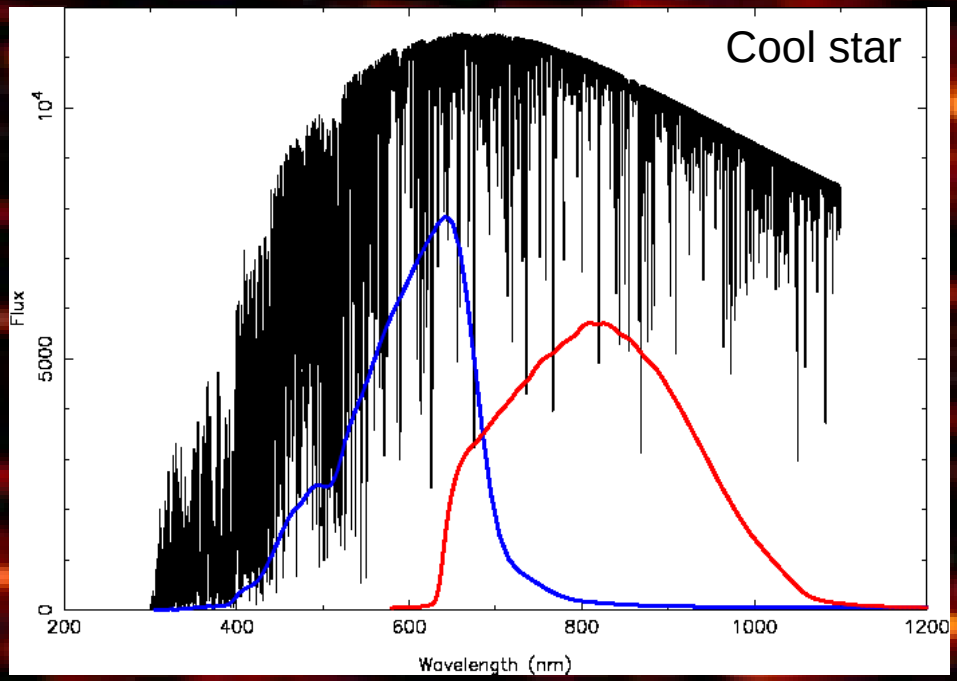


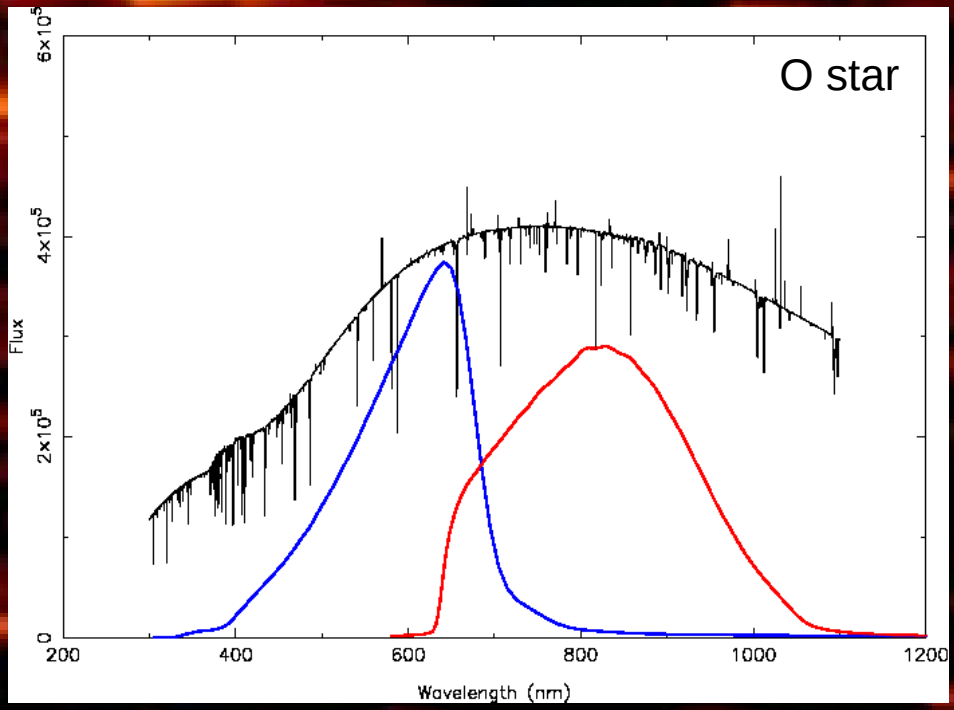

Cool star
Flux
Wavelength (nm)

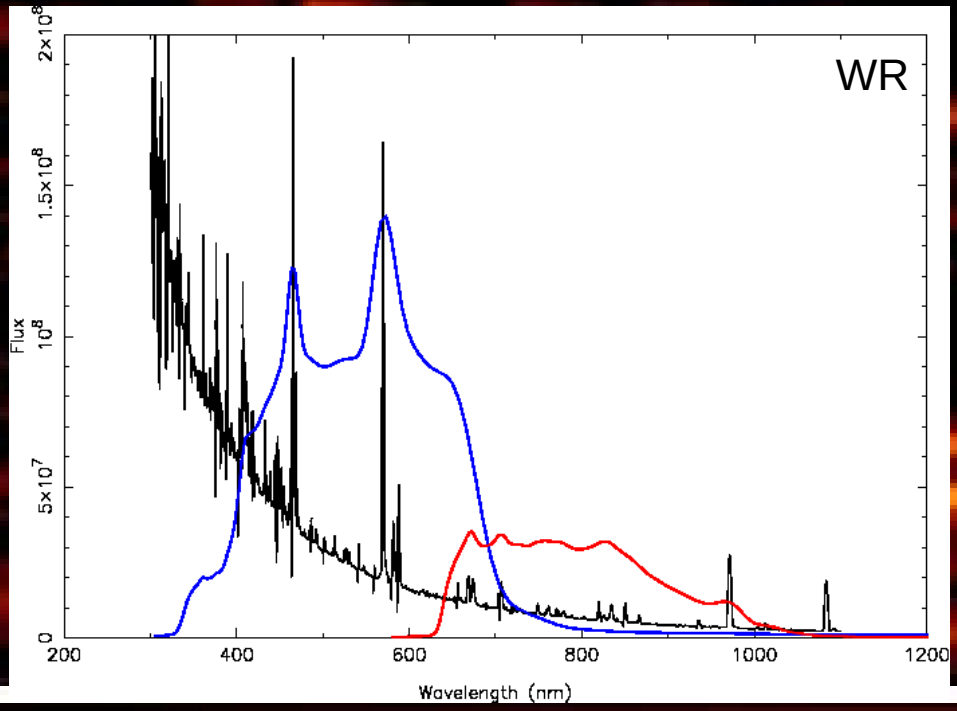

O star
Flux
Wavelength (nm)

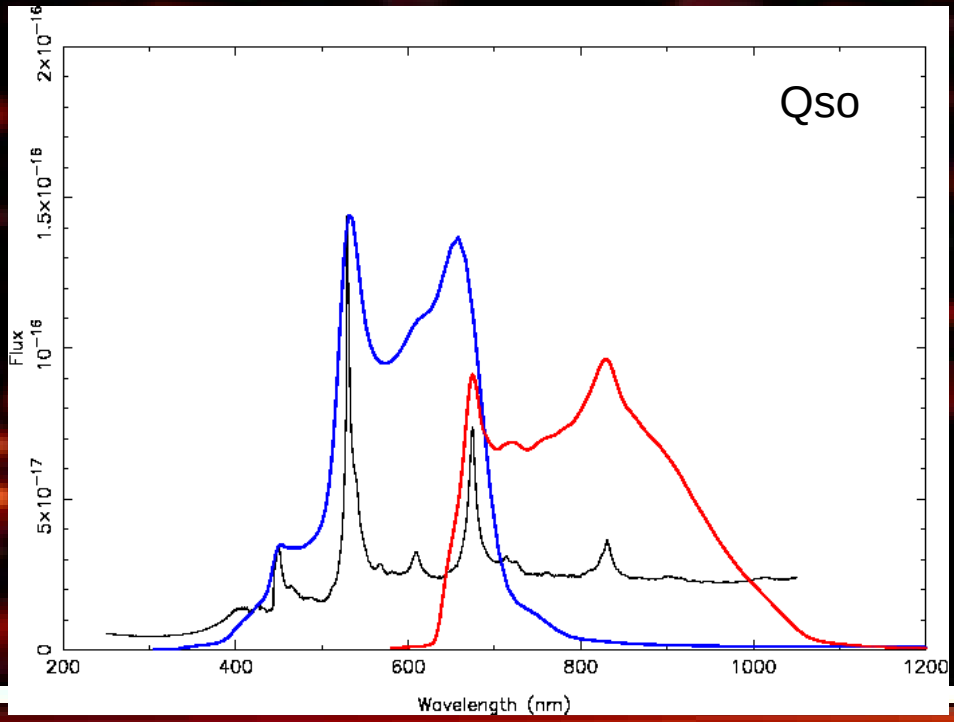

WR
Flux
Wavelength (nm)


Qso
Flux
Wavelength (nm)

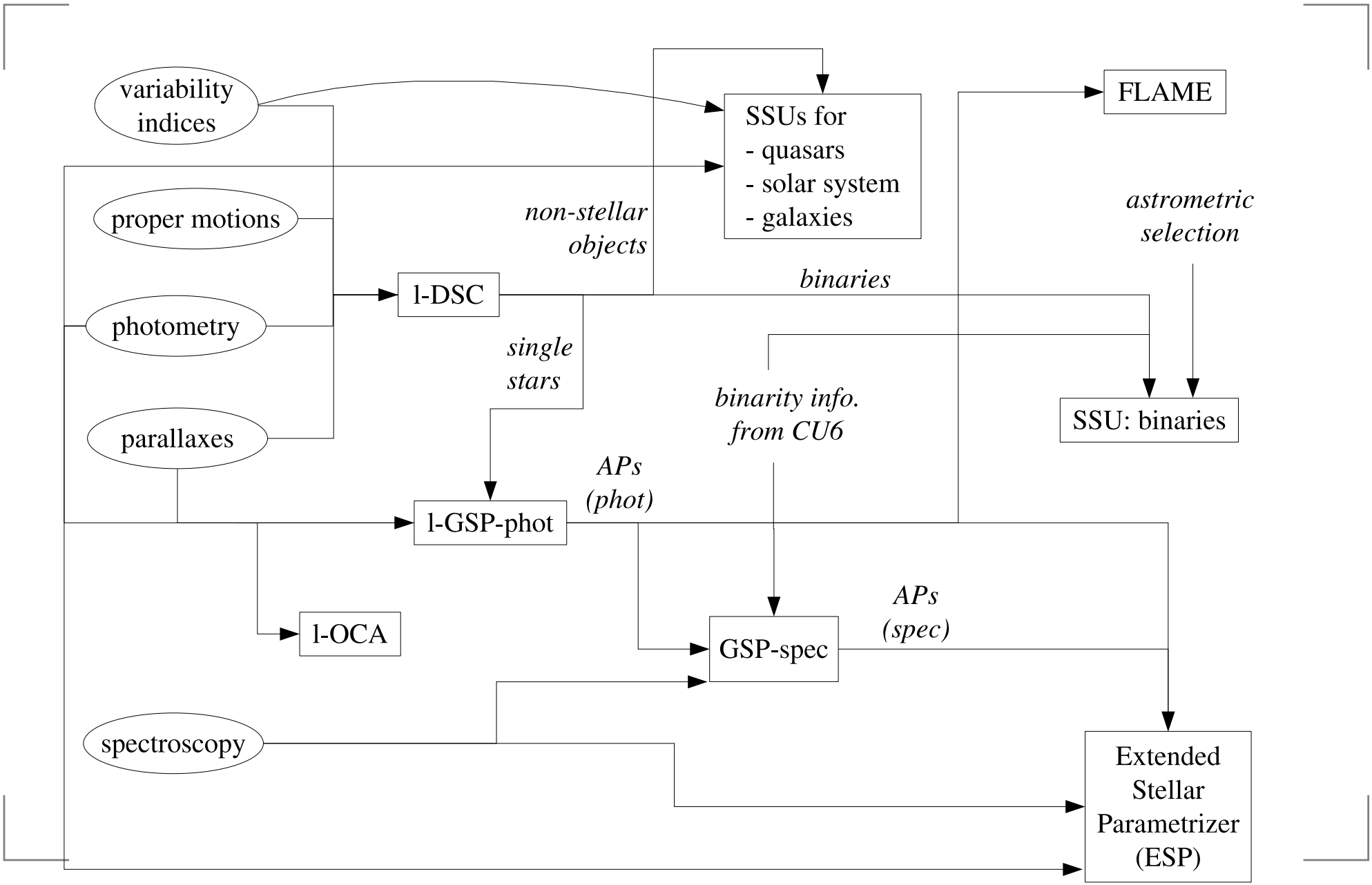

Cool star
Flux
Wavelength (nm)


O star
Flux
Wavelength (nm)

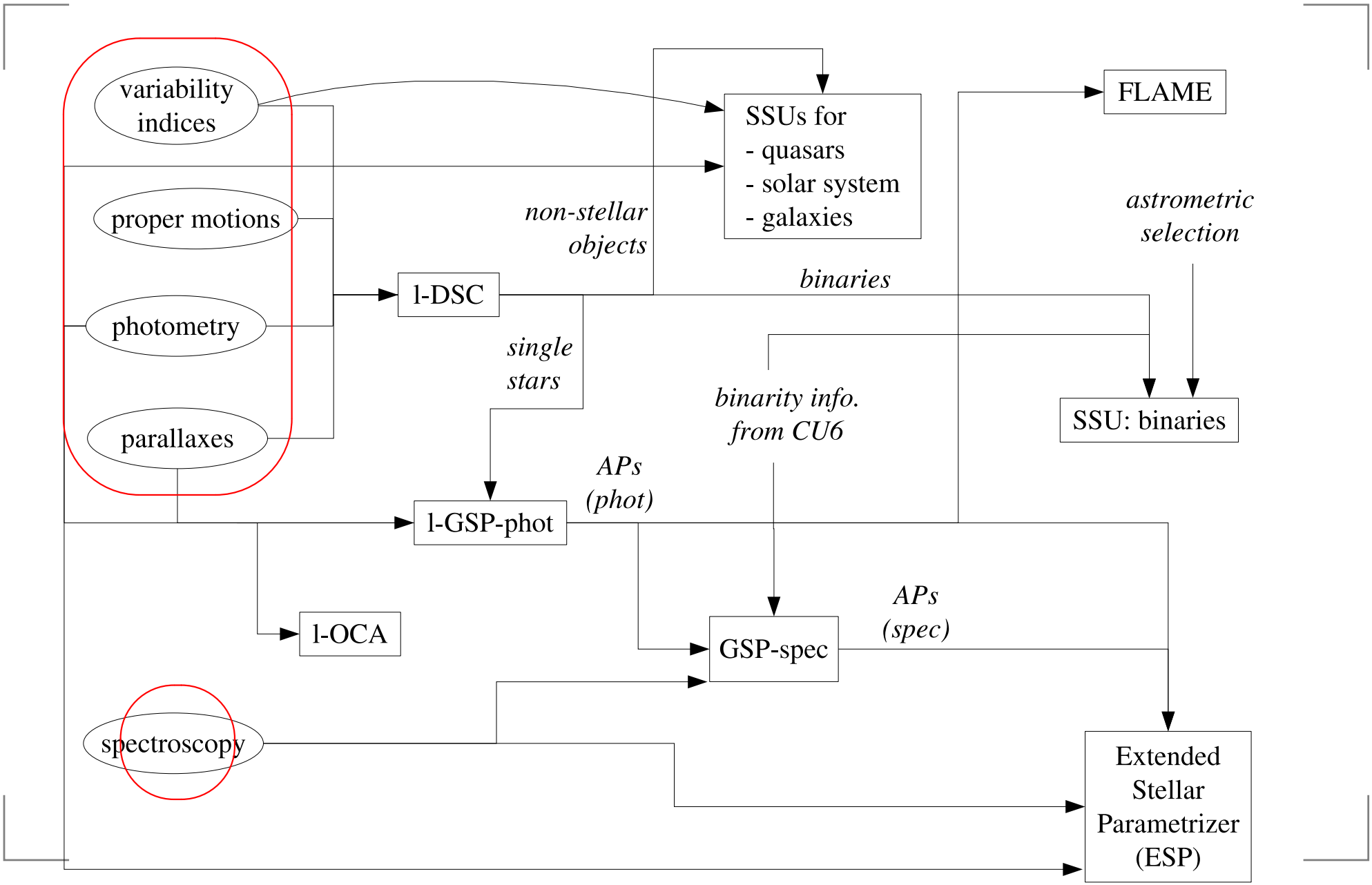

WR
Flux
Wavelength (nm)


Qso
Flux
Wavelength (nm)

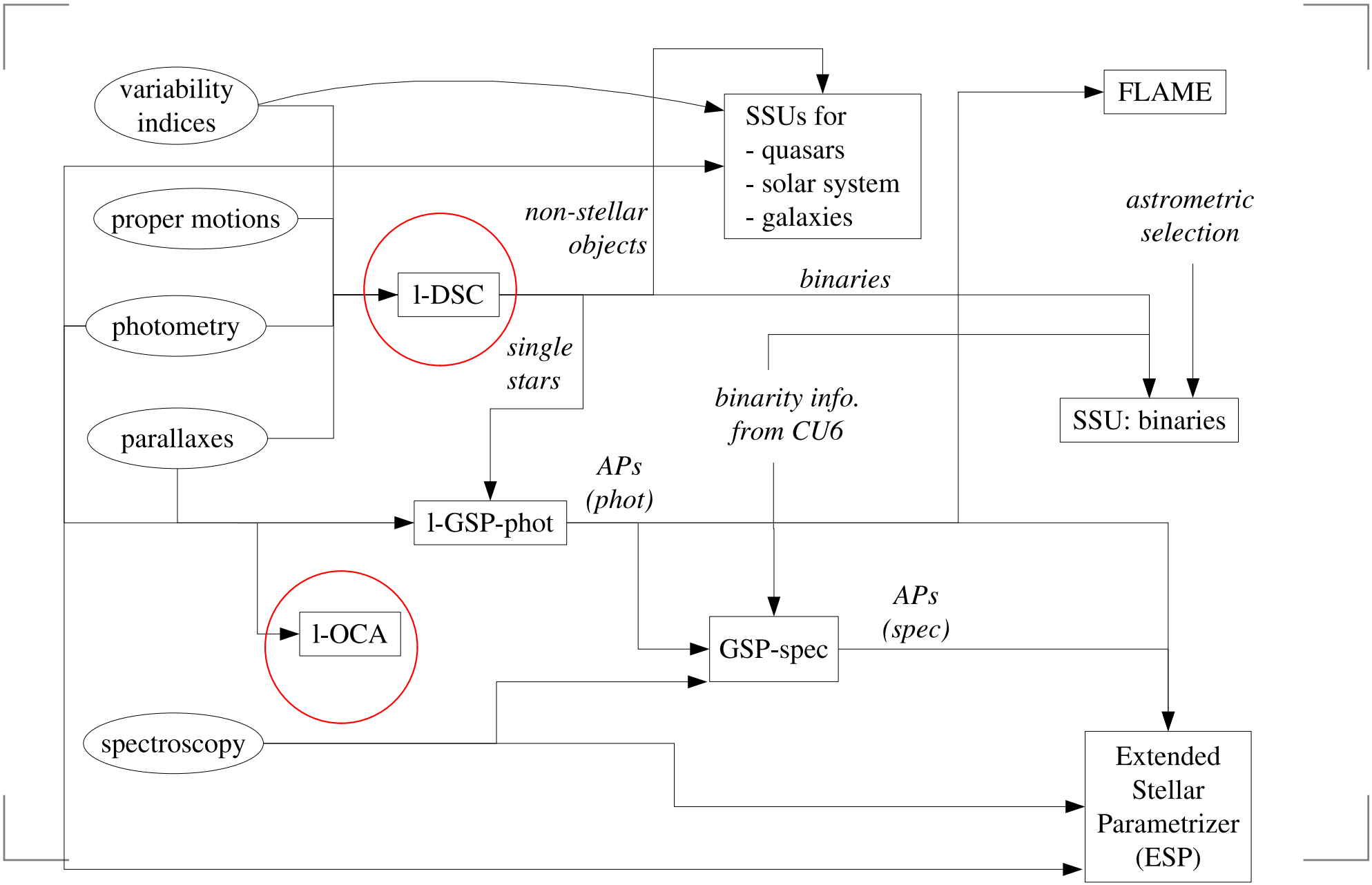
Astrophysical parameters



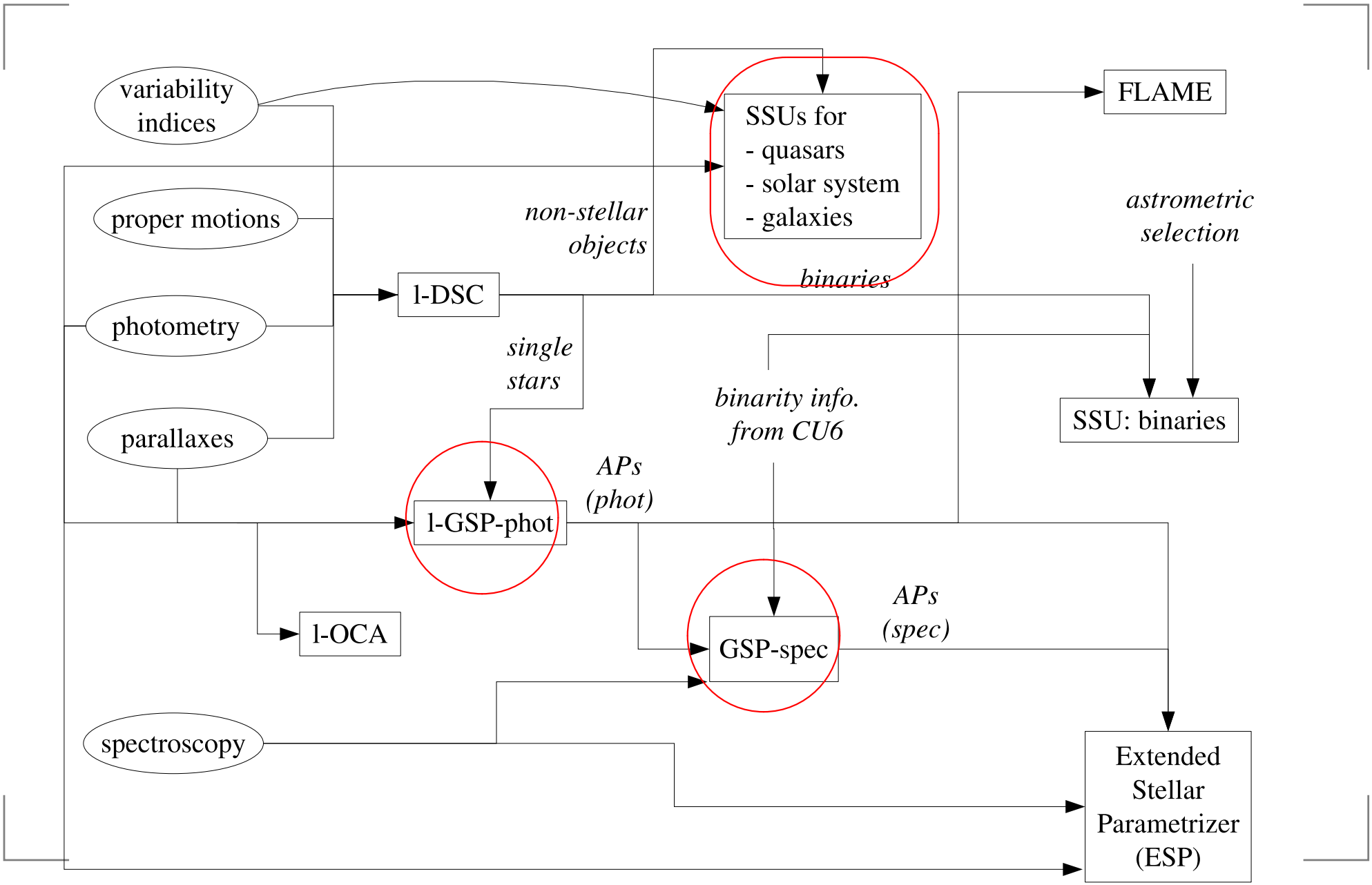
Astrophysical parameters



Astrophysical parameters



Astrophysical parameters



Machine learning

Machine learning: Definition is a little problematic.

Can be thought of as making *the computer perform a task for which the instructions have been derived from learning rules + examples, and are not provided directly by the programmer*.

Machine learning

Machine learning: Definition is a little problematic.

Can be thought of as making *the computer perform a task for which the instructions have been derived from learning rules + examples, and are not provided directly by the programmer*.

- **Data space:** The data you actually observe
- **Feature space:** Derived from your data

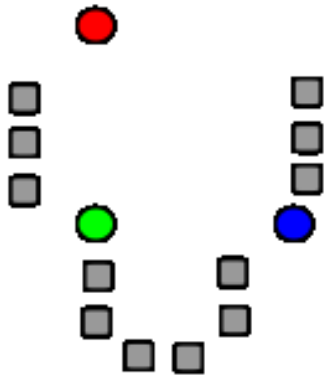
Can divide ML into

- Supervised learning
- Unsupervised learning
- Semi-supervised learning

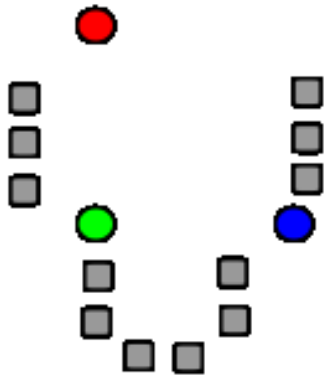
Can divide ML into

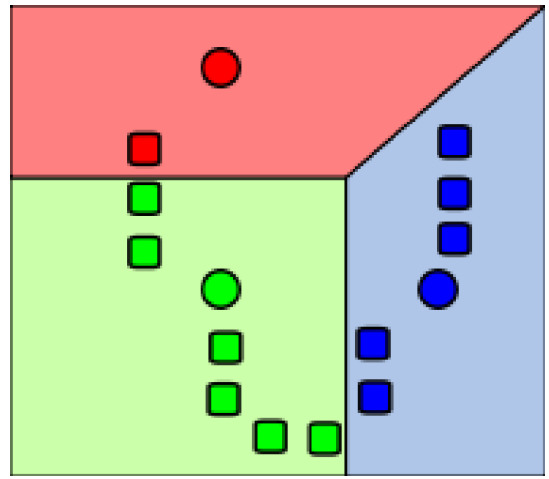
- Supervised learning - *DSC*
- Unsupervised learning - *OCA*
- Semi-supervised learning

Unsupervised: k-means

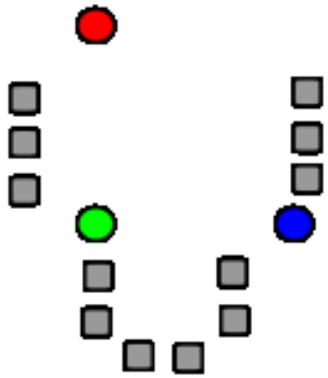


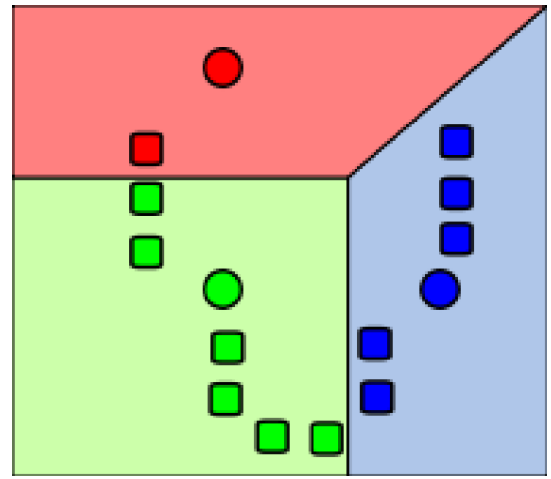
Unsupervised: k-means

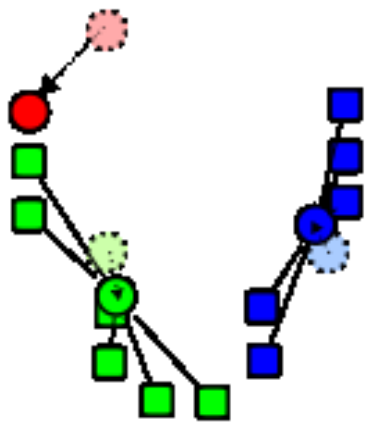




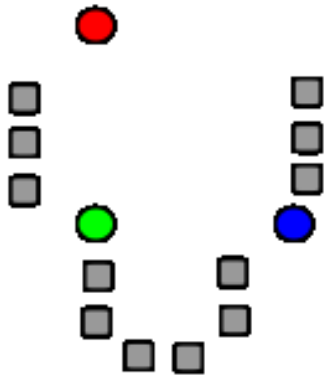
Unsupervised: k-means

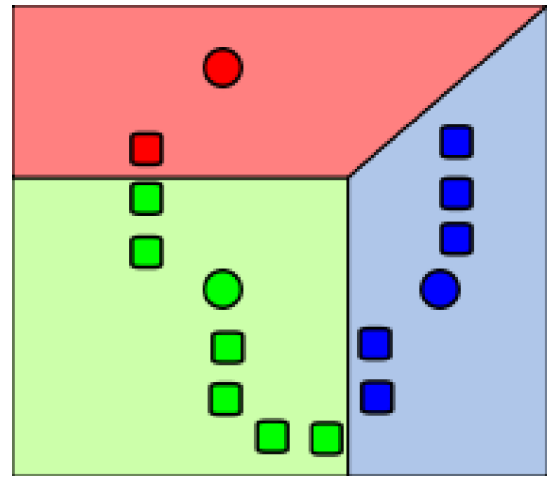


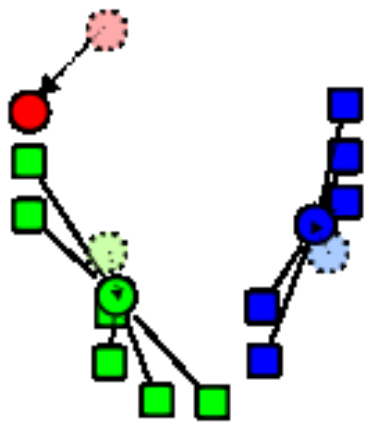


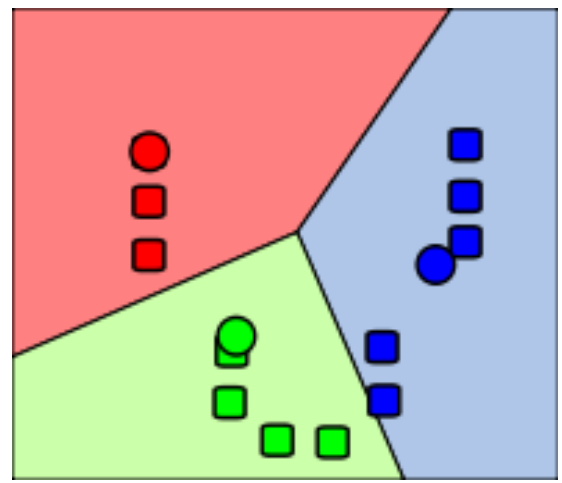


Unsupervised: k-means









Can also divide problems into

- Classification - discrete classes, e.g. Star, Galaxy etc.
- Regression - continuous variable, e.g. T_{eff}, redshift etc.

For classifiers, it's nice if output is *probabilistic*

Warning: overfitting and regularization

Some non-linear algorithms such as SVM and Neural networks, can fit anything perfectly.

Need to *regularize* the model and control by e.g. *cross validation*

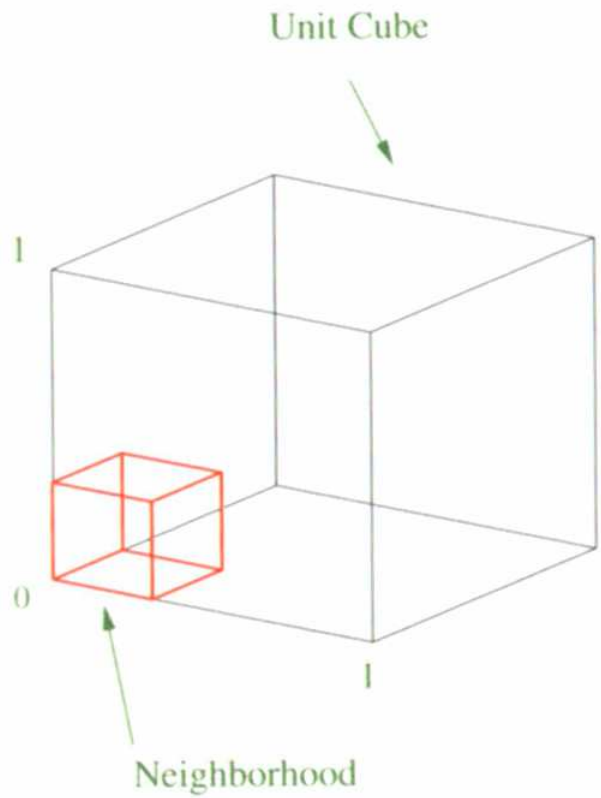
Gaia photometry problem

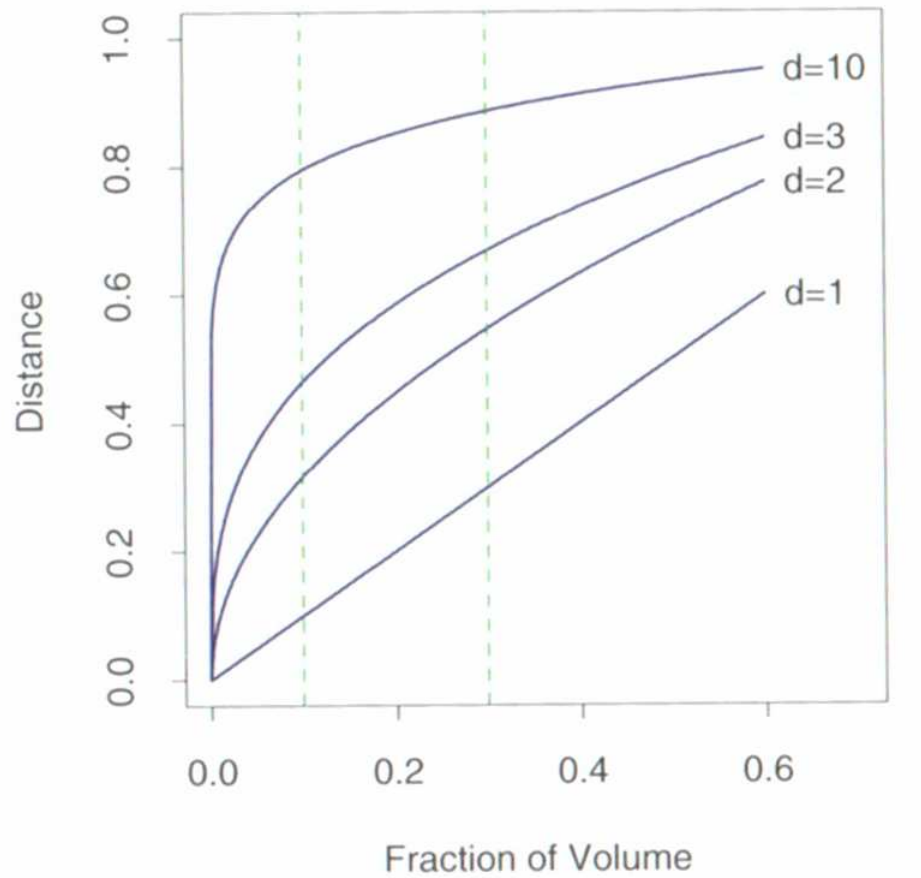
Low resolution spectrum - 2 $\times$ 180 'pixels'

(although intrinsic resolution $\sim$15)

Suffer from the *curse of dimensionality*

Curse


Unit Cube
1
0
1
Neighborhood


Distance
0.0
0.2
0.4
0.6
0.8
1.0
d=10
d=3
d=2
d=1
0.0
0.2
0.4
0.6
Fraction of Volume

Classifier specification

Need either;

- classifier that works well with a high number of data dimensions,

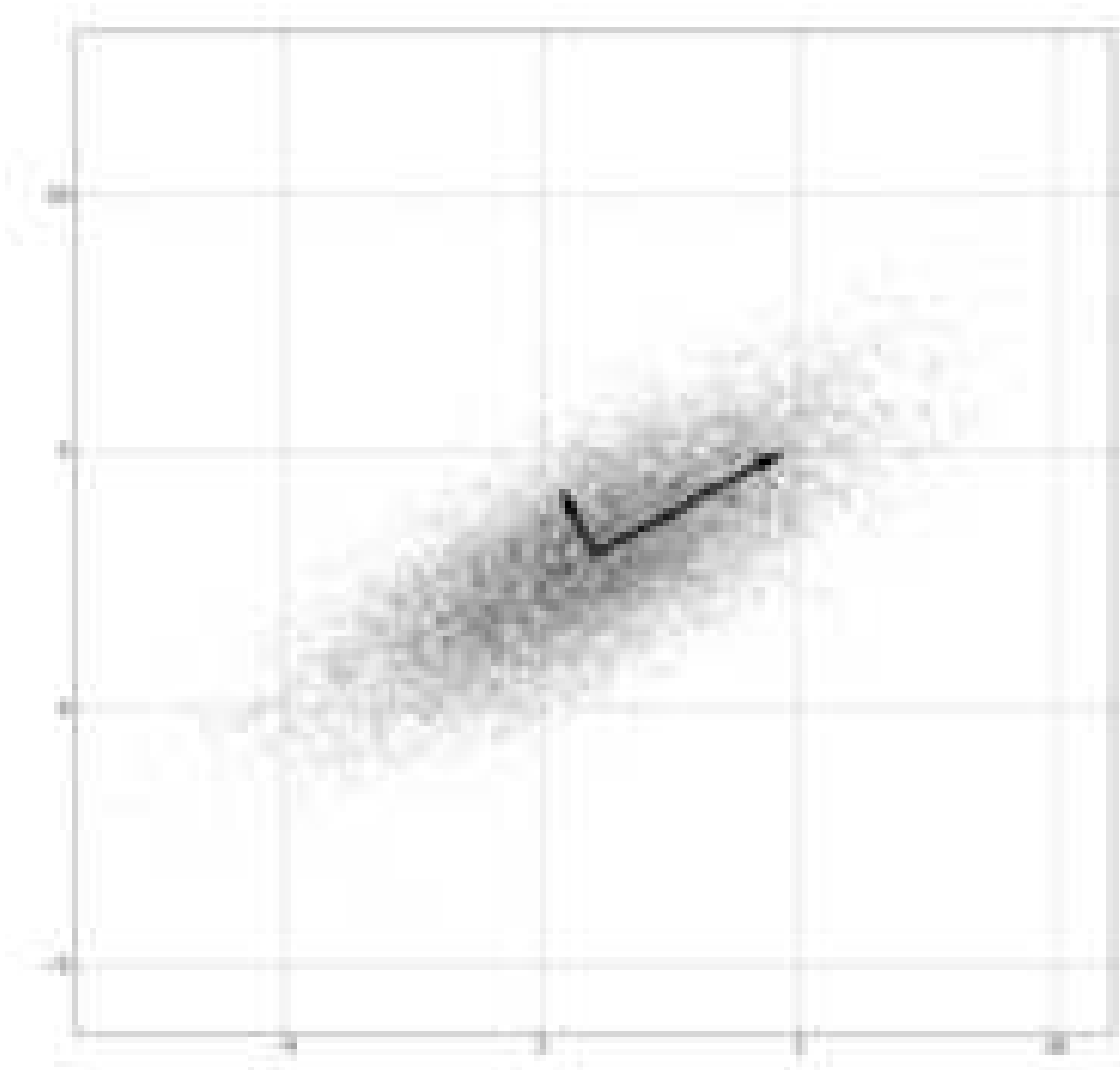
or

- reduce dimensions (PCA or similar).

also need to combine at least spectra and astrometry, and possibly other information.

$\rightarrow$ probabilistic subclassifiers

PCA



Classifier specification

fundamental choice;

- supervised
- unsupervised

We use both approaches.

	supervised	unsupervised
All sources	Support vector machine	k-means based method
Outliers		self-organizing map

Self organizing maps

Map nodes organized in a grid in map space

Each node has a random starting vector W

Self organizing maps

Map nodes organized in a grid in map space

Each node has a random starting vector W

For first data point, choose the 'nearest' node

Self organizing maps

Map nodes organized in a grid in map space

Each node has a random starting vector W

For first data point, choose the 'nearest' node

↓

Assign point to node

Self organizing maps

Map nodes organized in a grid in map space

Each node has a random starting vector W

For first data point, choose the 'nearest' node

↓

Assign point to node

↓

Adjust surrounding nodes to be 'closer'

Self organizing maps

Map nodes organized in a grid in map space

Each node has a random starting vector W

For first data point, choose the 'nearest' node

↓

Assign point to node

↓

Adjust surrounding nodes to be 'closer'

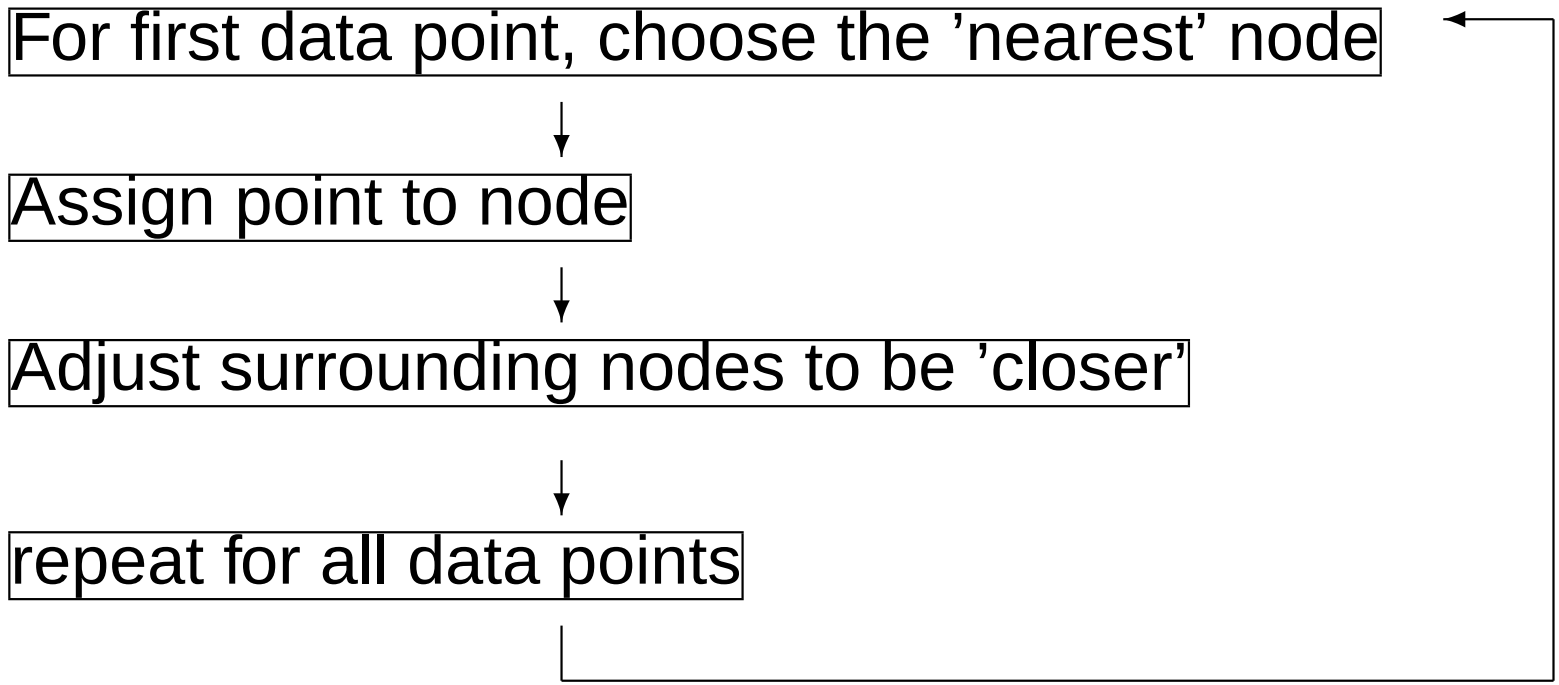
↓

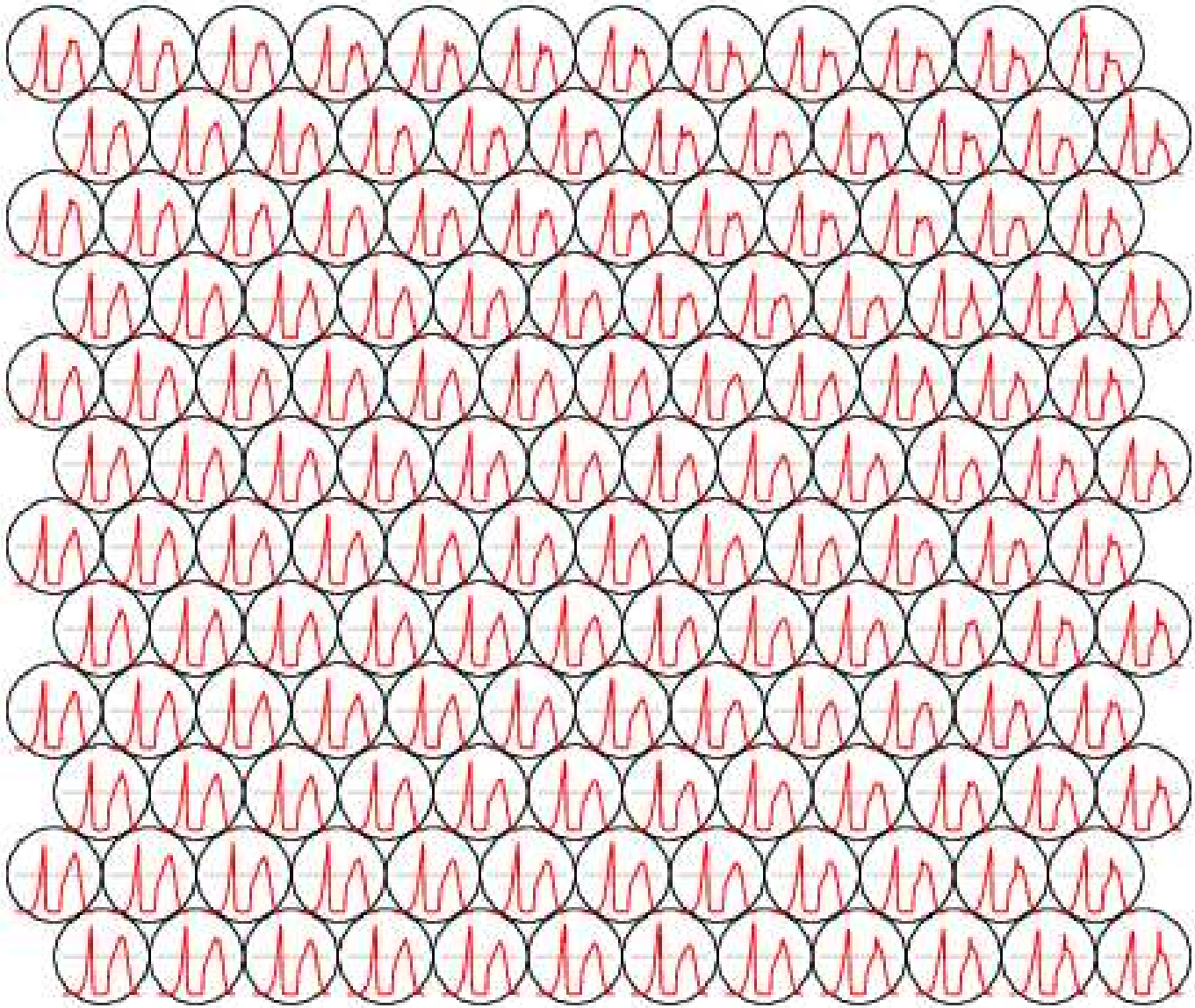
repeat for all data points

Self organizing maps

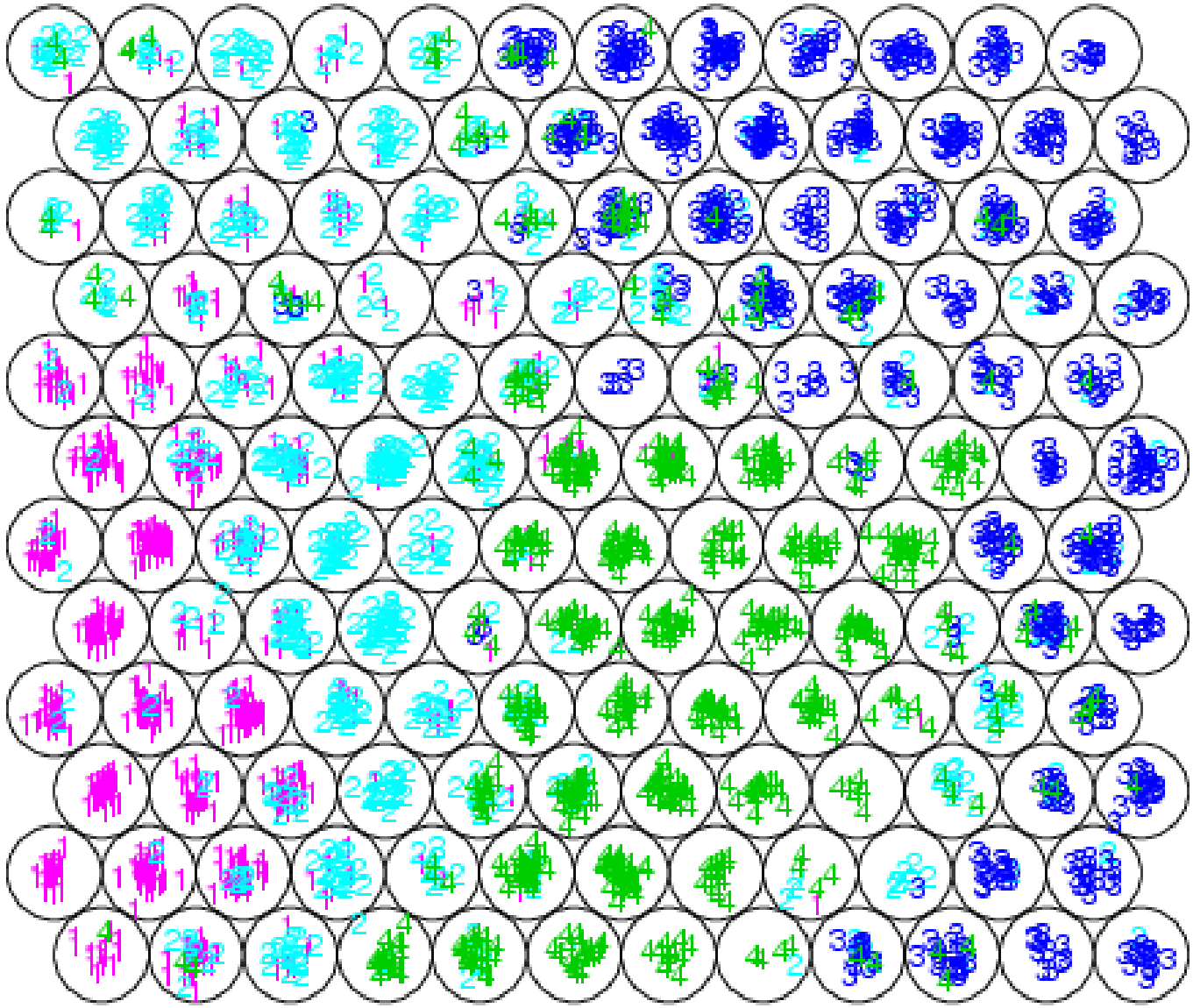
Map nodes organized in a grid in map space

Each node has a random starting vector W





Mapping plot

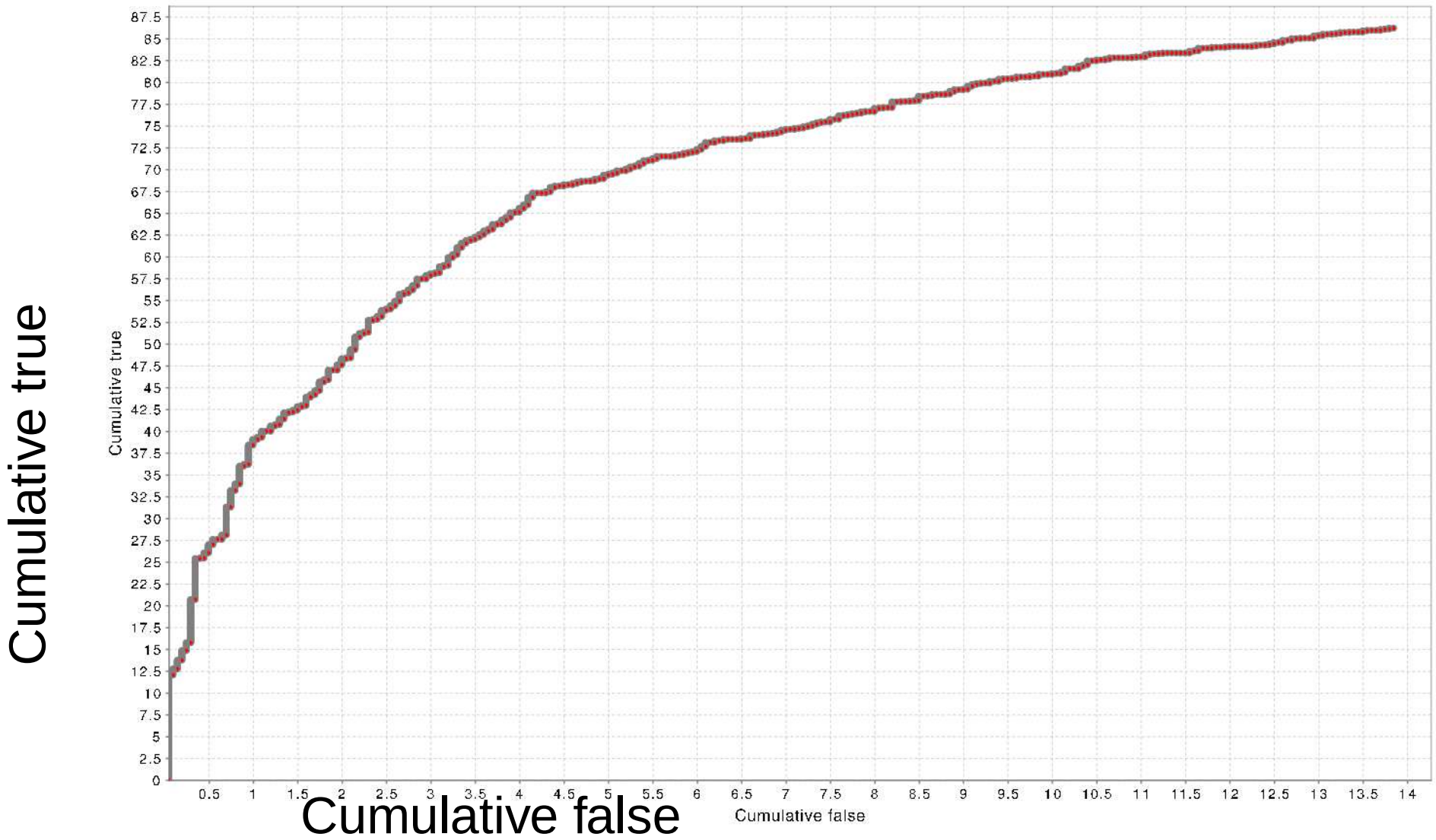


Classification quality

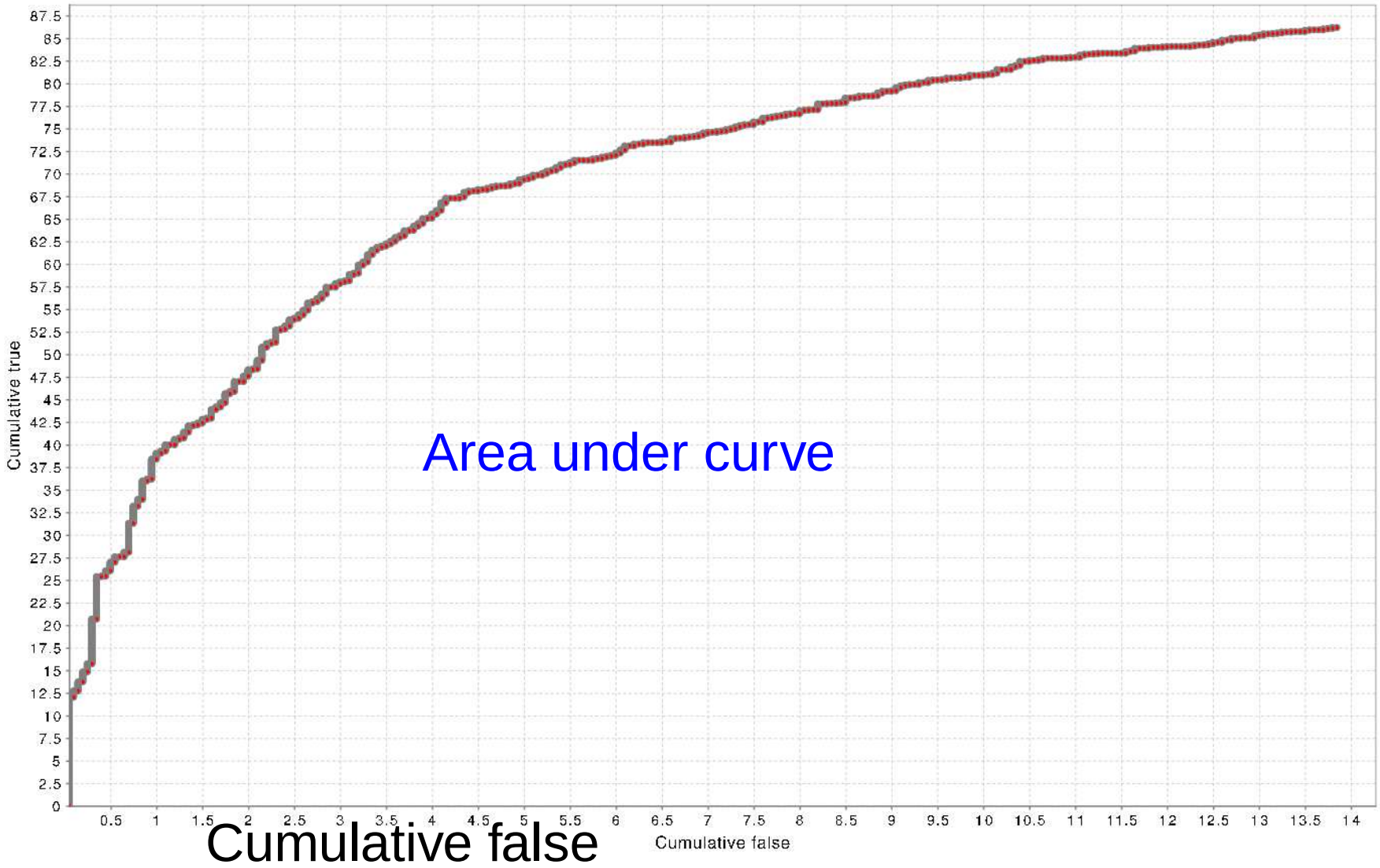
$$\text{Completeness}_j = \frac{n_{i=j,j}}{N_i}$$

$$\text{Contamination}_j = \frac{\sum_{i \neq j} n_{i,j}}{\sum_i n_{i,j}}$$

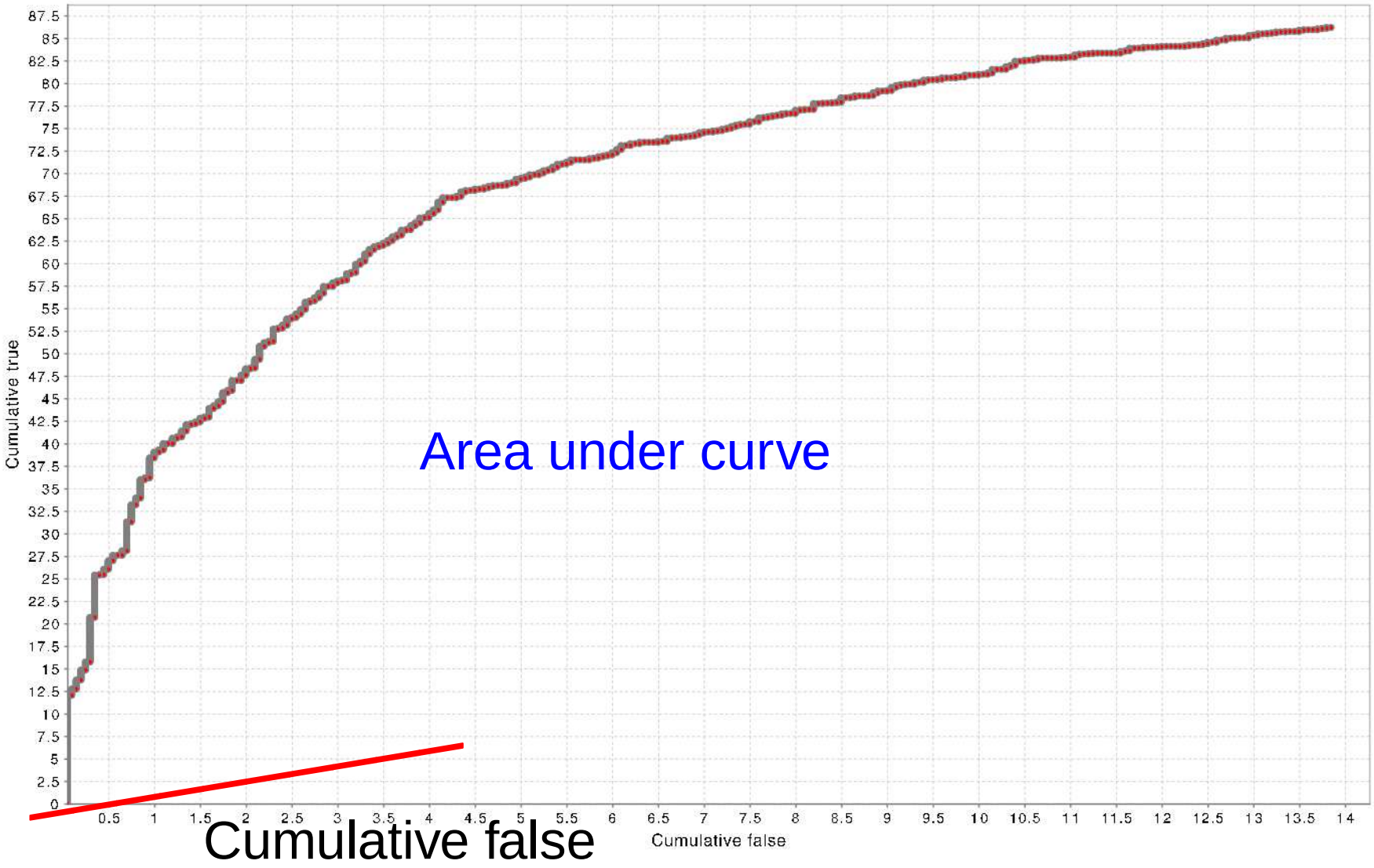
Classification performance



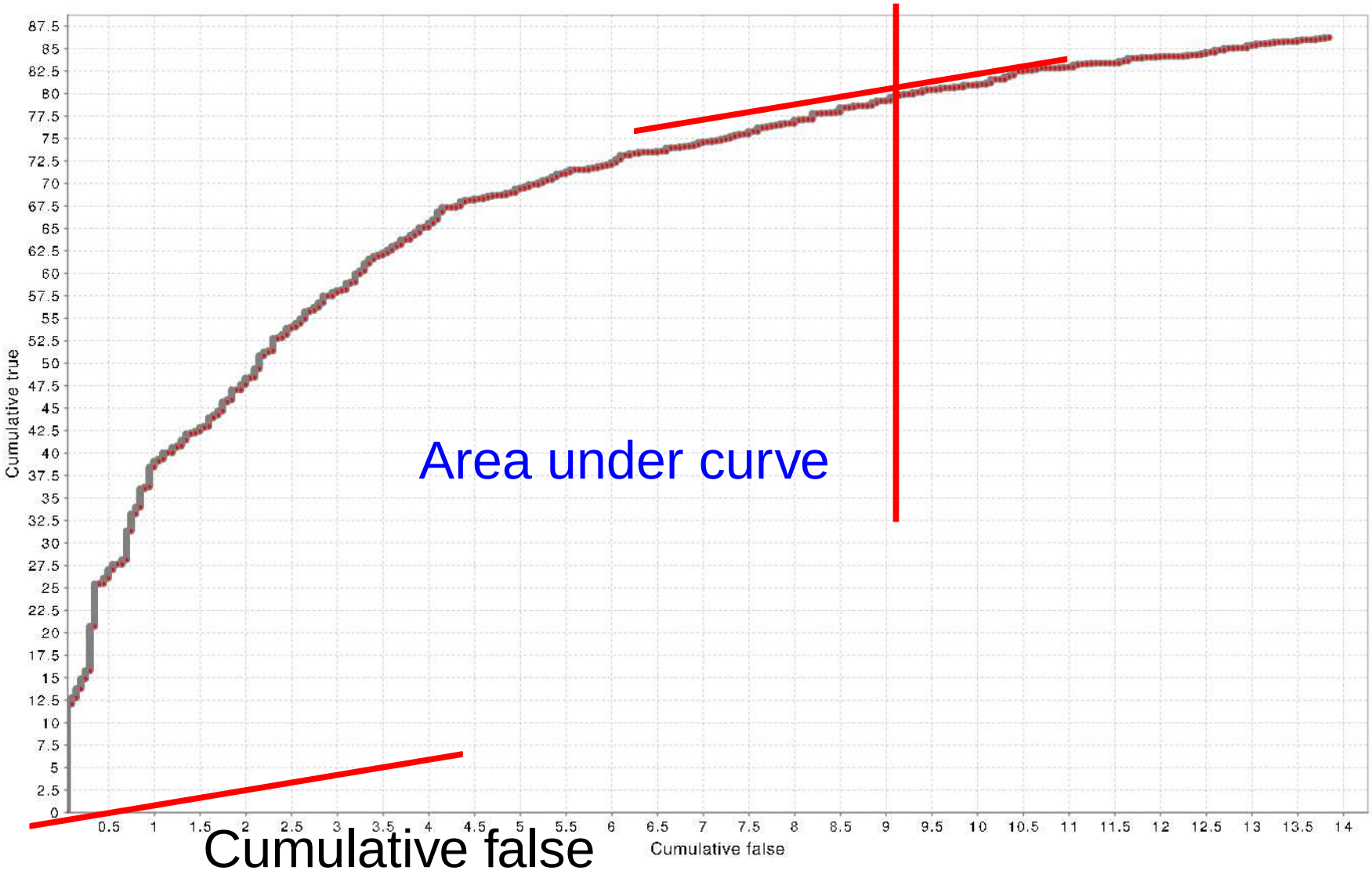
Classification performance



Classification performance



Classification performance

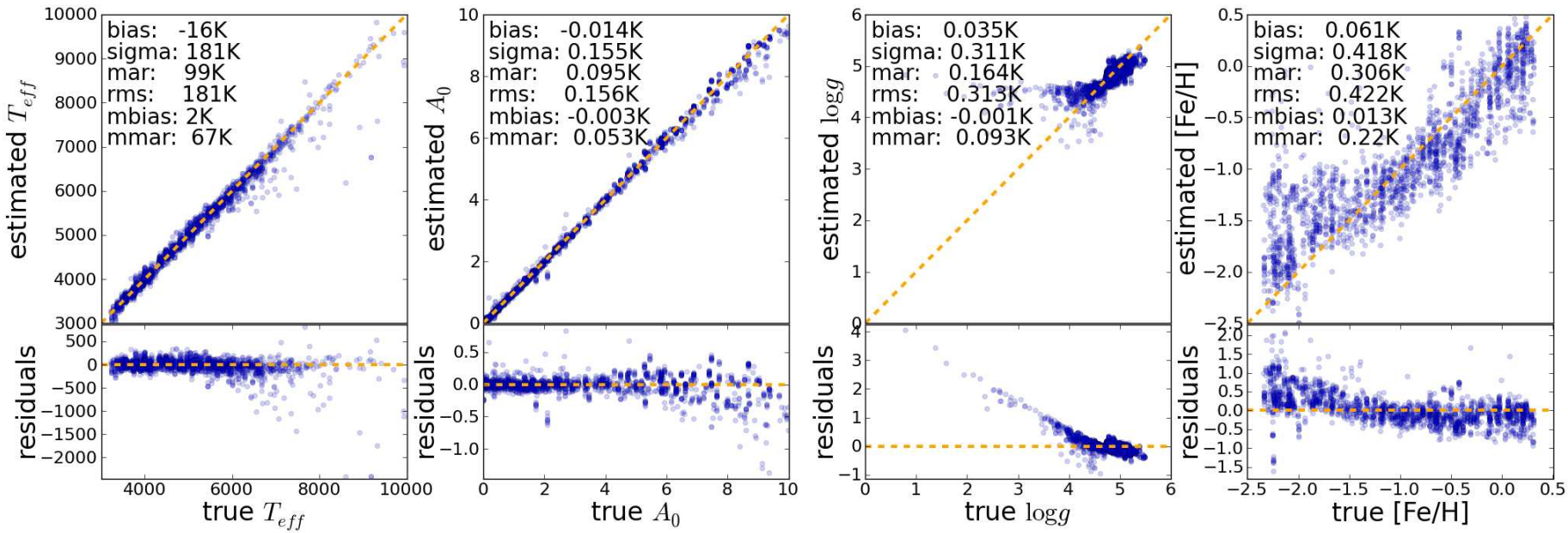


Parameterization

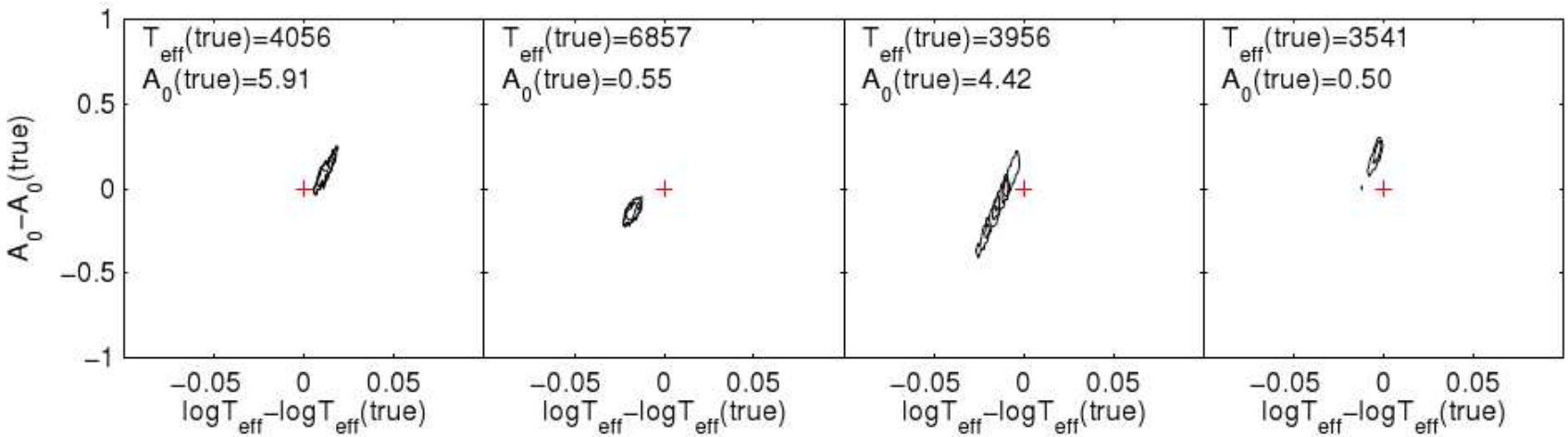
Uses a variety of methods.

- Support Vector Machines
- Neural Network
- Grid interpolation
- Matisse algorithm
- ...

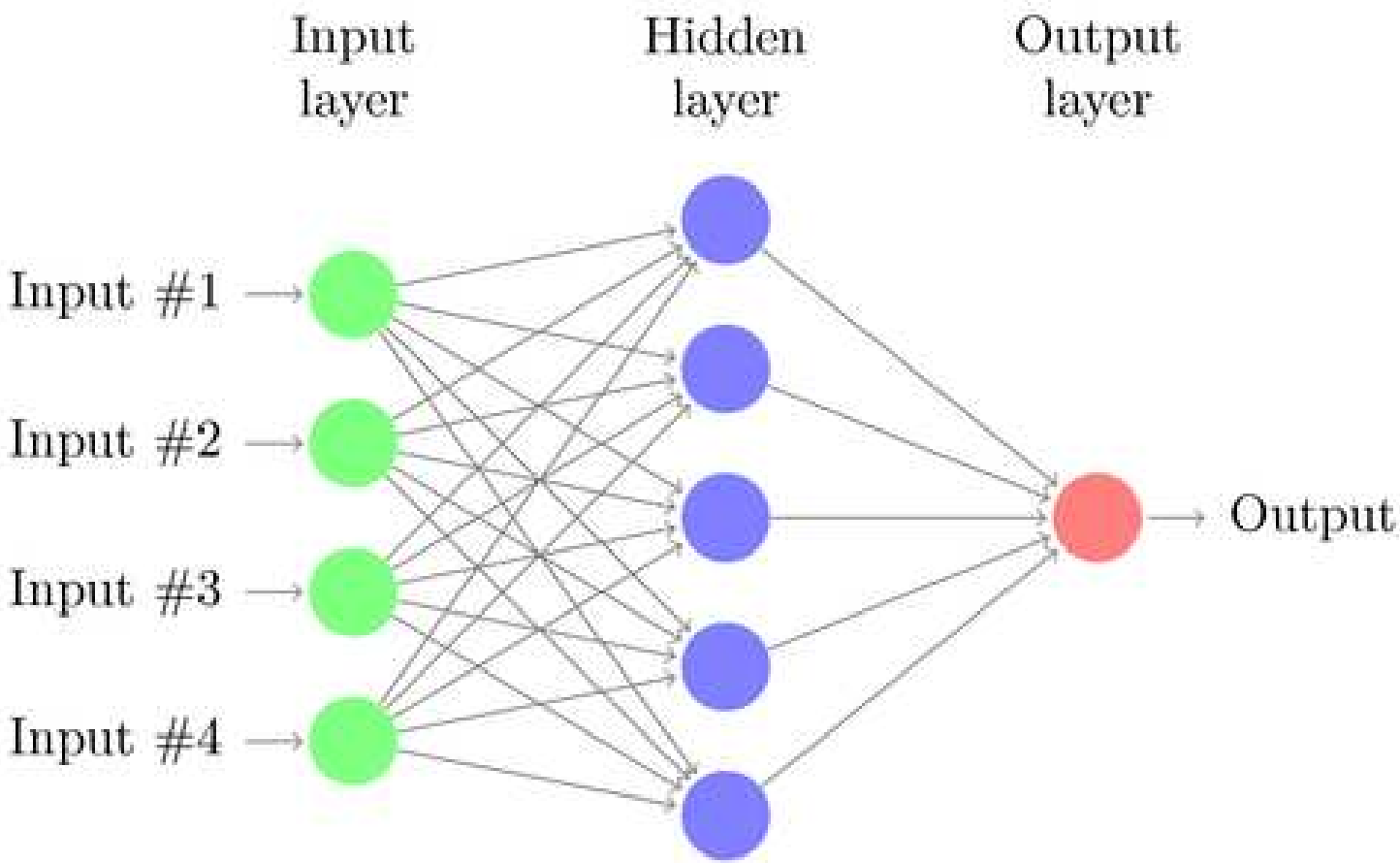
Regression - GSP-Phot



Regression - GSP-Phot



Neural network



Discrete Source Classifier (DSC)

- Task of DSC is classification of all 10^9 Gaia sources.
- Supervised approach
- Available data;
  - BPRP 'photometry'
  - Astrometry: π, proper motions
  - Variability
  - RVS spectrum
  - Possible morphology
  - Position, magnitude

Discrete Source Classifier (DSC)

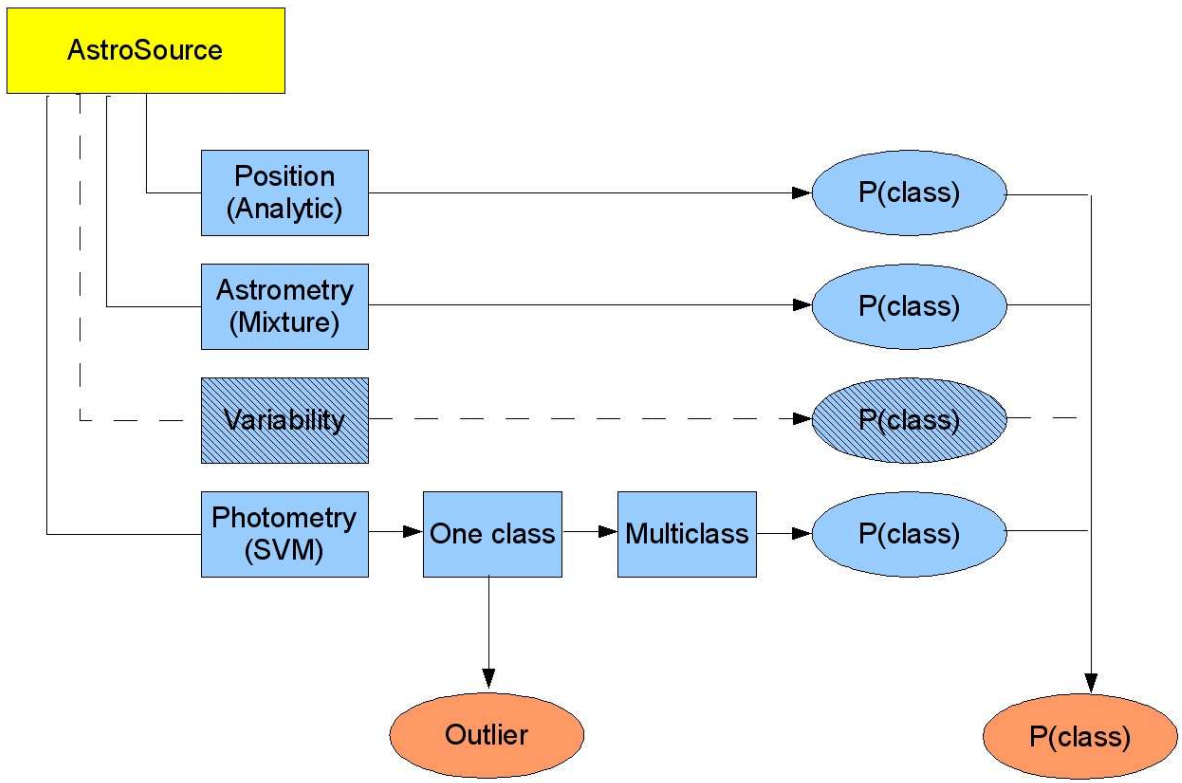
- Task of DSC is classification of all 10^9 Gaia sources.
- Supervised approach
- Available data;
  - BPRP photometry
  - Astrometry: π, proper motions
  - Variability
  - RVS spectrum
  - Possible morphology
  - Position, magnitude

Internal structure of DSC

- Modular design with subclassifiers

Internal structure of DSC

- Modular design with subclassifiers



Internal structure of DSC

- Modular design with subclassifiers
- 'Hierarchical' photometric classifier based on **Support Vector Machines (SVM)**

Internal structure of DSC

- Modular design with subclassifiers
- 'Hierarchical' photometric classifier based on **Support Vector Machines (SVM)**
- Astrometric subclassifier based on **Gaussian mixture model**

Internal structure of DSC

- Modular design with subclassifiers
- 'Hierarchical' photometric classifier based on **Support Vector Machines (SVM)**
- Astrometric subclassifier based on **Gaussian mixture model**
- Position G Mag subclassifier based on **Kernel Density Estimator**

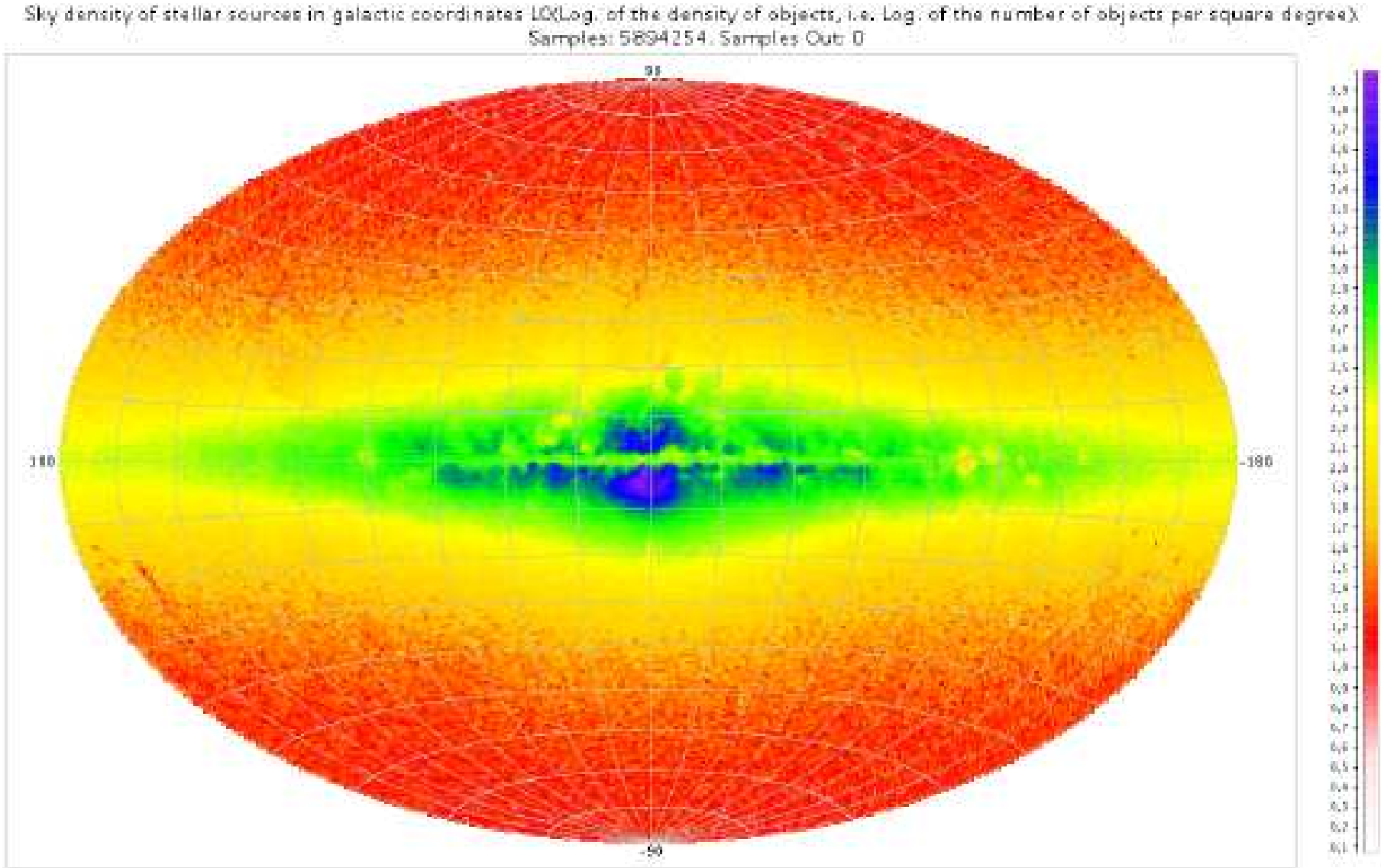
Internal structure of DSC

- Modular design with subclassifiers
- 'Hierarchical' photometric classifier based on **Support Vector Machines (SVM)**
- Astrometric subclassifier based on **Gaussian mixture model**
- Position G Mag subclassifier based on **Kernel Density Estimator**

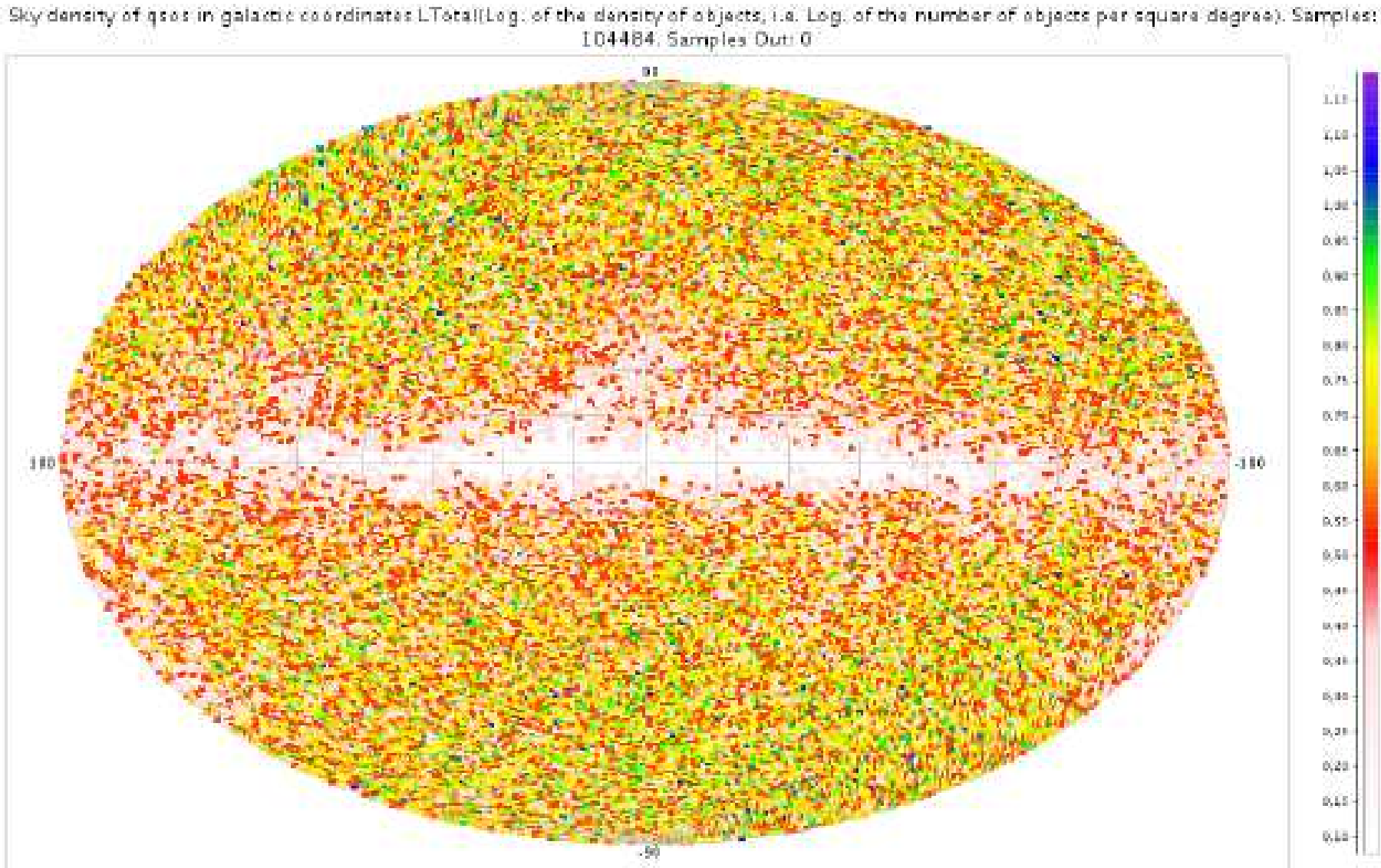
$$P(C|D_1 ... D_N) = \frac{\Pi_{n=1}^{n=N} P(C|D_n)}{P(C)^{N-1}}$$

[Bailer Jones & Smith, GAIA-C8-TN-MPIA-CBJ-053]

Position, G magnitude



Position, G magnitude



KDE

Estimate the underlying probability density fn. from which the data instances were presumably drawn.

$$f(x_0) = \frac{1}{Nh} \sum K_\lambda(\frac{x_i - x_0}{\lambda})$$

Estimate probability by

$Pr(\text{class} = j) = \frac{f_j}{\sum_{i=1}^{k} f_i}$ for k classes.

Astrometric Classification

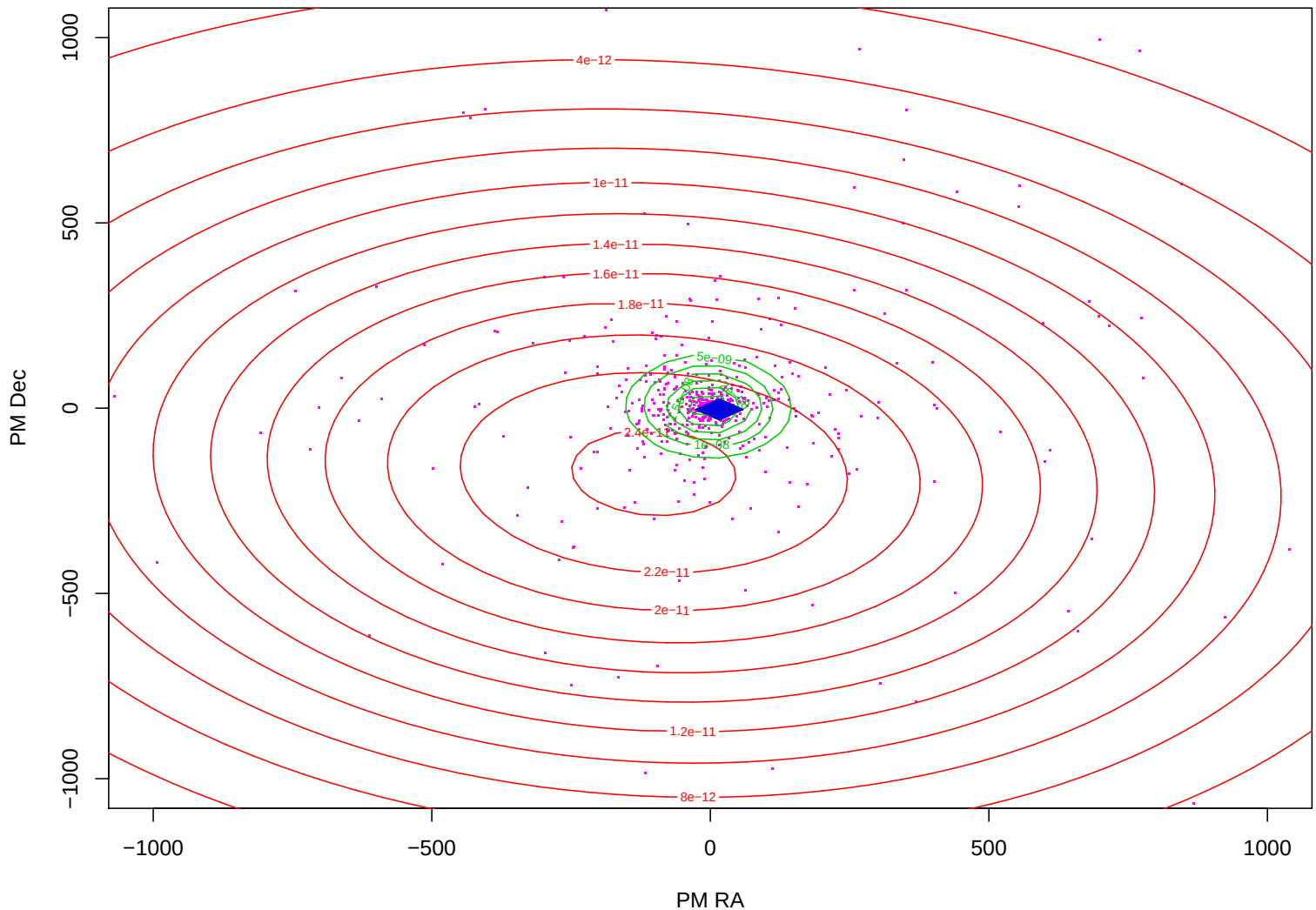
Based on mixture models

$$f(x_i) = \frac{A}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2} \times (x_i - \mu_i)^T \Sigma^{-1} (x_i - \mu_i))$$

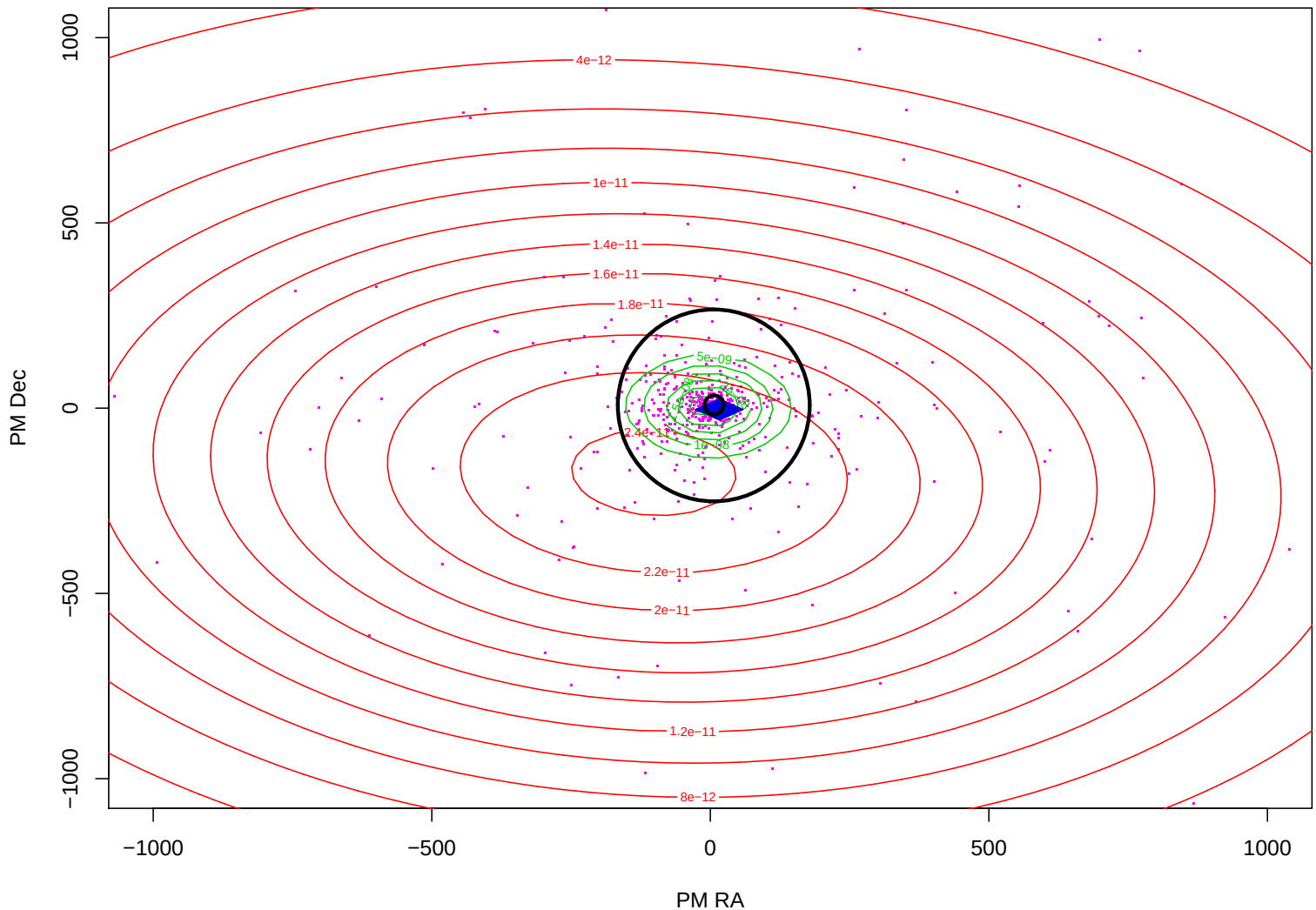
Effect of convolution;

$$\Sigma \rightarrow \Sigma + \Sigma_{noise}$$

Astrometric classifier



Astrometric classifier

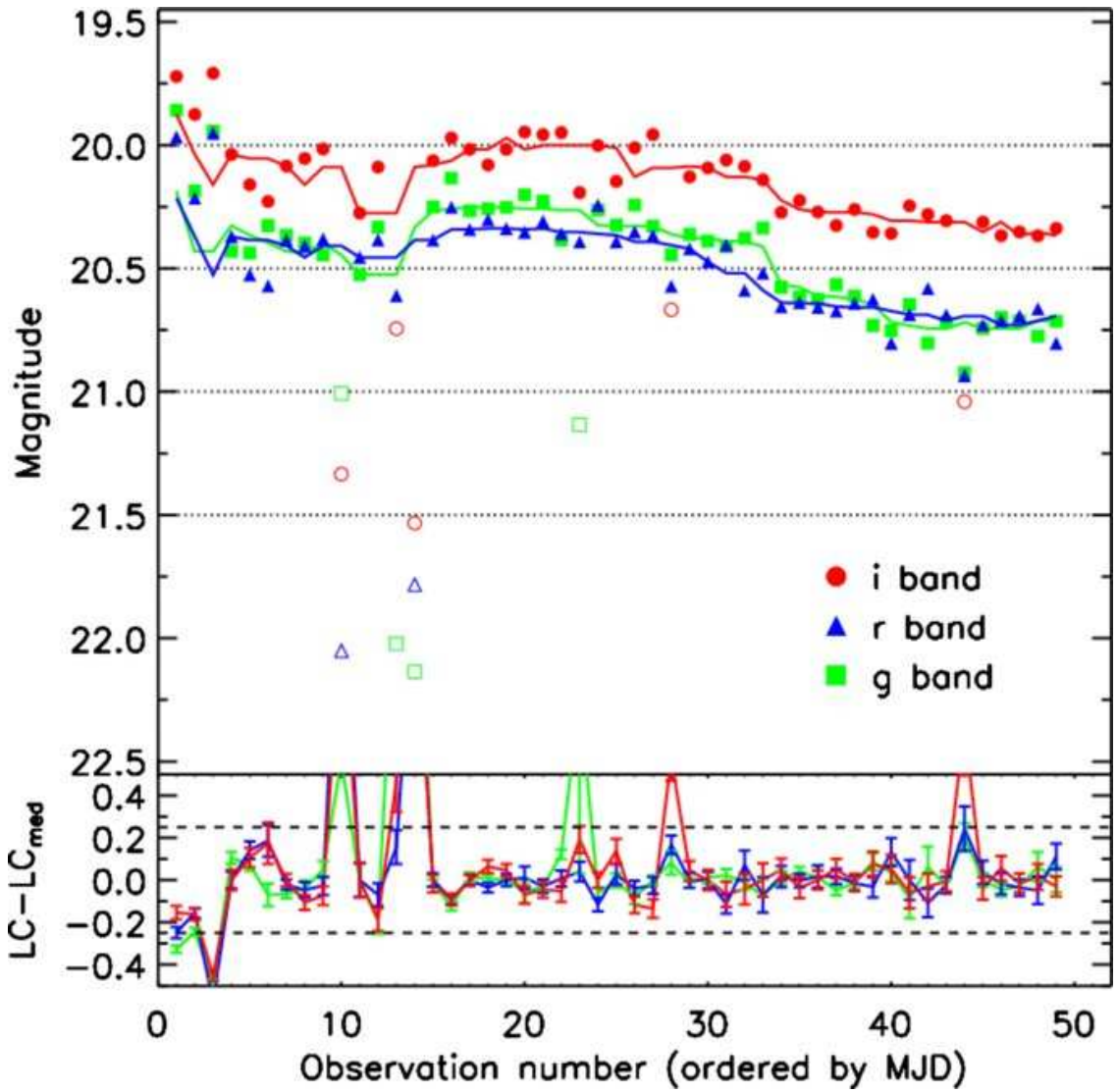


Model fitting

is done by Expectation maximization.

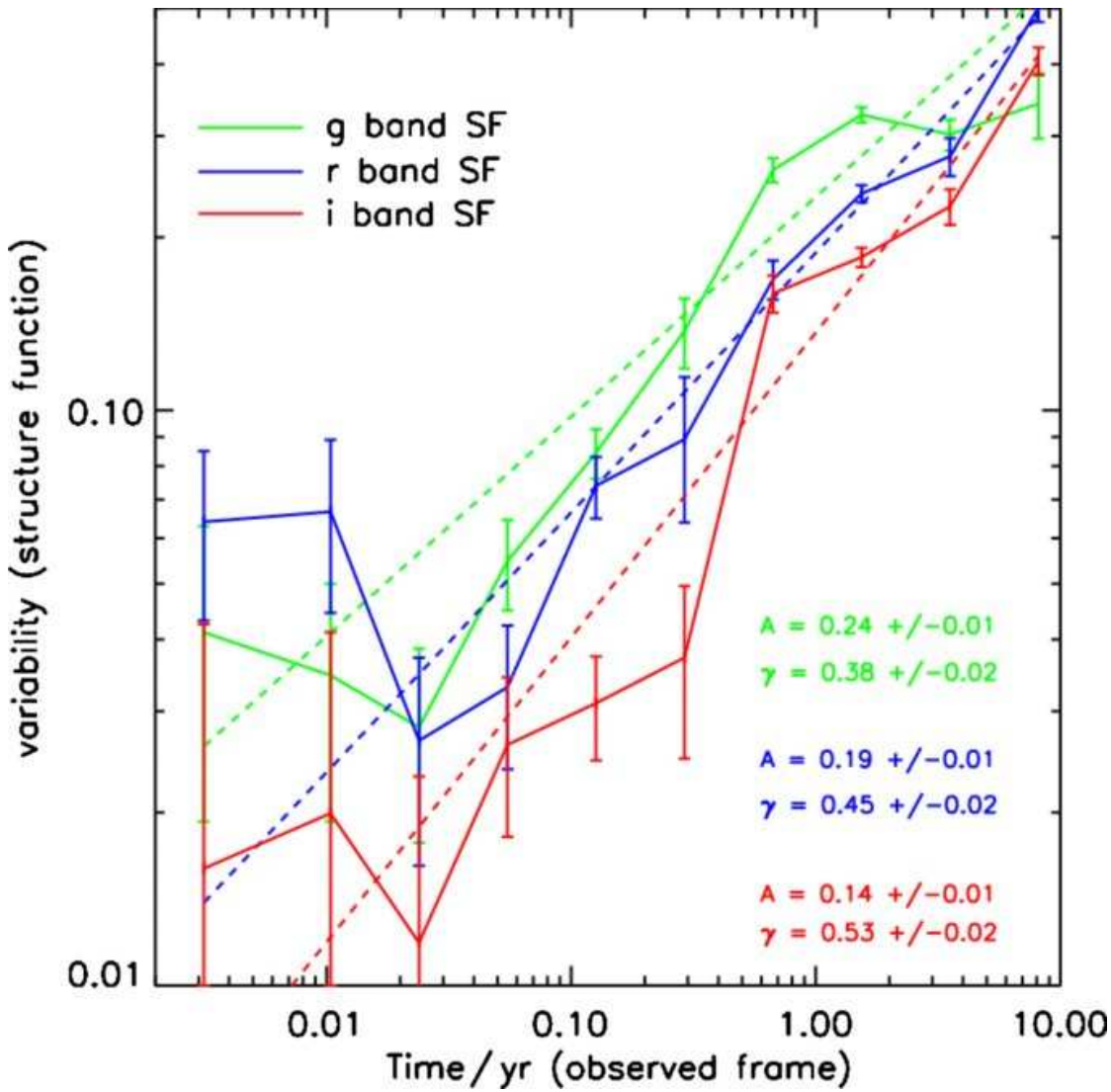
1. Determine the probabilities of the points for each existing model
2. Maximize the likelihood by adjusting model parameters
3. Repeat....

Variability classifier



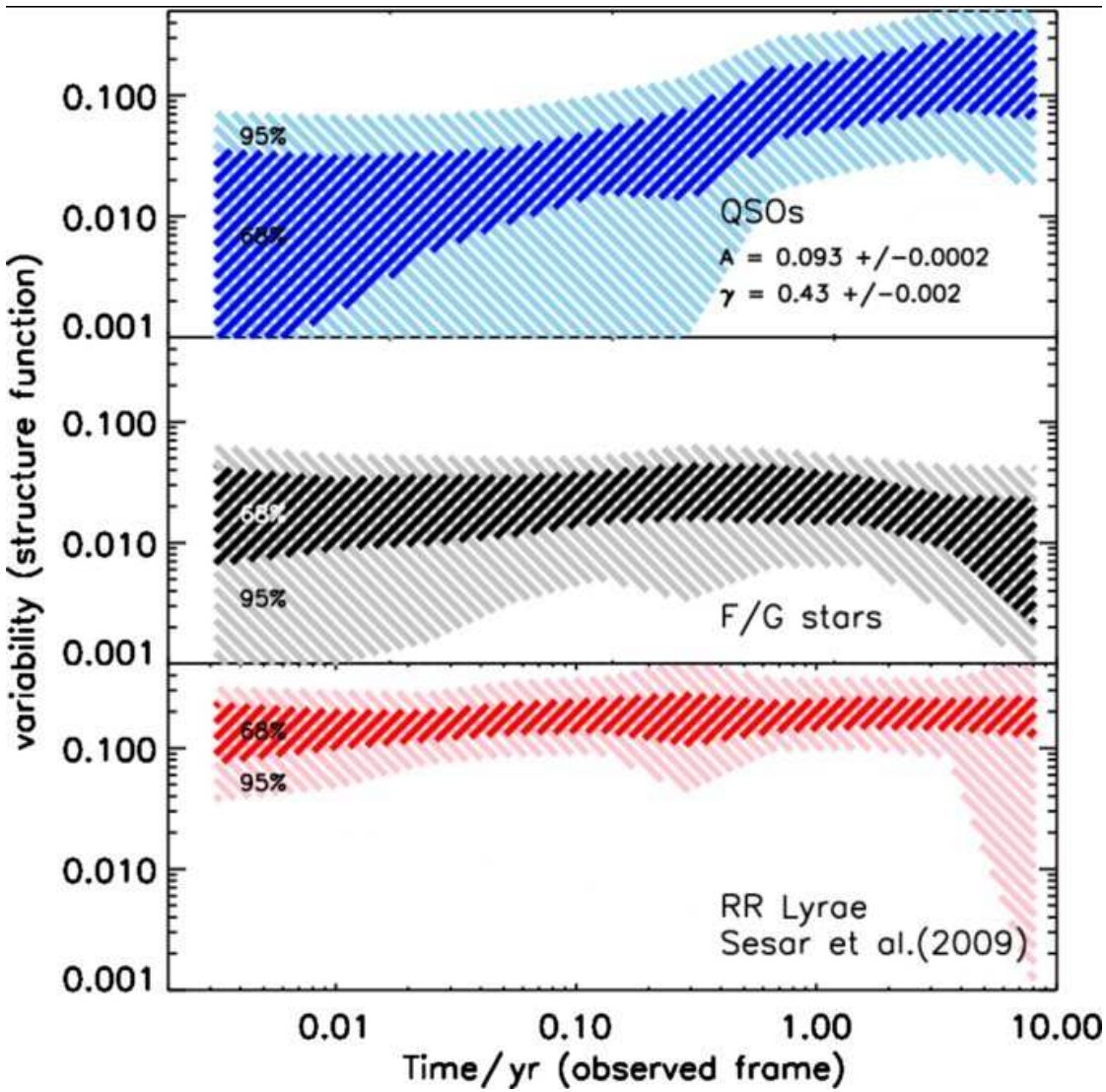
Schmidt et al. 2010
ApJ 714, 1194

Variability classifier



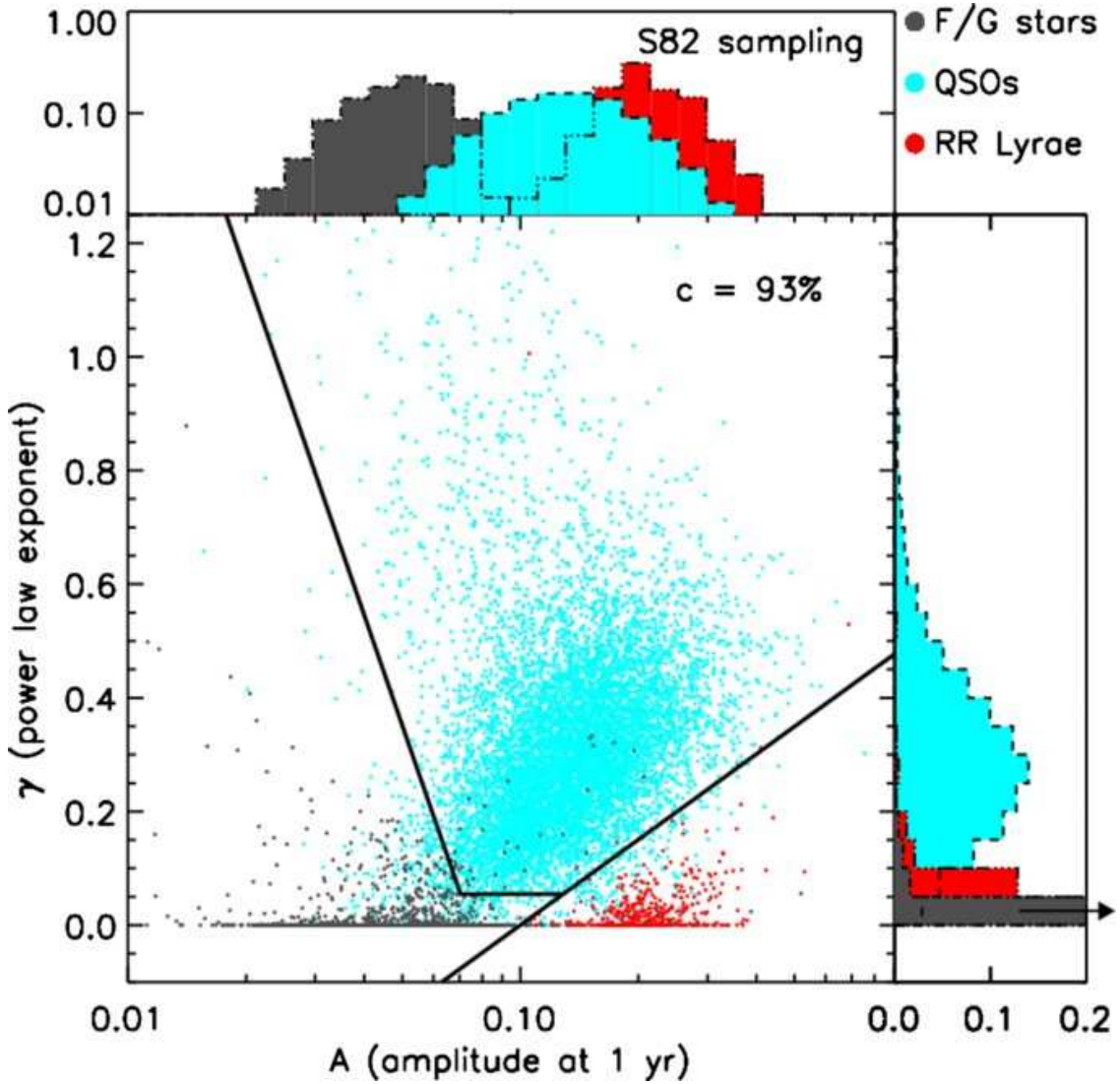
Schmidt et al. 2010
ApJ 714, 1194

Variability classifier



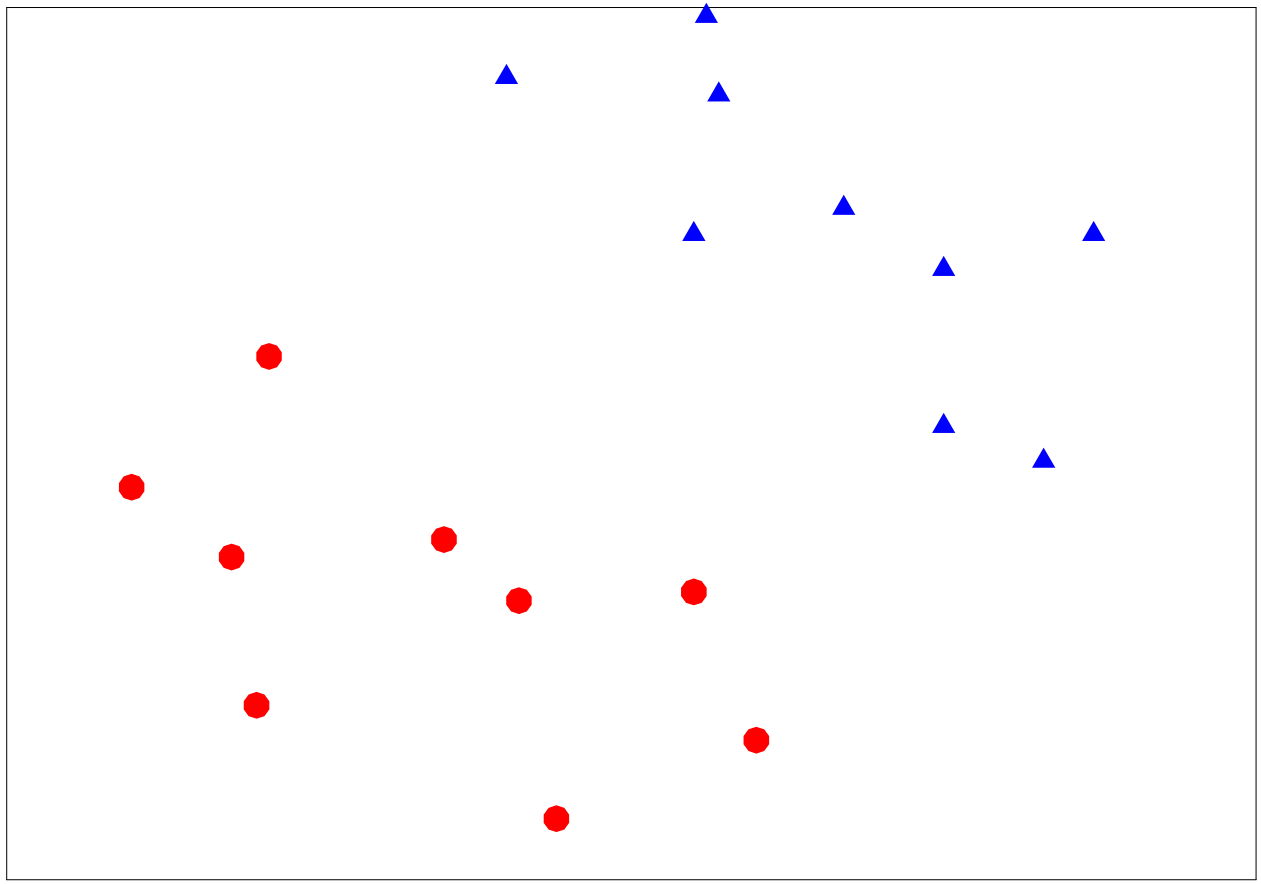
Schmidt et al. 2010
ApJ 714, 1194

Variability classifier

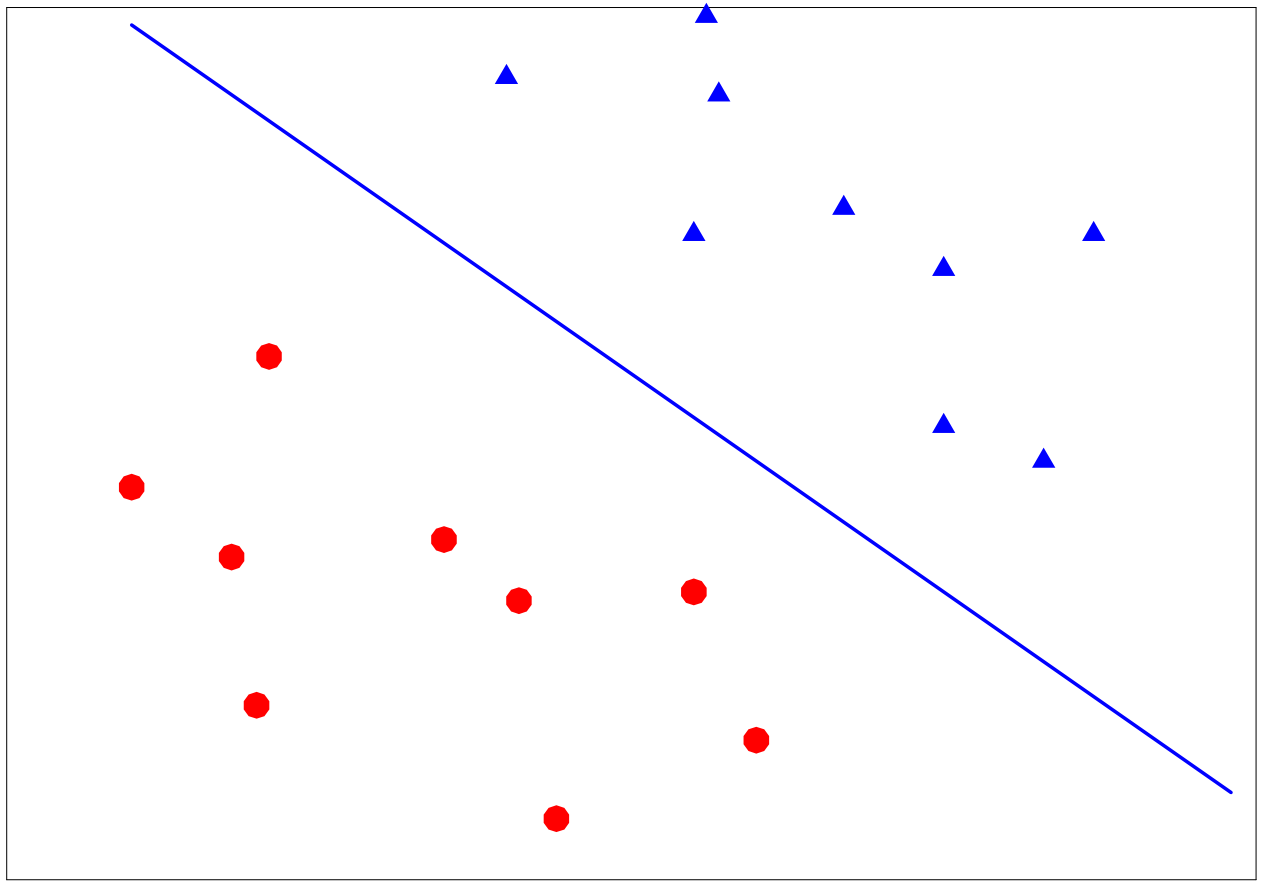


Schmidt et al. 2010
ApJ 714, 1194

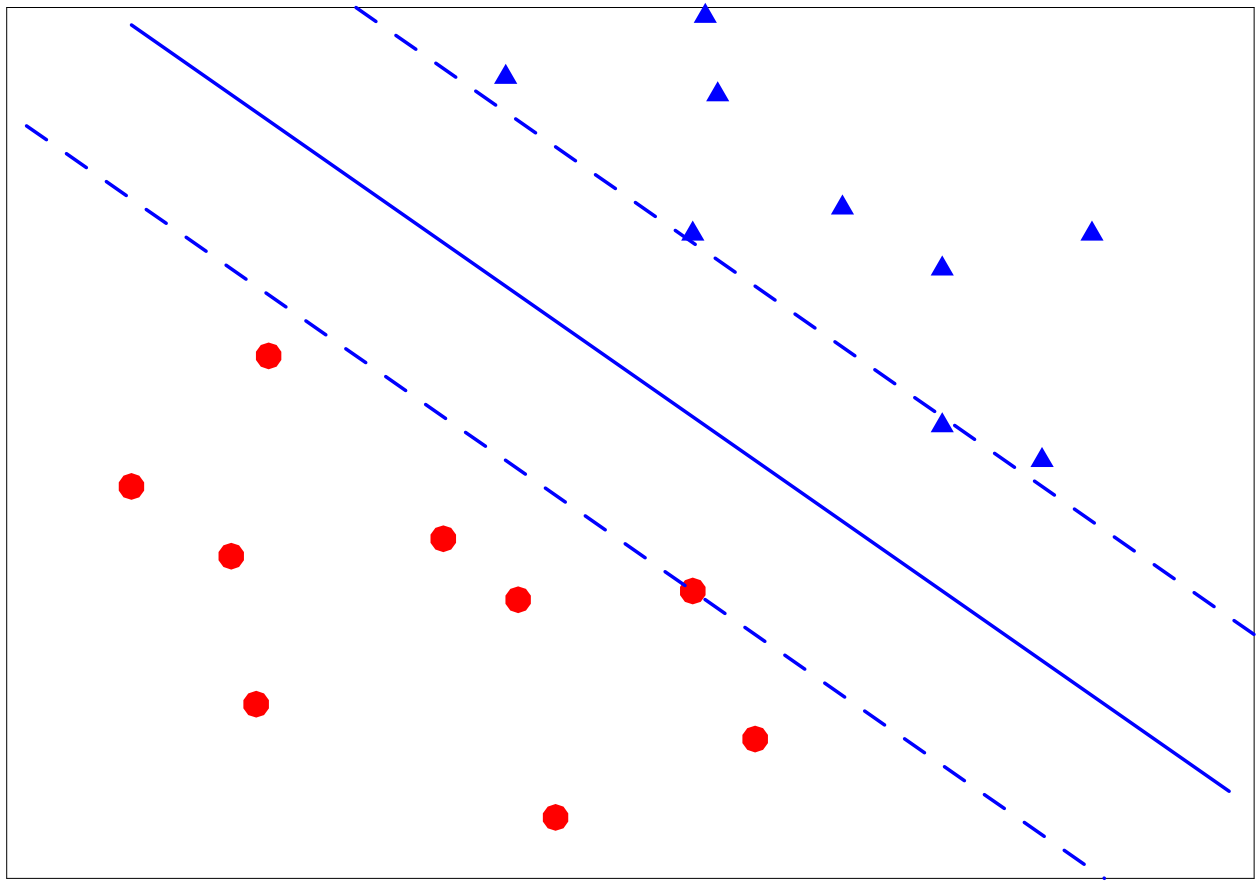
Photometric classifier: Svm



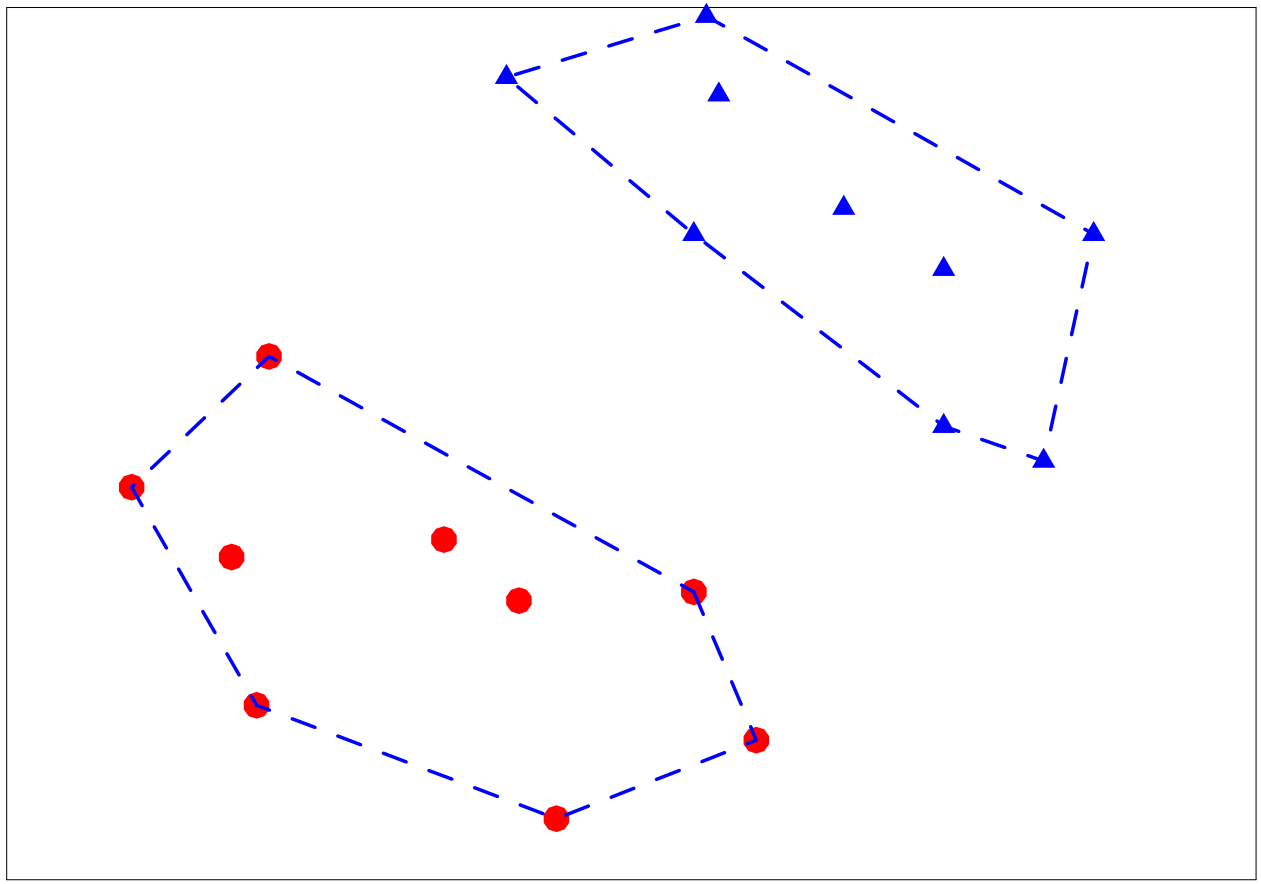
Photometric classifier: Svm



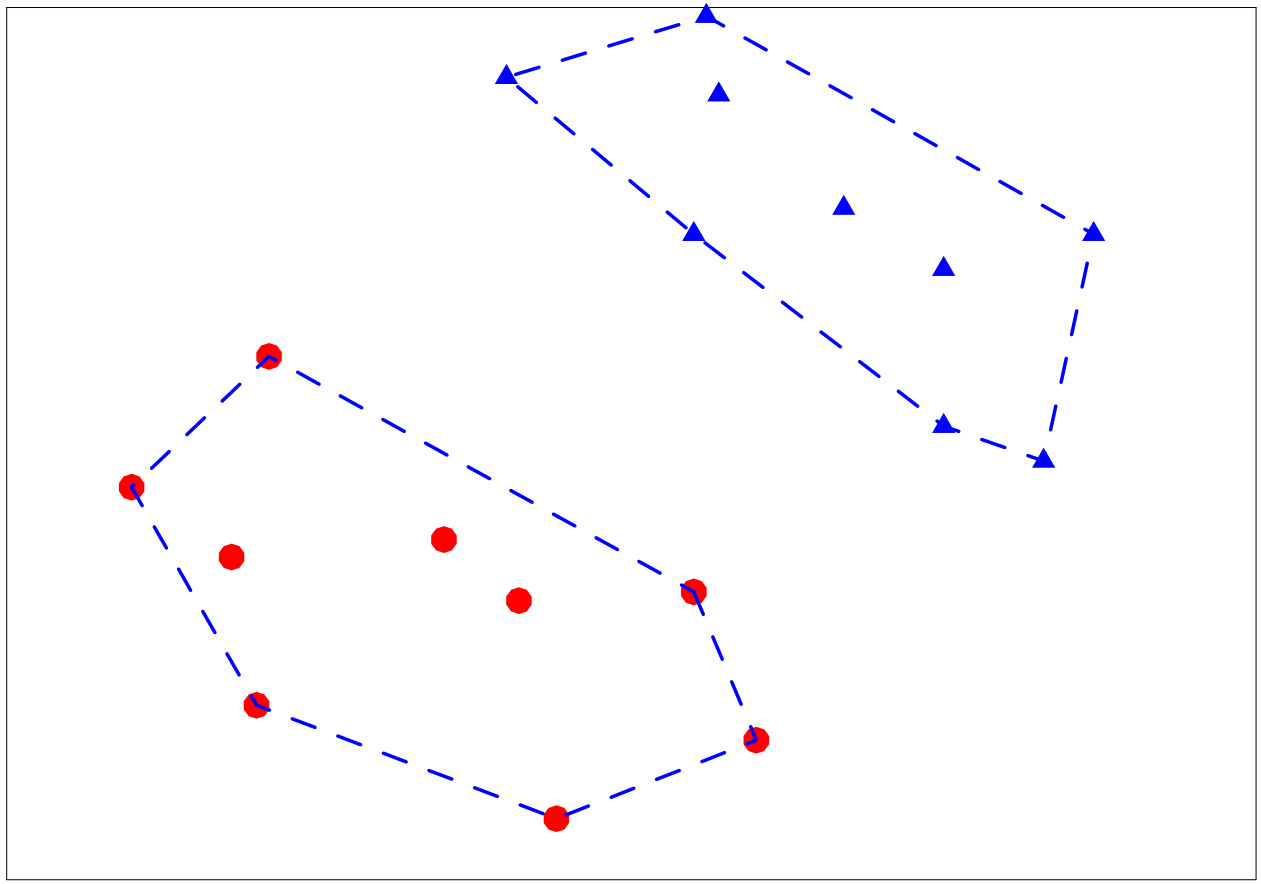
Photometric classifier: Svm



Svm concept: Convex hull



Svm concept: Convex hull



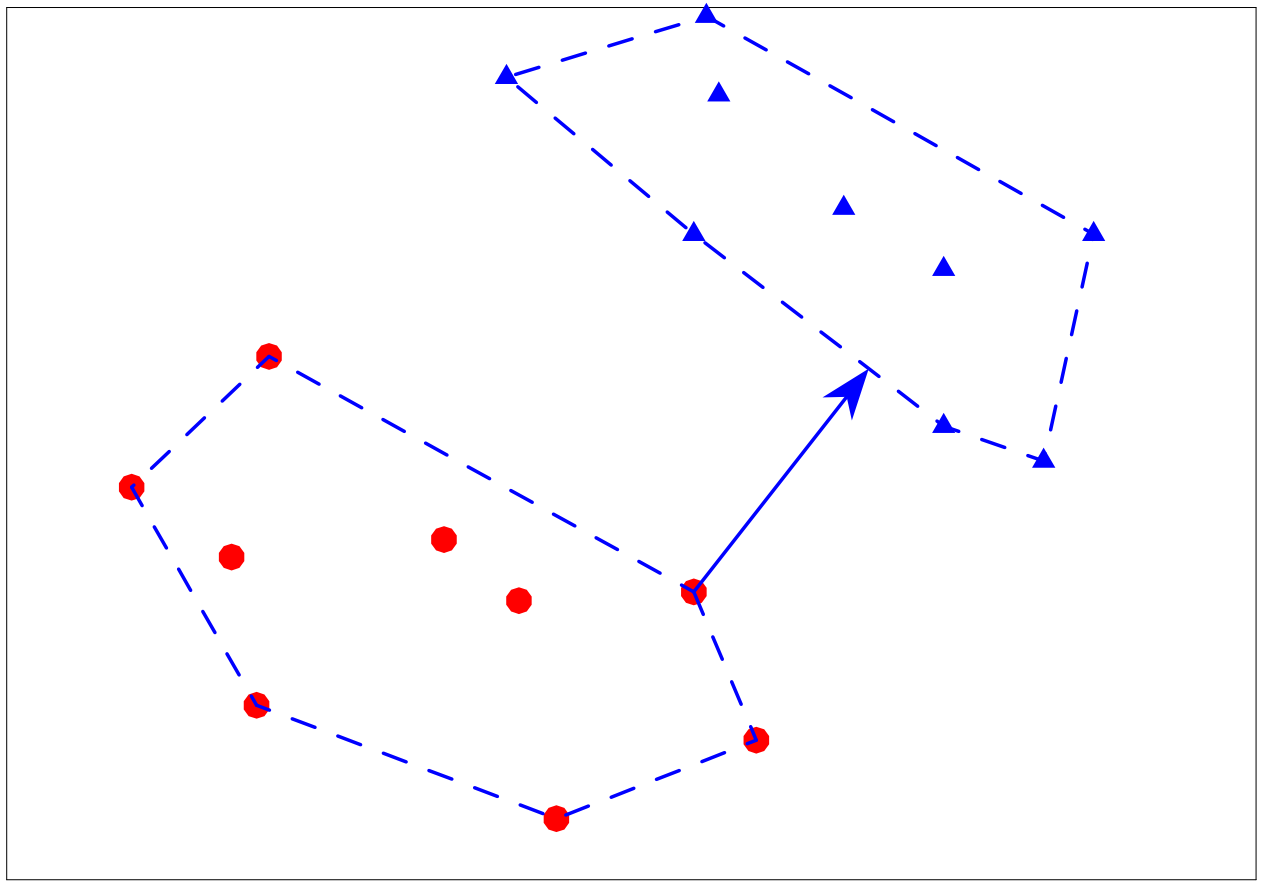
$\sum_i \alpha_i \mathbf{x}_i$

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

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



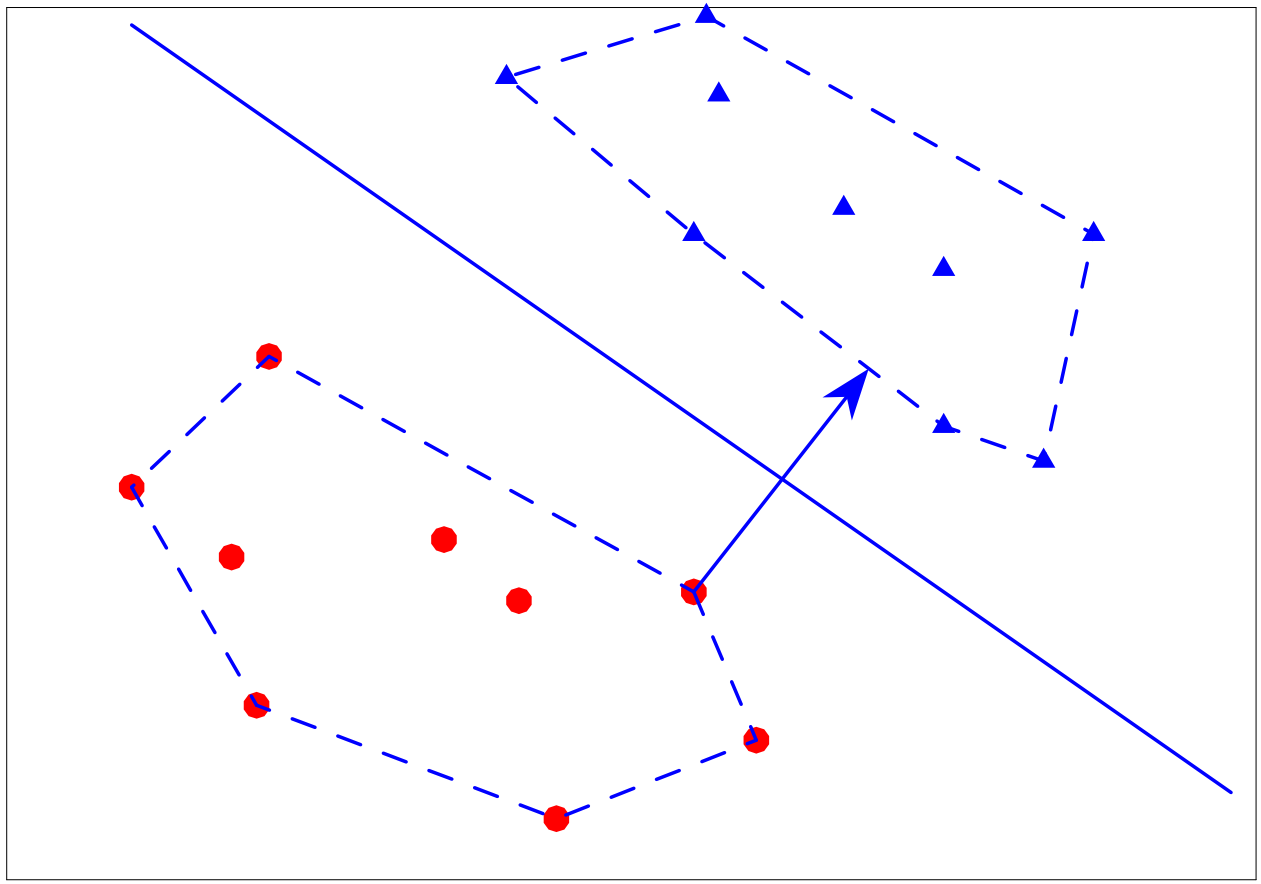
$\sum_i \alpha_i \mathrm{x}_i$

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

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



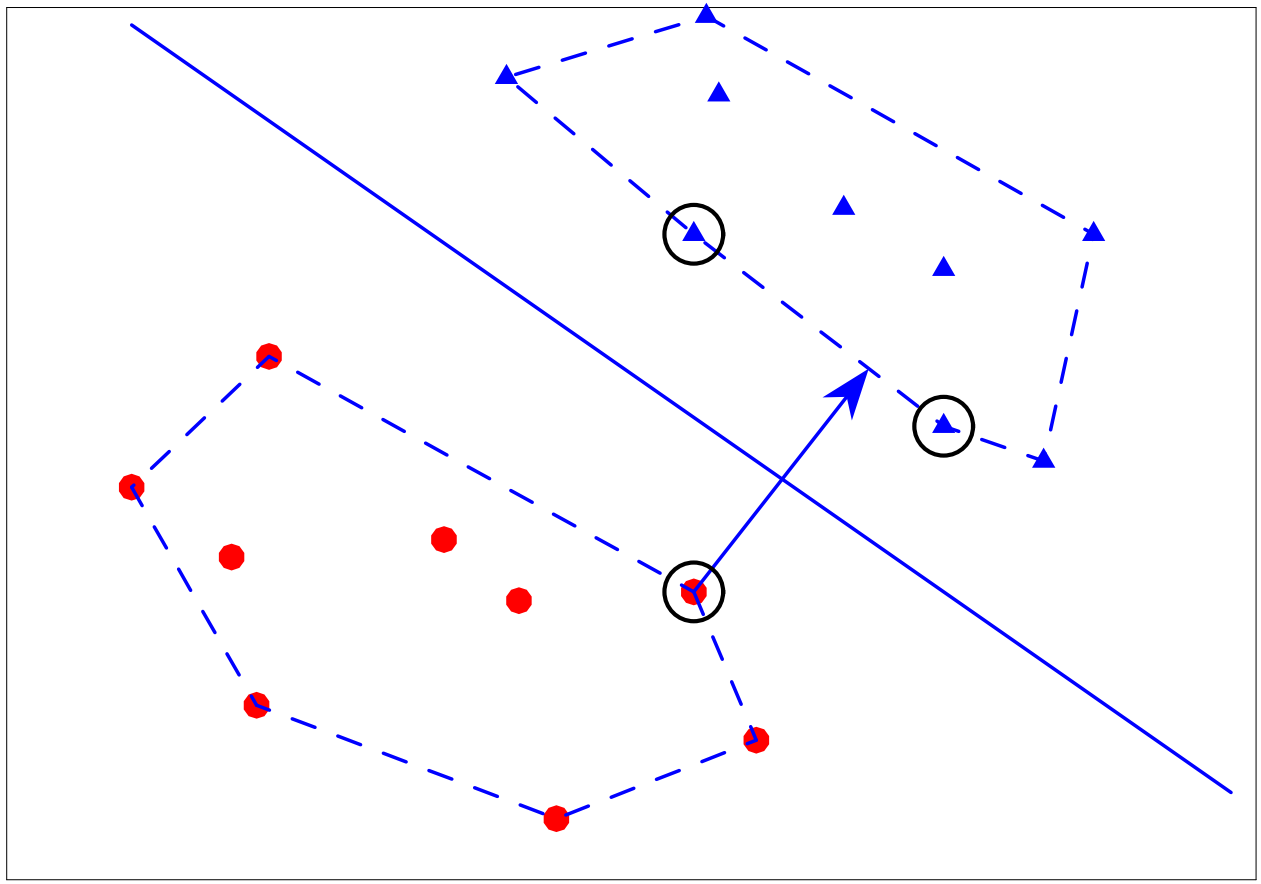
$\sum_i \alpha_i \mathbf{x}_i$

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

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



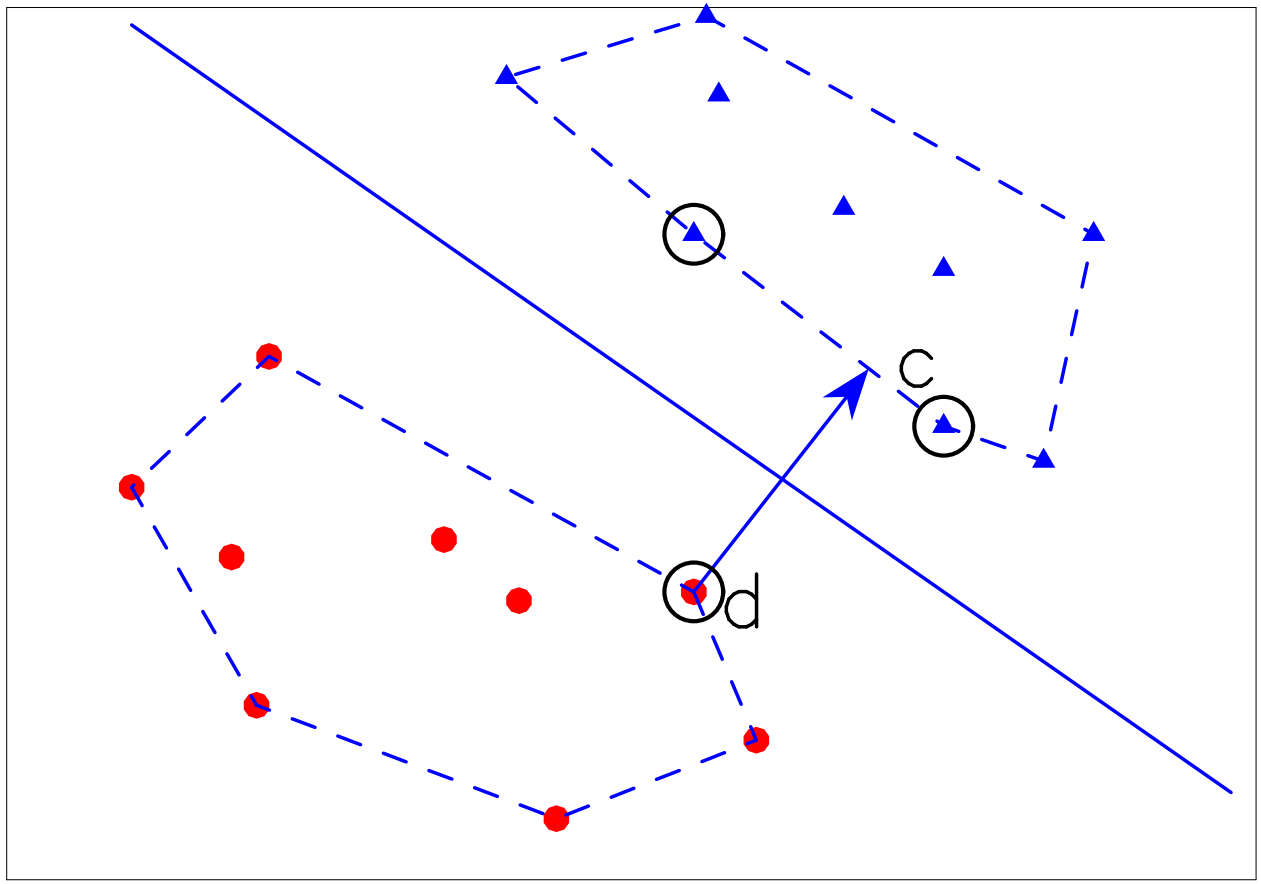
$\sum_i \alpha_i \mathbf{x}_i$

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

$\sum_i \alpha_i = 1$

$\alpha_i \geq 0$

Svm concept: Convex hull



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

c,d=$\sum_i \alpha_i x_i$

s.t. $\sum \alpha_i = 1$

$0 \leq \alpha_i \leq 1$

Applying the classifier

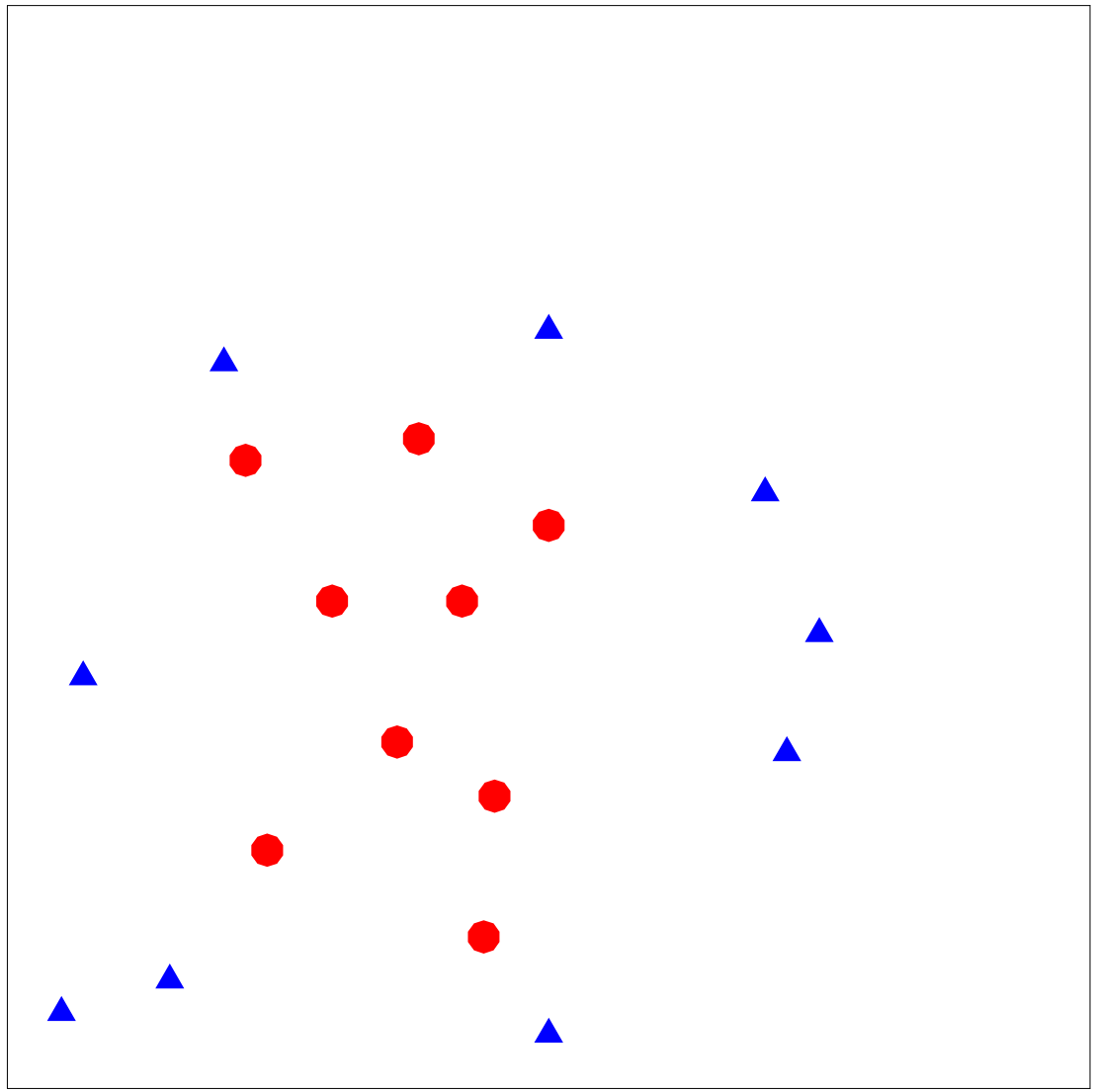
For a new point $\mathbf{s}$, the decision boundary is;

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

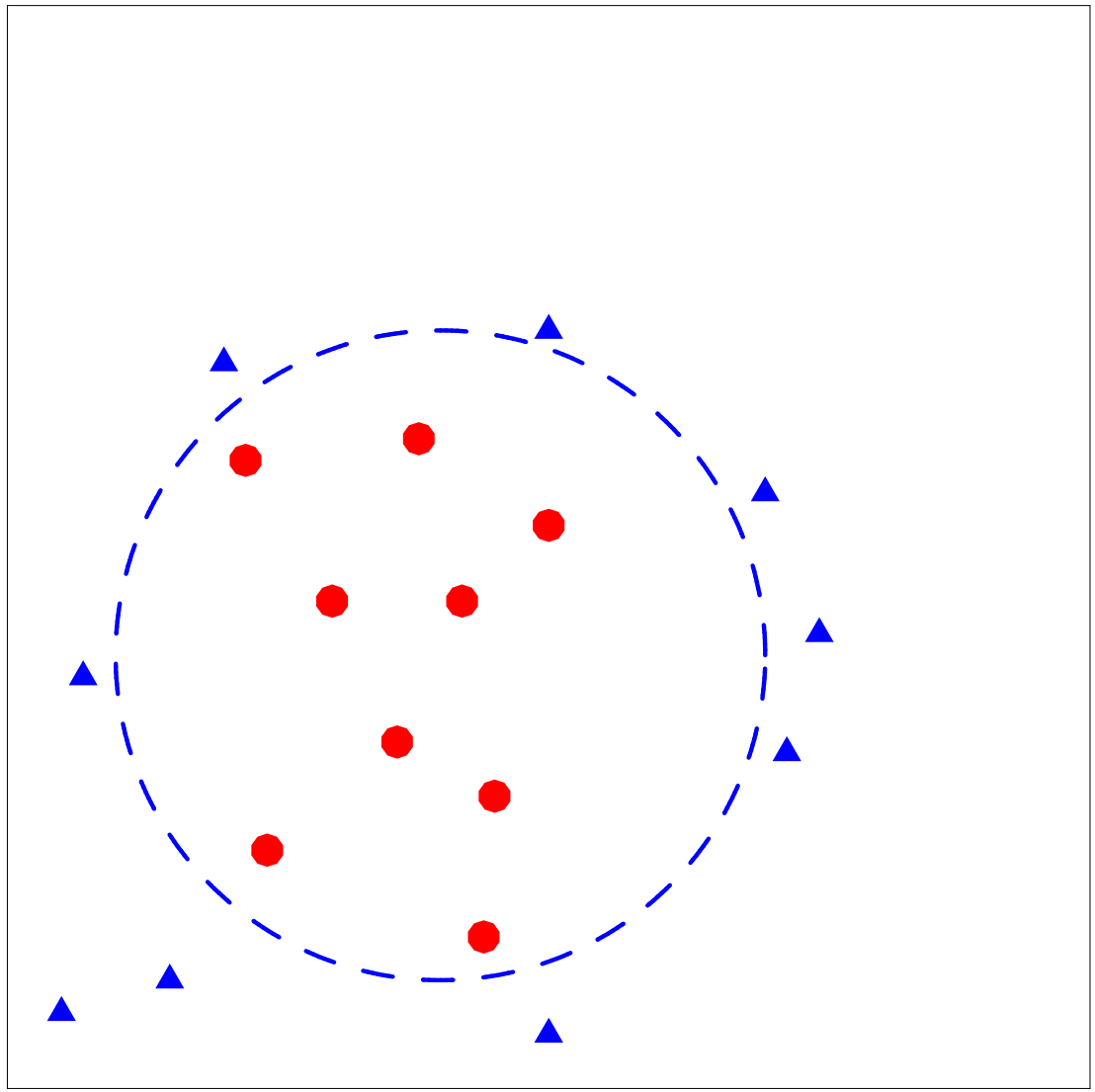
$$f > 0. \quad |\mathbf{s} \rightarrow class1$$

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

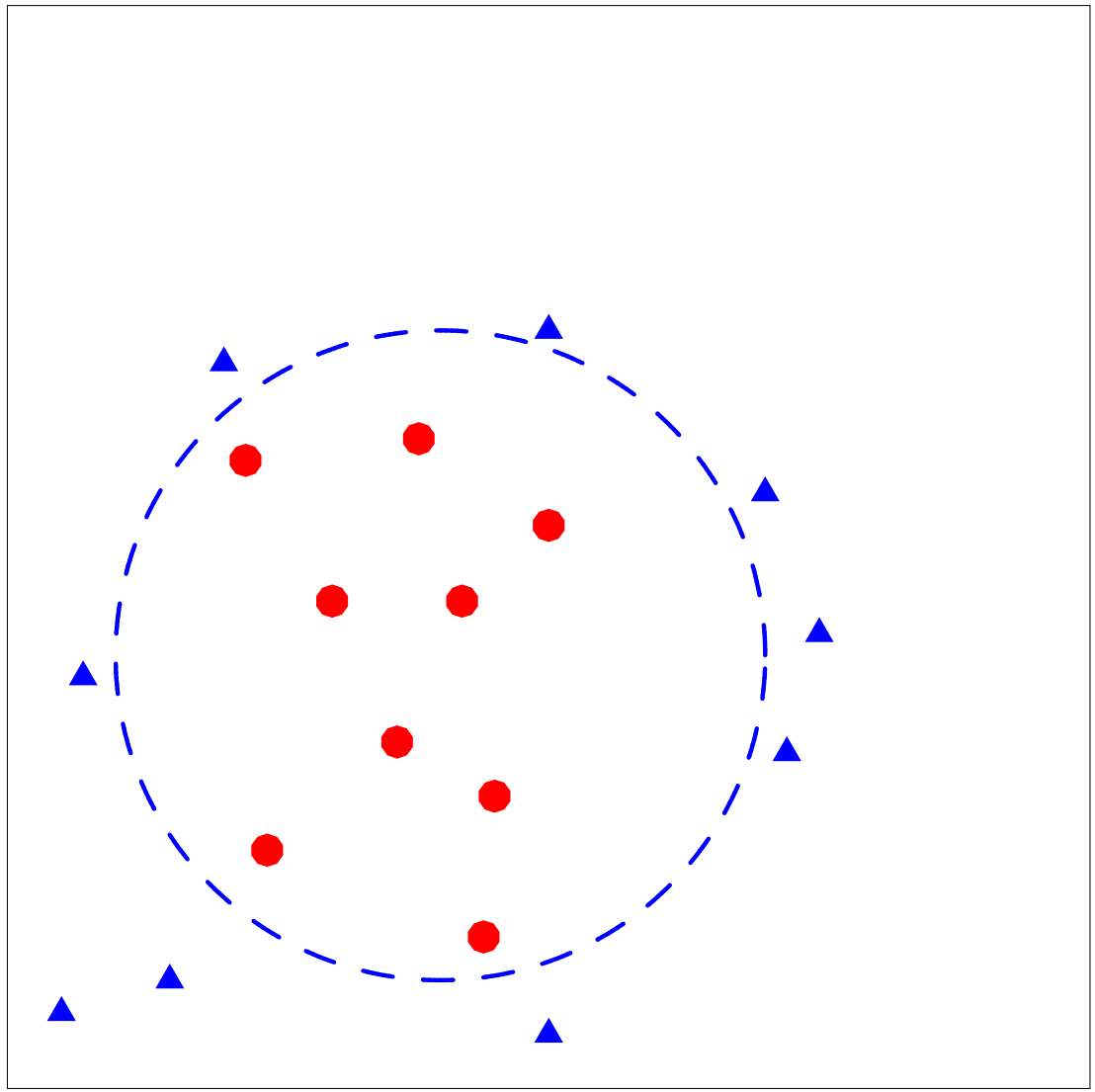
Linearly inseparable data



Linearly inseparable data

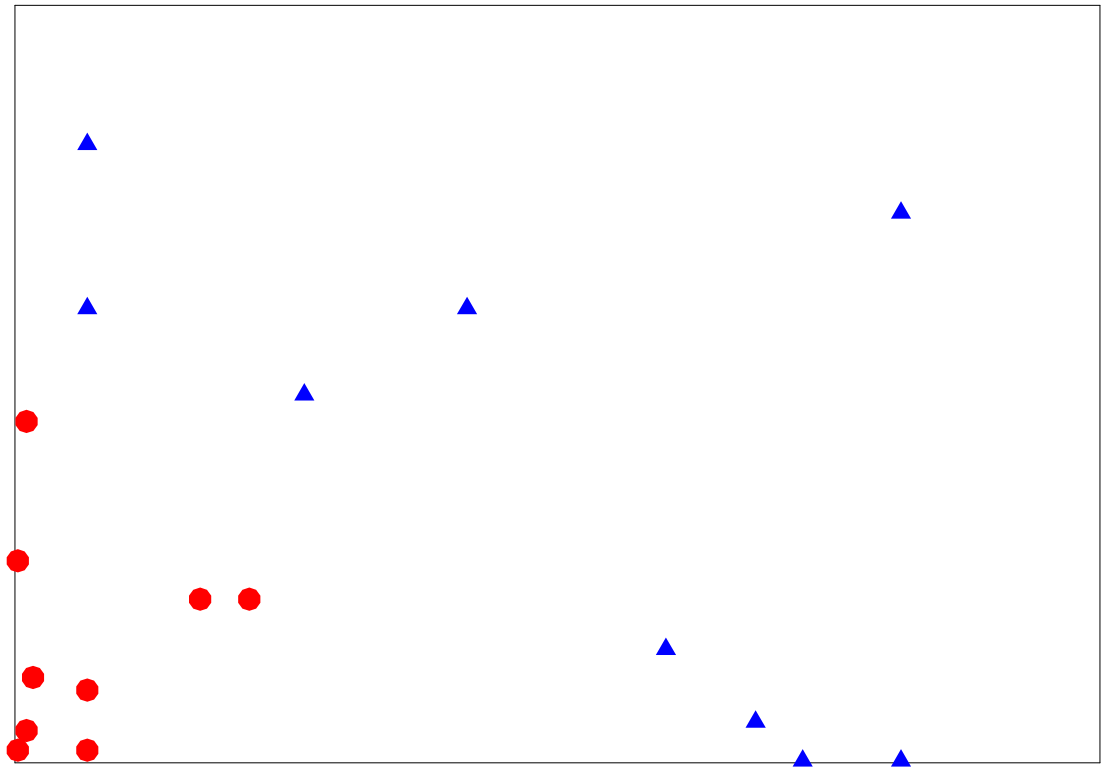


Linearly inseparable data



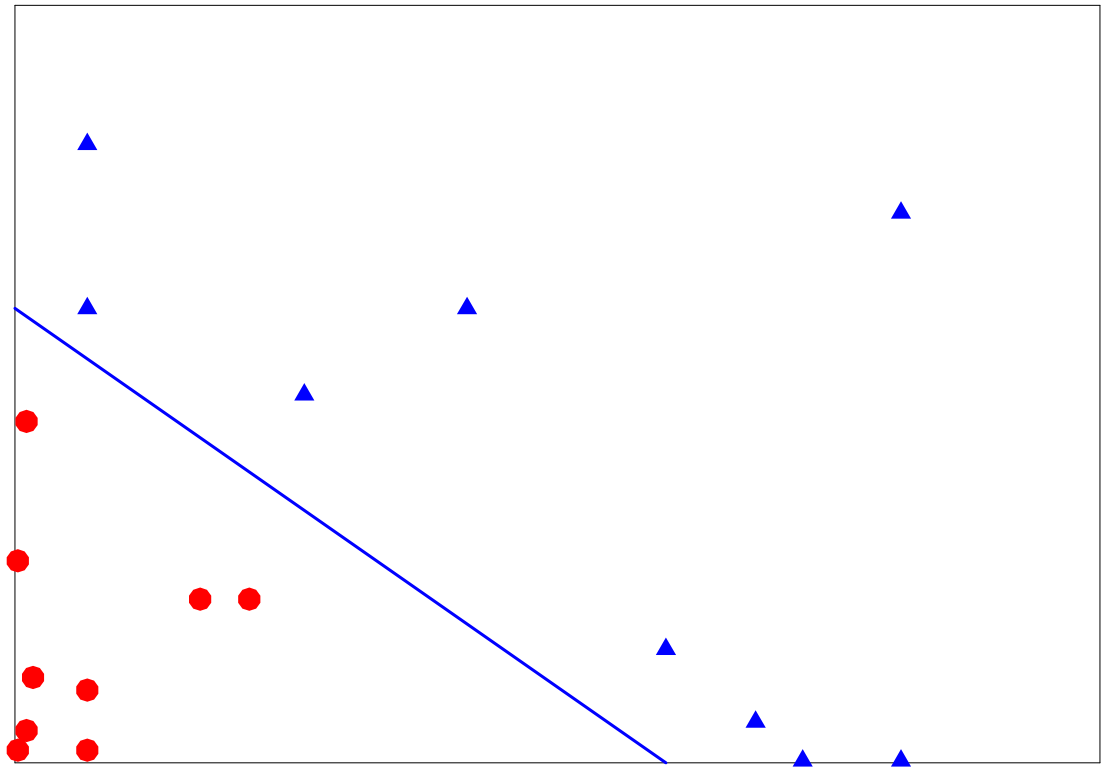
$$u, v \rightarrow u^2, v^2$$

Linearly inseparable data



$$u, v \rightarrow u^2, v^2$$

Linearly inseparable data



Kernel trick

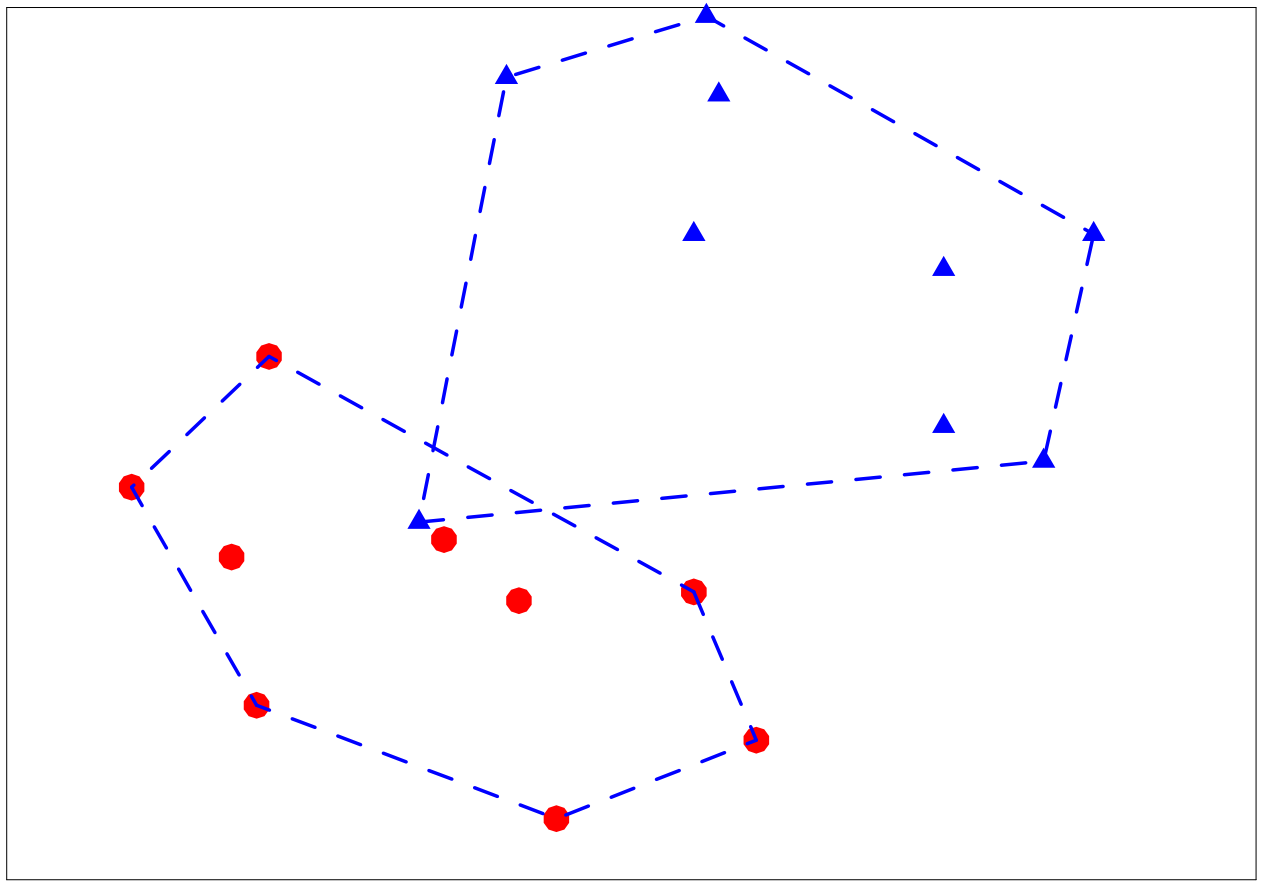
$$f = \sum_i \alpha_i y_i \mathbf{x}_i . \mathbf{s} + b,$$

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

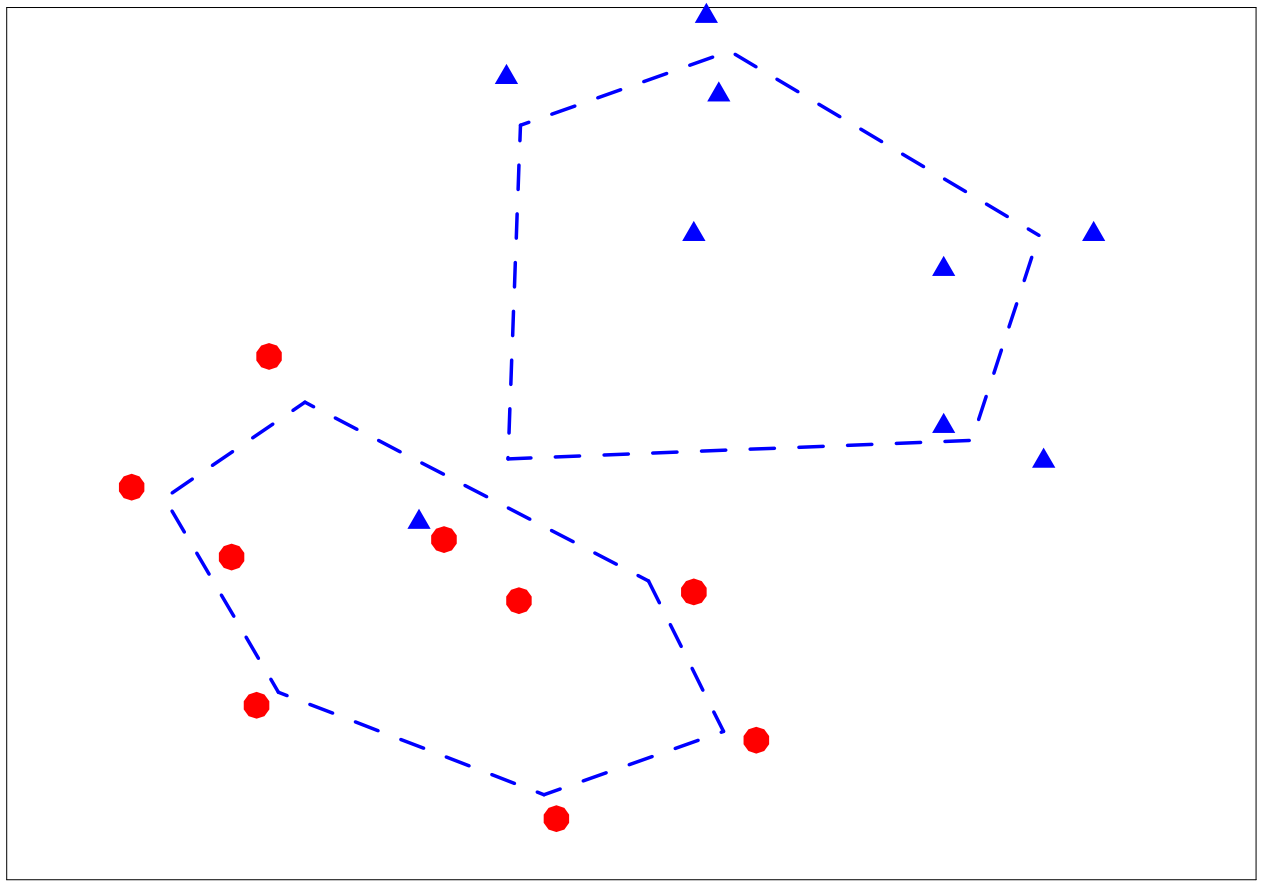
$$\exp\left(-\gamma \|x_i - s\|^2\right)$$

need to tune for γ.

Svm: Non separable cases

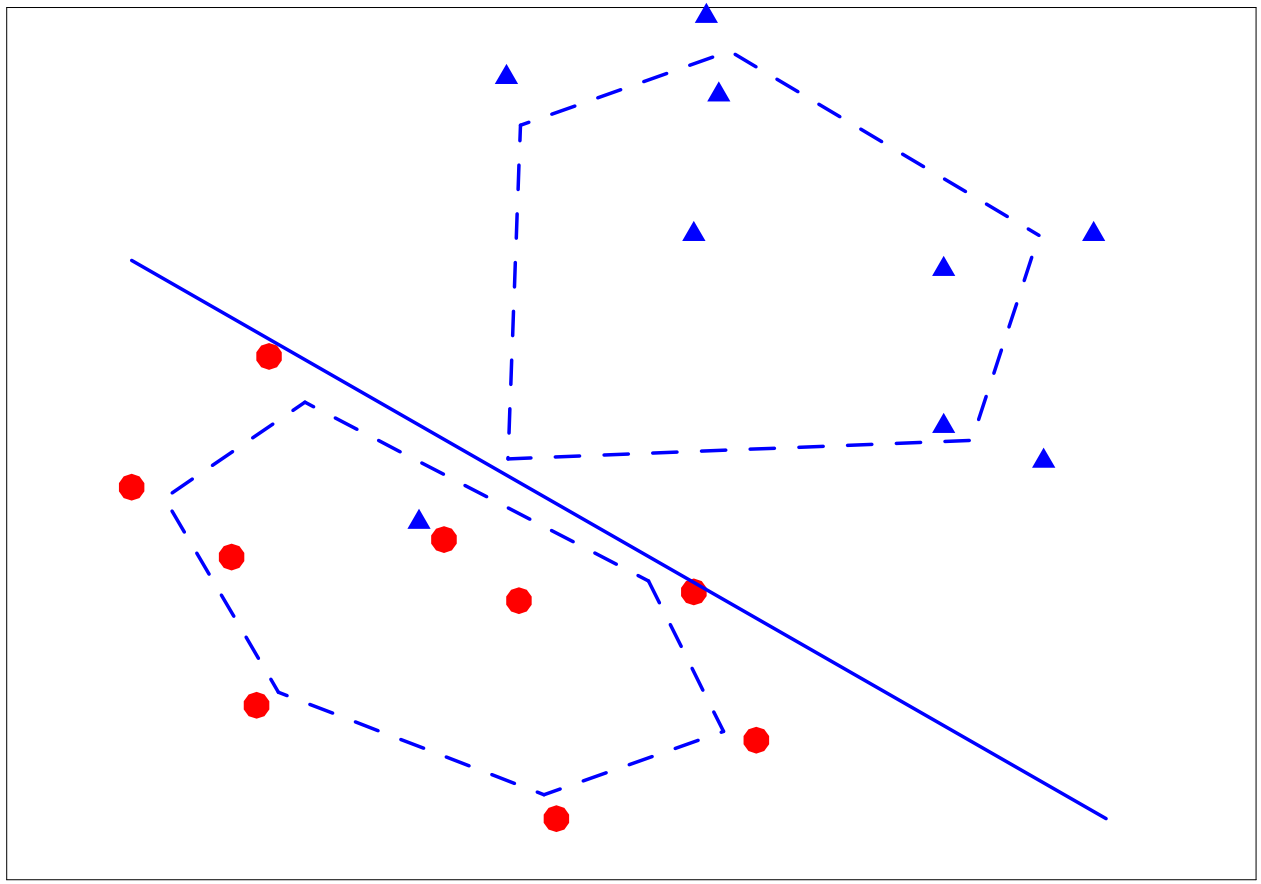


Svm: Non separable cases



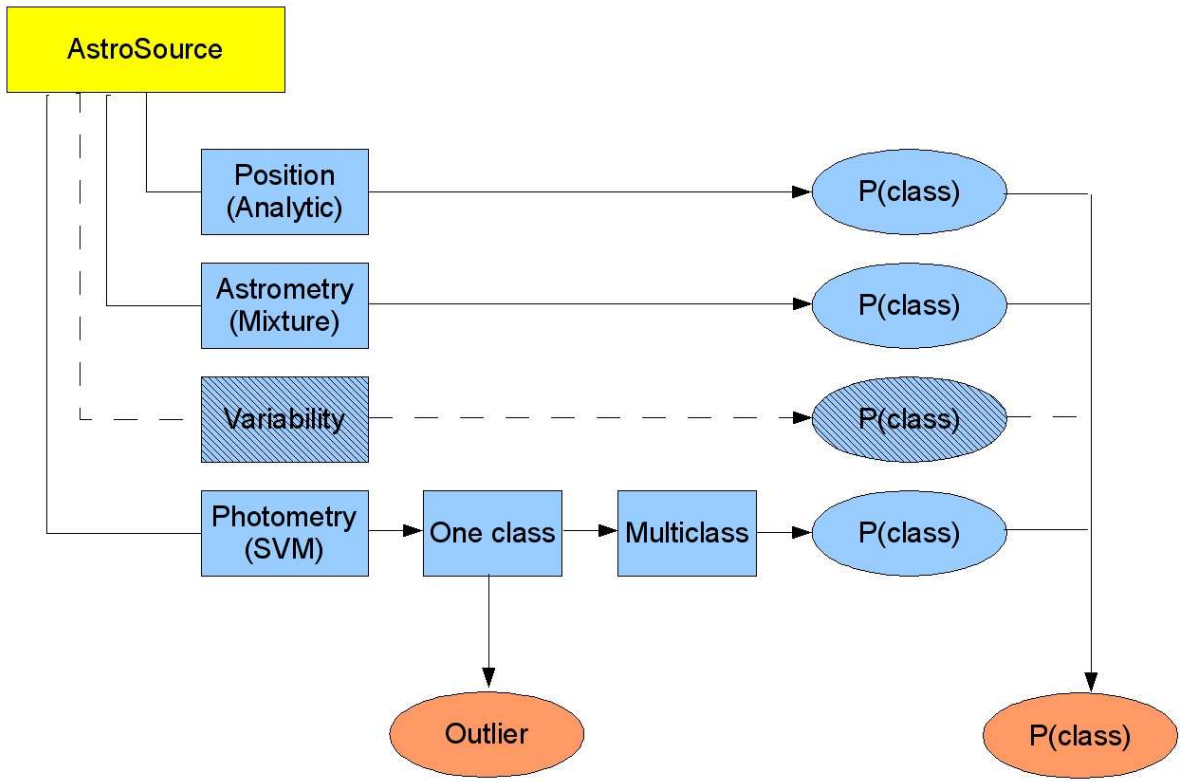
$$0 \leq \alpha_i \leq D$$

Svm: Non separable cases

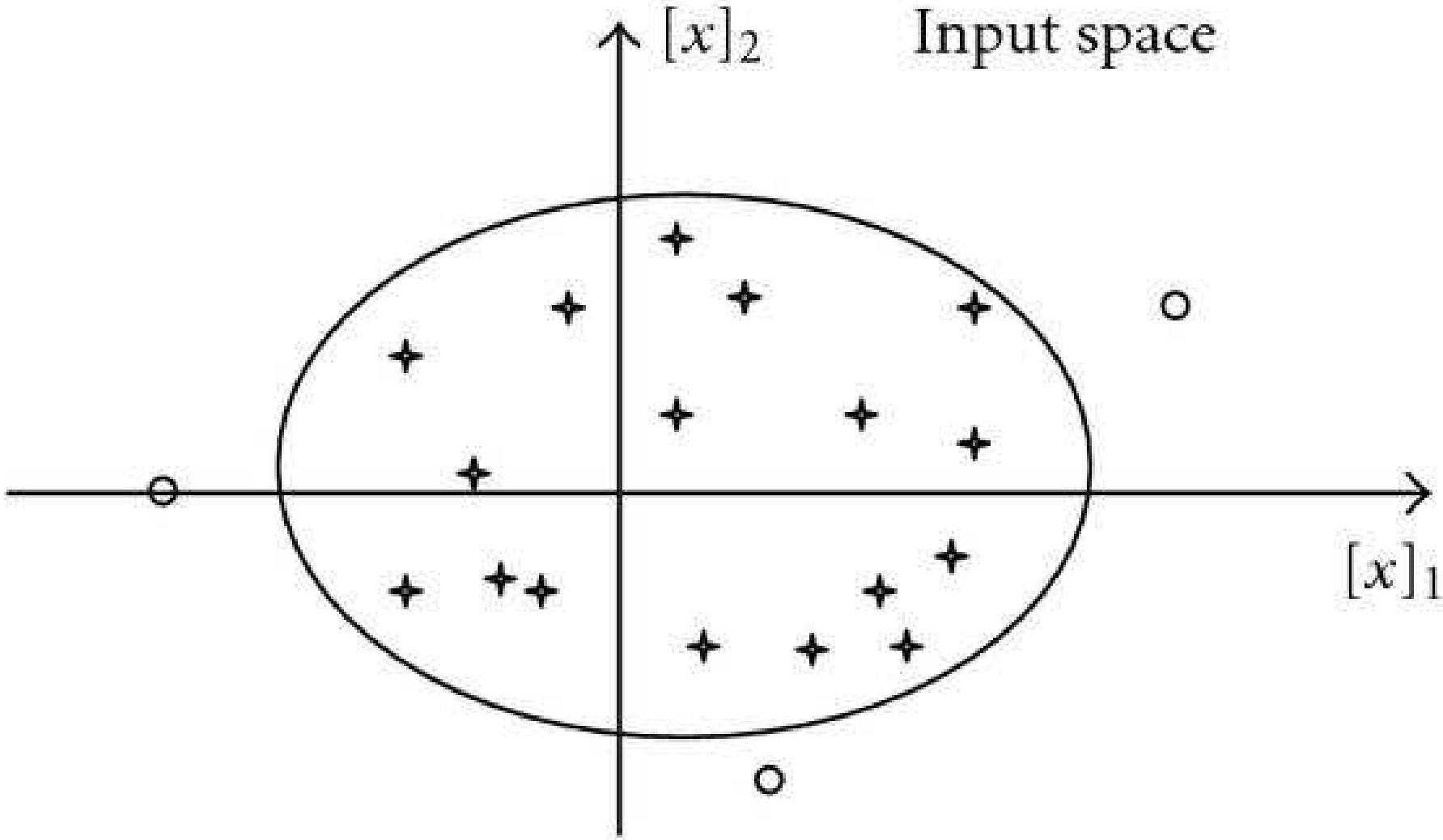


$$0 \leq \alpha_i \leq D$$

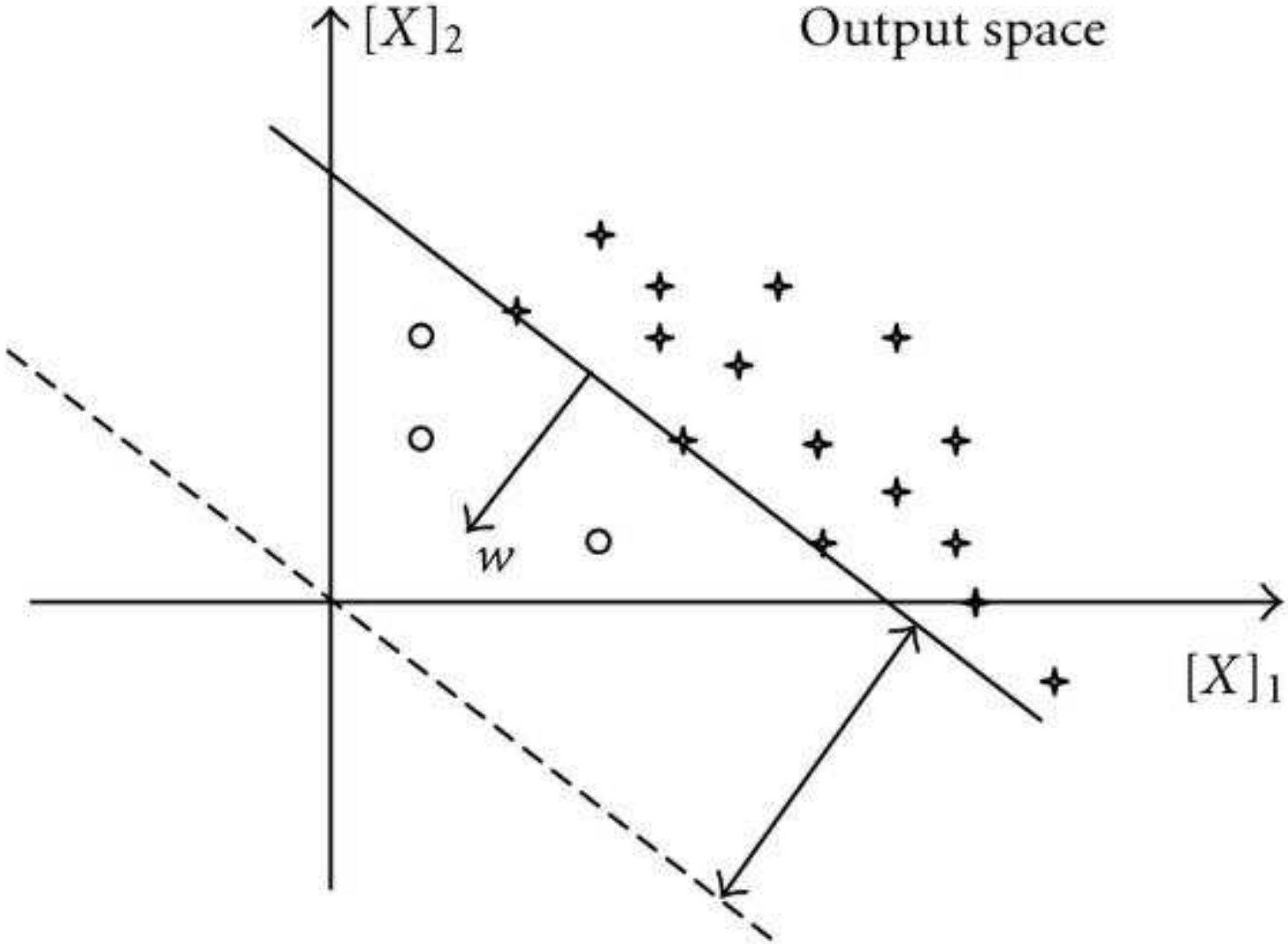
Internal structure of DSC



SVM - One Class

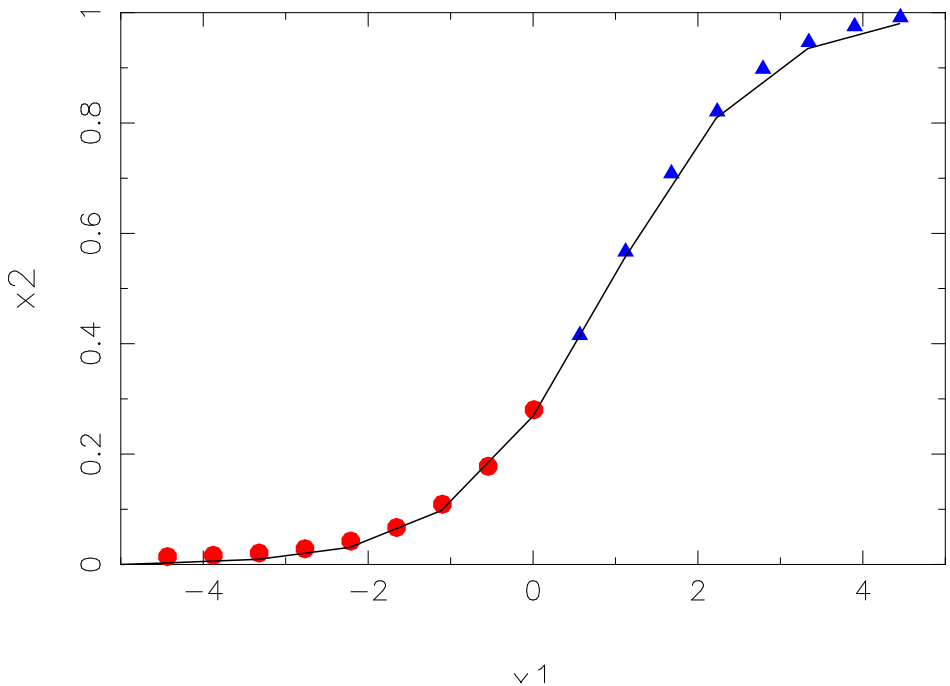


SVM - One Class



Probabilities

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



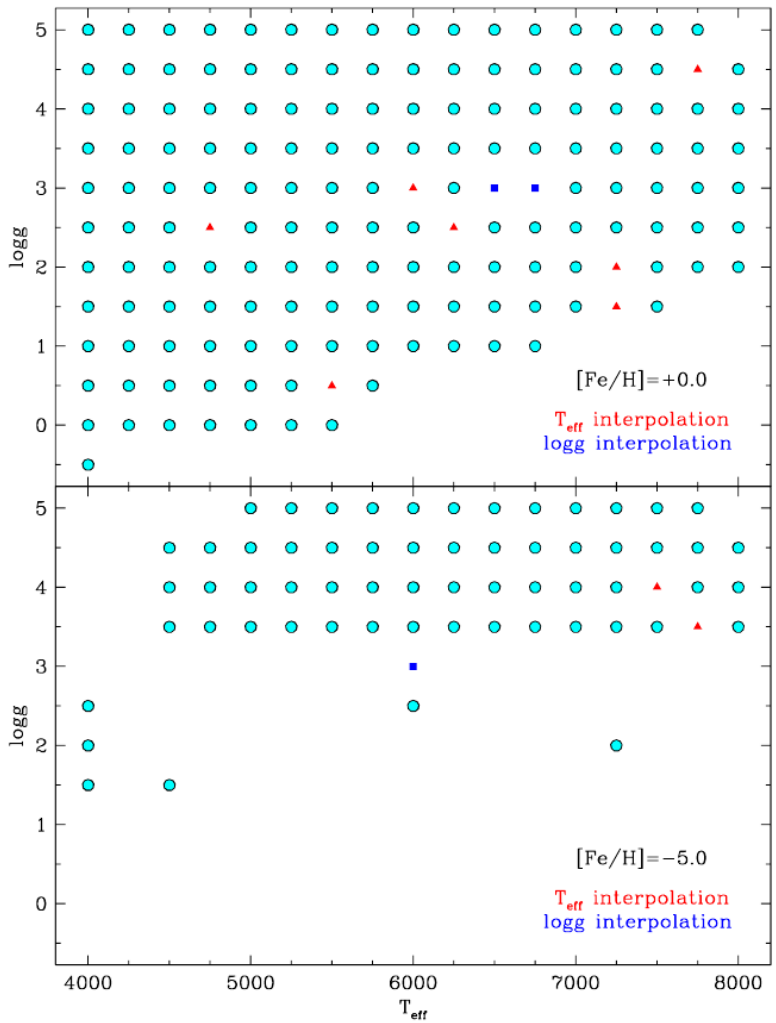
Probabilities: pairwise coupling

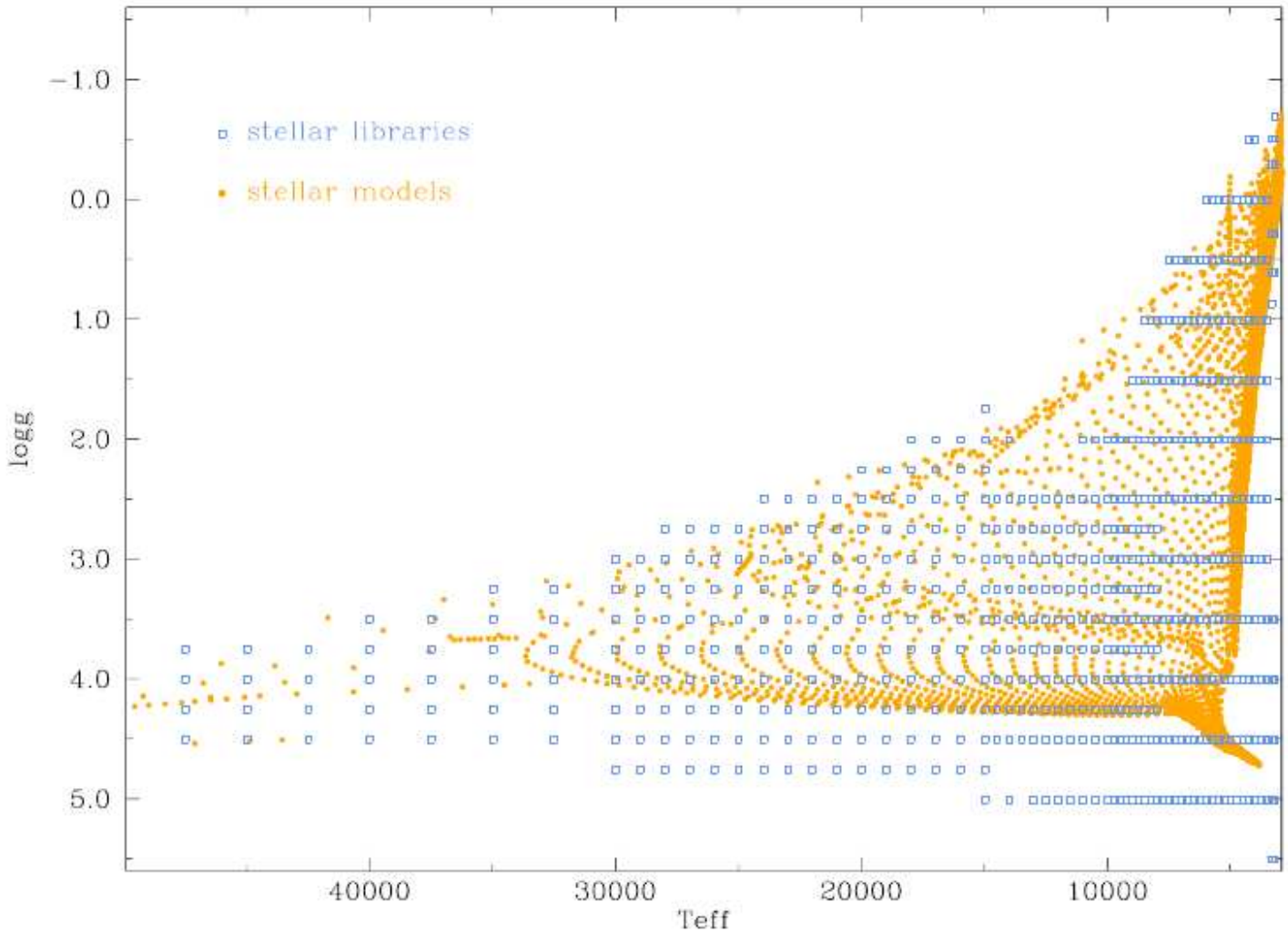
- classify pairwise (star vs. binary, star vs. quasar etc.)
- Can combine all pairwise probabilities to unique set of $P(C_m)$

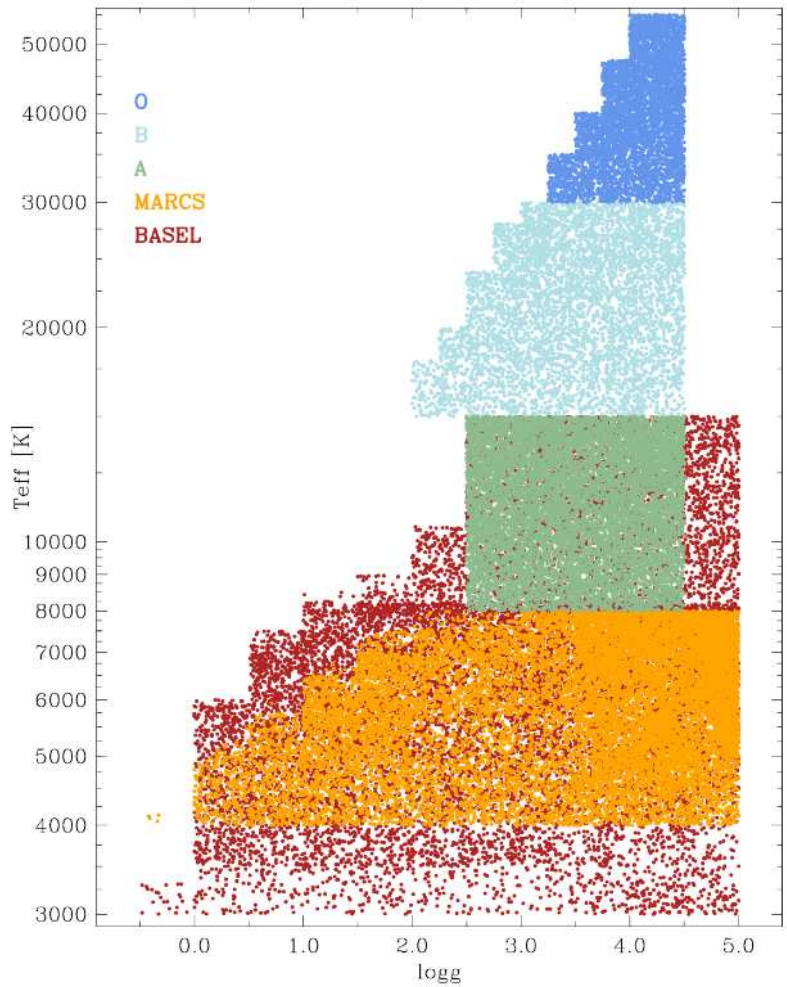
(Wu et al. 2004,Journal of Machine learning research)

Training data

Input spectra from codes or from empirical libraries+codes (semi-empirical data)
Semi-empirical = SDSS (usually)


logg
[Fe/H]=+0.0
T_eff interpolation
logg interpolation
logg
[Fe/H]=−5.0
T_eff interpolation
logg interpolation
T_eff
4000
5000
6000
7000
8000


stellar libraries
stellar models
−1.0
0.0
1.0
2.0
3.0
4.0
5.0
logg
40000
30000
20000
10000
Teff


O
B
A
MARCS
BASEL
Teff [K]
50000
40000
30000
20000
10000
9000
8000
7000
6000
5000
4000
3000
0.0
1.0
2.0
3.0
4.0
5.0
logg

Results overview

Grid	star	wd	bin	qso	gal	???
APec	94.44	–	–	5.55	0.00	0.00
Fastrot	98.26	–	–	0.69	0.00	1.04
Phoenix N	95.73	–	–	0.02	0.54	3.69
Phoenix R	98.67	–	–	0.15	0.85	0.33
SDSS Stars	99.58	–	–	0.20	0.21	0.00
UCD Cond	70.52	–	–	0.45	0.76	28.27
UCD Dust	98.30	–	–	0.00	0.00	1.70
WR	76.74	–	–	9.30	0.00	13.95
SDSS QSOs	0.20	–	–	95.85	2.33	1.60
SDSS galaxies	0.15	–	–	0.50	99.01	0.33
MARCS	82.05	2.80	5.75	4.05	1.90	3.45
O	78.80	11.00	0.40	5.80	1.20	2.80
B	86.20	5.80	1.00	4.40	0.80	1.80
A	89.10	3.20	0.90	3.60	1.00	2.20
Be	82.75	1.72	0.00	10.92	4.59	0.00
C stars	89.48	0.23	1.40	7.24	1.63	0.00
physBins	29.27	1.60	57.40	4.67	2.32	4.72
WDA	18.20	58.42	0.52	16.10	1.95	4.80
WDB	15.67	58.20	0.55	19.37	1.95	4.25

Results overview

Grid	star	wd	bin	qso	gal	???
APec	94.44	–	–	5.55	0.00	0.00
Fastrot	98.26	–	–	0.69	0.00	1.04
Phoenix N	95.73	–	–	0.02	0.54	3.69
Phoenix R	98.67	–	–	0.15	0.85	0.33
SDSS Stars	99.58	–	–	0.20	0.21	0.00
UCD Cond	70.52	–	–	0.45	0.76	28.27
UCD Dust	98.30	–	–	0.00	0.00	1.70
WR	76.74	–	–	9.30	0.00	13.95
SDSS QSOs	0.20	–	–	95.85	2.33	1.60
SDSS galaxies	0.15	–	–	0.50	99.01	0.33
MARCS	82.05	2.80	5.75	4.05	1.90	3.45
O	78.80	11.00	0.40	5.80	1.20	2.80
B	86.20	5.80	1.00	4.40	0.80	1.80
A	89.10	3.20	0.90	3.60	1.00	2.20
Be	82.75	1.72	0.00	10.92	4.59	0.00
C stars	89.48	0.23	1.40	7.24	1.63	0.00
Binaries	29.27	1.60	57.40	4.67	2.32	4.72
WDA	18.20	58.42	0.52	16.10	1.95	4.80
WDB	15.67	58.20	0.55	19.37	1.95	4.25

Results overview

Grid	star	wd	bin	qso	gal	???
APec	94.44	–	–	5.55	0.00	0.00
Fastrot	98.26	–	–	0.69	0.00	1.04
Phoenix N	95.73	–	–	0.02	0.54	3.69
Phoenix R	98.67	–	–	0.15	0.85	0.33
SDSS Stars	99.58	–	–	0.20	0.21	0.00
UCD Cond	70.52	–	–	0.45	0.76	28.27
UCD Dust	98.30	–	–	0.00	0.00	1.70
WR	76.74	–	–	9.30	0.00	13.95
SDSS QSOs	0.20	–	–	95.85	2.33	1.60
SDSS galaxies	0.15	–	–	0.50	99.01	0.33
MARCS	82.05	2.80	5.75	4.05	1.90	3.45
O	78.80	11.00	0.40	5.80	1.20	2.80
B	86.20	5.80	1.00	4.40	0.80	1.80
A	89.10	3.20	0.90	3.60	1.00	2.20
Be	82.75	1.72	0.00	10.92	4.59	0.00
C stars	89.48	0.23	1.40	7.24	1.63	0.00
physBins	29.27	1.60	57.40	4.67	2.32	4.72
WDA	18.20	58.42	0.52	16.10	1.95	4.80
WDB	15.67	58.20	0.55	19.37	1.95	4.25

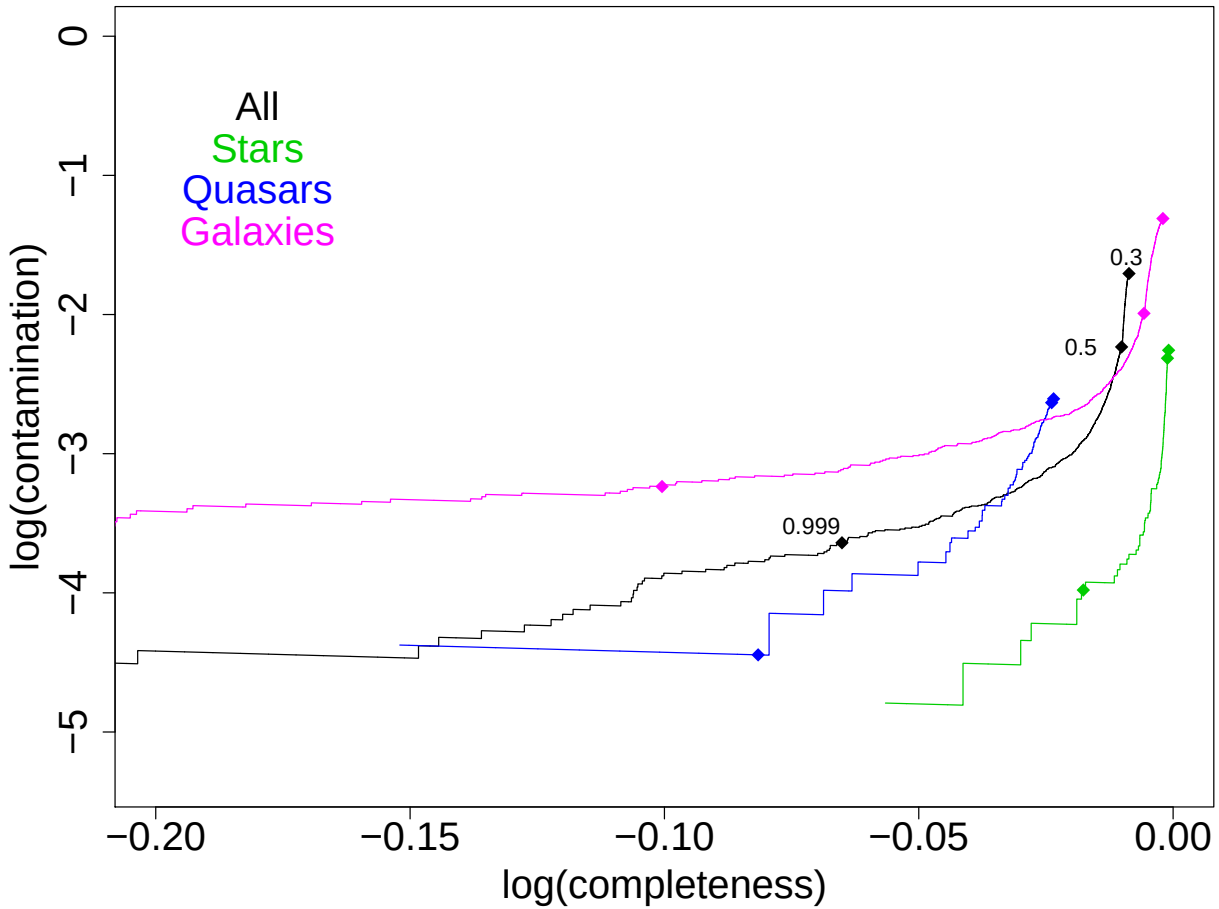
Results overview

Grid	star	wd	bin	qso	gal	???
APec	94.44	–	–	5.55	0.00	0.00
Fastrot	98.26	–	–	0.69	0.00	1.04
Phoenix N	95.73	–	–	0.02	0.54	3.69
Phoenix R	98.67	–	–	0.15	0.85	0.33
SDSS Stars	99.58	–	–	0.20	0.21	0.00
UCD Cond	70.52	–	–	0.45	0.76	28.27
UCD Dust	98.30	–	–	0.00	0.00	1.70
WR	76.74	–	–	9.30	0.00	13.95
SDSS QSOs	0.20	–	–	95.85	2.33	1.60
SDSS galaxies	0.15	–	–	0.50	99.01	0.33
MARCS	82.05	2.80	5.75	4.05	1.90	3.45
O	78.80	11.00	0.40	5.80	1.20	2.80
B	86.20	5.80	1.00	4.40	0.80	1.80
A	89.10	3.20	0.90	3.60	1.00	2.20
Be	82.75	1.72	0.00	10.92	4.59	0.00
C stars	89.48	0.23	1.40	7.24	1.63	0.00
physBins	29.27	1.60	57.40	4.67	2.32	4.72
WDA	18.20	58.42	0.52	16.10	1.95	4.80
WDB	15.67	58.20	0.55	19.37	1.95	4.25

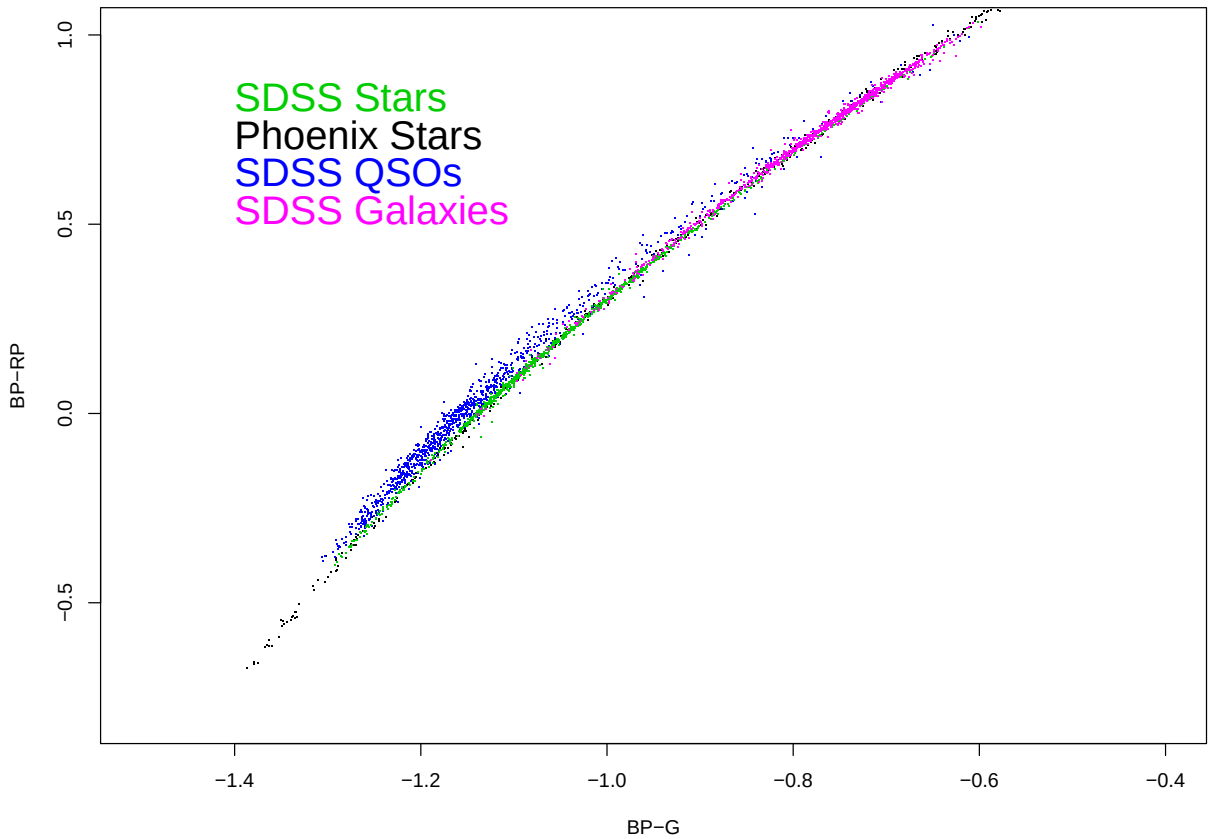
Results overview

Grid	star	wd	bin	qso	gal	???
APec	94.44	–	–	5.55	0.00	0.00
Fastrot	98.26	–	–	0.69	0.00	1.04
Phoenix N	95.73	–	–	0.02	0.54	3.69
Phoenix R	98.67	–	–	0.15	0.85	0.33
SDSS Stars	99.58	–	–	0.20	0.21	0.00
UCD Cond	70.52	–	–	0.45	0.76	28.27
UCD Dust	98.30	–	–	0.00	0.00	1.70
WR	76.74	–	–	9.30	0.00	13.95
SDSS QSOs	0.20	–	–	95.85	2.33	1.60
SDSS galaxies	0.15	–	–	0.50	99.01	0.33
MARCS	82.05	2.80	5.75	4.05	1.90	3.45
O	78.80	11.00	0.40	5.80	1.20	2.80
B	86.20	5.80	1.00	4.40	0.80	1.80
A	89.10	3.20	0.90	3.60	1.00	2.20
Be	82.75	1.72	0.00	10.92	4.59	0.00
C stars	89.48	0.23	1.40	7.24	1.63	0.00
physBins	29.27	1.60	57.40	4.67	2.32	4.72
WDA	18.20	58.42	0.52	16.10	1.95	4.80
WDB	15.67	58.20	0.55	19.37	1.95	4.25

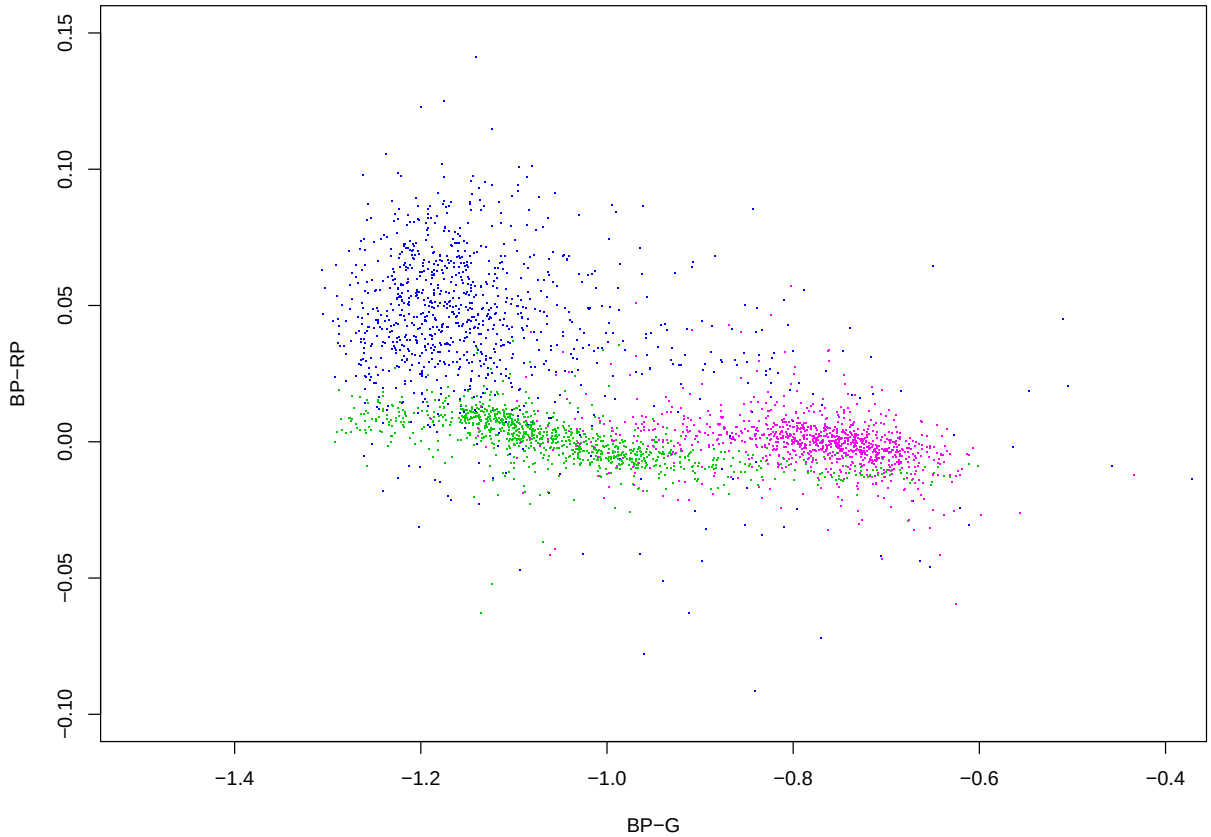
Dsc results


All
Stars
Quasars
Galaxies
0.3
0.5
0.999
log(contamination)
0
−1
−2
−3
−4
−5
−0.20
−0.15
−0.10
−0.05
0.00
log(completeness)

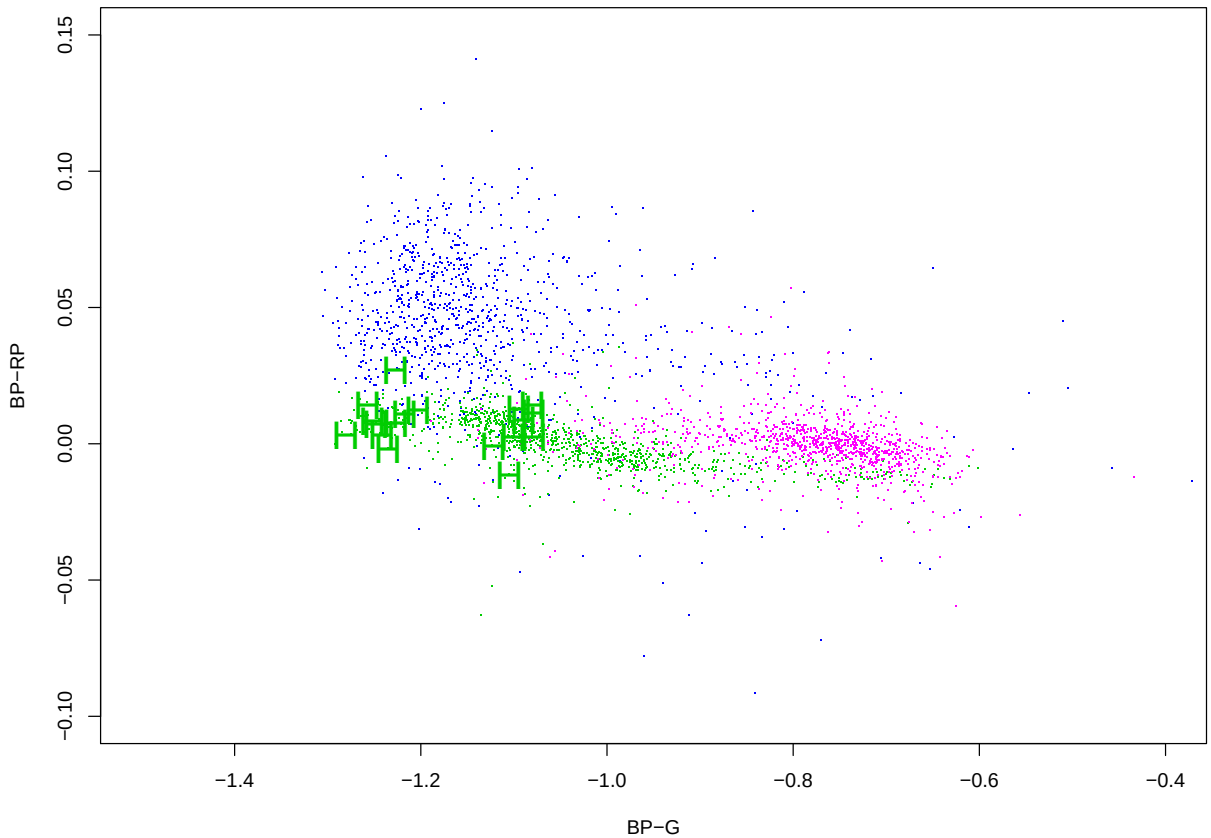
Colours


SDSS Stars
Phoenix Stars
SDSS QSOs
SDSS Galaxies
BP−RP
1.0
0.5
0.0
−0.5
−1.4
−1.2
−1.0
−0.8
−0.6
−0.4
BP−G

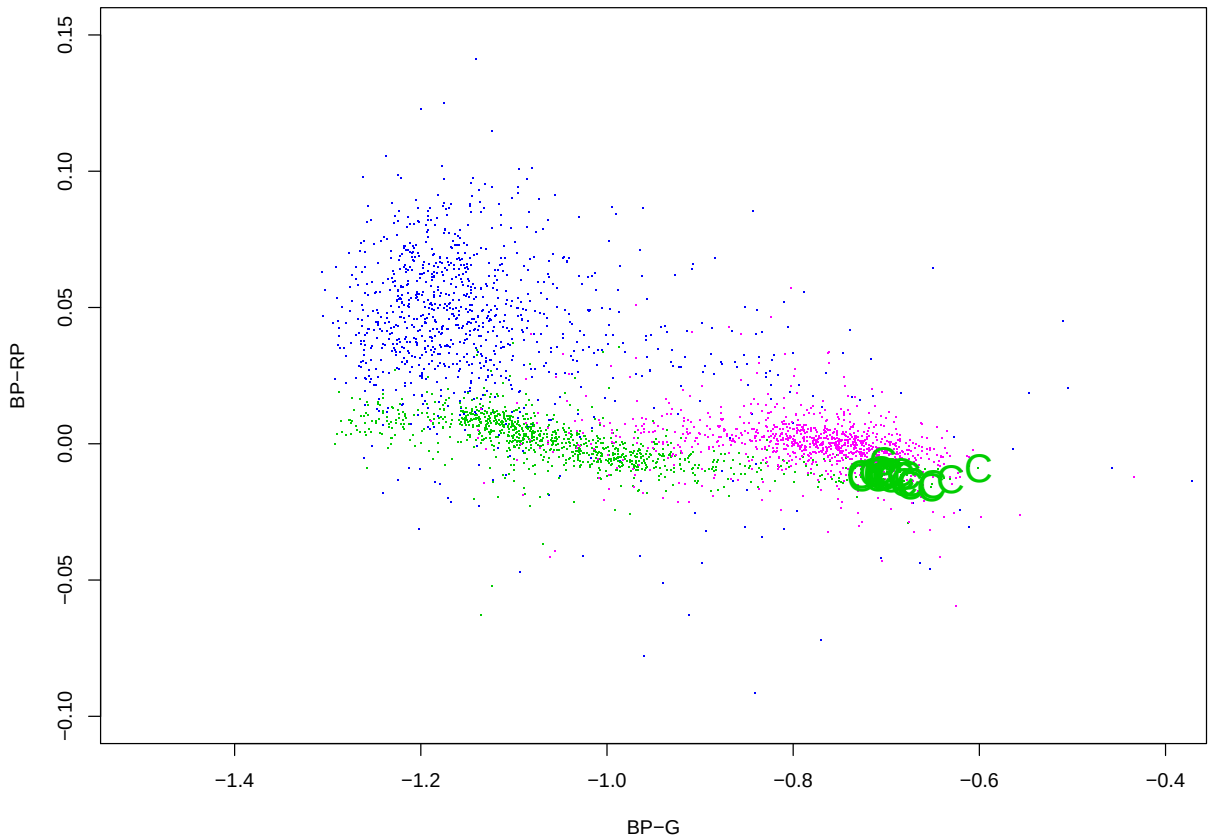
Colours



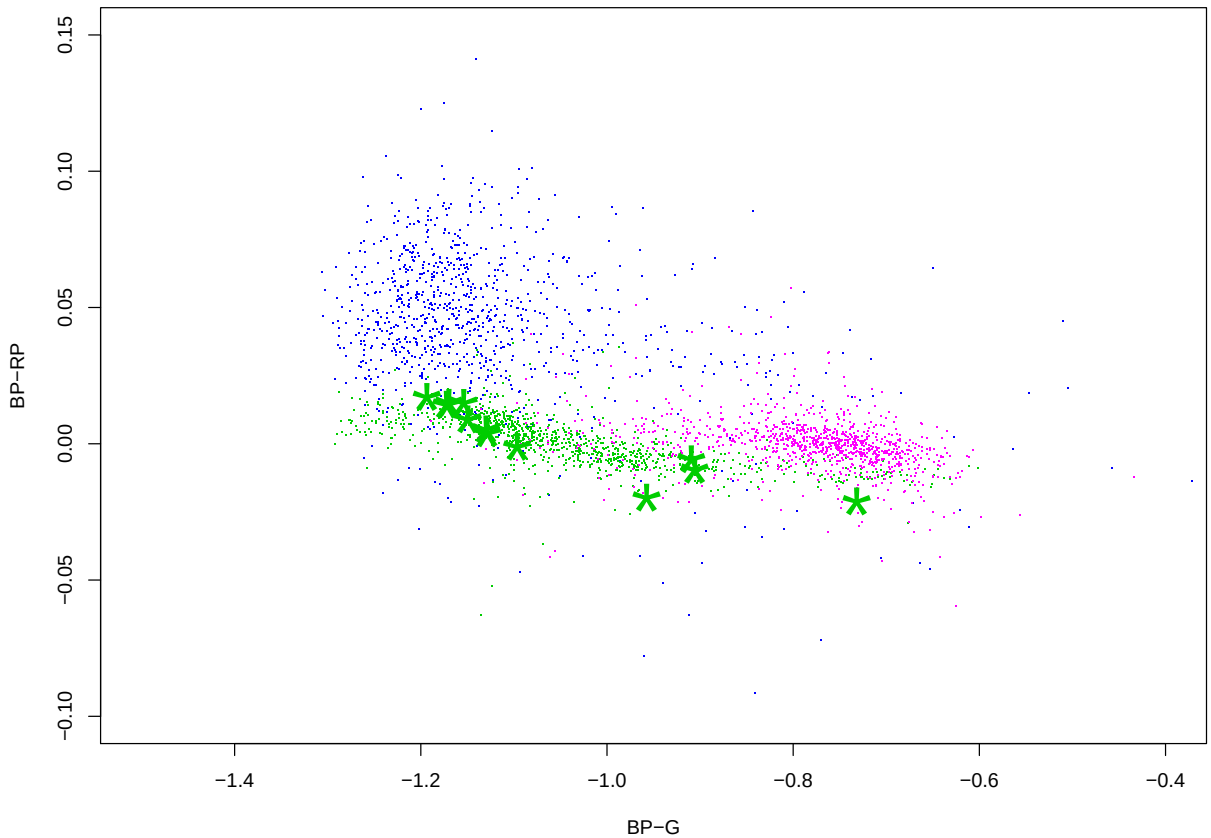
Colours



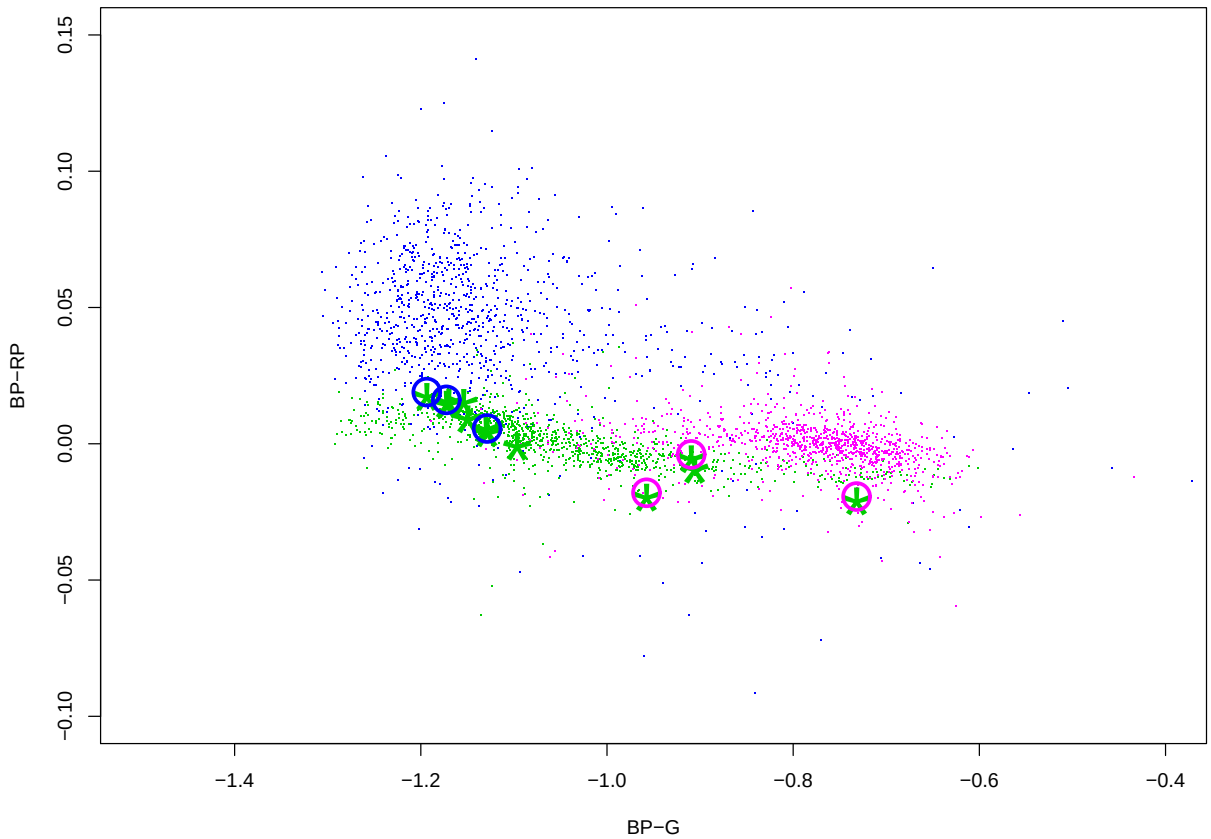
Colours



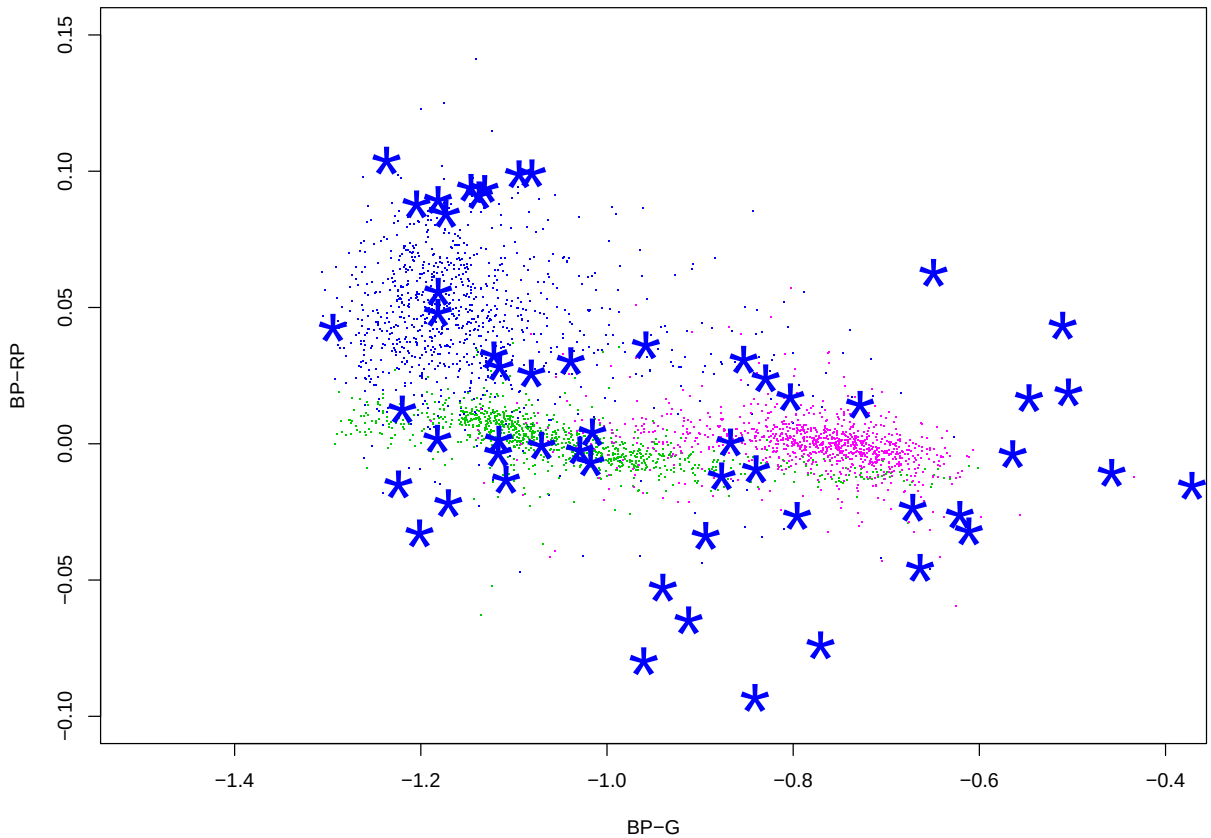
Colours



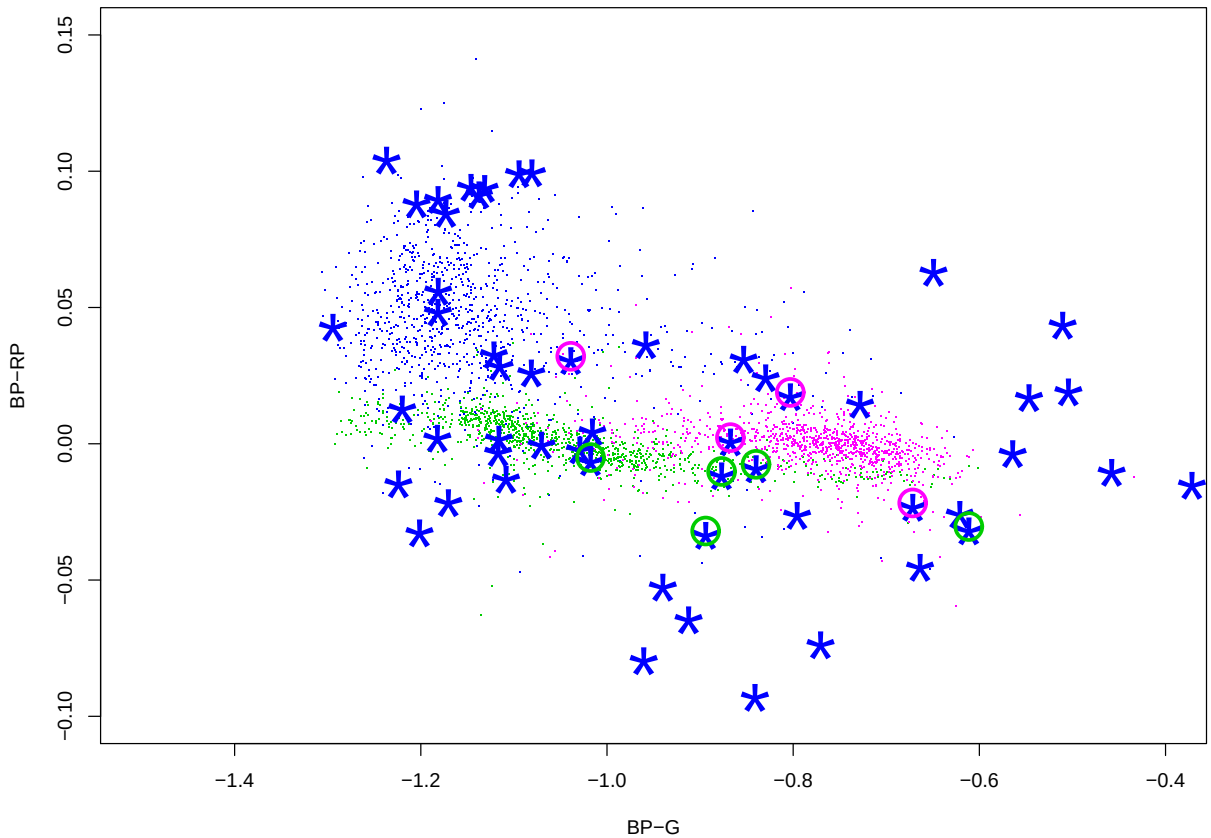
Colours



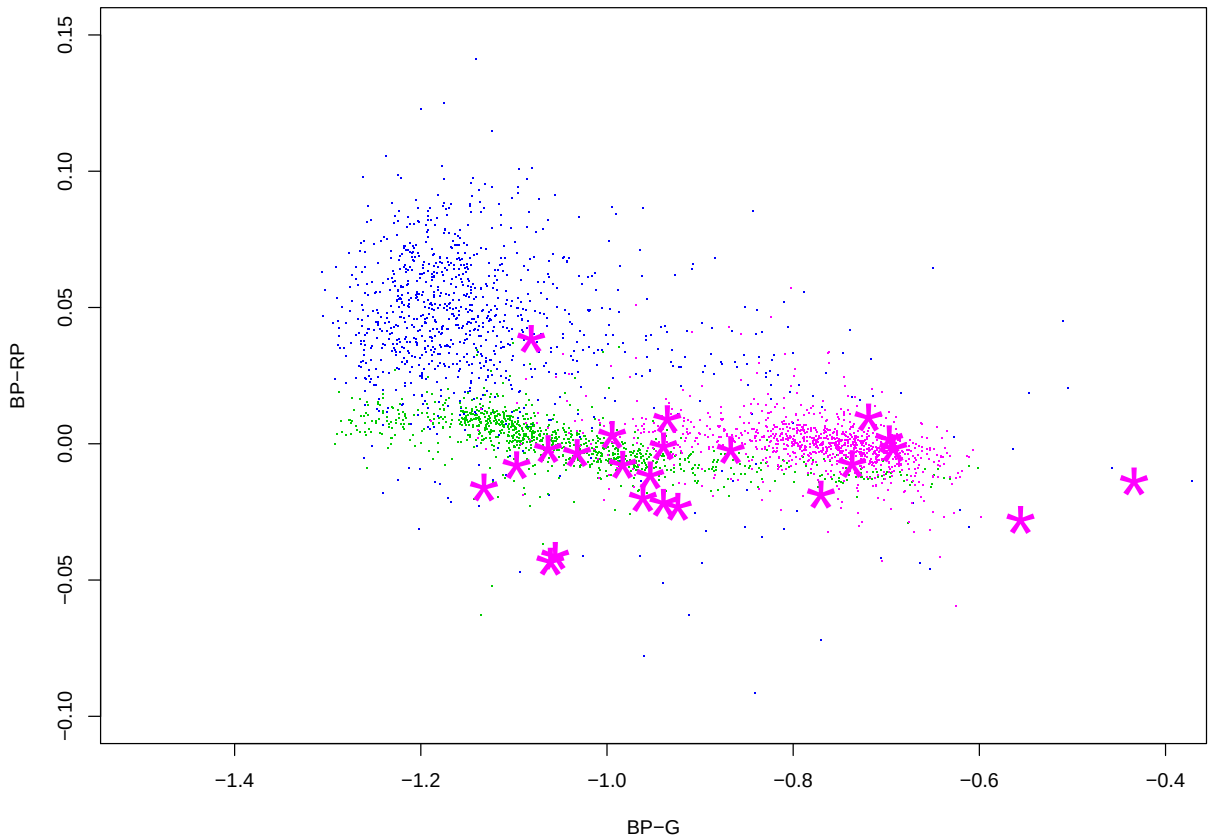
Colours



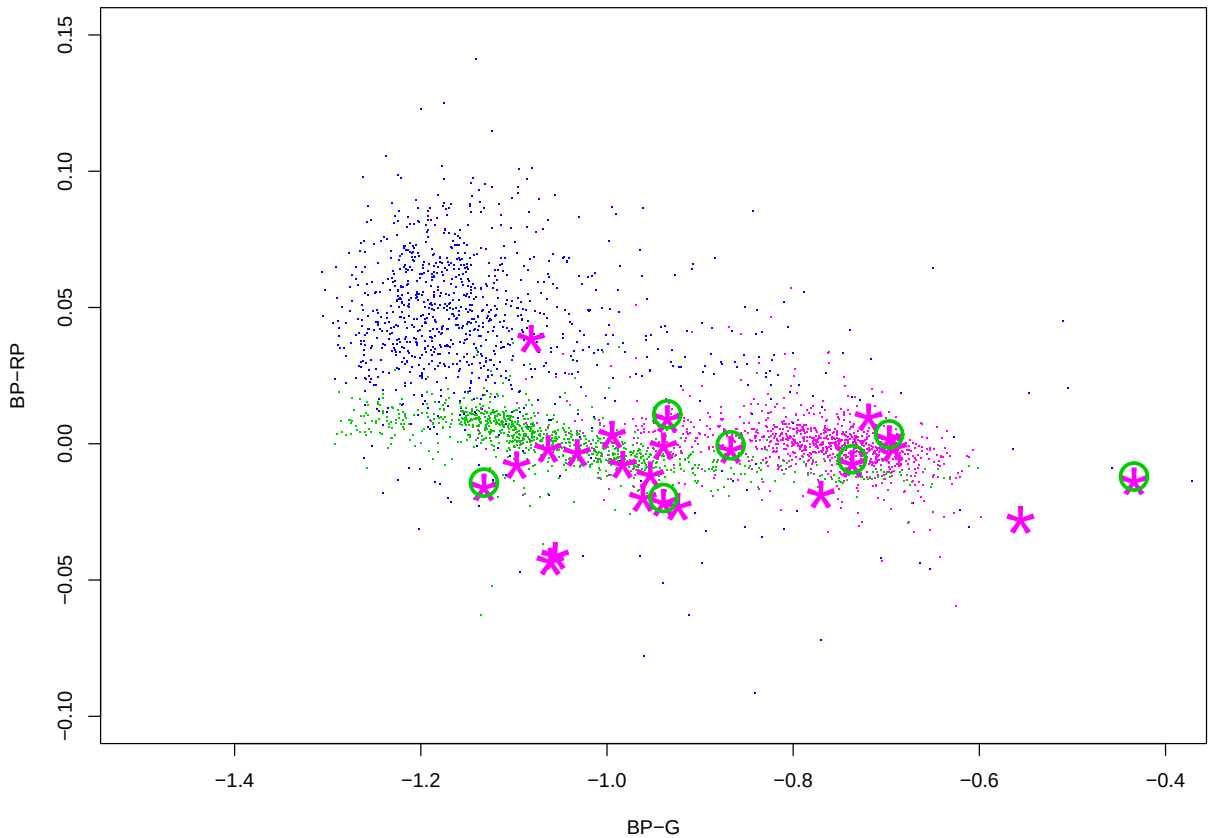
Colours



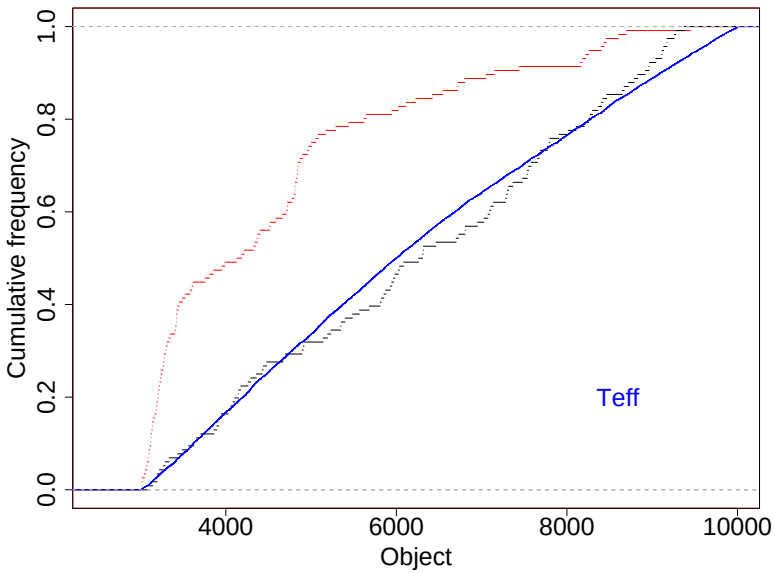
Colours

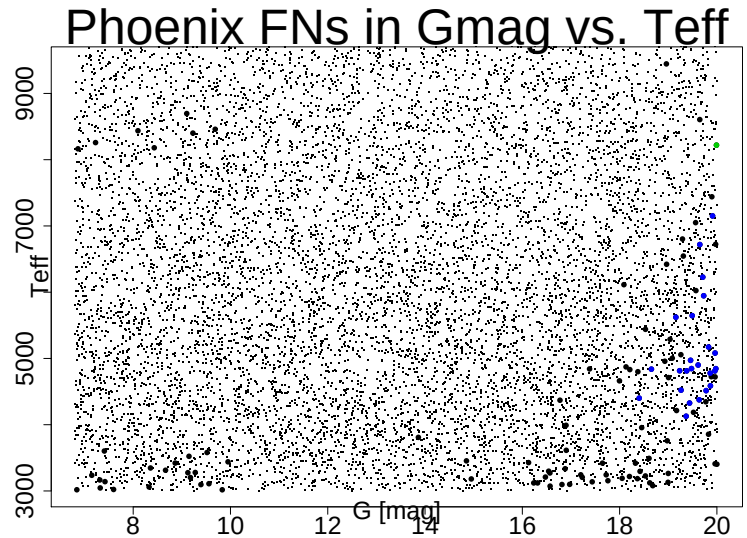


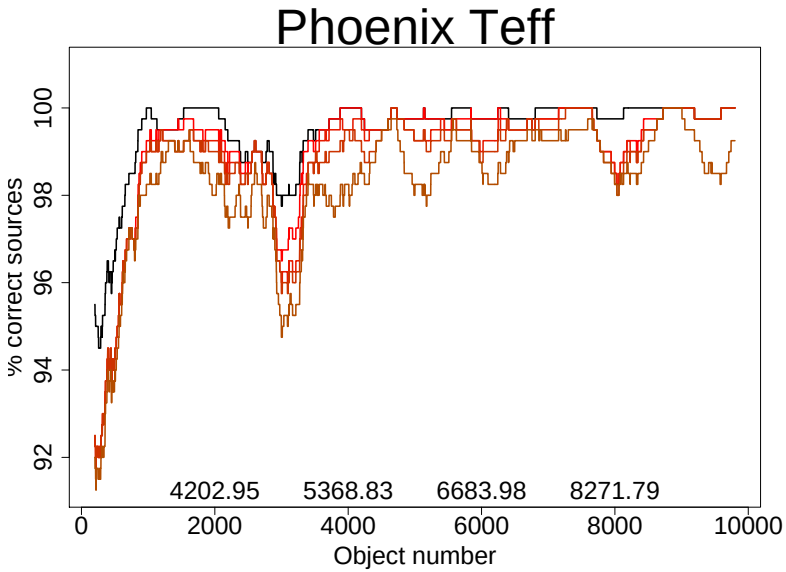
Colours

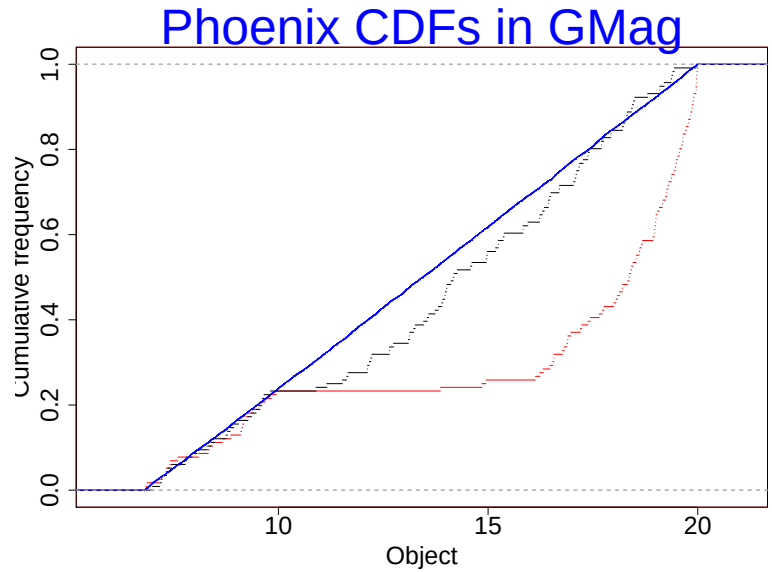


Phoenix results

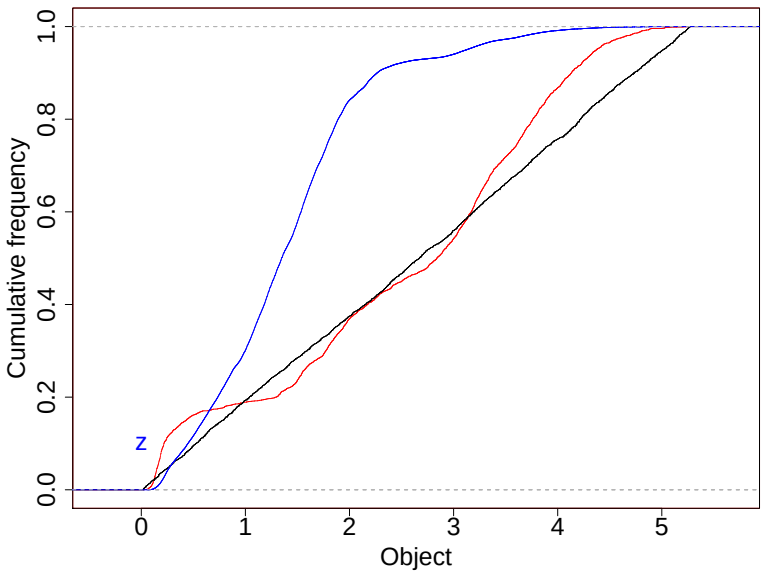


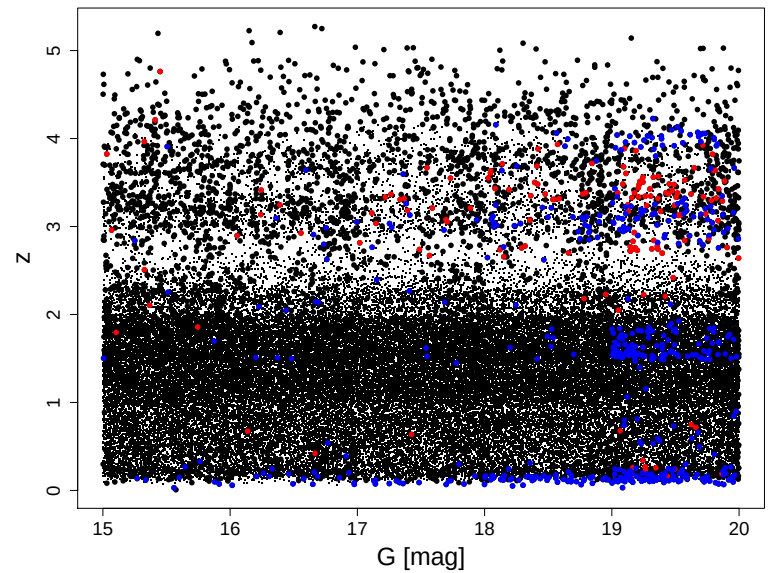


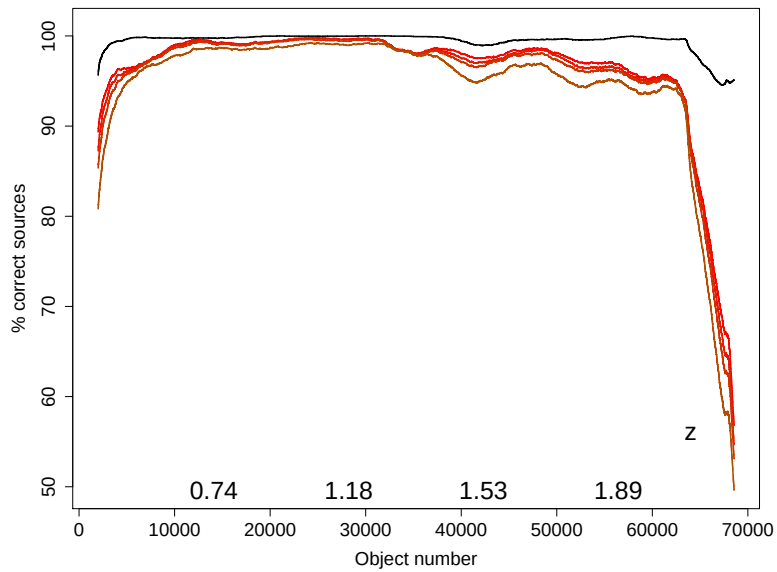


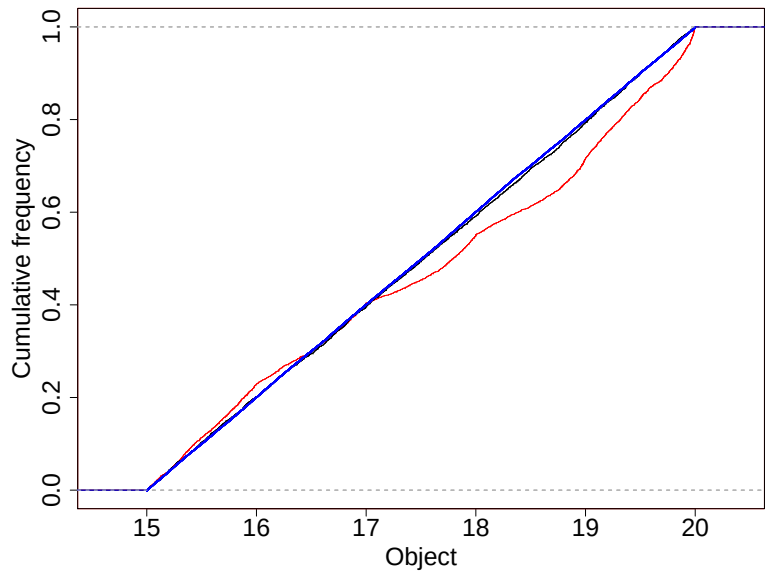


QSOs

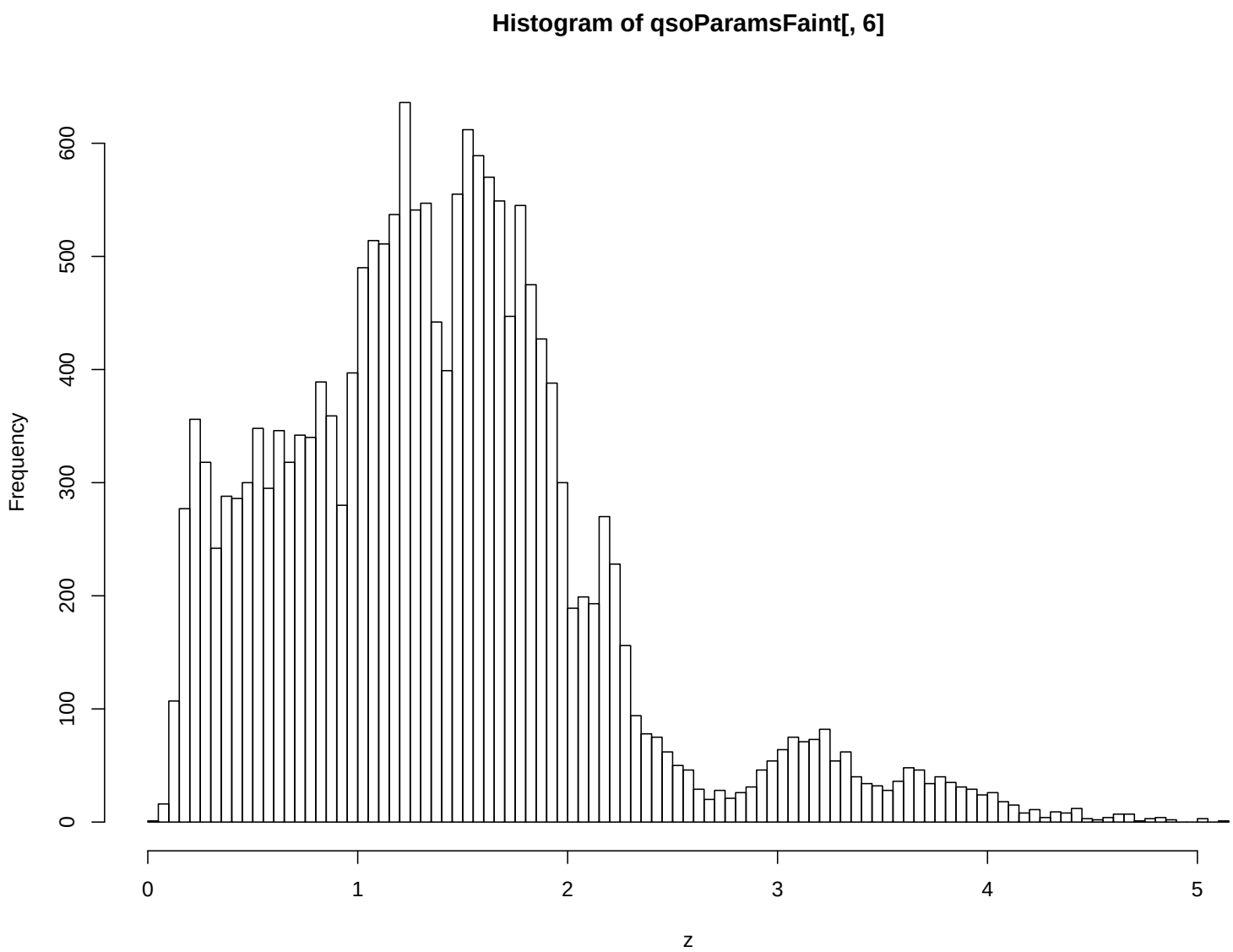

Cumulative frequency
0.0
0.2
0.4
0.6
0.8
1.0
z
0
1
2
3
4
5
Object


z
0
1
2
3
4
5
15
16
17
18
19
20
G [mag]

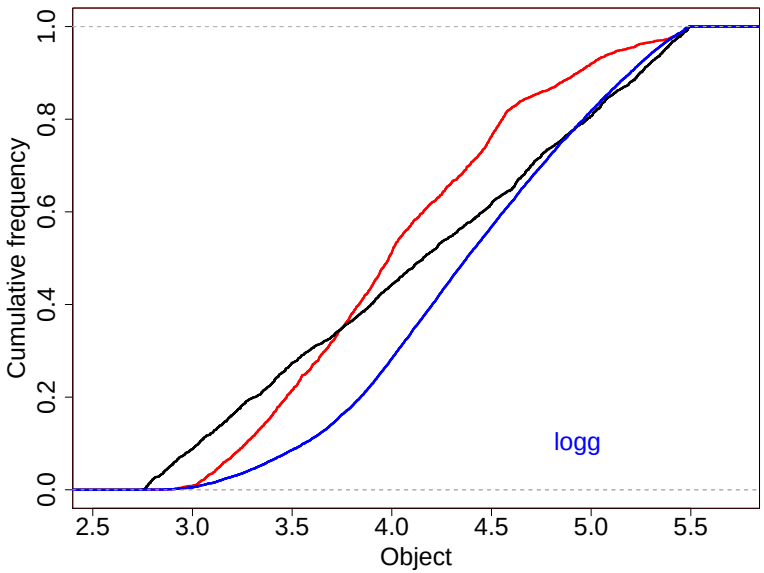

% correct sources
50
60
70
80
90
100
z
0.74
1.18
1.53
1.89
0
10000
20000
30000
40000
50000
60000
70000
Object number

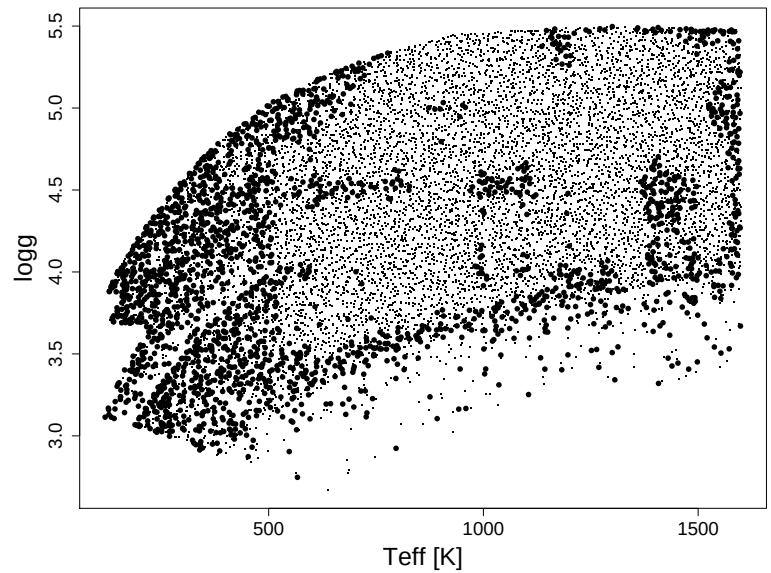

Cumulative frequency
0.0
0.2
0.4
0.6
0.8
1.0
15
16
17
18
19
20
Object

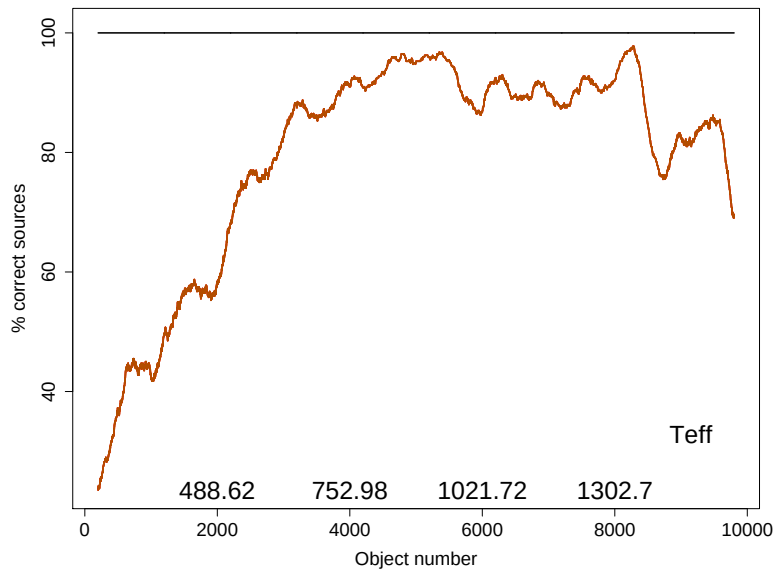
Qso z distribution

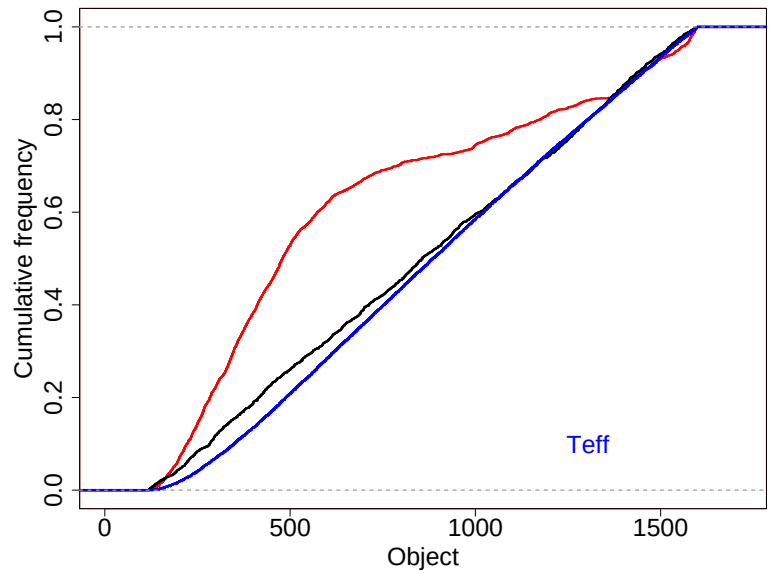


Ultra-cool dwarfs









Summary

- Gaia data analysis relies heavily on Machine learning algorithms
- These can provide powerful tools for analyzing large data sets (data mining)
- Increasingly used in astronomy, with larger surveys
- But still probably under-utilized

Exercise

- simulated Gaia data set with several classes
- Use R package, which implements many ML algorithms
- Exercise 1: Choose a ML algorithm and classify the sources
- Exercise 2: Choose a ML algorithm and parameterize a class
- Exercise 3: Choose an unsupervised algorithm to cluster the data

Some simple R scripts are provided as a starting point, including a knn classifier and parameterizer.

Algorithms

Algorithm	Classification	Regression	Supervised	Unsupervised
kNN	✓	✓	✓	X
SVM	✓	✓	✓	X
SoM	✓	✓	✓	✓
kMeans	✓	✓	✓	✓
Neural Network	✓	✓	✓	X
PCA+KDE	✓	X	✓	X
Tree methods	✓	✓	✓	✓
Boosting	✓	✓	✓	✓
...				