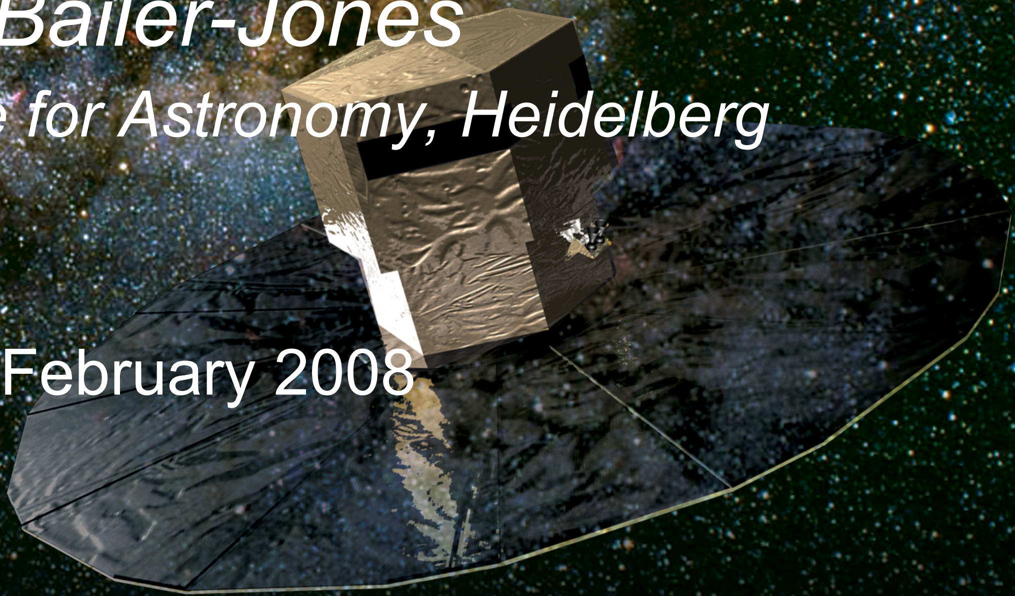
Classification and parameter estimation in astronomy

Coryn Bailer-Jones

Max Planck Institute for Astronomy, Heidelberg



UCL, 4 February 2008

Overview

- learning from astronomical data
- Gaia Galactic survey mission
- object classification
- astrophysical parameter estimation
- optimizing the survey: heuristic filter design
- summary

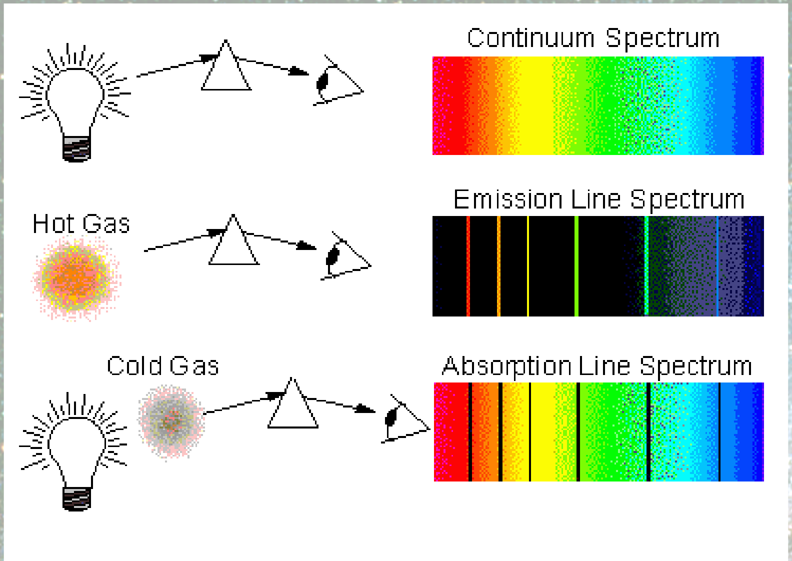
http://www.mpia.de/homes/calj

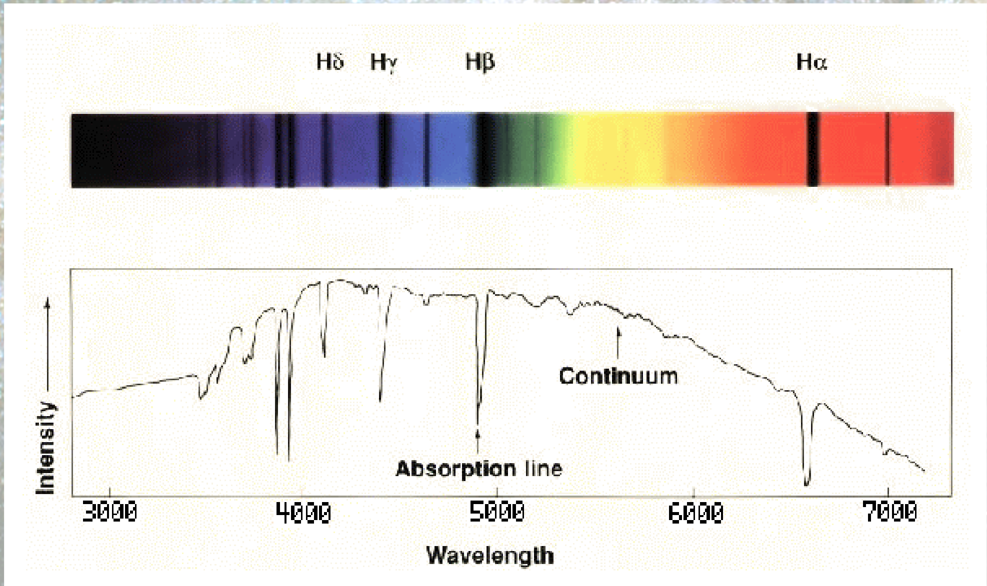
Acknowledgements

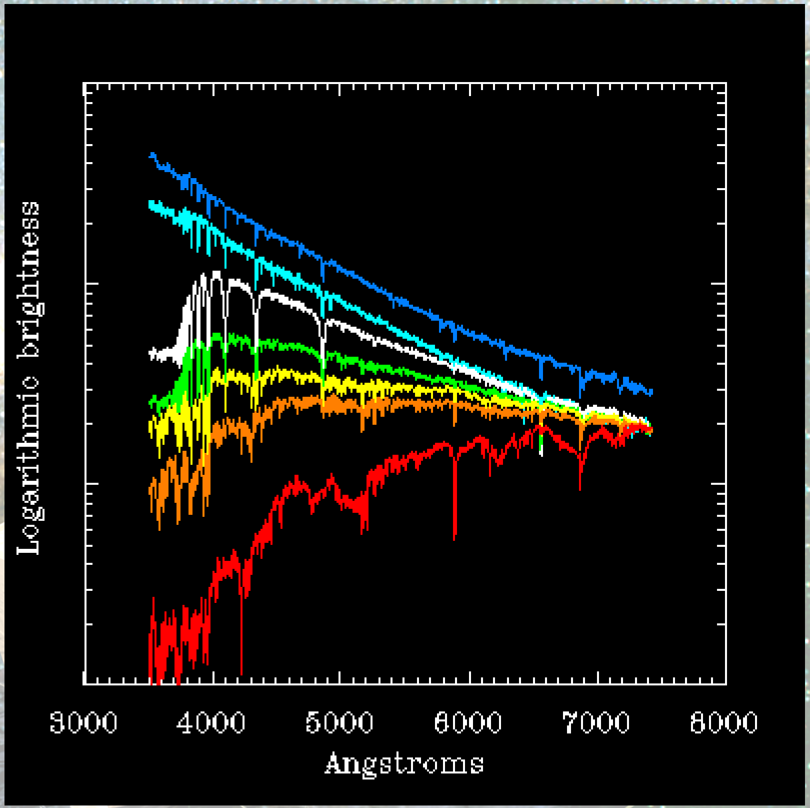
- Gaia @ MPIA group
  - Christian Elting
  - Paola Re Fiorentin
  - Kester Smith
  - Carola Tiede
  - Vivi Tsalmantza

- More information on Gaia:
  http://sci.esa.int/Gaia

Astronomical spectra





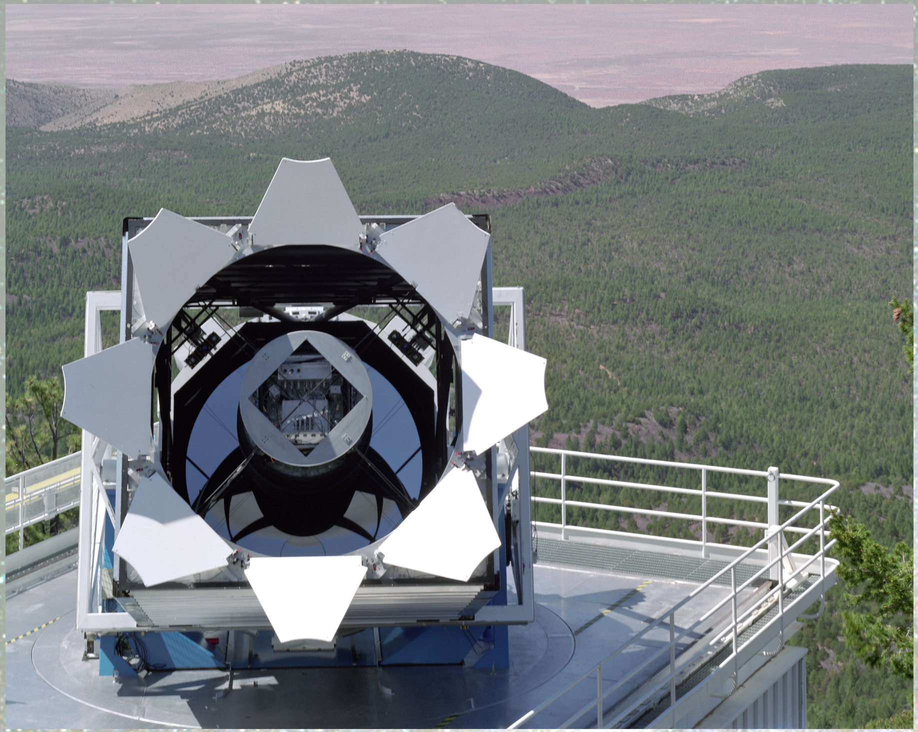


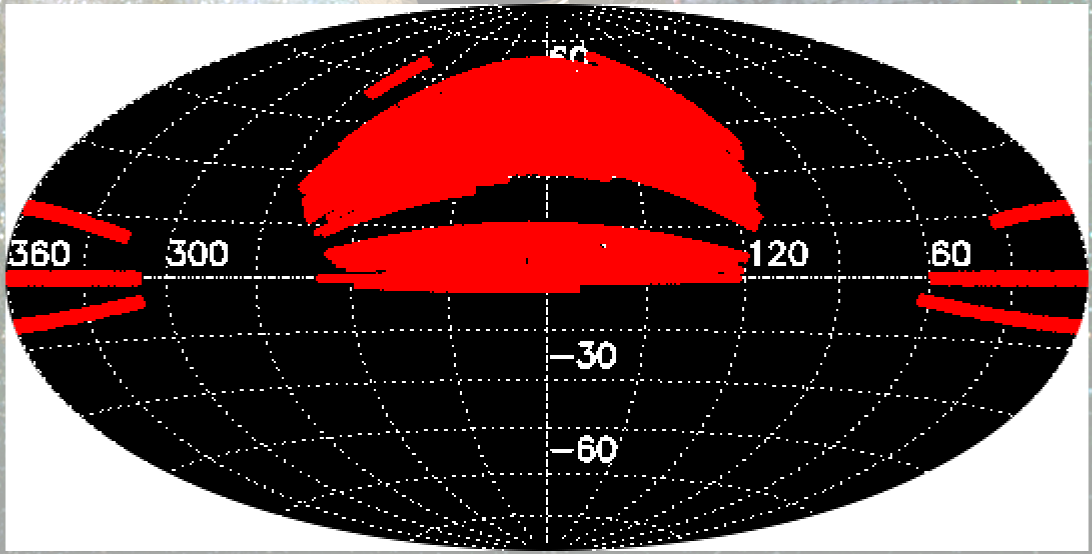
images from G. Smith, UCSD and J. Dalcanton, Washington

What can we learn from spectra?

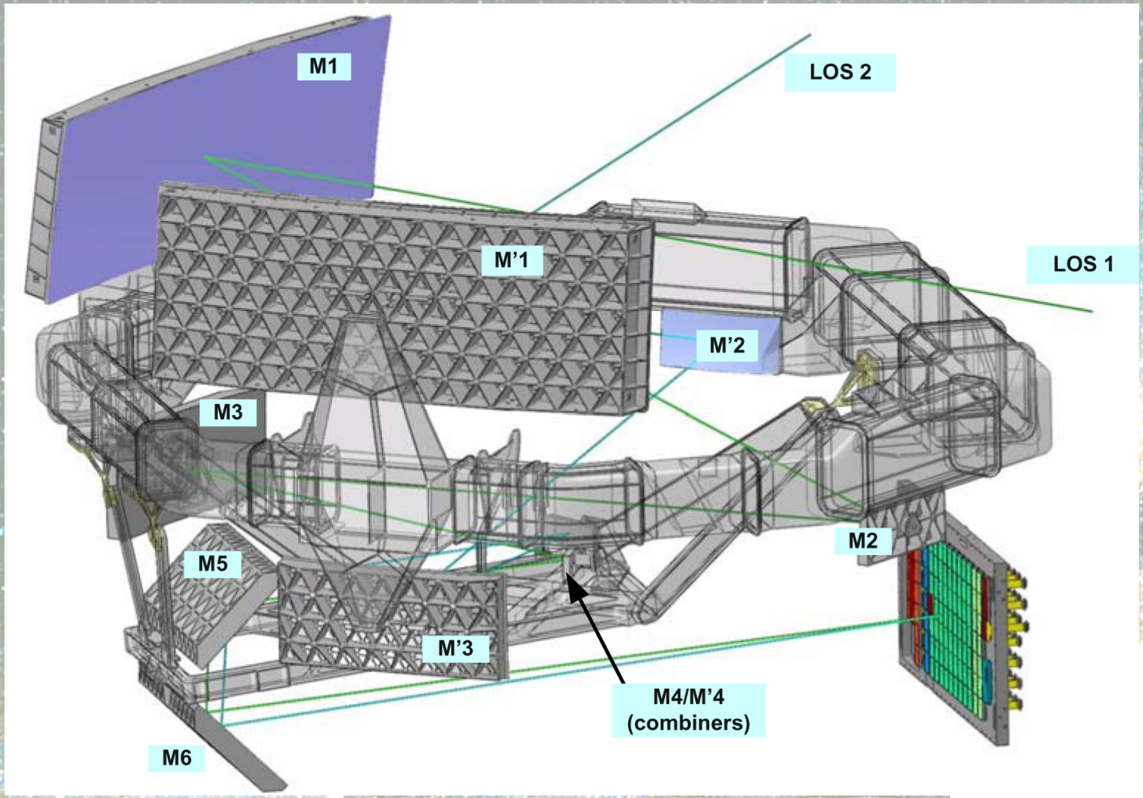
- object type (classification)
  - overall spectral shape
  - characteristic features
- stars
  - atmospheric temperature, pressure, abundances
  - derive mass, age, radius (via evolutionary models)
  - interstellar extinction (along line-of-sight)
- galaxies
  - stellar and gas composition of galaxy
  - infer evolution
- quasars
  - redshift (cosmological distance)
  - use as background lights to study cosmology

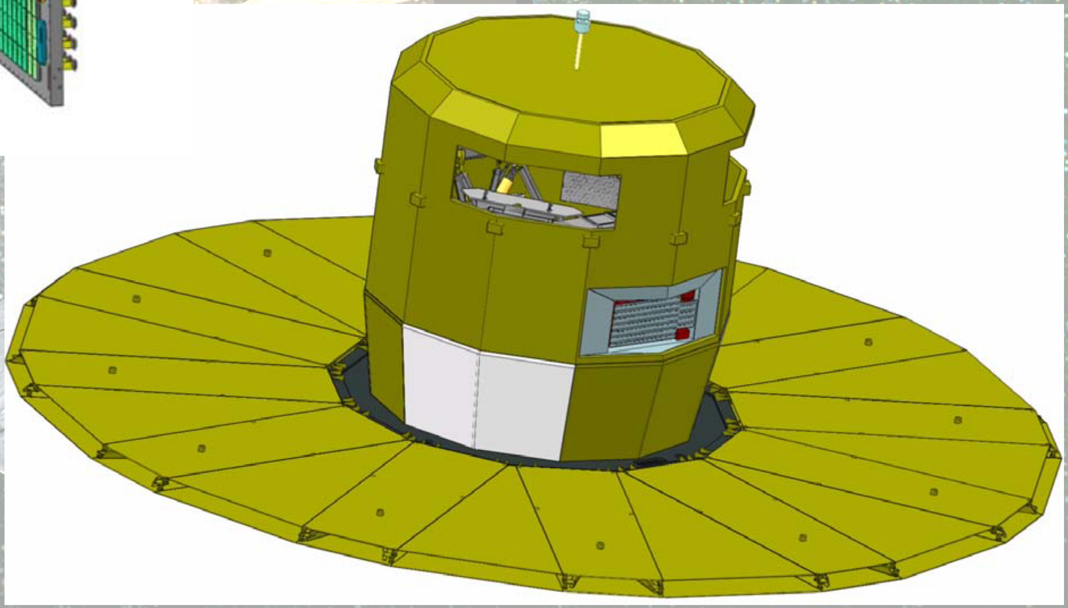
Astronomical surveys





The Gaia Galactic survey mission



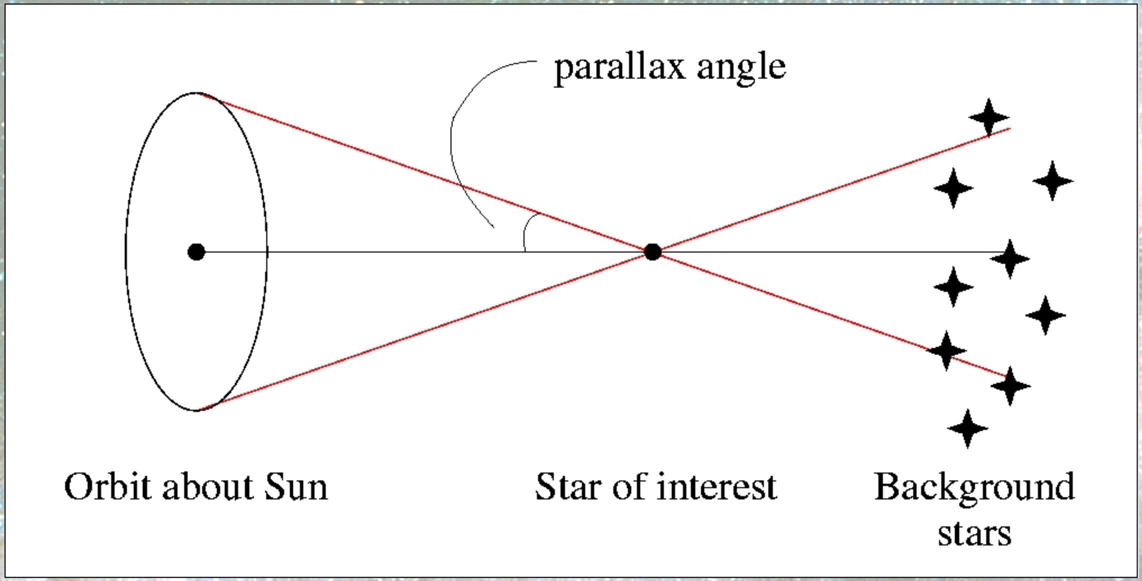


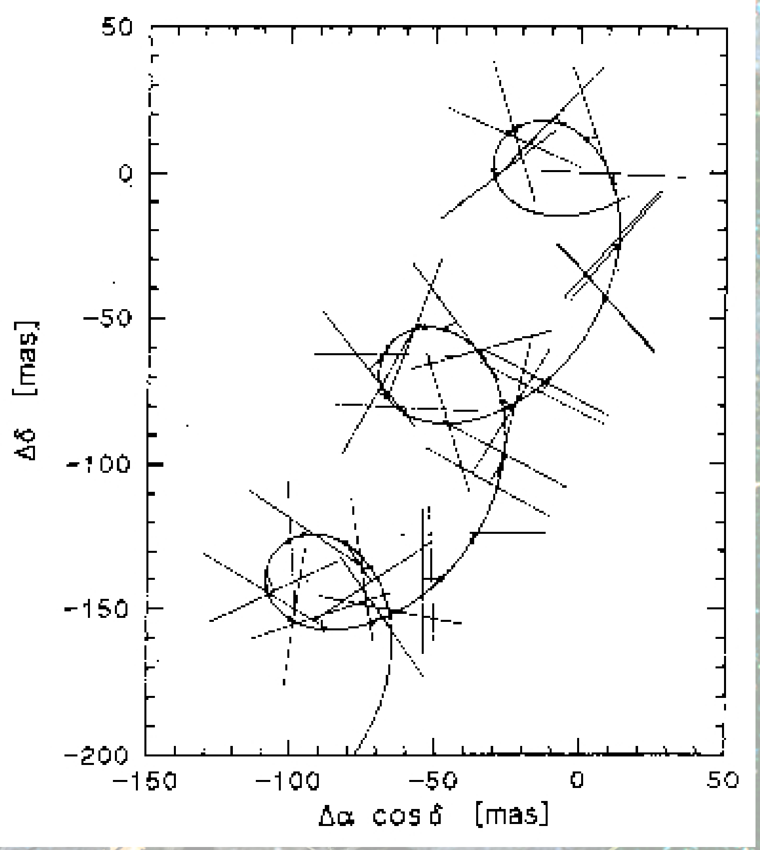
images © EADS Astrium

The Gaia Galactic survey mission

- five-year all sky survey of one billion objects
- detect objects in real time and measure
  - distances and space velocities
  - spectra (to determine physical properties)
- Science goals
  - structure, origin and evolution of our Galaxy
  - nature of dark matter
  - better understand stars and star clusters
  - find planets outside of the solar system
- strength of Gaia is both accuracy and large number statistics
  - distances to accuracy of 1% for over 10 million stars

Measuring distances via parallax

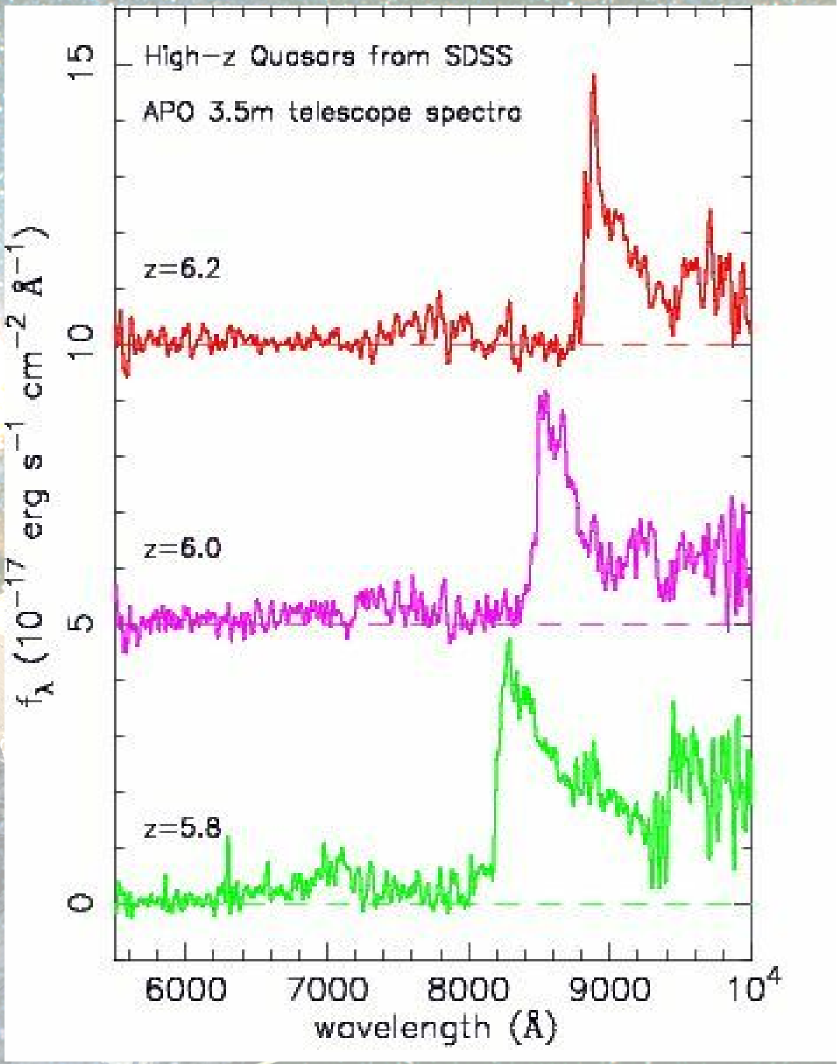




parallax measurements are relative to reference (background) objects

Quasar identification

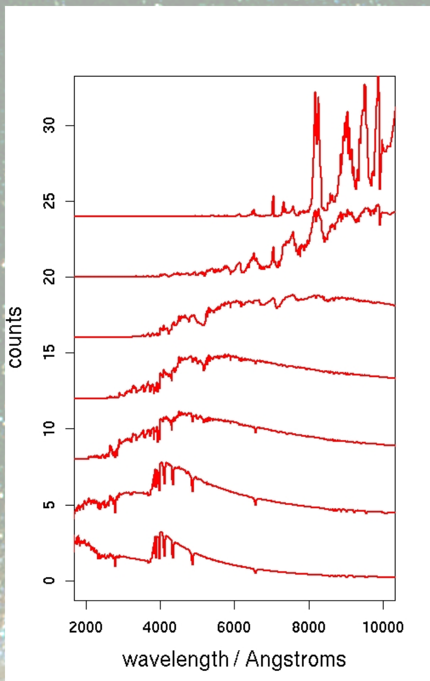
- extragalactic objects at large distances
- used to define an inertial reference frame ("fixed background")
- characteristic broad emission lines
- position of lines in spectrum varies due to redshift



The classification problem

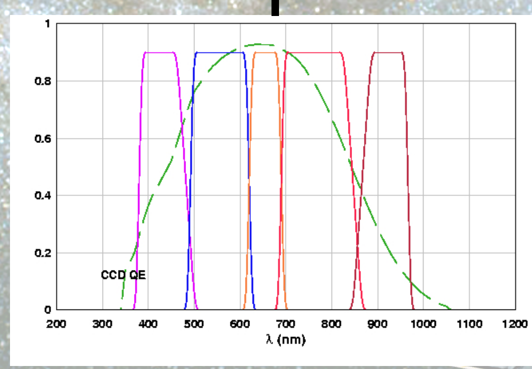
- “blind” survey
- object (spectrum) classification via pattern recognition
- four class problem
  - star, quasar, galaxy, physical binary star
- support vector machine classifier (libsvm)
  - RBF kernel
  - multiclass using the one-against-one method
- 8000 objects (~2000 of each class) in training set, same for test

Data simulation

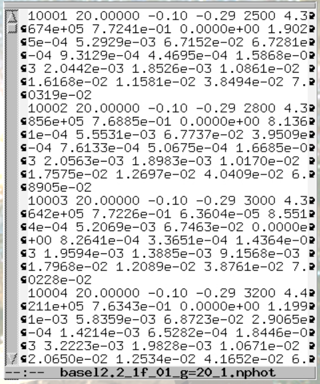


synthetic spectra

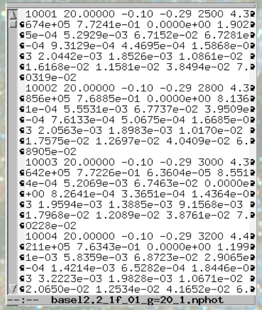
instrument simulator including noise



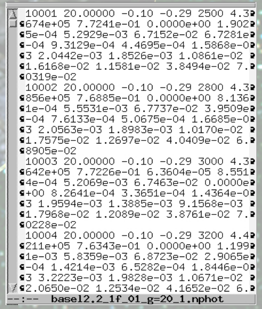
instrument model



fluxes and estimated covariances



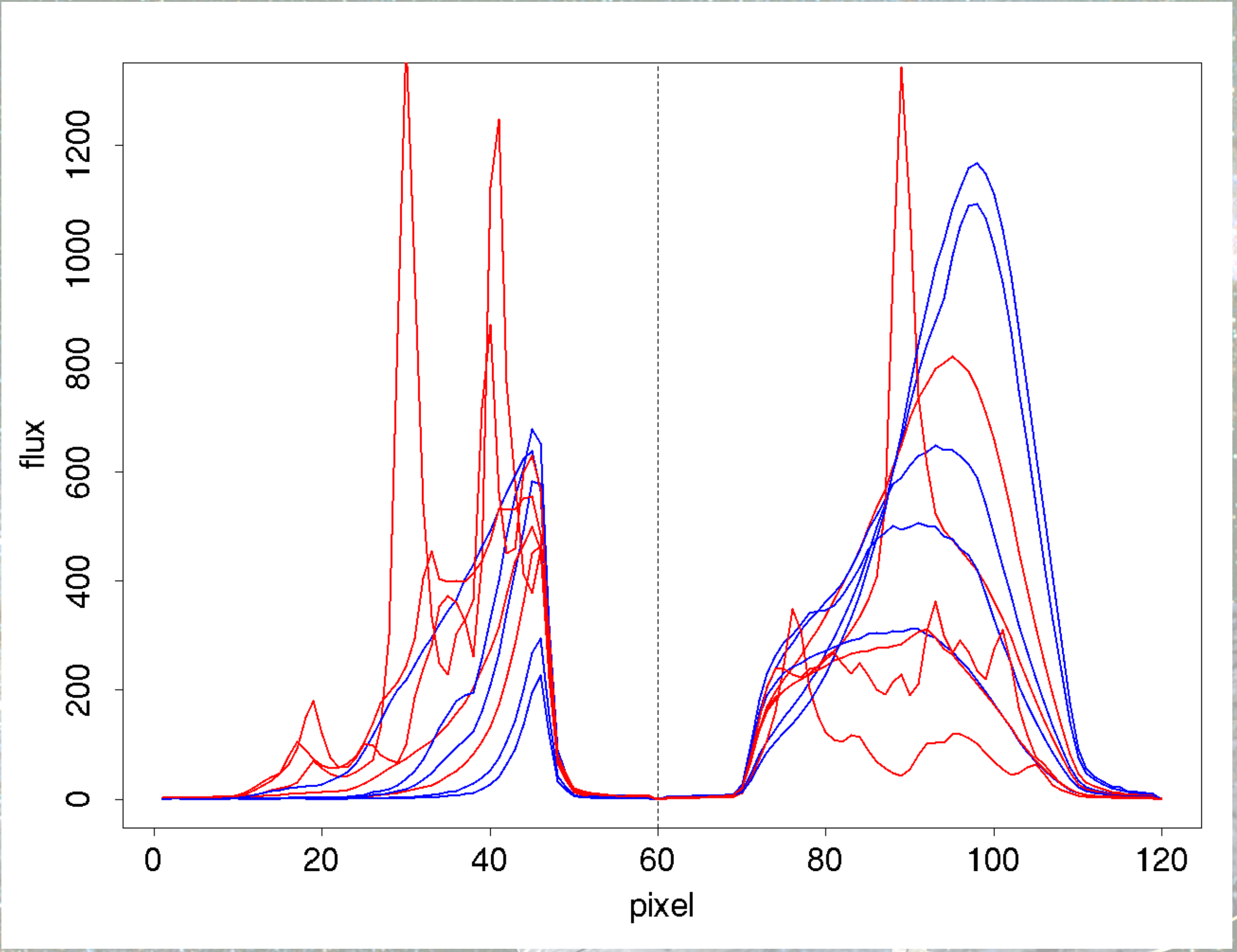
training set



test set

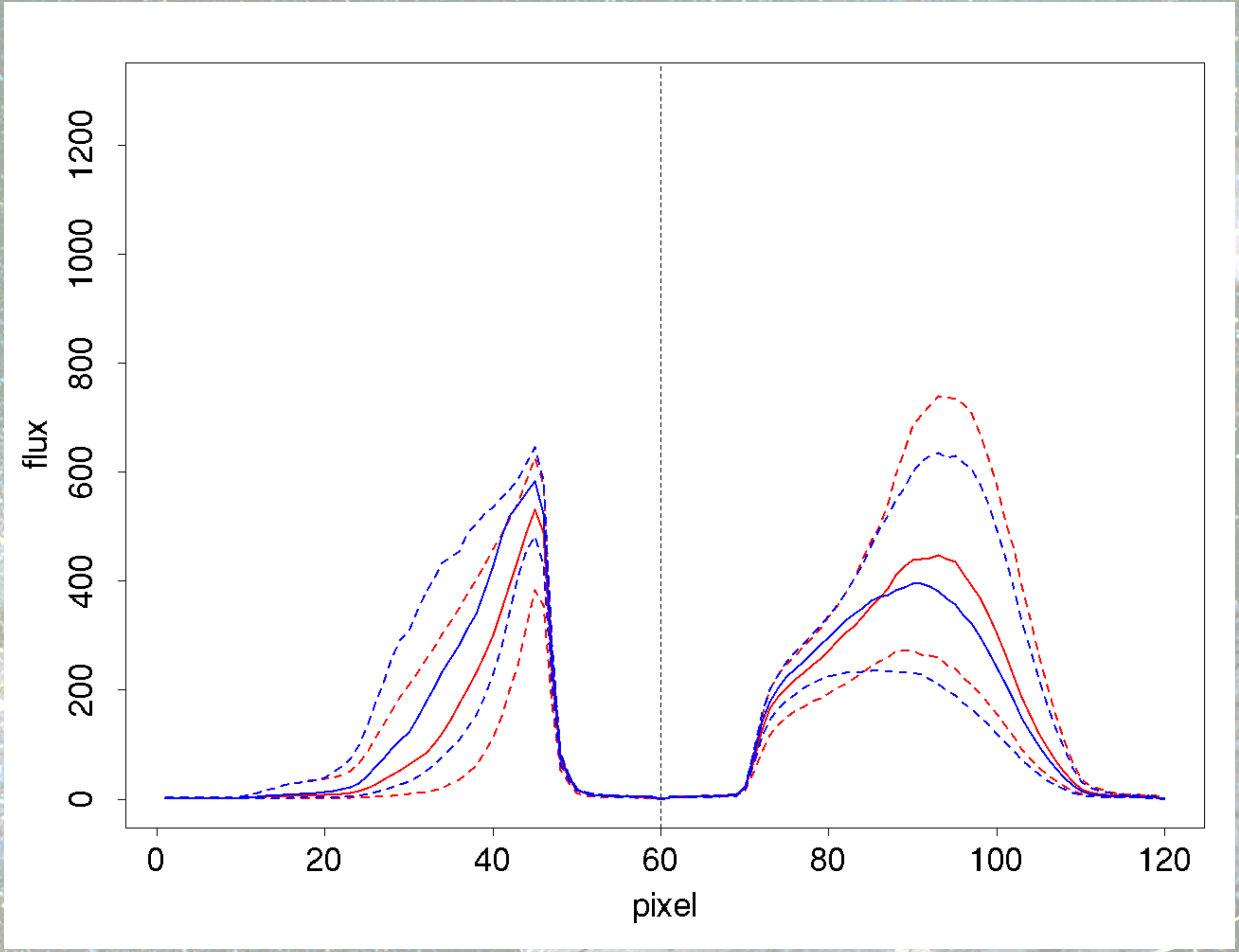
Coryn Bailer-Jones, MPIA, Heidelberg

Example Gaia spectra



blue: star red: quasar

Gaia spectral data



blue: star   red: quasar

solid: median
dashed: upper and lower quartiles

Coryn Bailer-Jones, MPIA, Heidelberg

Classification results: contingency table

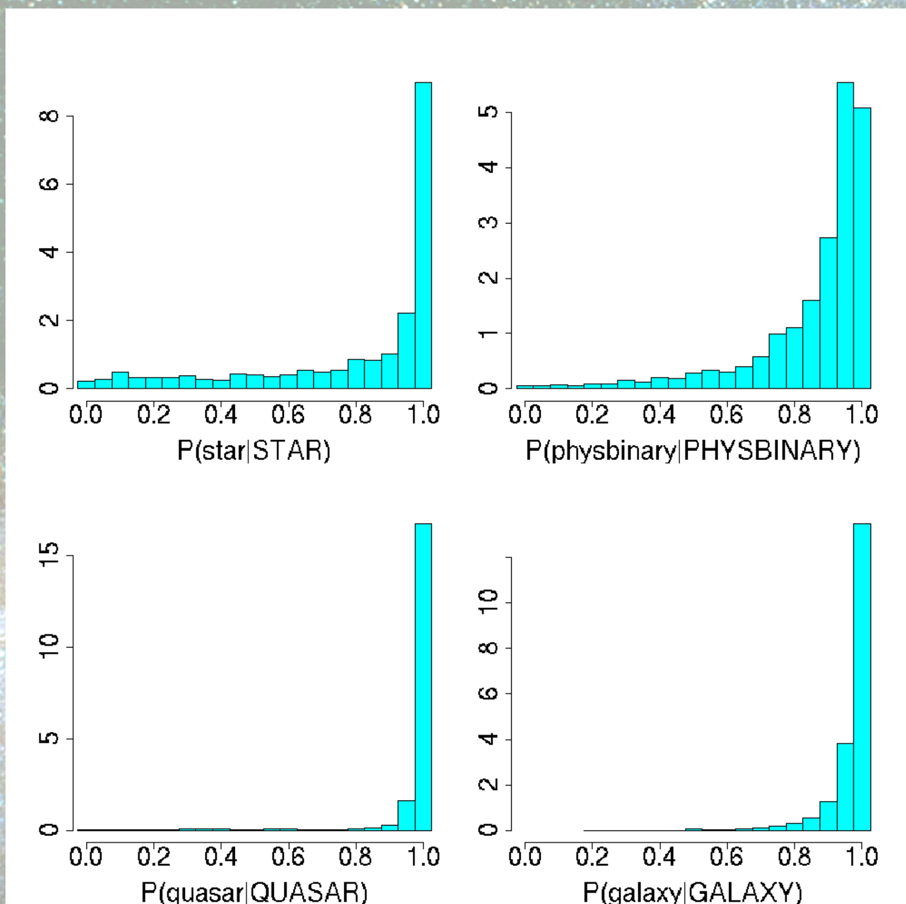
	galaxy	physbin	quasar	star
GALAXY	99.65	0.00	0.35	0.00
PHYSBIN	0.00	94.38	0.15	5.46
QUASAR	2.93	0.05	96.81	0.21
STAR	0.10	13.99	3.16	82.75

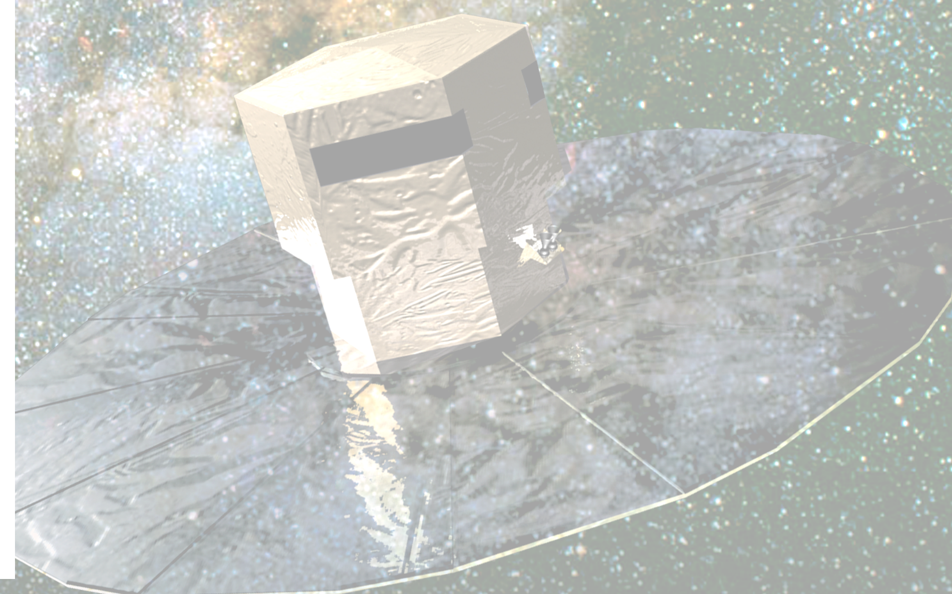
- stars and physbin (physical binaries) are hard to distinguish, even in principle
- use output probabilities to vary *completeness* and *contamination* of a sample

Classifier confidence

how well does it find the correct objects?

P(class | CLASS)





classifier output probability →

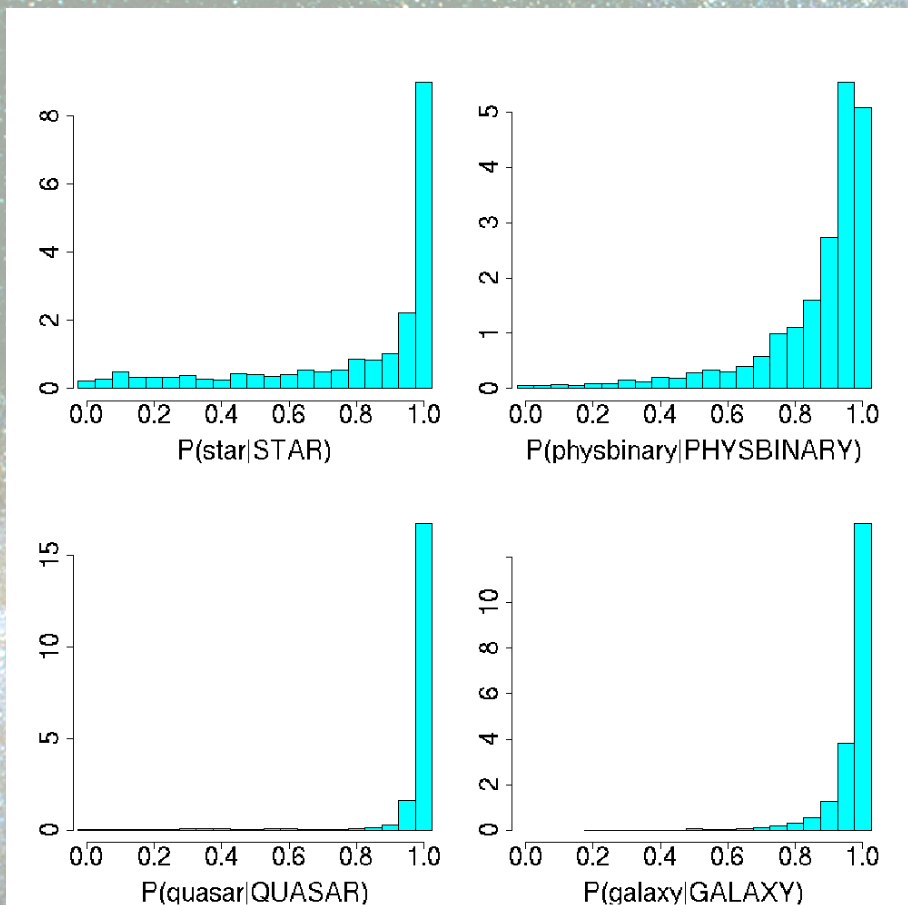
Classifier confidence

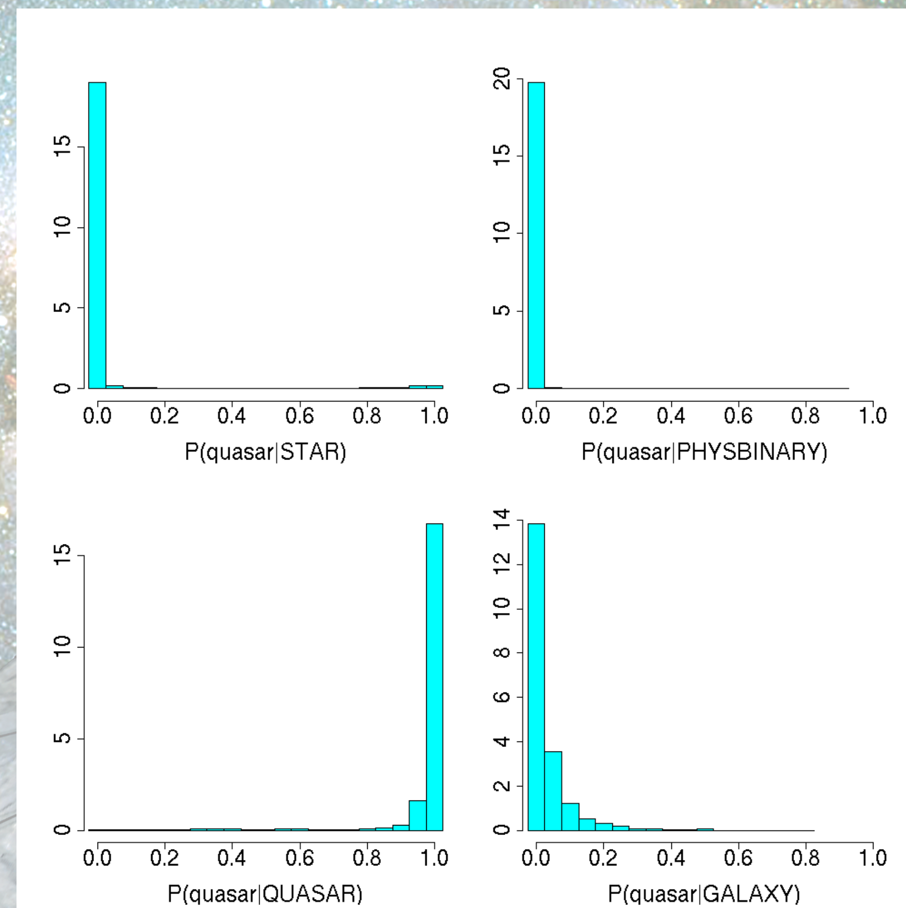
how well does it find the correct objects?

P(class | CLASS)

how much many false positives in the quasar sample?

P(quasar | CLASS)



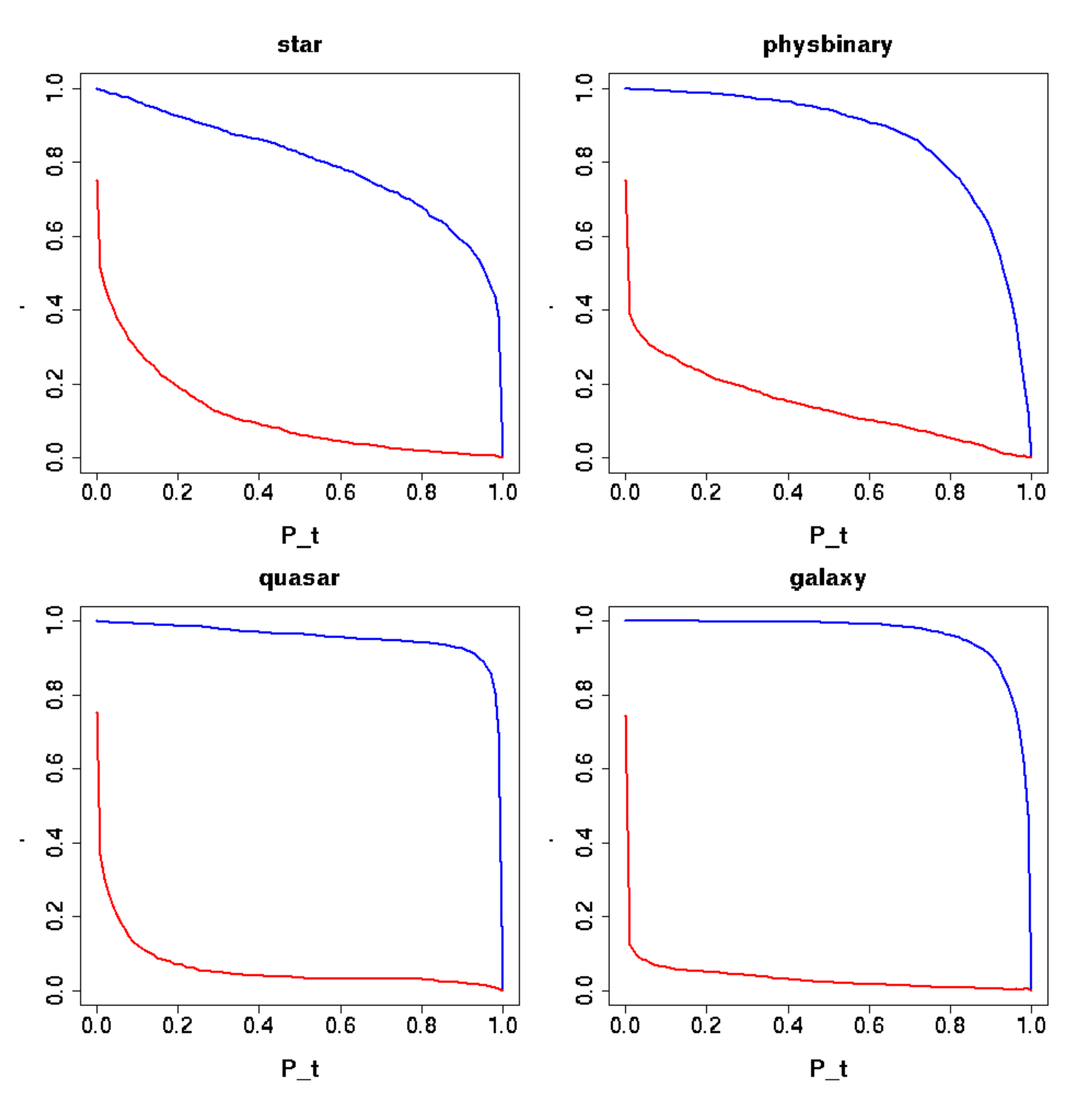


classifier output probability →

Sample completeness & contamination

select threshold to build a sample

blue = completeness
red = contamination

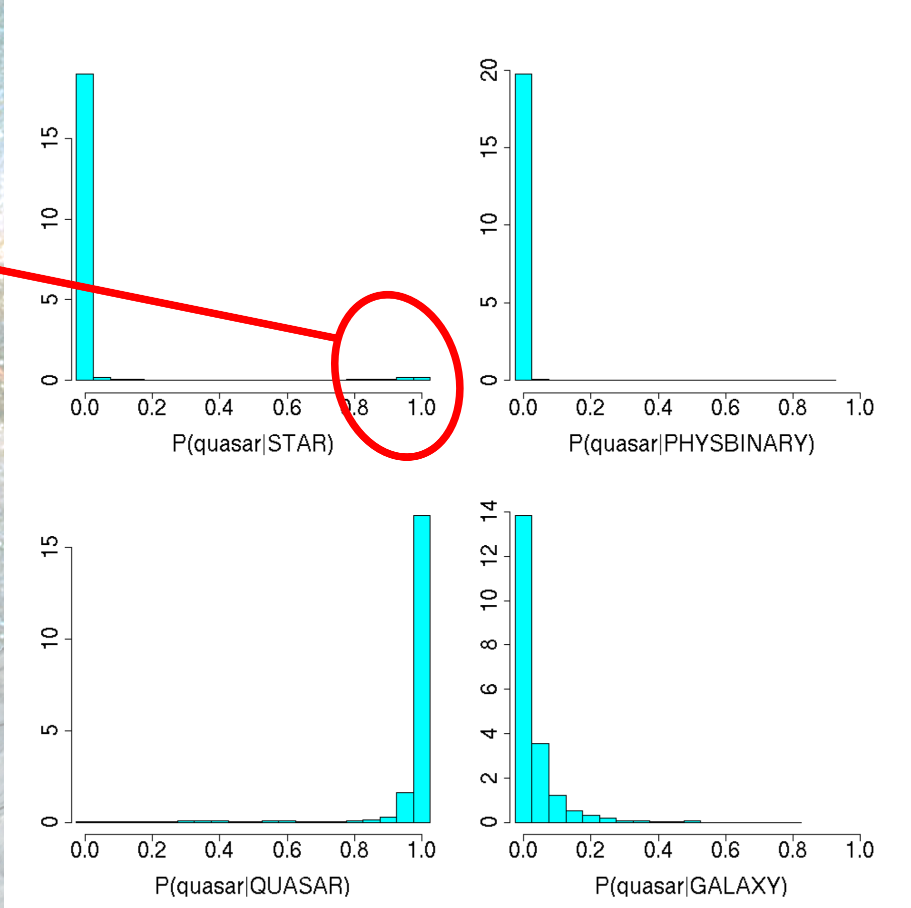


threshold probability →

Classifier confidence

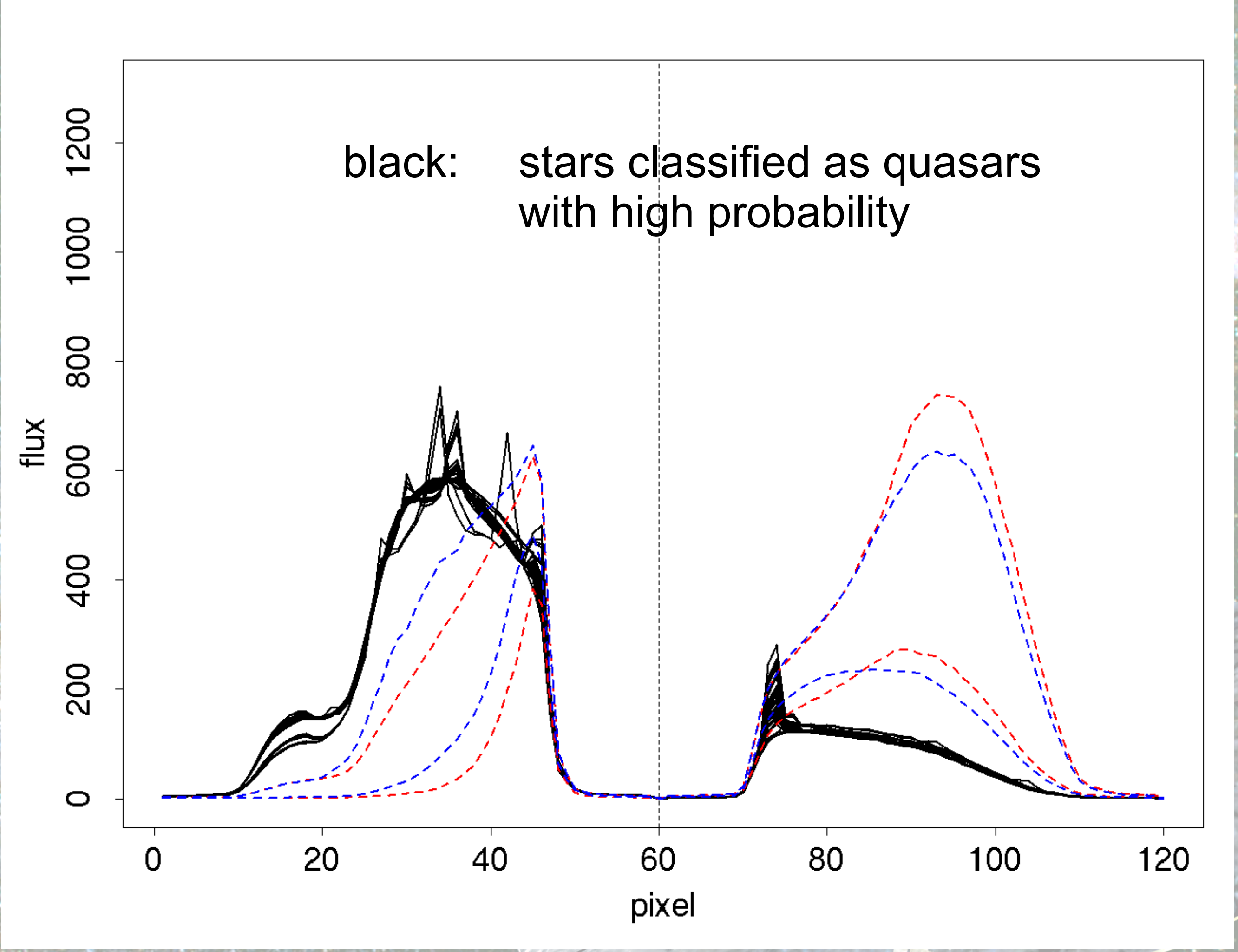
P(quasar | CLASS)

small but high confidence contamination of quasars by stars



classifier output probability

What is contaminating the quasar sample?

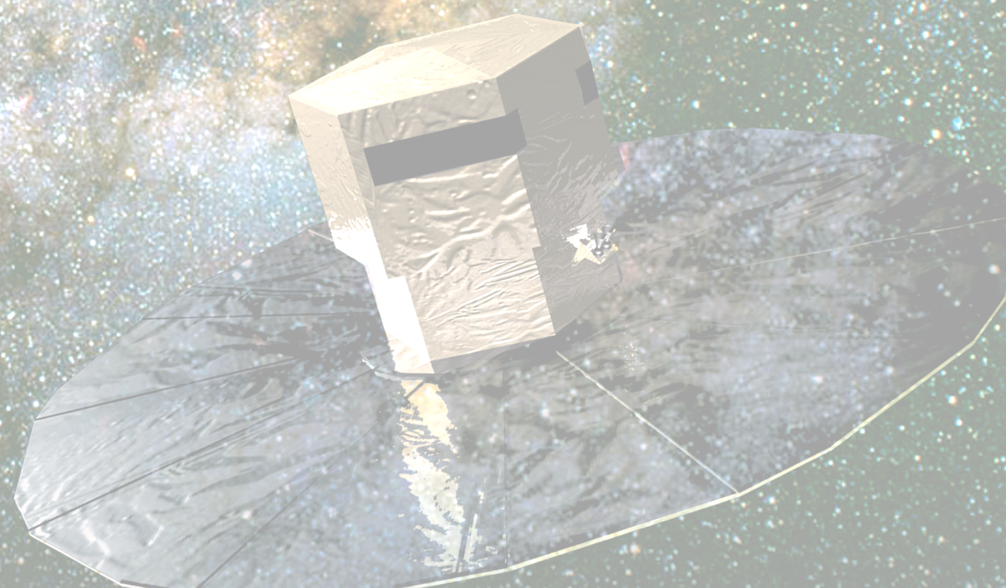


blue: star red: quasar

solid: median
dashed: upper and lower quartiles

Meet reality: population fractions

- in reality quasars are rare compared to stars (ca. 1000 times)
- don't modify training sample distribution/re-train
- instead
  - modify priors
  - treat SVM outputs as likelihoods



Meet reality: population fractions

- in reality quasars are rare compared to stars (ca. 1000 times)
- don't modify training sample distribution/re-train
- instead
  - modify priors
  - treat SVM outputs as likelihoods

$$P^{mod}(i) \propto P^{nom}(i) \frac{f_j^{mod}}{f_j^{nom}}$$

$P(i) =$ posterior probabilty that object is in class i

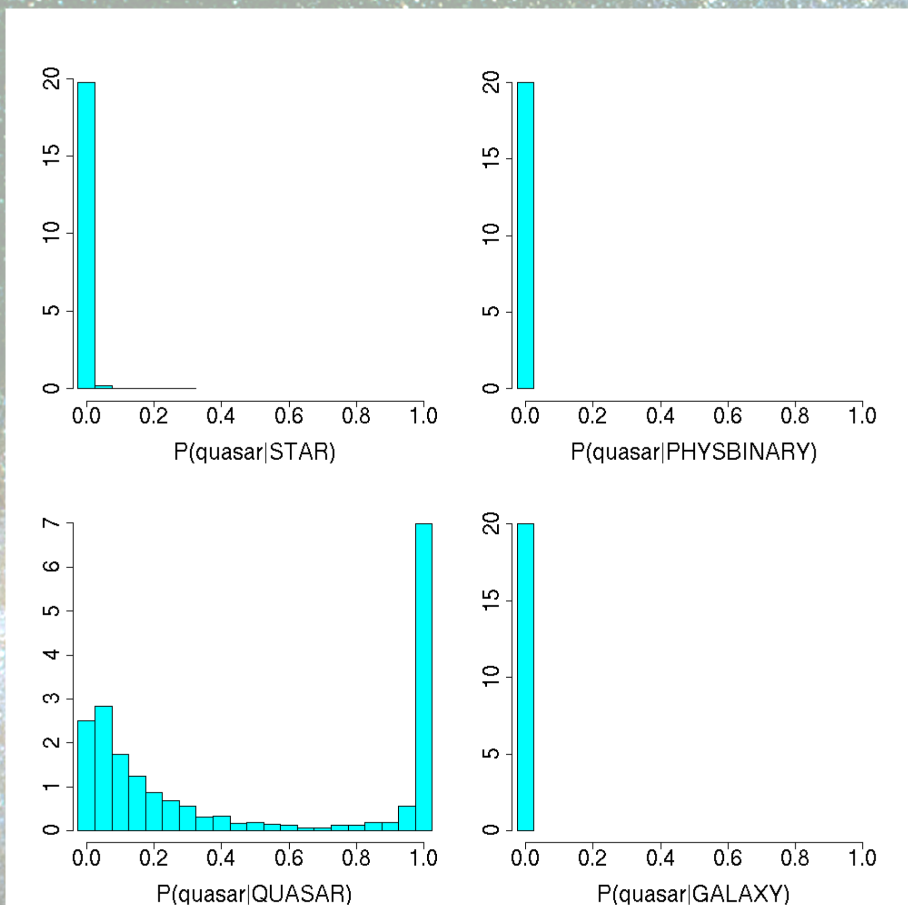
$f_j =$ prior fraction of objects in true class j

$mod =$ modified, $nom =$ nominal

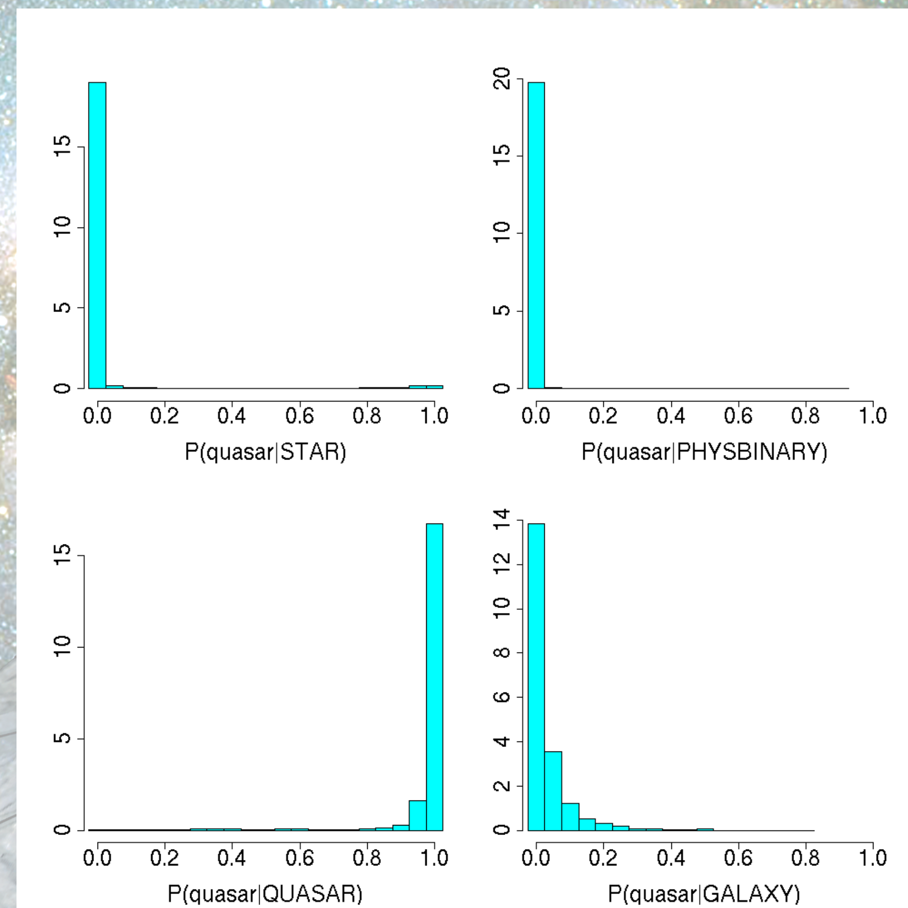
Classifier confidence: nominal vs. modified

P(quasar | CLASS)

modified population



nominal population



classifier output probability →

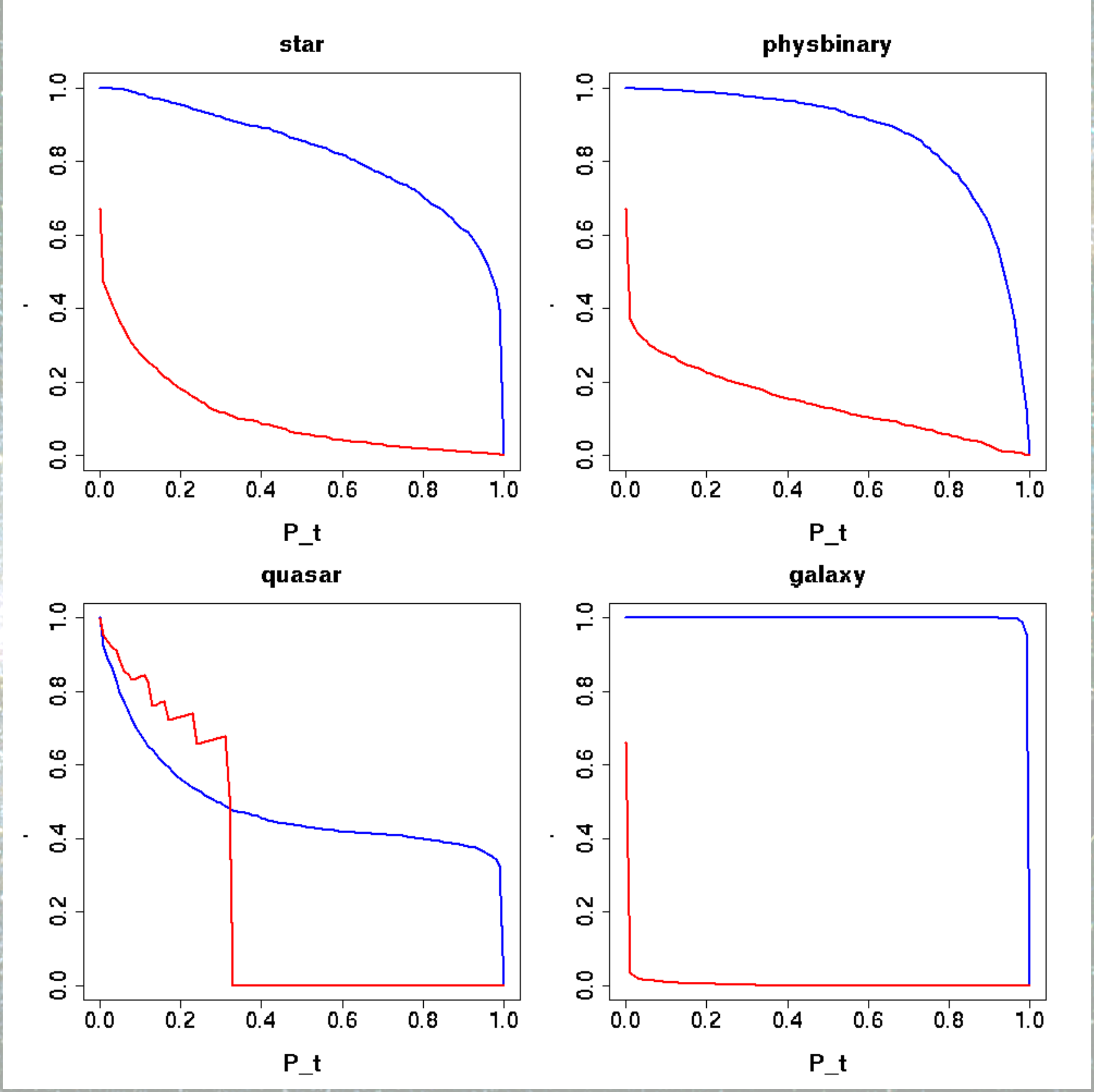
Modified population fractions (priors)

blue = completeness

red = contamination

relative fractions:

star	1
physbinary	1
quasar	0.001
galaxy	1



threshold probability

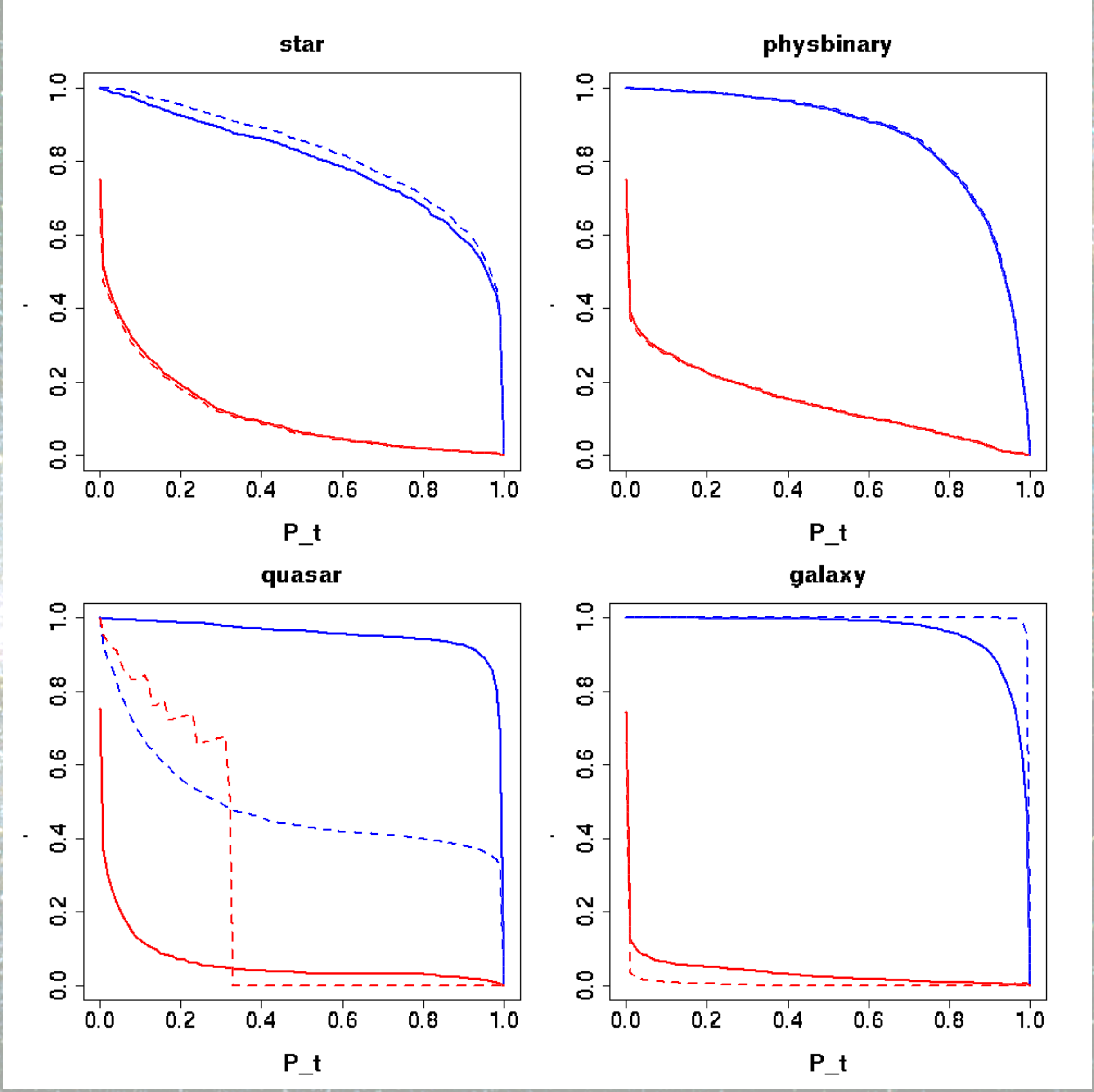
Modified population fractions (priors)

blue = completeness

red = contamination

solid = nominal fractions (equal)

dashed = modified fractions (quasars 1000 times rarer)



threshold probability →

Gaia DPAC

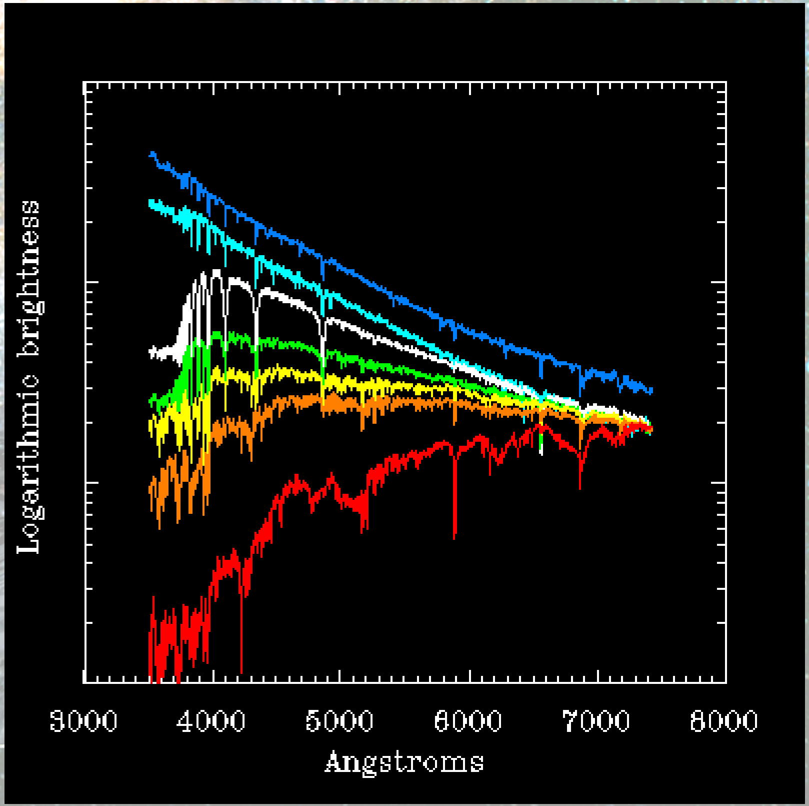
Comparison of different ML models

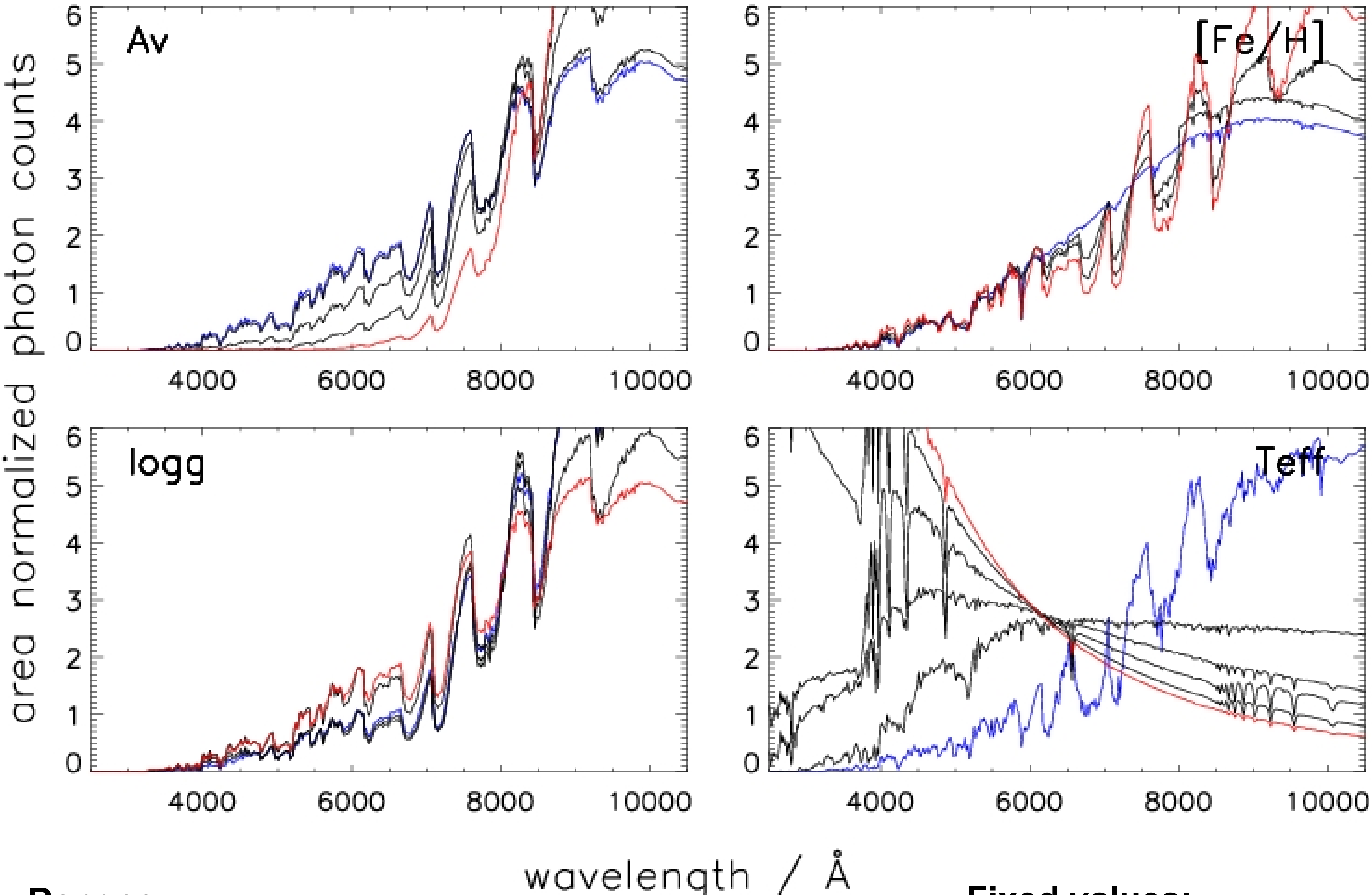
	spectra only		spectra + astrometry	
	K=4	K=3	K=4	K=3
SVM	10.4	2.5	8.3	0.6
Boosting	37.8	33.9	13.0	1.5
MLP	9.4	1.8	7.5	0.2
Mclust	10.9	0.8	9.4	0.1
RBF	38.5	24.0	27.1	15.3

overall classification error (percent)
class assignment via highest probability
K=3: 3 classes only (star and physbin merged)

Estimating stellar parameters

- several astrophysical parameters (APs) of interest
  - effective temperature, Teff
  - surface gravity, logg
  - abundance, [Fe/H]
  - interstellar extinction, Av
- different nature and degree of impact on the spectra





Ranges:

A_v = {0,10} [Fe/H] = {-2.5, 0.5}

logg = {0.5, 4.5} T_{eff} = {3350, 35000}

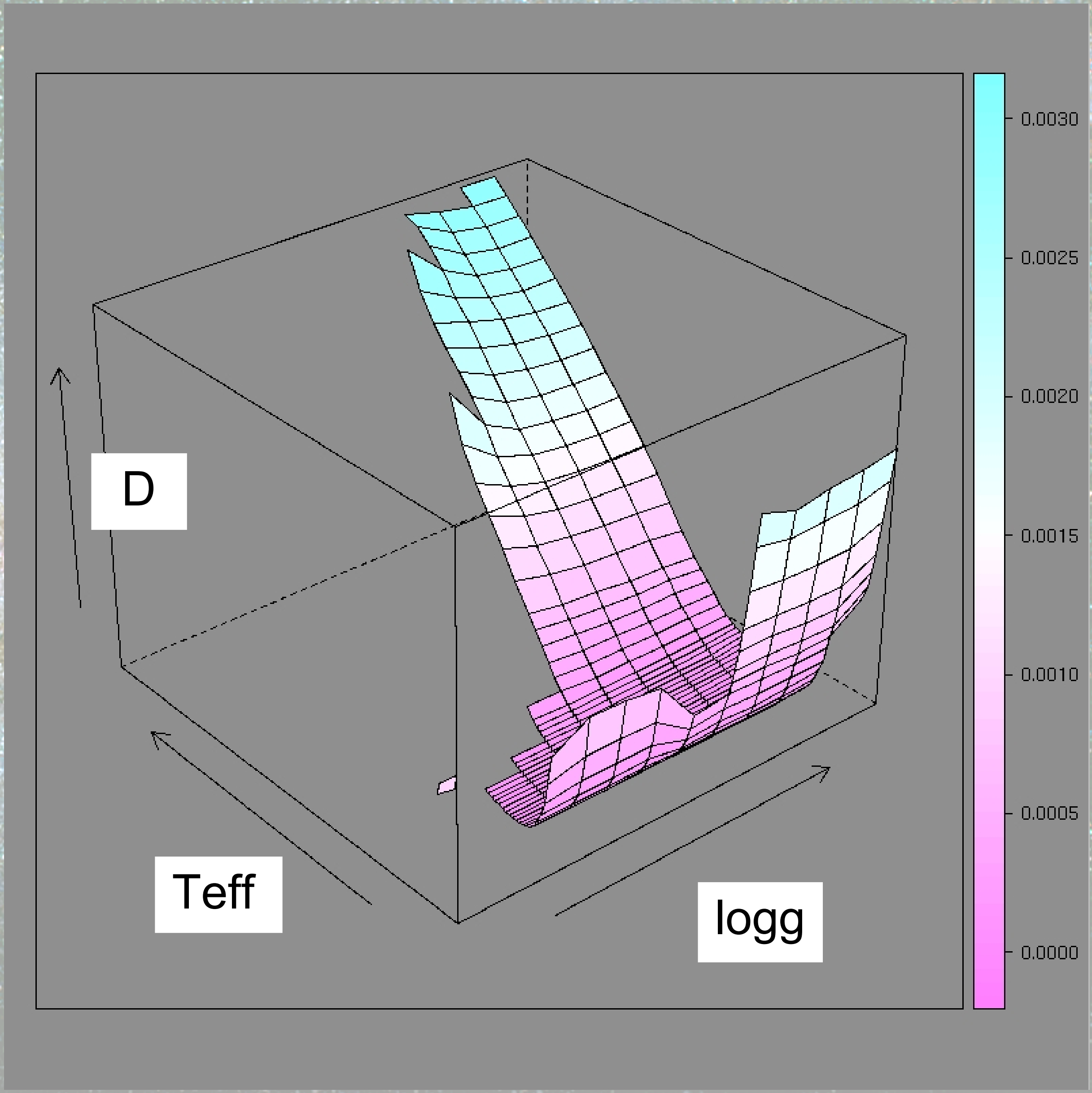
Fixed values:

A_v = 0, [Fe/H] = 0.0

logg = 4.5, T_{eff} = 3500

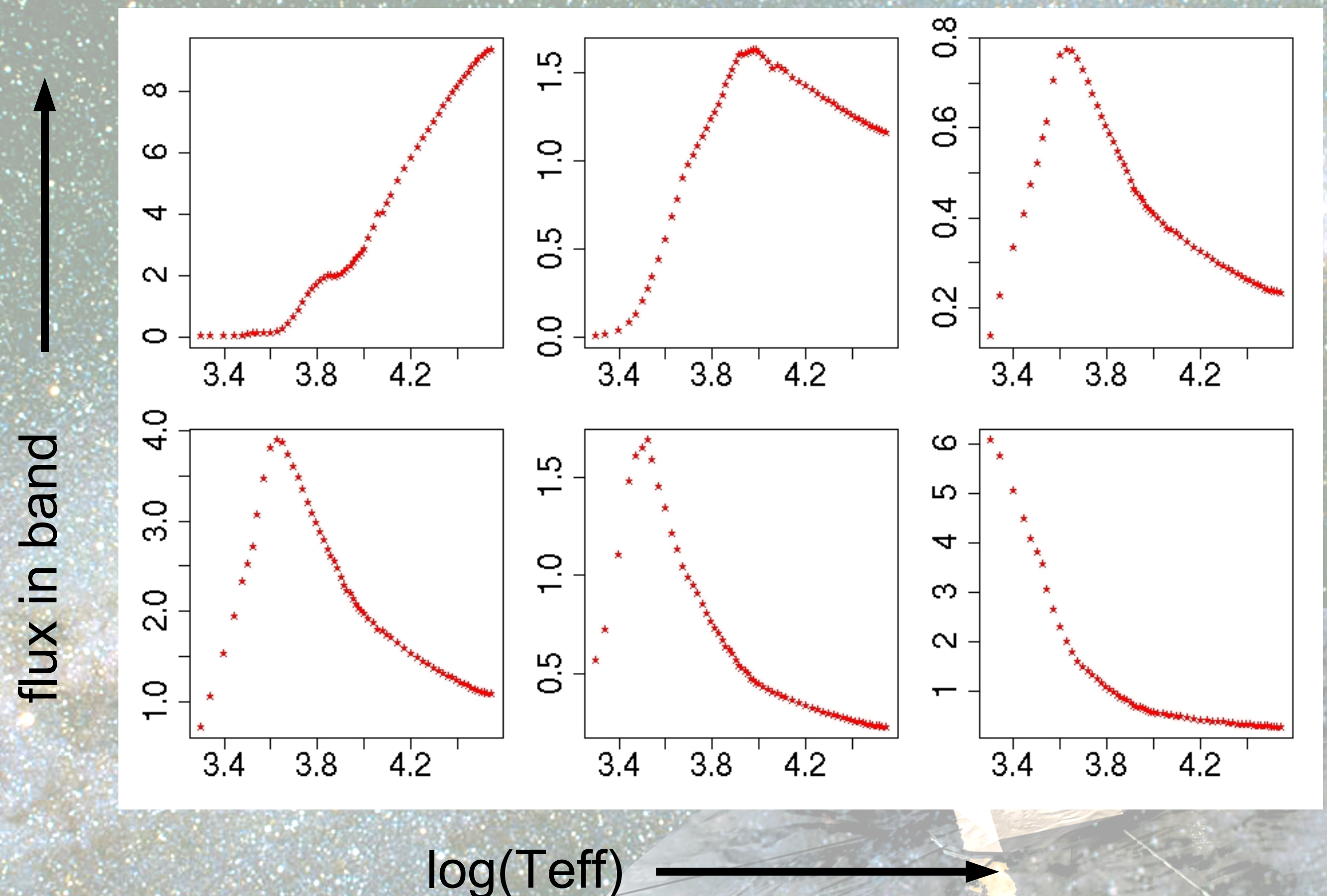
Strong vs. weak parameters

Minimum distance (nearest neighbour) method

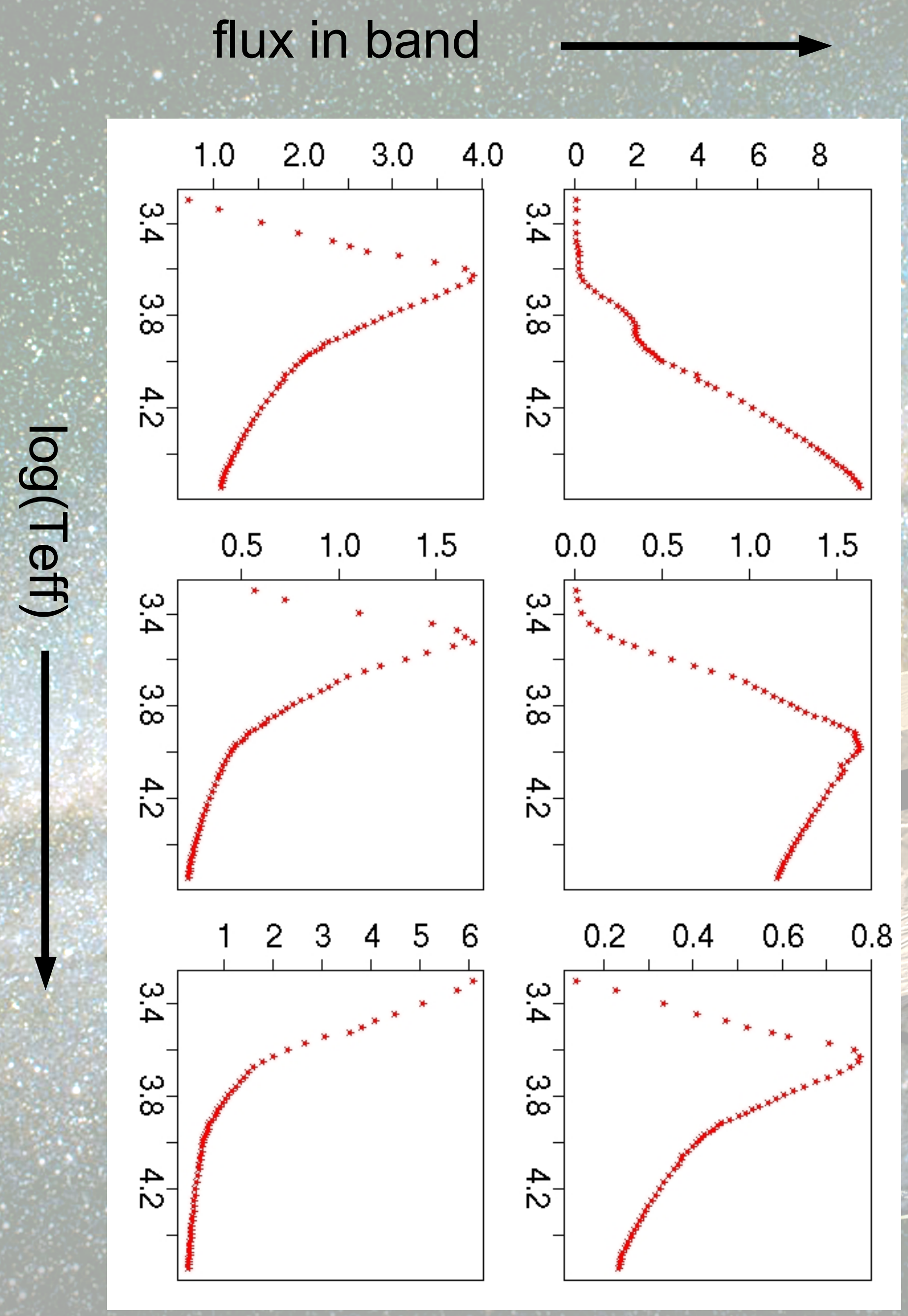


AP ranges:
$A_v = 0$
[Fe/H] = 0
logg = {-0.5, 5.5}
T_{eff} = {2000, 29000}

An inverse problem



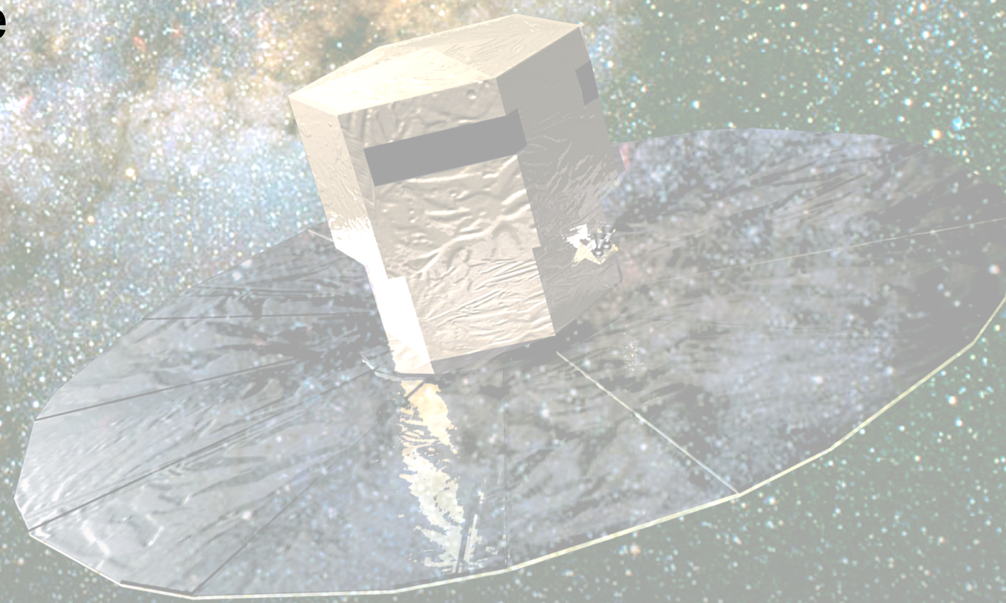
An inverse problem



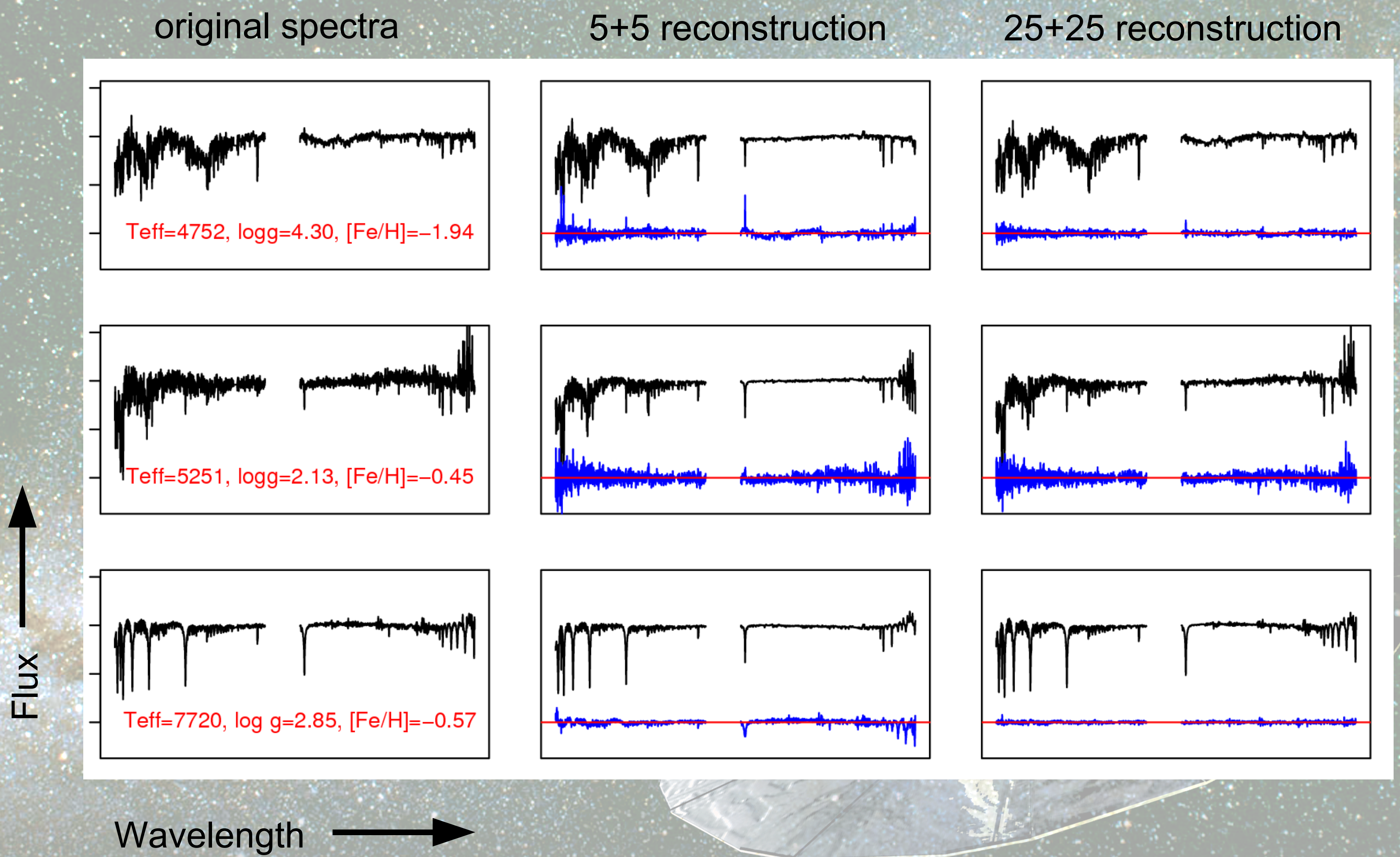
Coryn Bailer-Jones, MPIA, Heidelberg

Stellar parameters from SDSS spectra

- train regression model using simulated spectra (true physical parameters are known)
- predict (unknown) parameters of real objects
- model: MLP neural network (statnet)
- large input space (thousands of correlated pixels)
  - dimension reduction via PCA
  - also removes some noise
  - can act as a junk filter

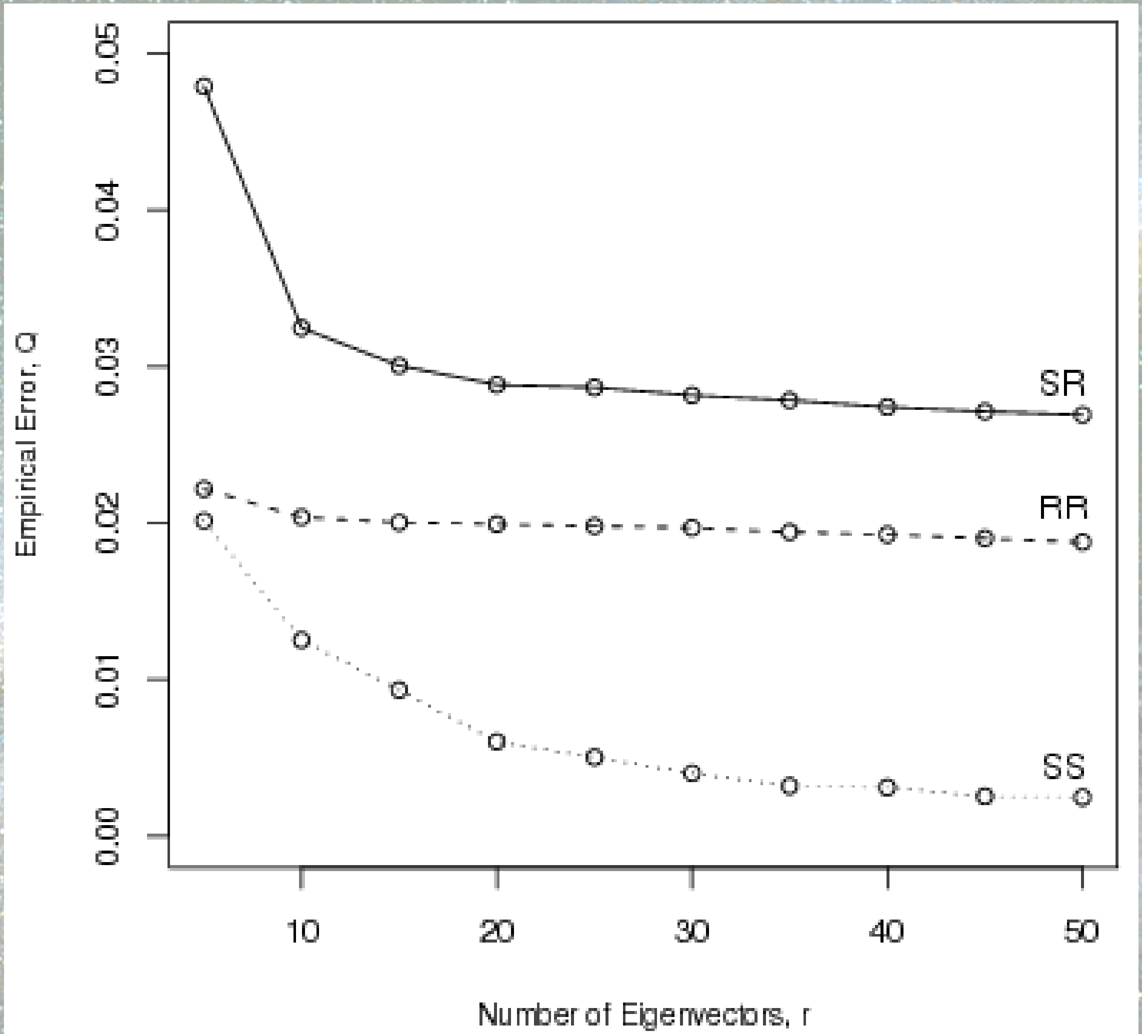


PCA



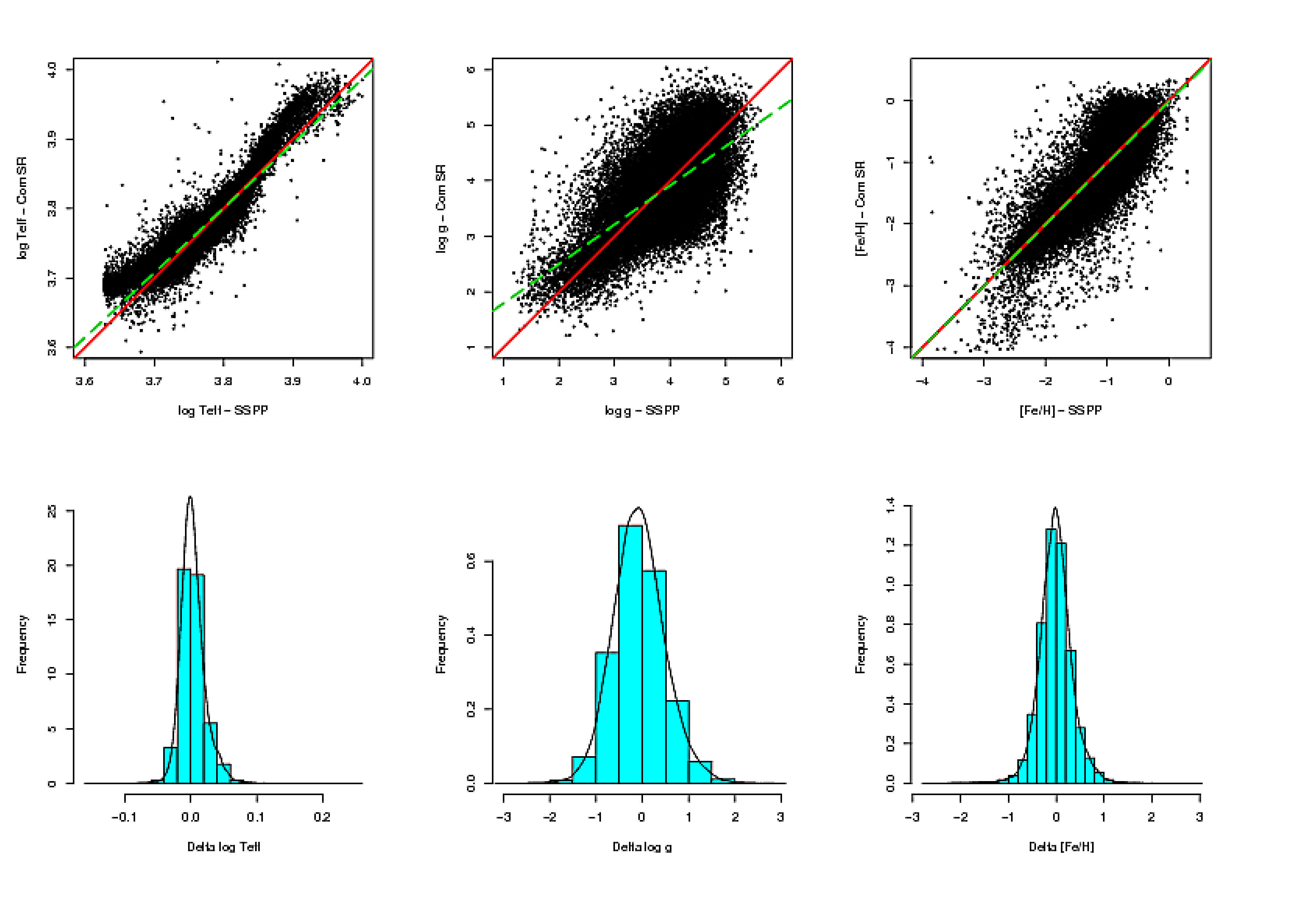
Re Fiorentin, Bailer-Jones et al. 2007

PCA



Re Fiorentin, Bailer-Jones et al. 2007

Results: comparison of parameter determinations with independent estimates



Re Fiorentin, Bailer-Jones et al. 2007

Exploiting the results

Physical parameter estimates are the basis for detailed physical studies

nature
Vol 450|13 December 2007|doi:10.1038/nature06460

ARTICLES

Two stellar components in the halo of the Milky Way

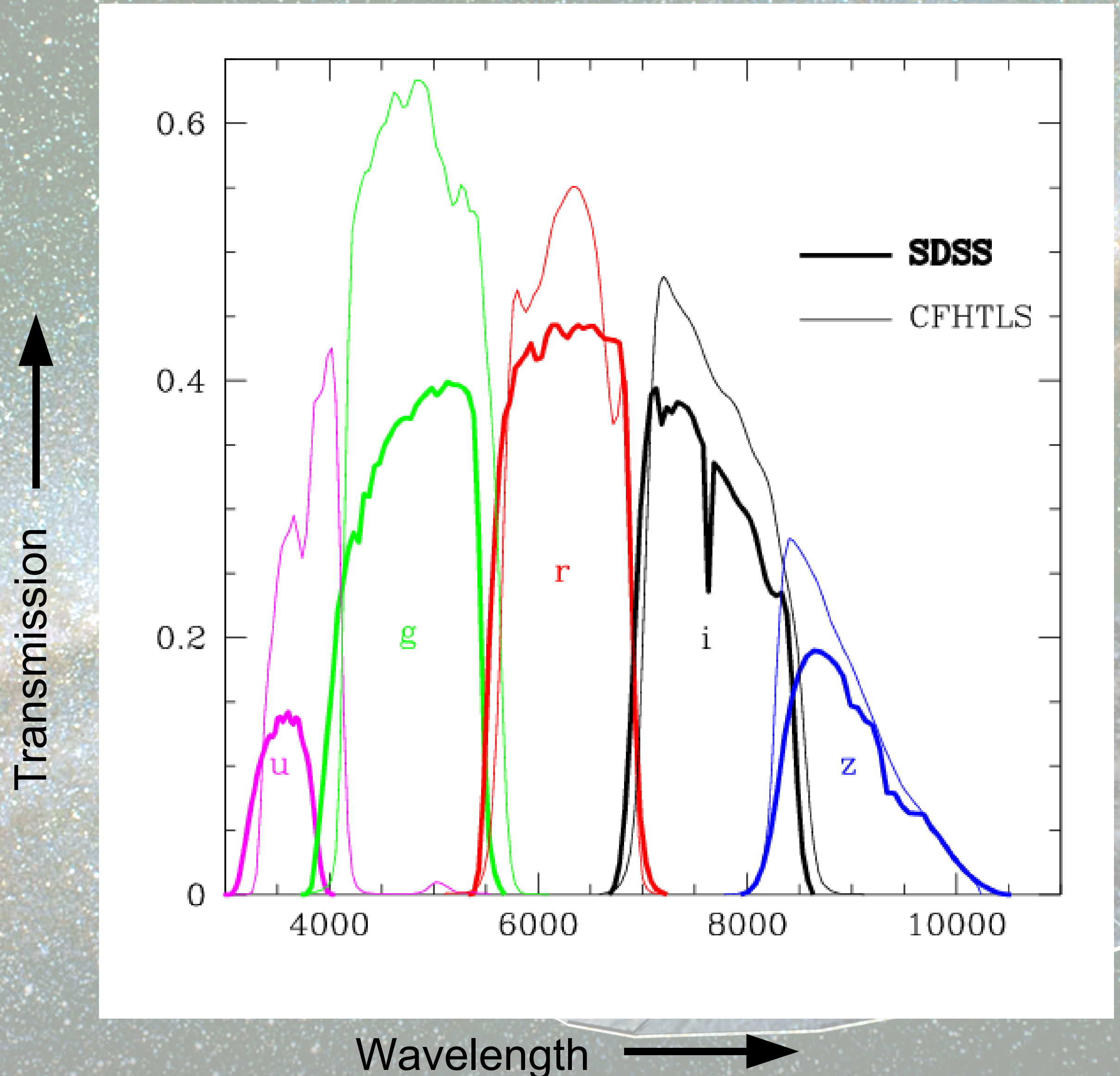
Daniela Carollo[1,2,3,5], Timothy C. Beers[2,3], Young Sun Lee[2,3], Masashi Chiba[4], John E. Norris[5], Ronald Wilhelm[6], Thirupathi Sivarani[2,3], Brian Marsteller[2,3], Jeffrey A. Munn[7], Coryn A. L. Bailer-Jones[8], Paola Re Fiorentin[8,9] & Donald G. York[10,11]

The halo of the Milky Way provides unique elemental abundance and kinematic information on the first objects to form in the Universe, and this information can be used to tightly constrain models of galaxy formation and evolution. Although the halo was once considered a single component, evidence for its dichotomy has slowly emerged in recent years from inspection of small samples of halo objects. Here we show that the halo is indeed clearly divisible into two broadly overlapping structural components—an inner and an outer halo—that exhibit different spatial density profiles, stellar orbits and stellar metallicities (abundances of elements heavier than helium). The inner halo has a modest net prograde rotation, whereas the outer halo exhibits a net retrograde rotation and a peak metallicity one-third that of the inner halo. These properties indicate that the individual halo components probably formed in fundamentally different ways, through successive dissipational (inner) and dissipationless (outer) mergers and tidal disruption of proto-Galactic clumps.

Incorporating domain knowledge

- much redundant information in spectra
- can directly extract sensitivity of inputs (pixels) from simulated data (physical models)
  - soft dimension reduction
- use in various ways
  - data weighting (global)
  - local regression via local weighting (iterative) *[in progress...]*
- ideal case
  - design optimal filters to use in telescope in first place

Photometric filters



Coryn Bailer-Jones, MPIA, Heidelberg

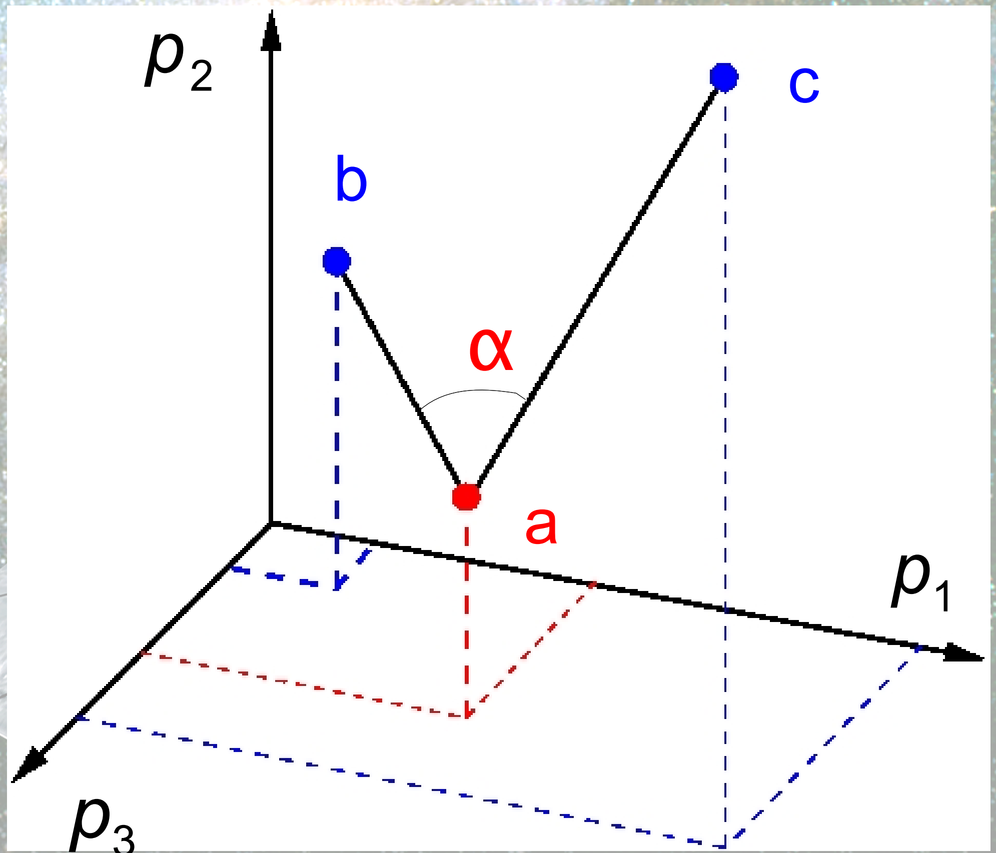
image from R. Sharp, IoA

Heuristic filter design

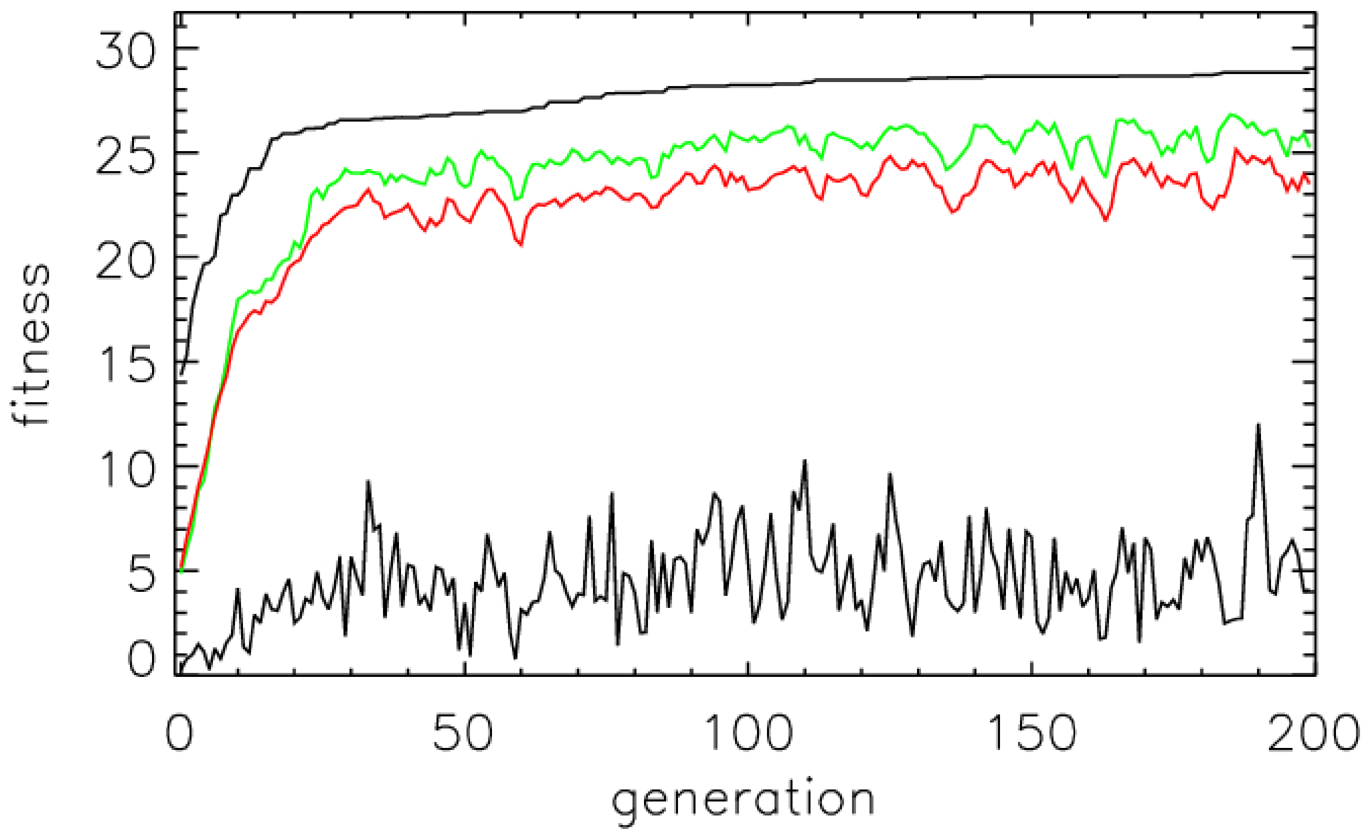
- algorithm
  - parametrize filter system (three per fiter)
  - establish a figure-of-merit (FoM) of filter system performance
  - maximize FoM with respect to filter system parameters
- optimizer
  - cannot define derivatives
  - local minima guaranteed!
  - evolutionary algorithm (genetic algorithm)

Figure of merit (fitness)

- space: flux in each filter of a filter system
- maximum fitness: local directions of variance of the different astrophysical parameters are orthogonal
- calculated from labelled data set

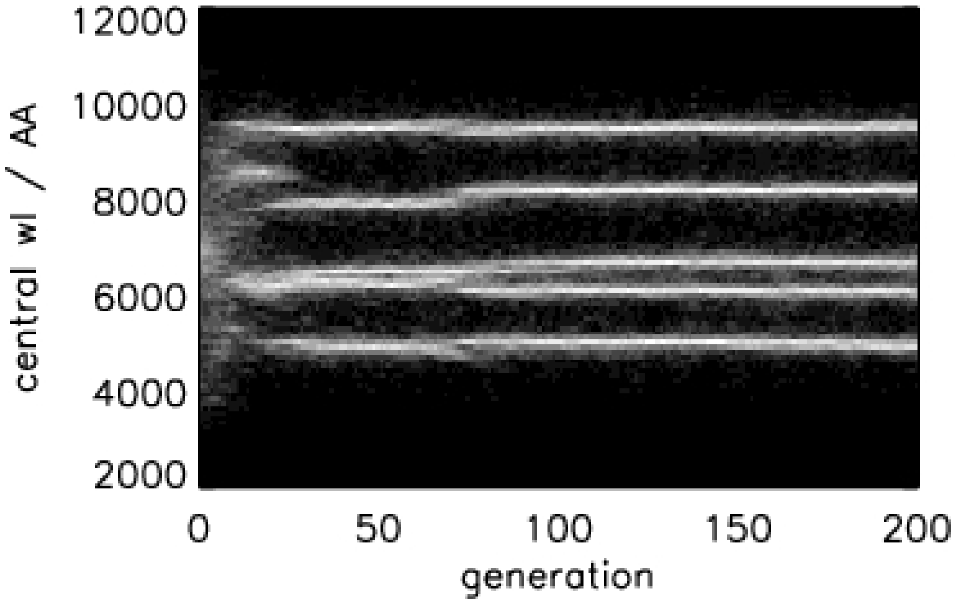


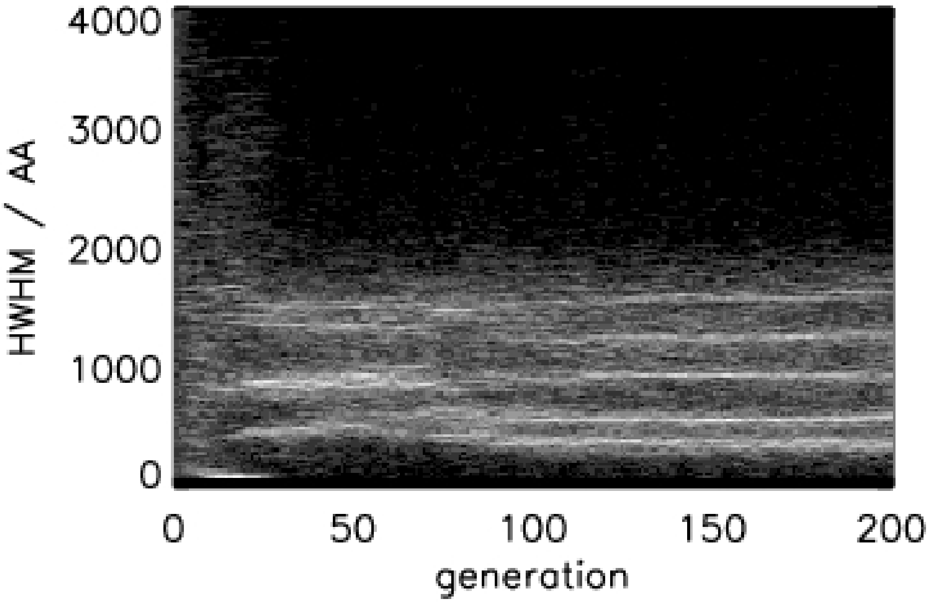
Fitness evolution

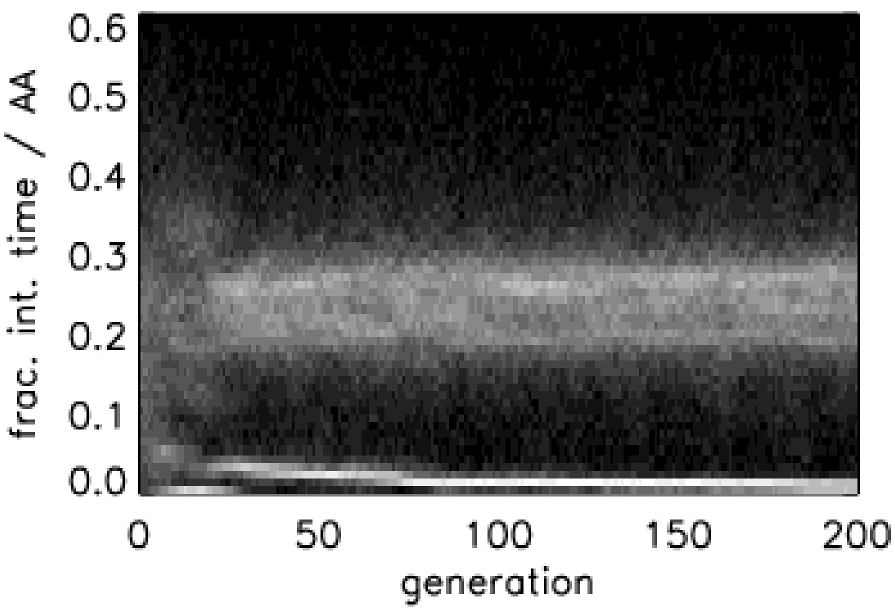


Bailer-Jones 2004

Filter system evolution



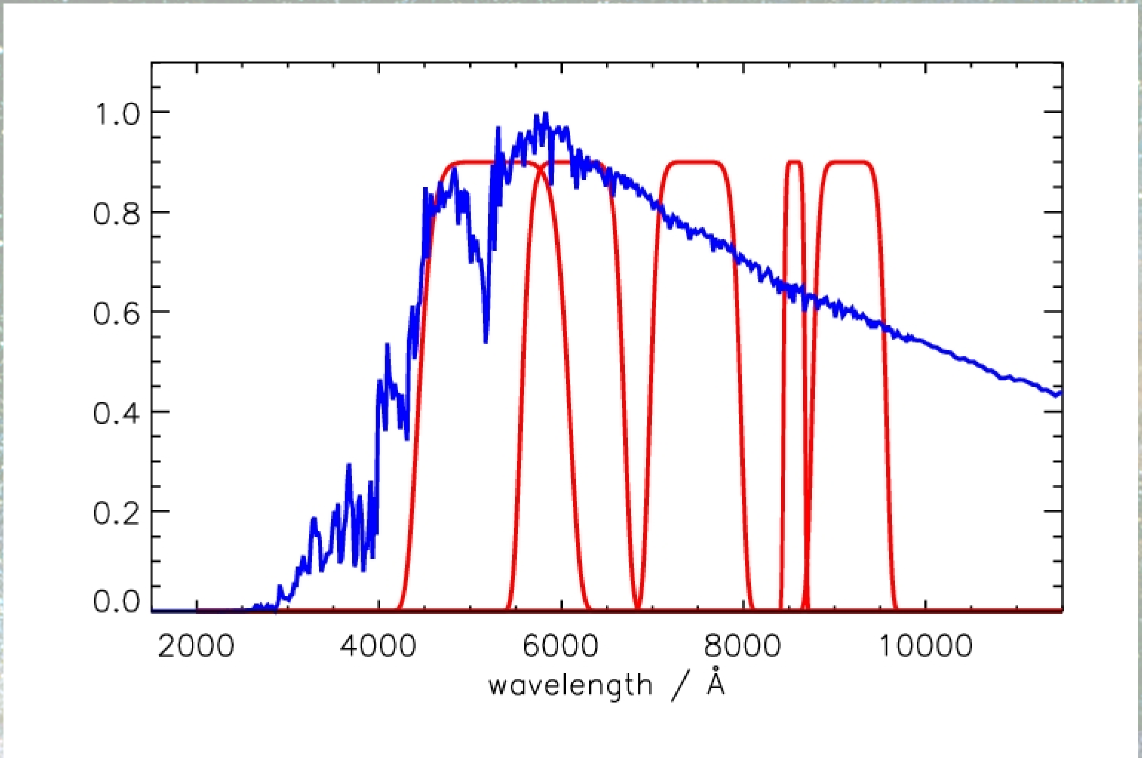




Evolution of all filter system parameters
(200*5 for each parameter type at each generation)

Bailer-Jones 2004

Discovered filter system



- unconventional: overlapping filters
- learned filter system competitive in parameter estimation work
- method could be extended to design “software filters” in algorithms (e.g. hierarchical classification system)

Summary

- astronomy driven by what we can learn from data (surveys)
- extract optimal information from spectra
  - classification (star, quasar etc.)
  - estimate physical parameters
- use of machine learning methods
  - probabilistic system for classification (SVM)
  - issues of inverse problem and “weak” parameters
- exploit domain knowledge
  - post hoc adjustment based on population fractions (priors)
  - deriving sensitivities from simulated data
  - EA-based heuristic filter design
- future work
  - nonlinear dimensionality reduction
  - degenerate solutions and error estimation