Finding rare objects in astronomical surveys

Coryn Bailer-Jones
Max Planck Institute for Astronomy, Heidelberg
http://www.mpia.de/~calj

MPI für biologische Kybernetik
19 November 2008

Overview

1. The classification problem
2. Probability modification method
3. Classification engine (SVM)
4. Application to the future Gaia mission
5. Questions and open issues

Acknowledgements

- Gaia @ MPIA group
  - Kester Smith
  - Vivi Tsalmantza
  - Rainer Klement
  - formerly: Christian Elting, Carola Tiede
- Various data providers within Gaia DPAC
- http://www.mpia.de/GAIA

DLR

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

esa

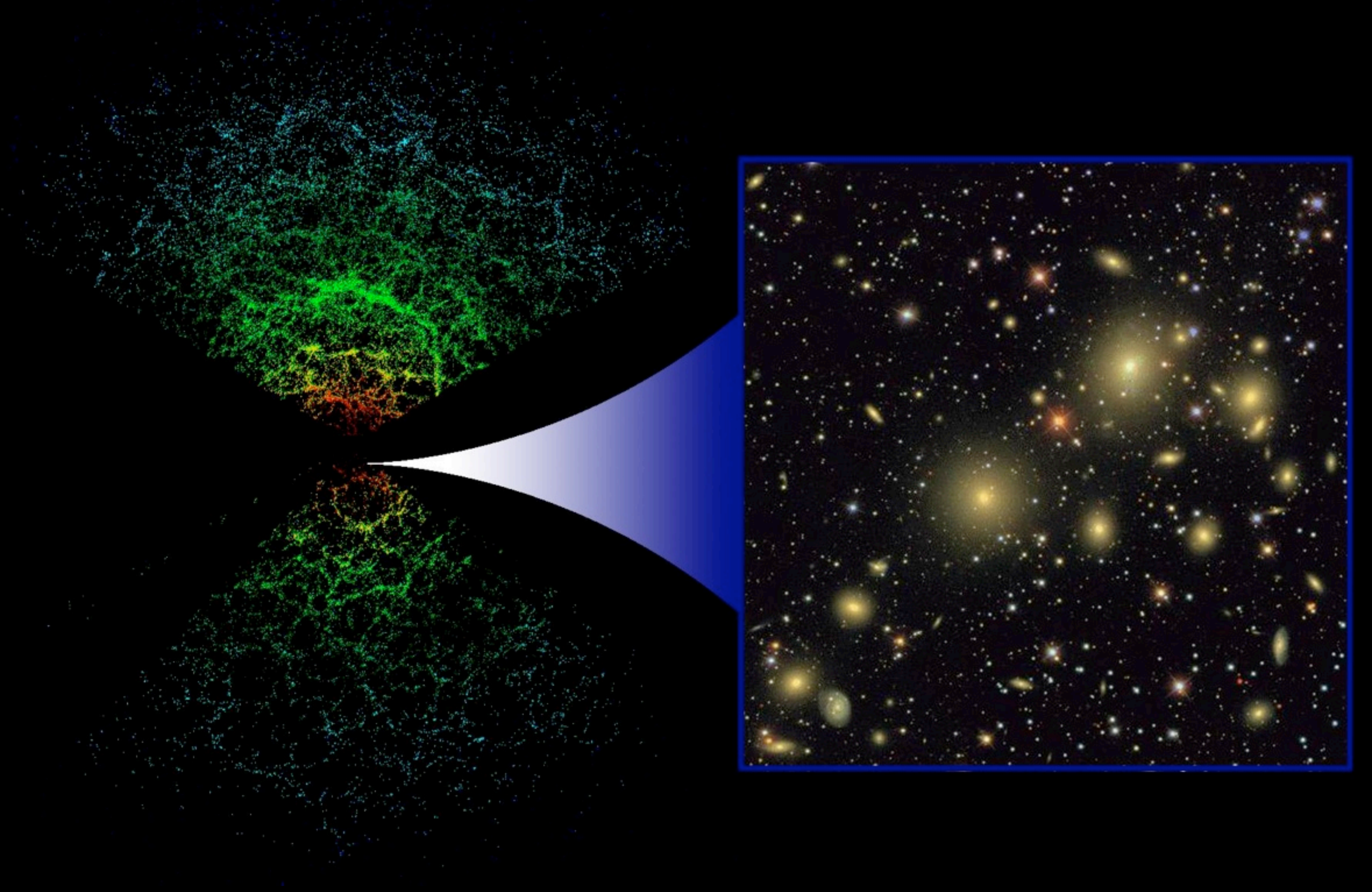


image credit: SDSS

The classification problem

- Astronomical surveys
  - “blind”
  - large, multidimensional data sets
- Have (good) physical models for some of the objects
  - can simulate objects
  - can do supervised classification and derive astrophysical parameters

simulation
(physical model)

Astronomical data → Classes or APs

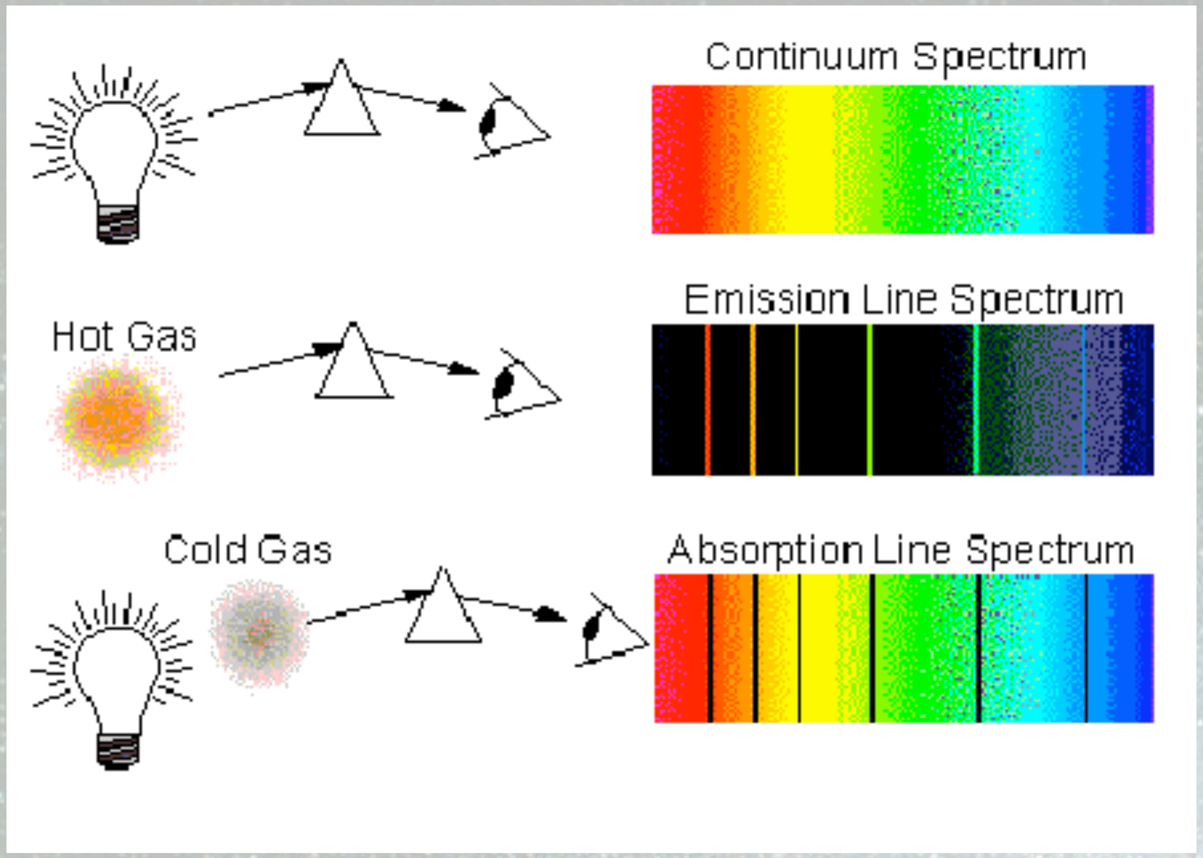
classification
(ML model)

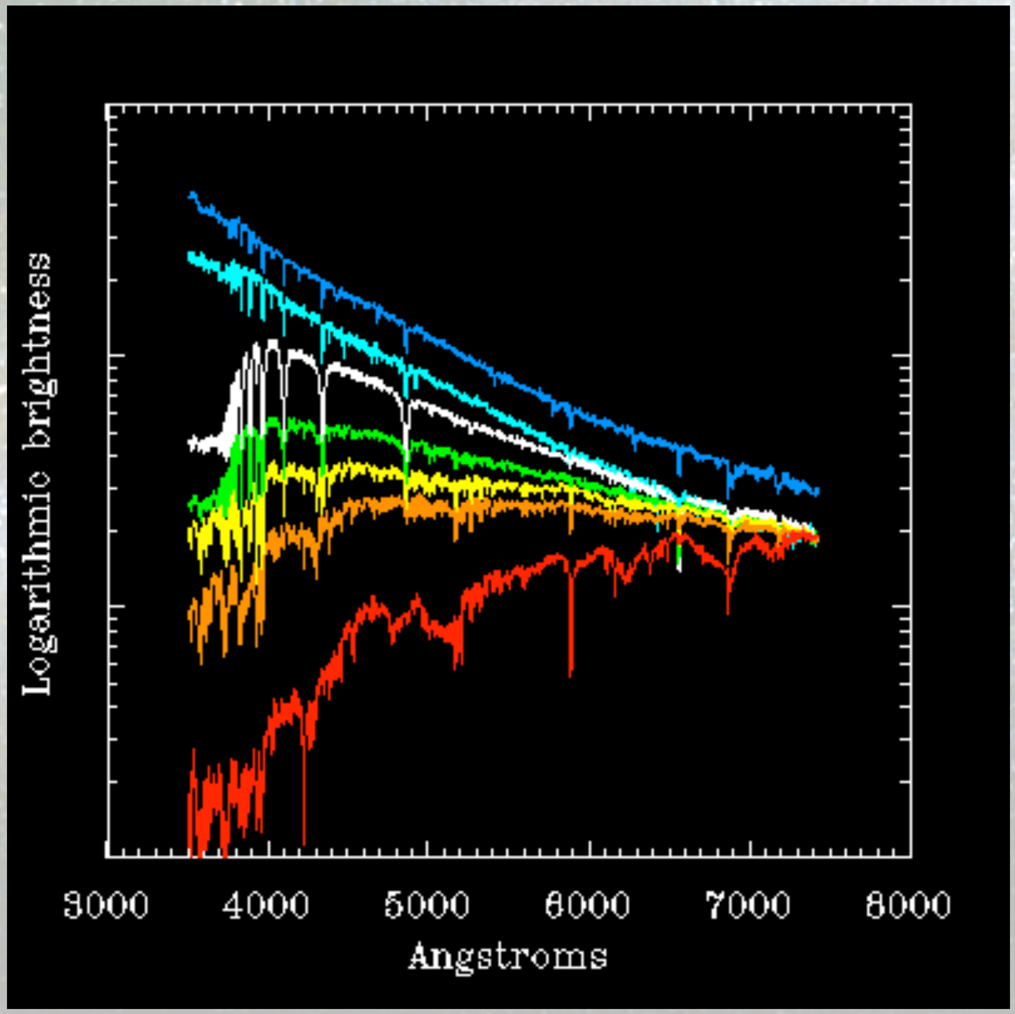
Stars

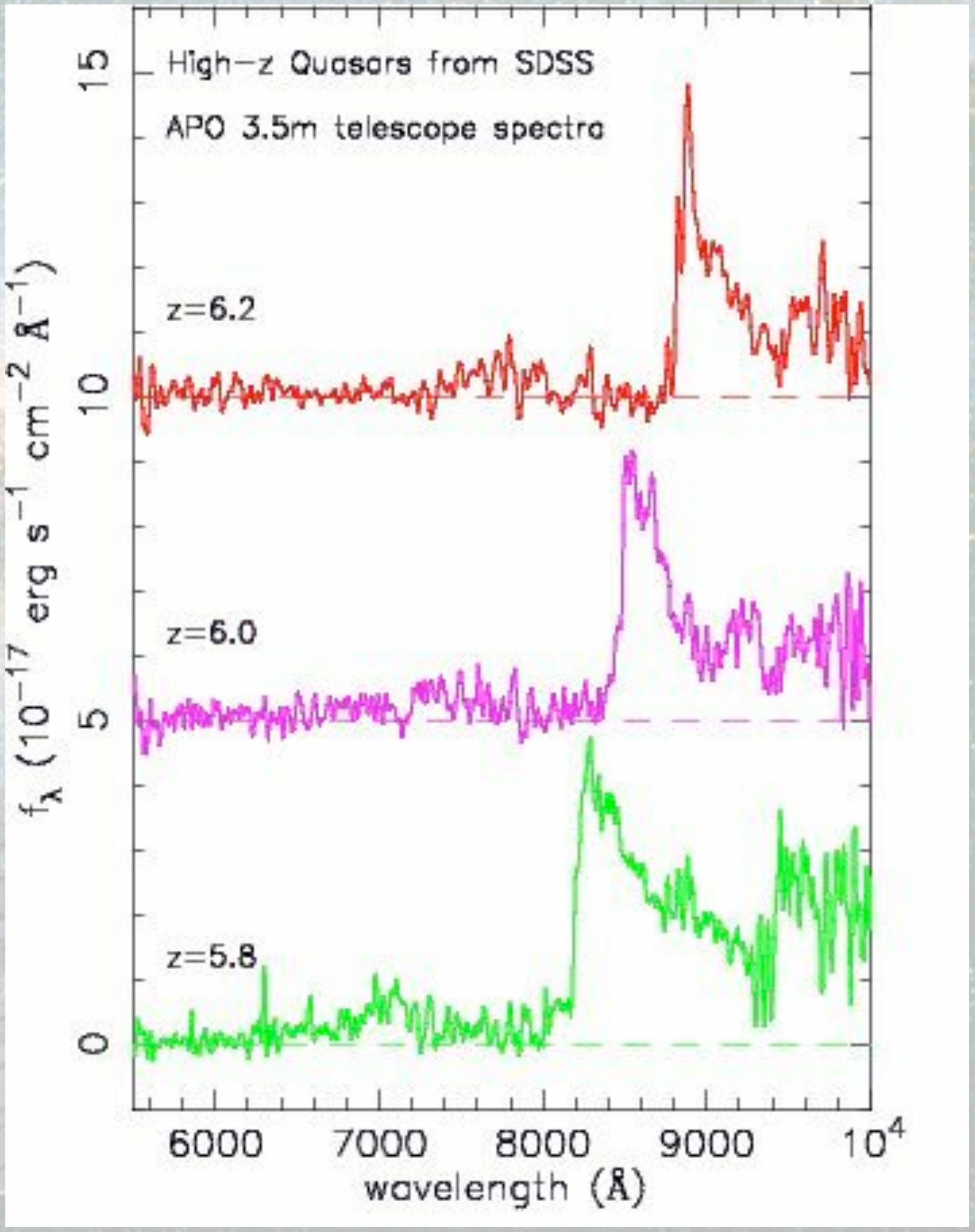
Galaxies

Quasars
(QSO)

Astronomical spectra







Coryn Bailer-Jones, MPIA Heidelberg

Bayesian learning

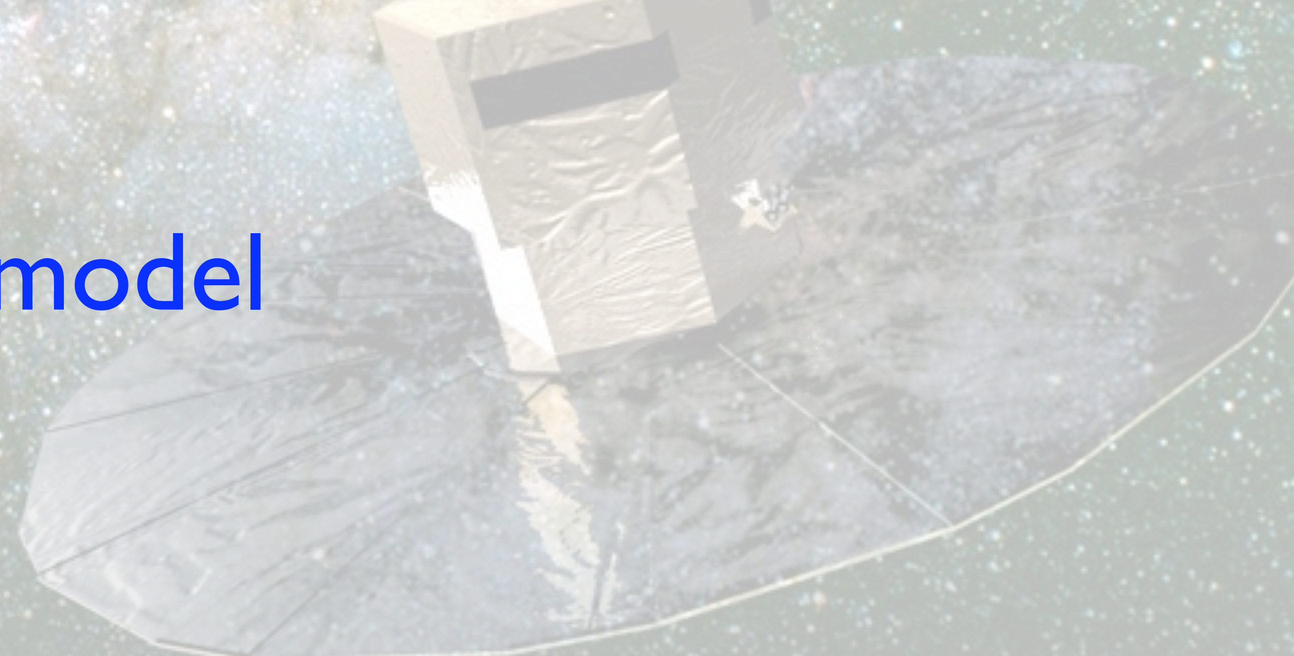
posterior likelihood prior

$$P(C_j|x_n,\theta) = \frac{P(x_n|C_j,\theta)P(C_j|\theta)}{P(x_n|\theta)}$$

class
e.g. star, galaxy

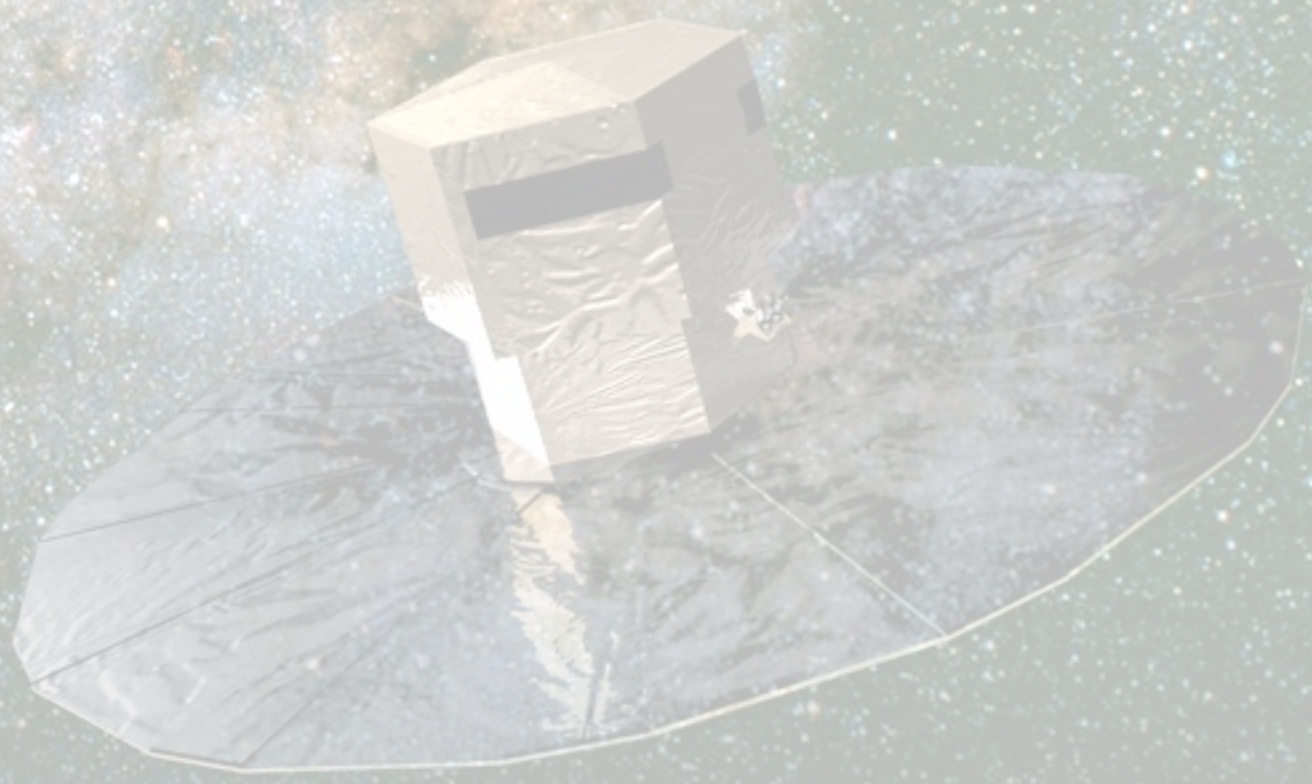
data
(spectrum)

model

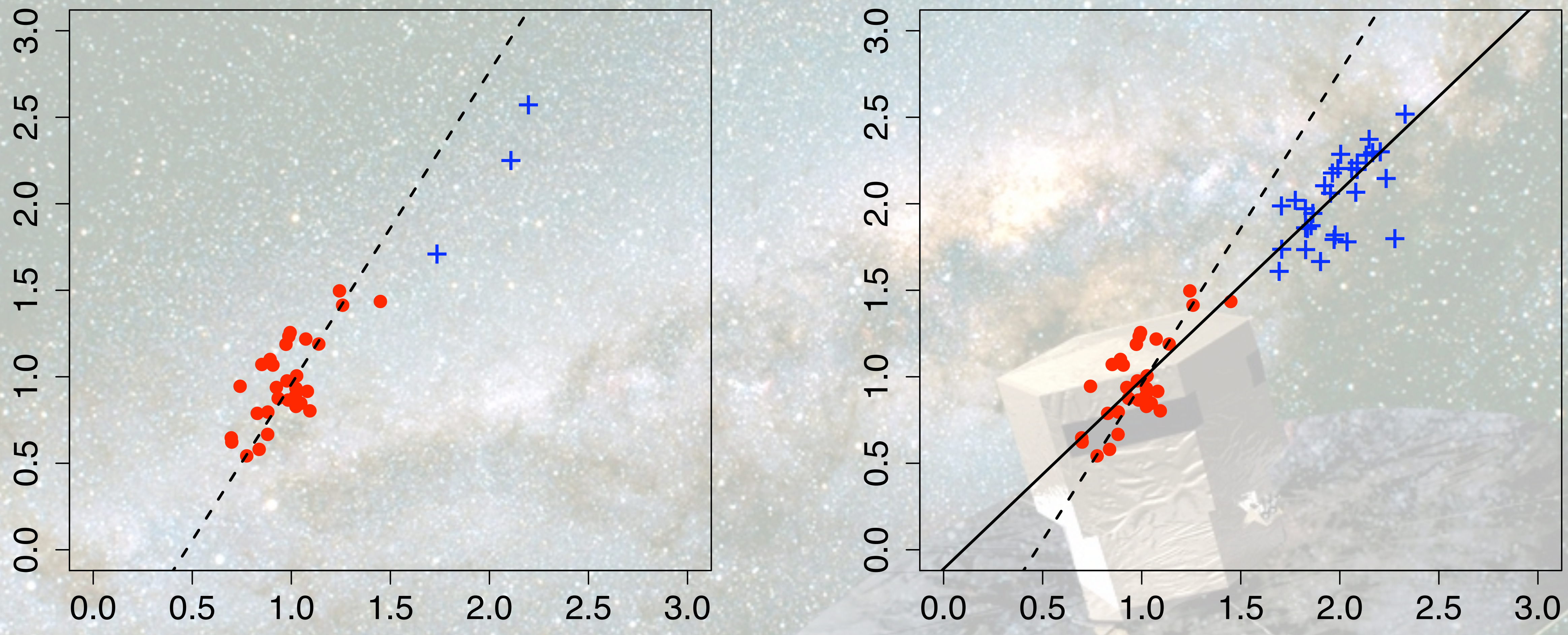


What is the prior?

- All classification models have a prior (maybe implicit)
- We always have *some* prior
- Prior influenced by distribution in training data



Training data distribution influences model fit



Coryn Bailer-Jones, MPIA Heidelberg

What is the prior?

- All classification models have a prior (maybe implicit)
- We always have *some* prior
- Prior influenced by distribution in training data
- Motivation behind method
  - remove influence of training data distribution (“class imbalance”)
  - avoid rebuilding/retraining model to match target population
  - actively control priors

Class fractions

- Relative fraction of objects of each class in a data set

$$\mathbf{f} = (f_{\mathrm{galaxy}}, f_{\mathrm{quasar}}, f_{\mathrm{star}})$$

- Training set typically has equal class fractions

$$\mathbf{f}^{train} = (1, 1, 1)$$

- *Target population* examined here has quasars rare

$$\mathbf{f}^{target} = (1, 0.001, 1)$$

written here unnormalized

The modified model

$$P(C_j|x_n,\theta) = \frac{P(x_n|C_j,\theta)P(C_j|\theta)}{P(x_n|\theta)}$$

define *modified model:*

$$P^{mod}(C_j|x_n,\theta^{mod}) = a_n\, P^{nom}(C_j|x_n,\theta^{nom}) \times \frac{P^{mod}(C_j|\theta^{mod})}{P^{nom}(C_j|\theta^{nom})}$$

approximate priors using class fractions:

$$P^{nom}(C_j|\theta^{nom}) = f_{i=j}^{train} \qquad P^{mod}(C_j|\theta^{mod}) = f_{i=j}^{target}$$

$$P^{mod}(C_j|x_n,\theta^{mod}) = a_n\, P^{nom}(C_j|x_n,\theta^{nom})\, \frac{f_{i=j}^{target}}{f_{i=j}^{train}}$$

Model-based priors

$$P(C_j|x_n, \theta) = \frac{P(x_n|C_j, \theta)P(C_j|\theta)}{P(x_n|\theta)}$$

Can calculate the *model-based priors* from a trained model

$$P(C_j|\theta) = \sum_{n=1}^{n=N_{test}} P(C_j|x_n, \theta)P(x_n|\theta)$$

posterior
(model outputs)

$$= \frac{1}{N_{test}}$$

Consistency of priors

1. $$P^{mod}(C_j|x_n, \theta^{mod}) = a_n\, P^{nom}(C_j|x_n, \theta^{nom})\, \frac{f_{i=j}^{target}}{f_{i=j}^{train}}$$

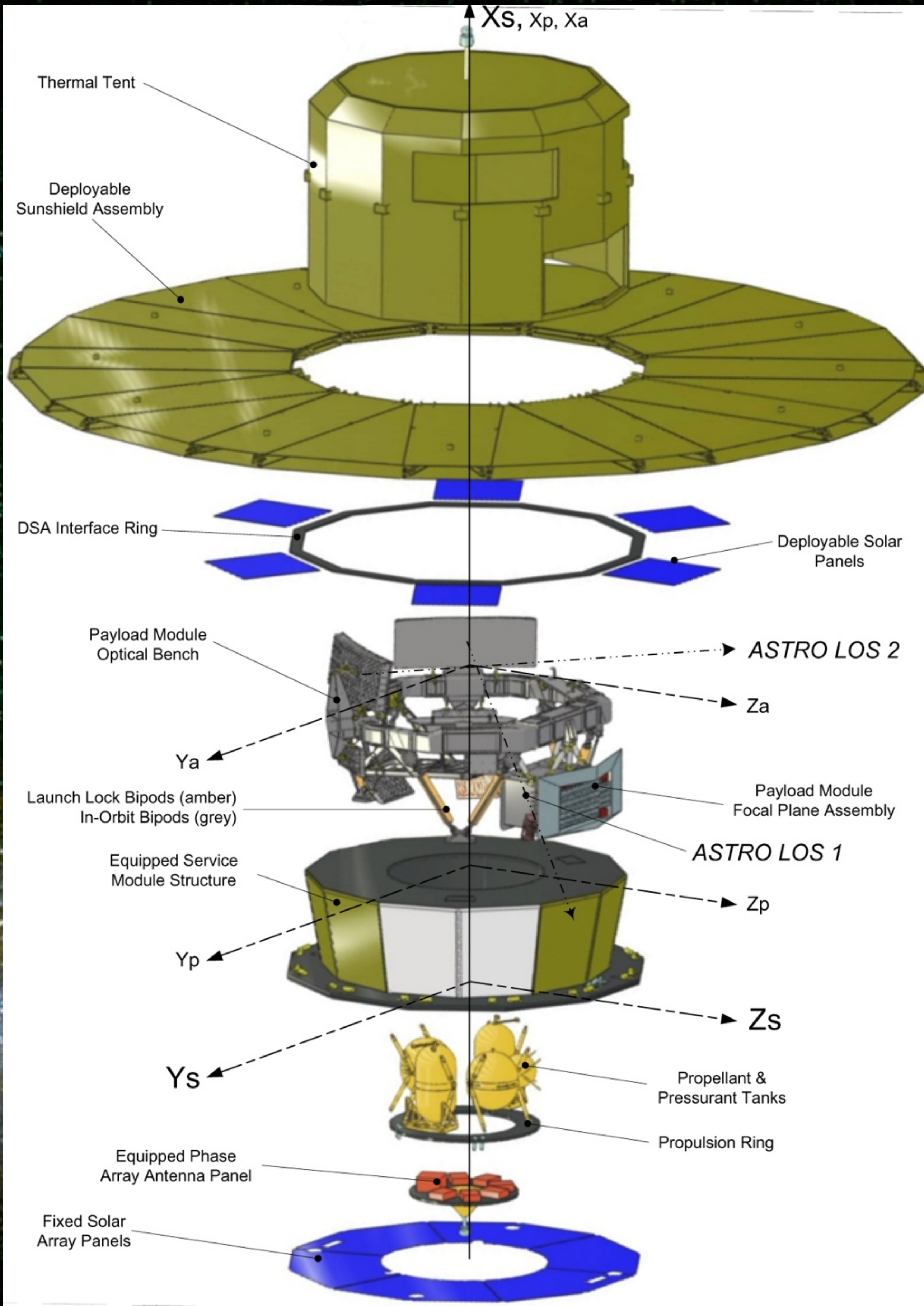
2. $$P(C_j|\theta) = \sum_{n=1}^{n=N_{test}} P(C_j|x_n, \theta) P(x_n|\theta)$$ for *nom* and *mod*

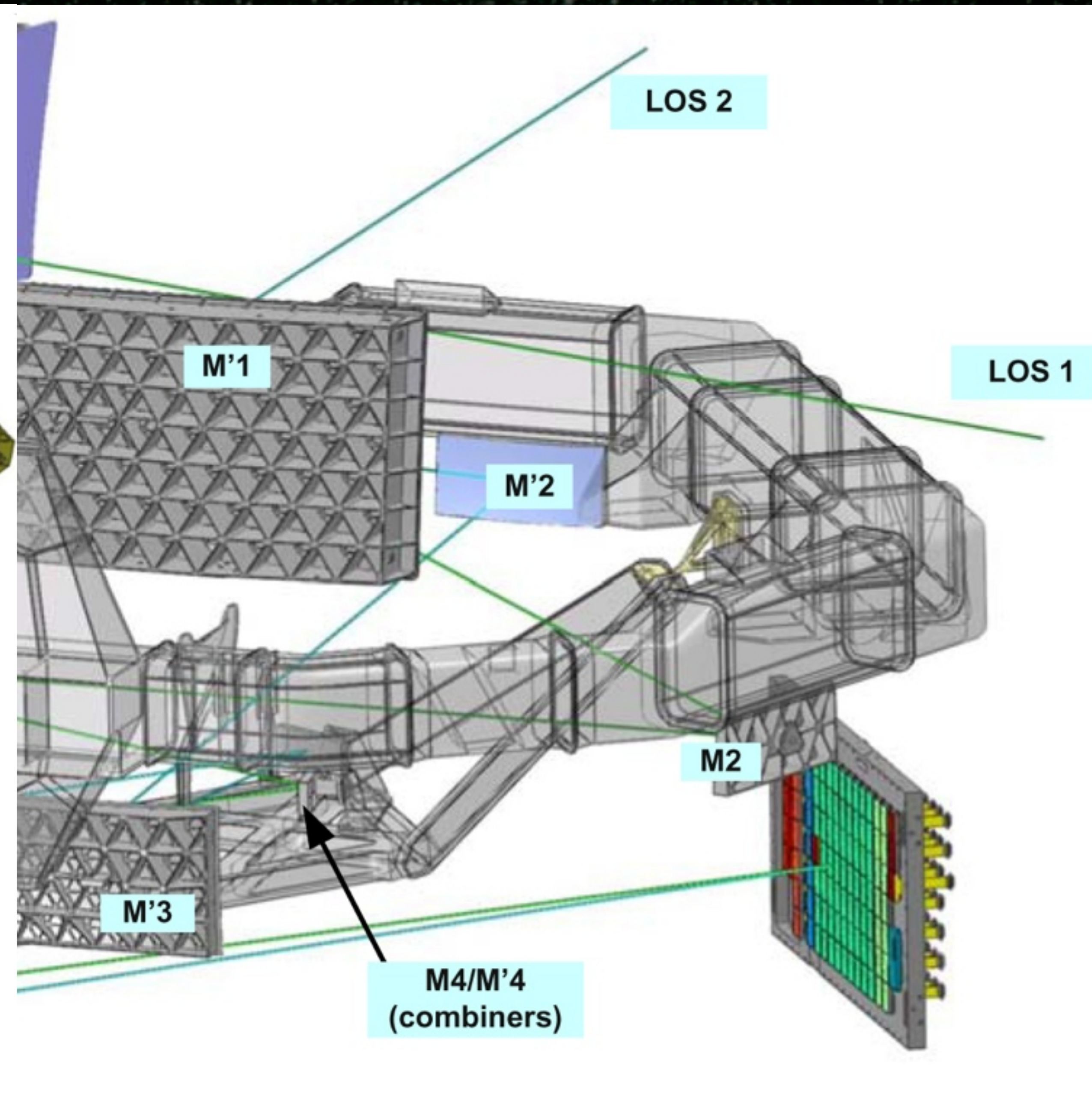
3. $$P^{mod}(C_j|x_n, \theta^{mod}) = a_n\, P^{nom}(C_j|x_n, \theta^{nom}) \times \frac{P^{mod}(C_j|\theta^{mod})}{P^{nom}(C_j|\theta^{nom})}$$

Iterate stage 2 and 3 to achieve consistent priors

Comparison of model-based priors and class fractions

	data	G	star	quasar	galaxy
$P(C_j \vert \theta^{nom})$	full	18.5	0.3380	0.3279	0.3341
f_i^{train}	full	18.5	0.3333	0.3333	0.3333
$P(C_j \vert \theta^{mod})$	full	18.5	0.4965	0.002514	0.5010
f_i^{mod}	full	18.5	0.4998	0.000500	0.4998
$P(C_j \vert \theta^{nom})$	nlEW	18.5	0.367	0.283	0.350
$P(C_j \vert \theta^{nom})$	nlEW	20.0	0.368	0.260	0.372
f_i^{train}	nlEW	both	0.388	0.225	0.388
$P(C_j \vert \theta^{mod})$	nlEW	18.5	0.4983	0.000328	0.5013
$P(C_j \vert \theta^{mod})$	nlEW	20.0	0.4762	0.000277	0.5234
f_i^{mod}	nlEW	both	0.4998	0.000500	0.4998





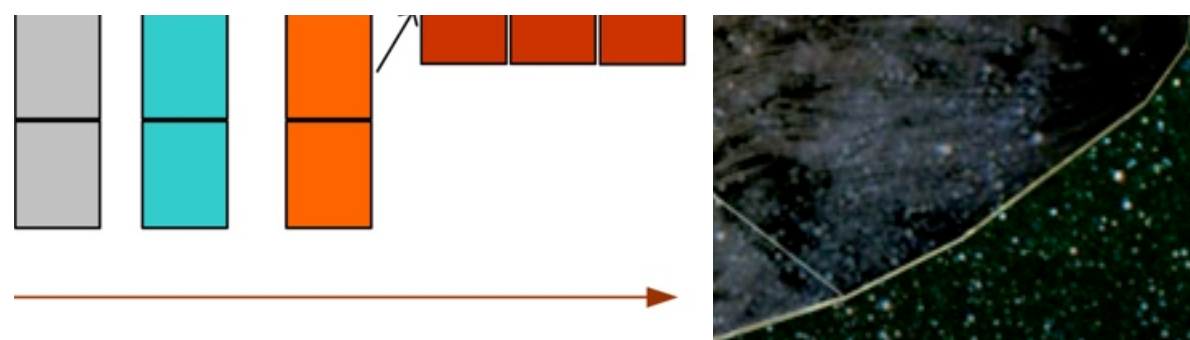
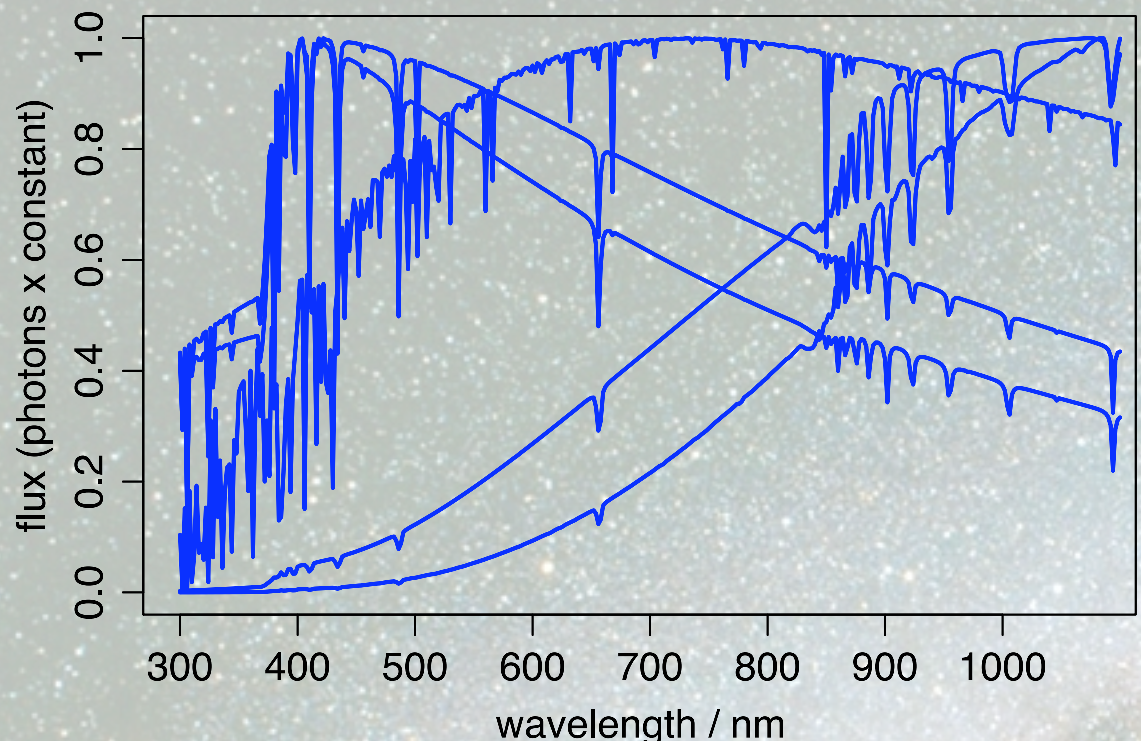


image credit: EADS Astrium

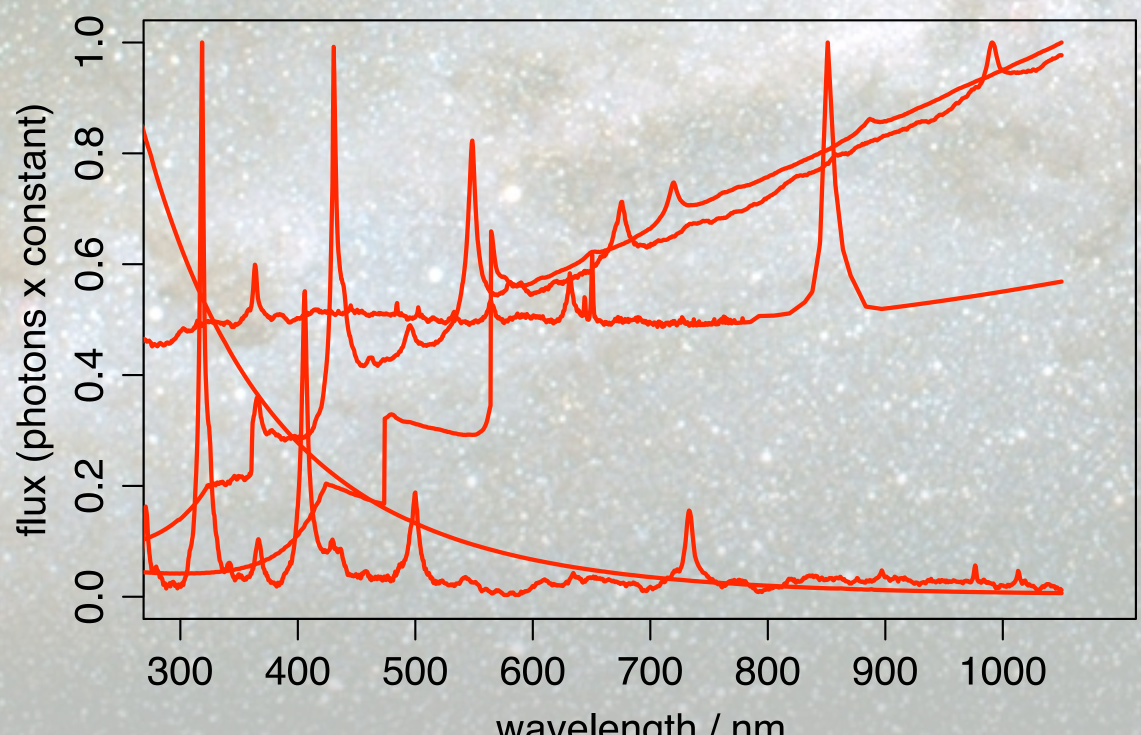
Gaia Galactic survey

- 10^9 objects
- large data variance
- variable noise
- multidimensional data on each object (~80 element spectrum)
- Build classification models with simulated data (test too, for now)

Optical spectra



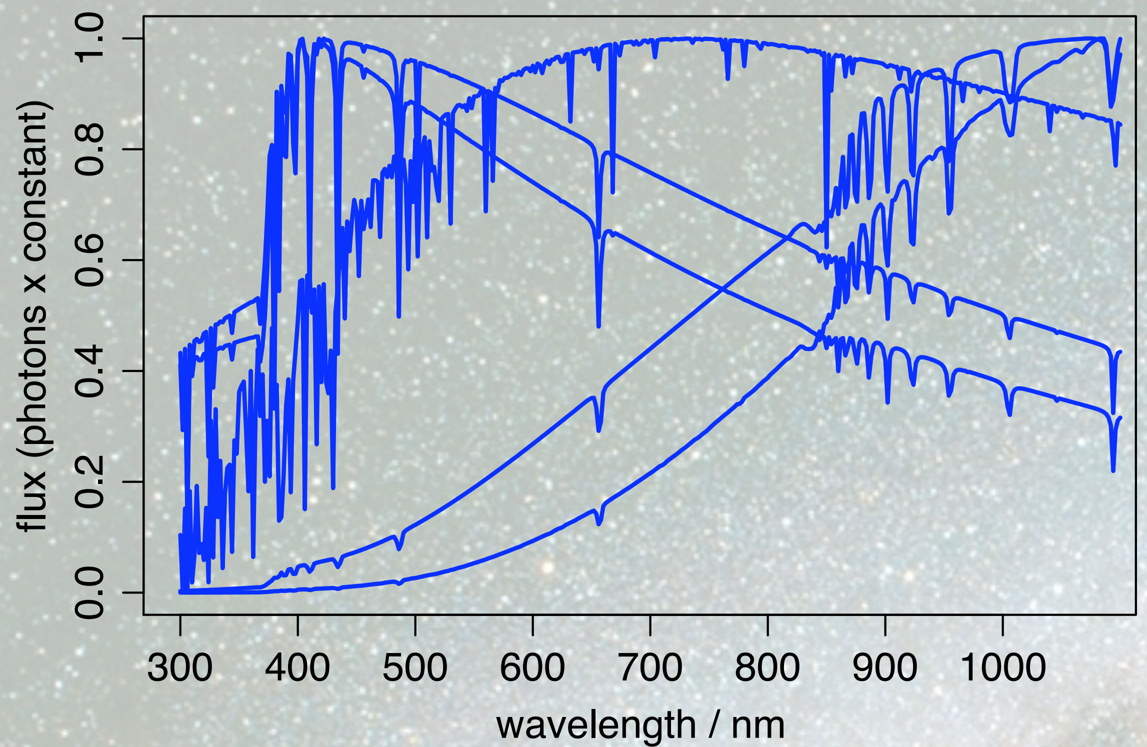
Stars
Variation due to temperature, age and composition

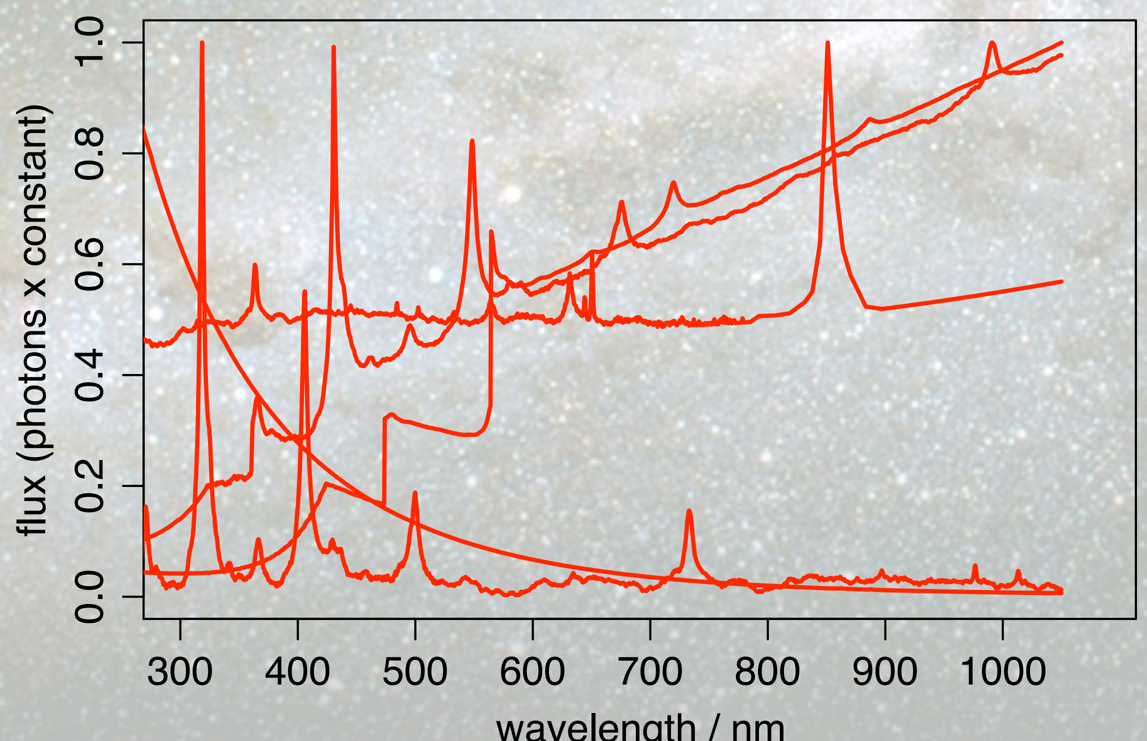


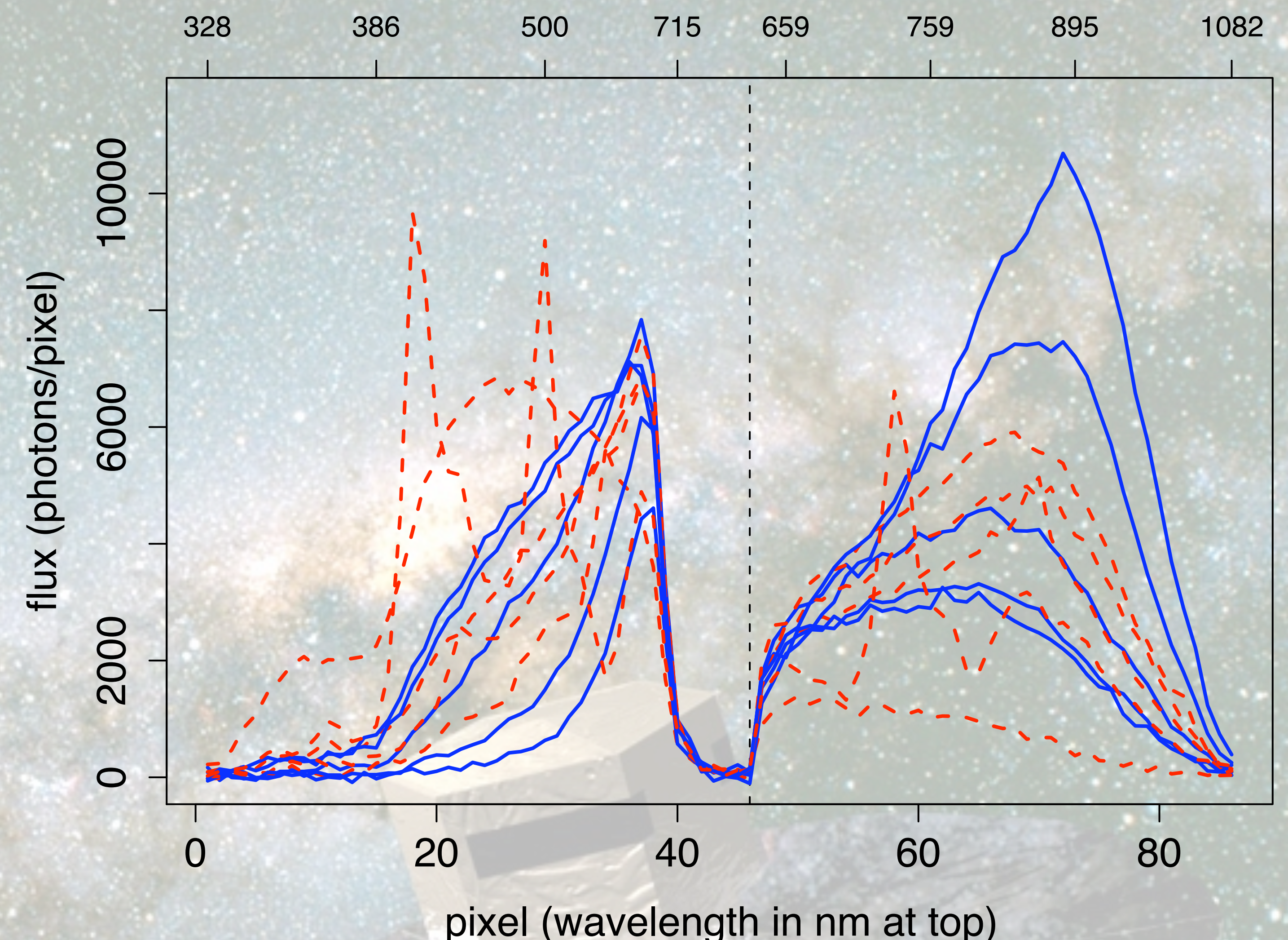
Quasars
Variation in slope, line strength, redshift

Input spectra

Gaia spectra







86 pixels per object

blue = stars

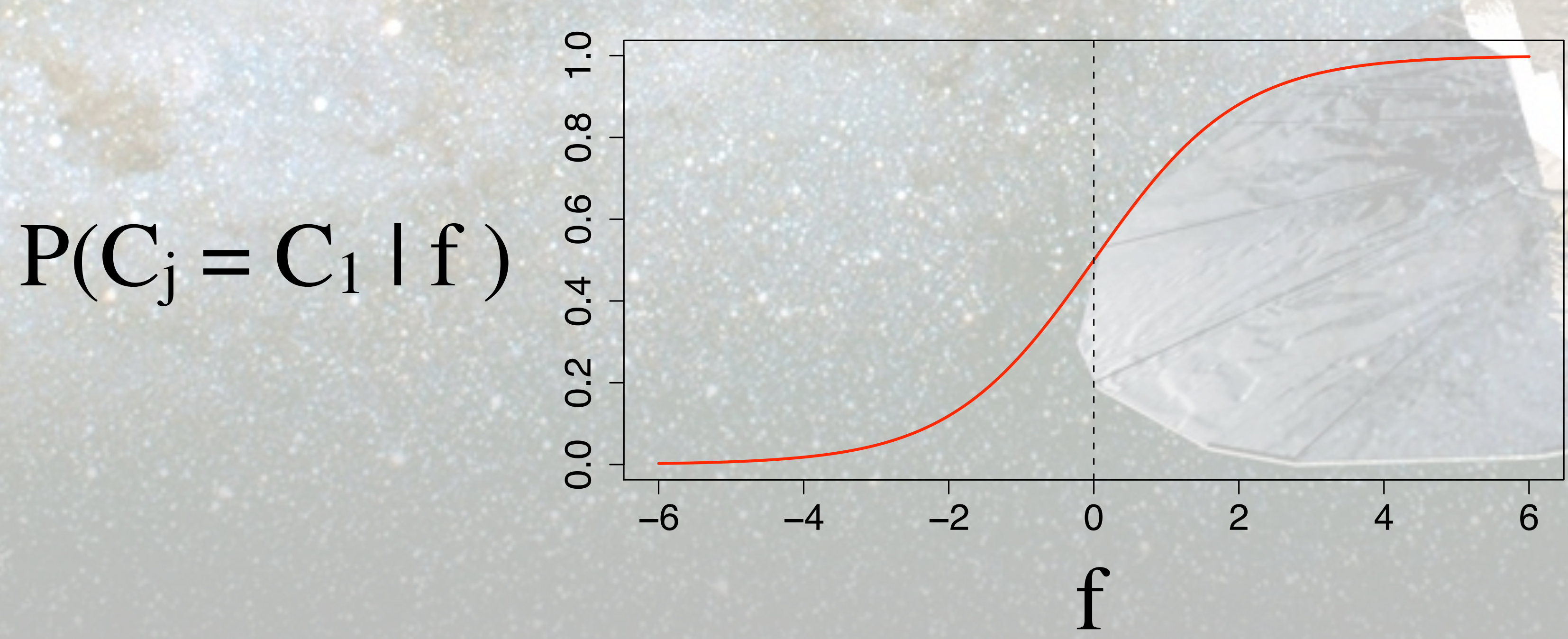
red / dashed = quasars

Coryn Bailer-Jones, MPIA Heidelberg

Classification engine: SVM

libSVM (Java)

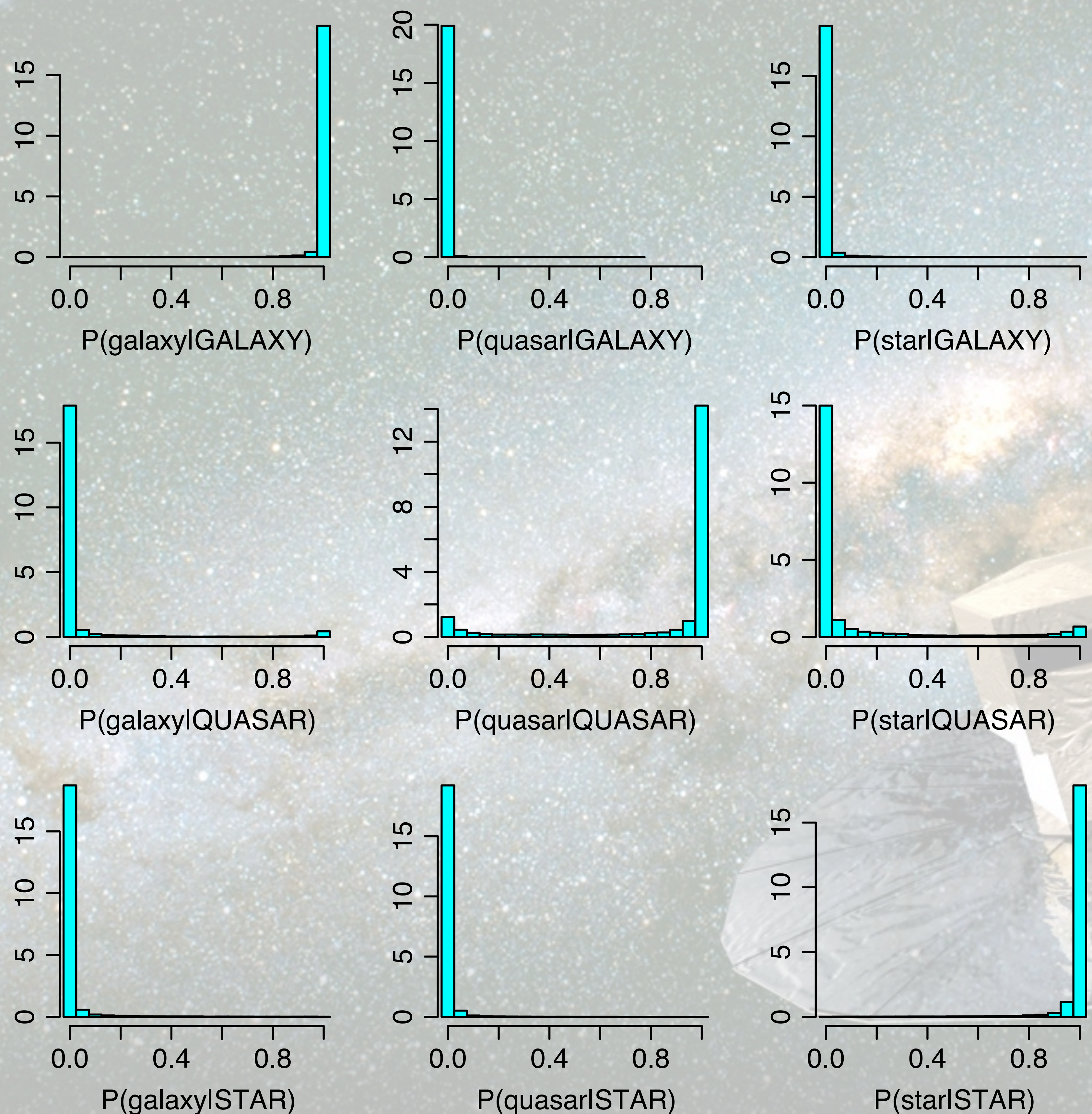
- RBF kernel, scale length γ
- probabilities from sigmoidal fit (Platt 2000)
- multiple classes from pairwise coupling (Wu et al. 2004)
- tune C (regularizer) and γ using CV and Nelder-Mead
- train: 5000 of each class test: 60 000 of each class



Predicted classes

Output probabilities

Nominal model



True classes

Confusion matrix

	galaxy	quasar	star
GALAXY	99.37	0.00	0.63
QUASAR	4.22	85.59	10.19
STAR	0.68	0.13	99.19

Assign objects to class with largest probability

Performance metrics

- Build a sample by setting a probability threshold, P_t
- Sample completeness for class j

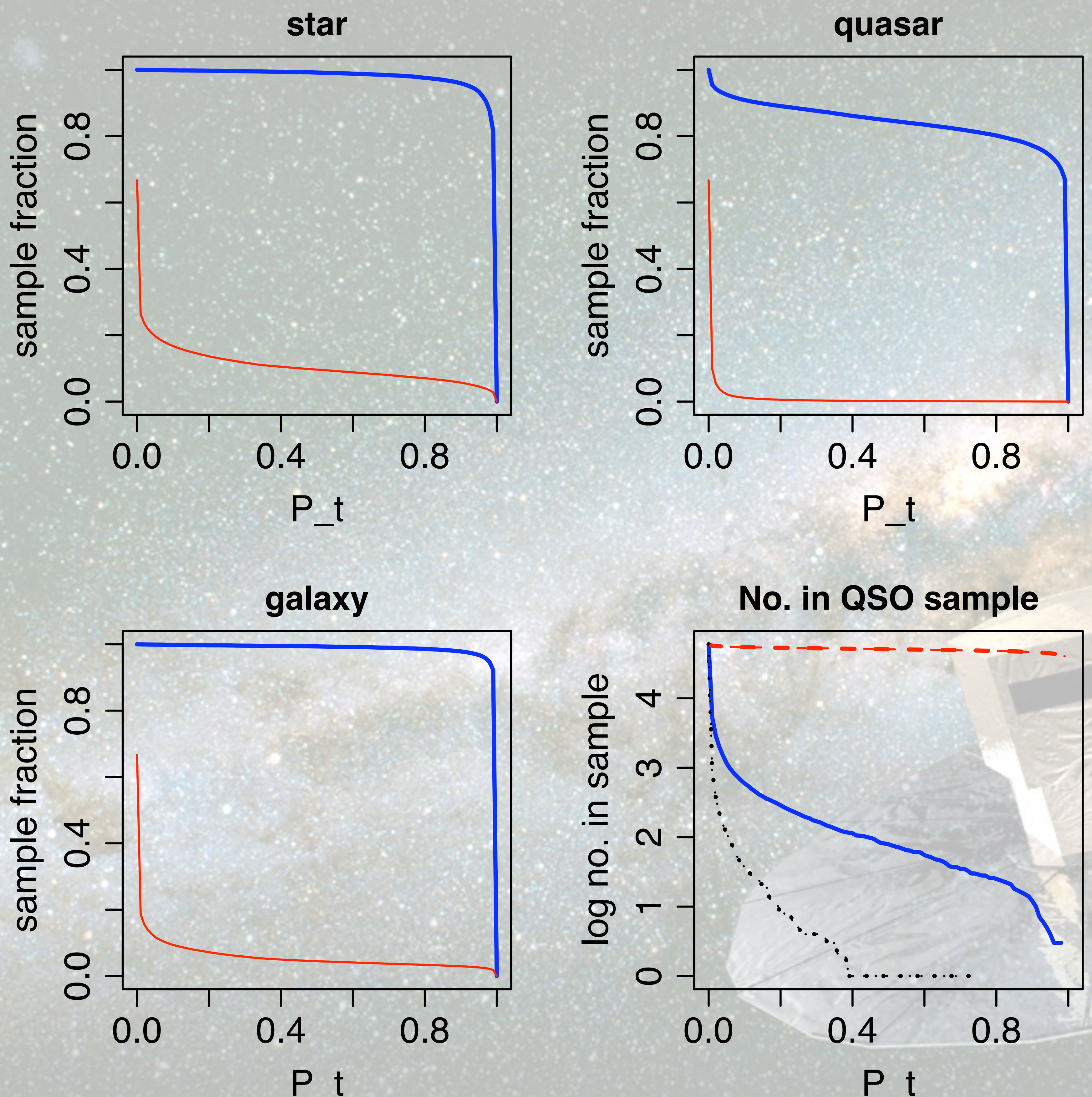
$$\frac{\text{N(truly of class j in sample)}}{\text{N(class j in test set)}}$$

- Sample contamination for class j

$$\frac{\text{N(of all other classes in sample)}}{\text{N(in sample)}}$$

Sample building

Nominal model



blue line is completeness

red line is contamination

Modified model

Target population has quasars rare (prior)

$$\mathbf{f} = (f_{\mathrm{galaxy}}, f_{\mathrm{quasar}}, f_{\mathrm{star}})$$

$$\mathbf{f}^{target} = (1, 0.001, 1)$$

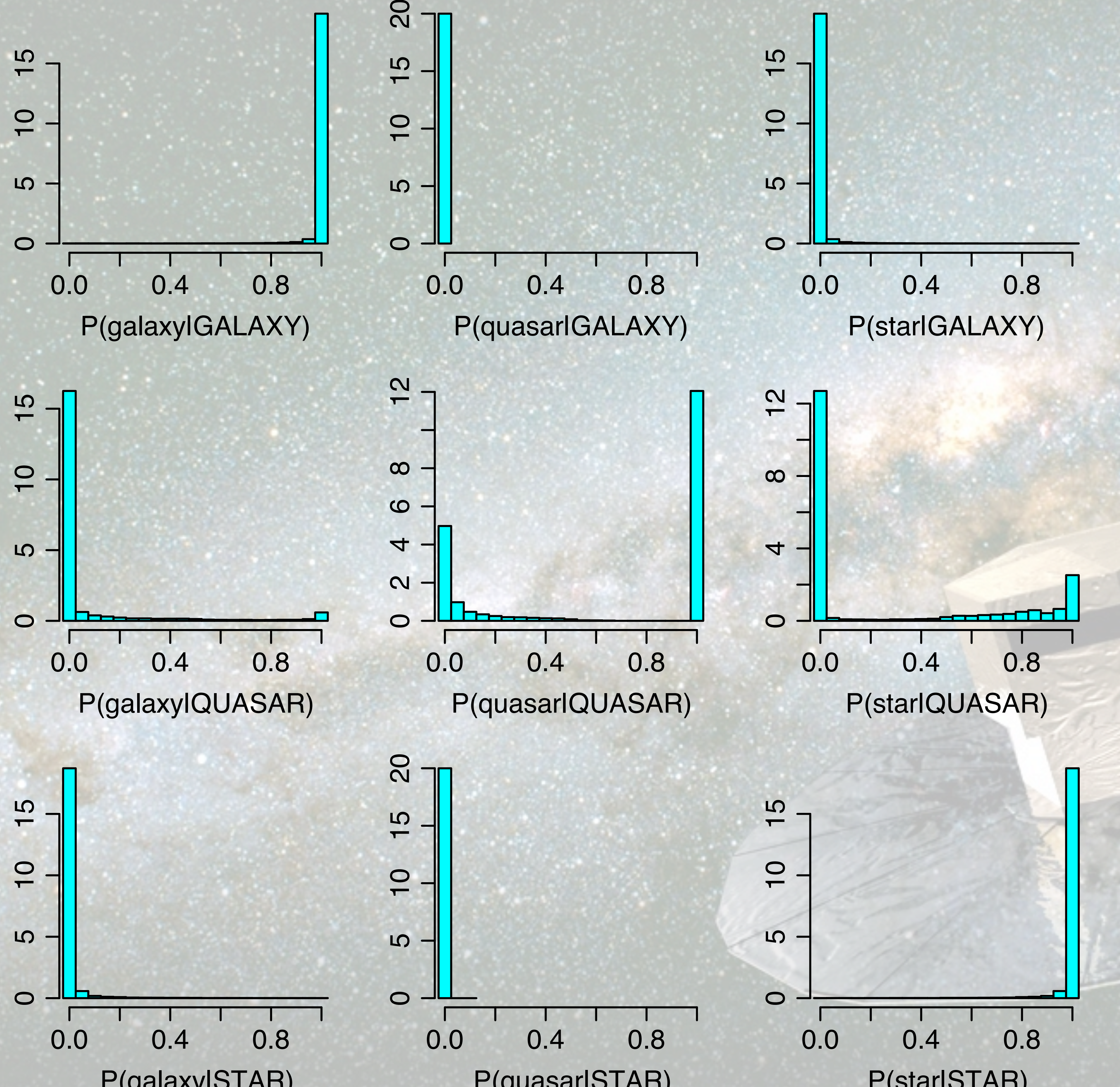
$$P^{mod}(C_j|x_n, \theta^{mod}) = a_n \, P^{nom}(C_j|x_n, \theta^{nom}) \, \frac{f_{i=j}^{target}}{f_{i=j}^{train}}$$

*Test data set not changed, but calculations for completeness and contamination are modified to account for changed **f***

Output probabilities

Modified model

Predicted classes

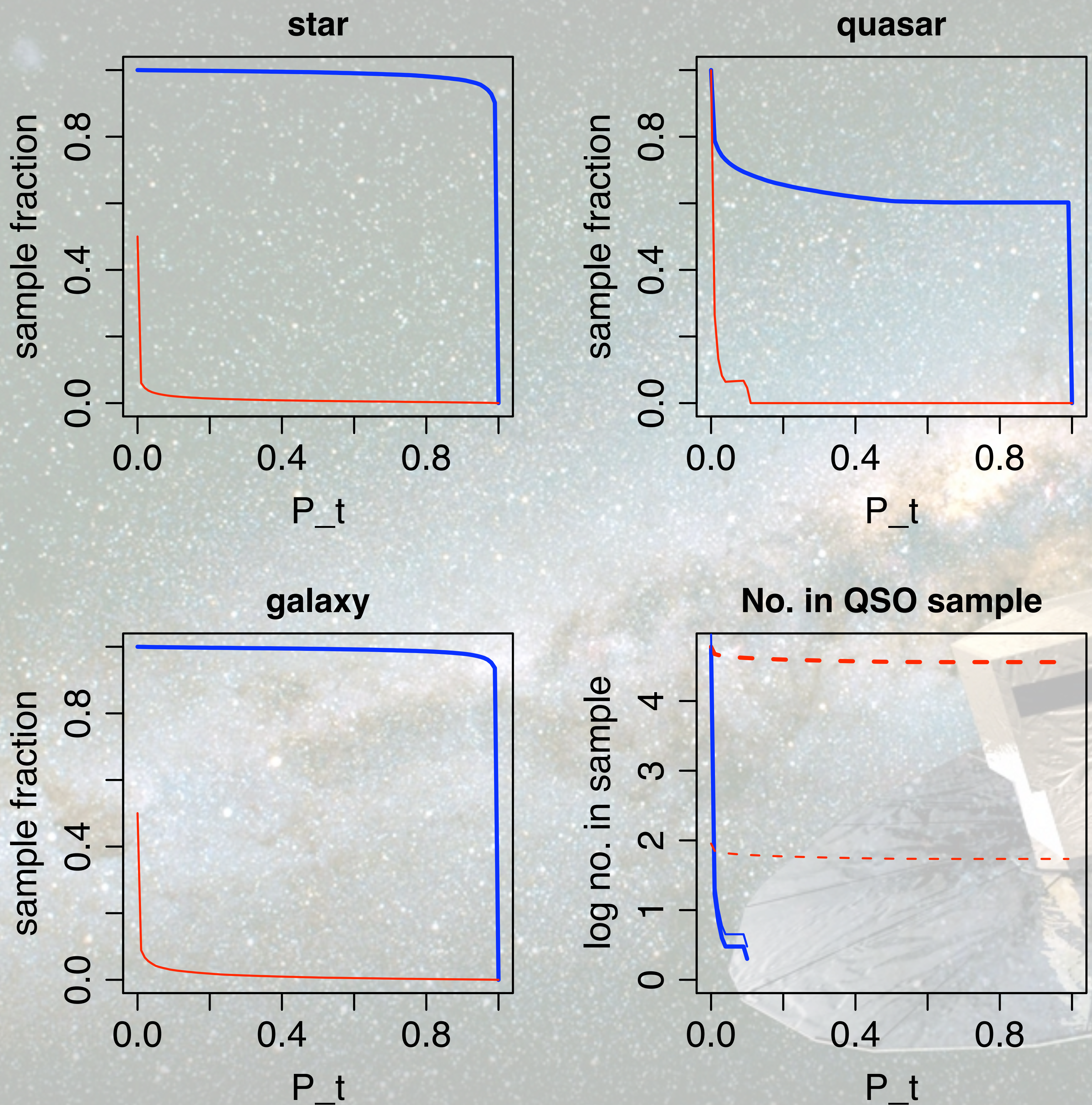


True classes

Coryn Bailer-Jones, MPIA Heidelberg

Sample building

Modified model



blue line is completeness

red line is contamination

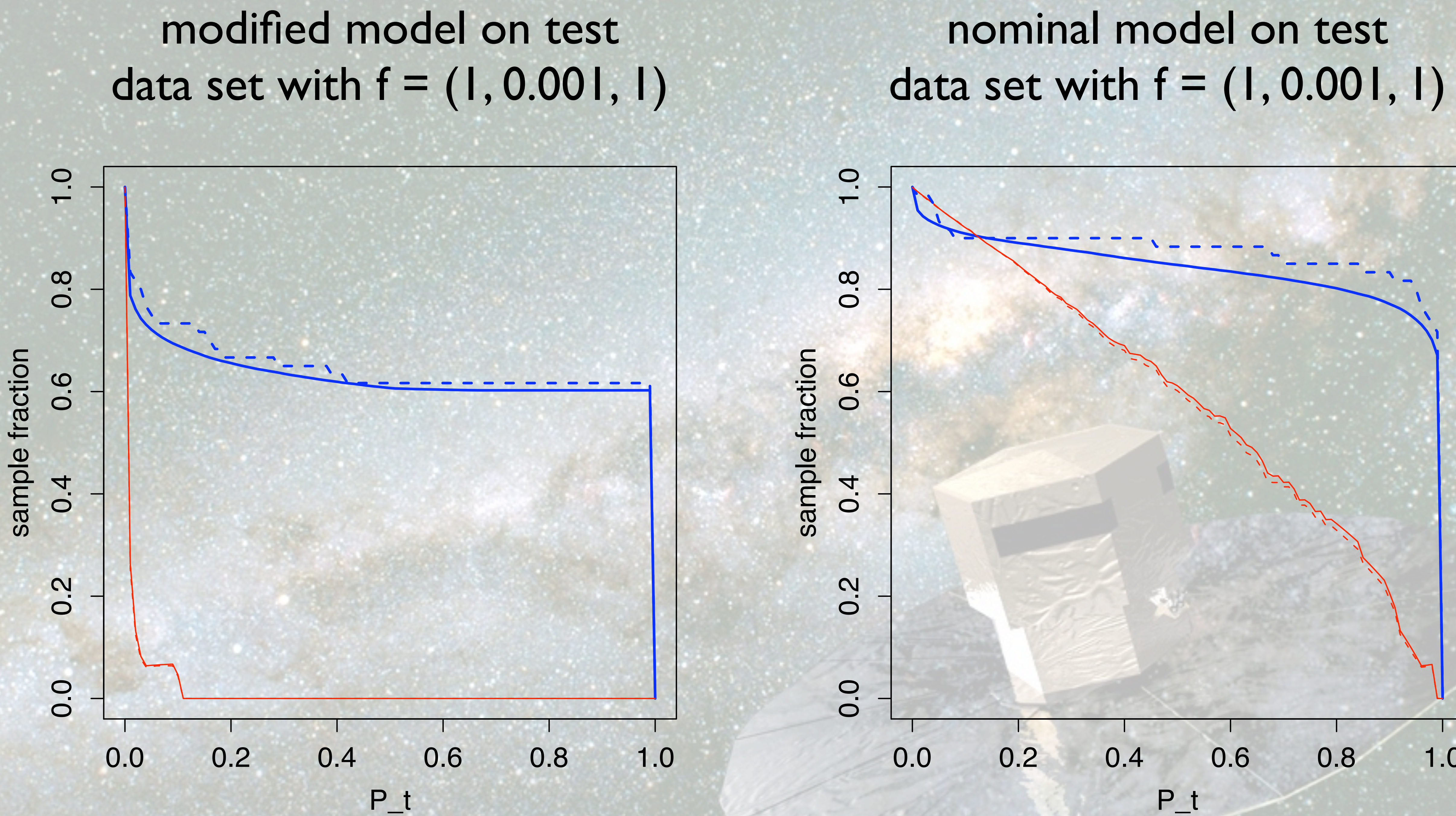
Coryn Bailer-Jones, MPIA Heidelberg

Thresholded confusion matrix

	galaxy	quasar	star	unclassified	Effective fraction
GALAXY	98.97	0.00	0.64	0.73	1.0
QUASAR	6.82	62.00	26.37	8.73	0.001
STAR	0.78	0.00	98.69	1.09	1.0

threshold of P = 0.8 for stars and galaxies
P = 0.2 for quasars

Checking and comparing the models



blue = quasar completeness   red = quasar contamination

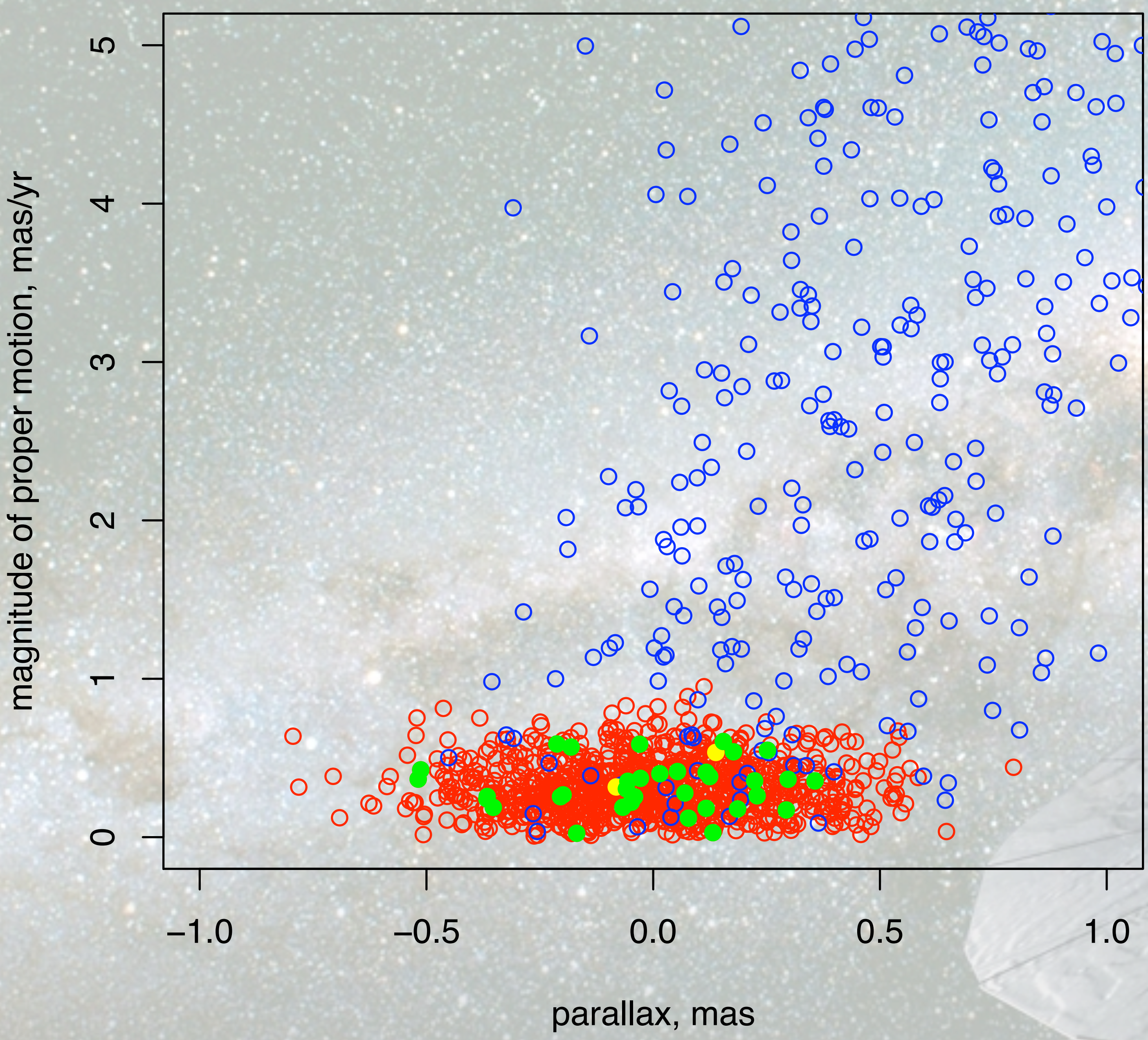
solid = predicted   dashed = measured

Coryn Bailer-Jones, MPIA Heidelberg

The advantages of the modified model

- Zero contamination of the quasar sample with a completeness of 62%
  - simultaneously star and galaxy sample completeness of 99% with low contamination (0.7%)
- Can apply to any target population without retraining
- Using nominal model on a population in which quasars really are rare gives poor results

Further development: add other data



- 2D astrometric data
- Could fit with a mixture model

blue = Galactic objects
red = extragalactic objects

Further development: subclassifiers

BP/RP $\rightarrow$ $P(C_k \mid BP/RP)$ $\rightarrow$ STAR, QSO, GALAXY

μ,ϖ $\rightarrow$ $P(C_k \mid \mu,\varpi)$ $\rightarrow$ GALACTIC, EXTRAGALACTIC

Galactic latitude $\rightarrow$ $P(C_k \mid lat)$ $\rightarrow$ STAR, QSO, GALAXY

G-band magnitude $\rightarrow$ $P(C_k \mid G)$ $\rightarrow$ STAR, QSO, GALAXY

external data $\rightarrow$ $P(C_k \mid ext)$ $\rightarrow$ STAR, QSO, GALAXY

$$P_i(C_j|data) \;=\; P_{bprp}(C_j|x_i)\,P_{astro}(C_j|\varpi_i,\mu_i)\,P_G(C_j|G_i)\,P_{lat}(C_j|l_i)\,P_{ext}(C_j|\cdot)$$

Open issues and questions

1. Is a SVM optimal for this?
    - Which kernel? Tuning? No genuine probabilities
    - KDE works okay for lower-D problems
2. How best to do initial outlier detection?
    - currently use 1-class SVM
3. Data-model mismatch
    - lots of simulated data but imperfect models $\Rightarrow$ missing variance (calibration problem); *covariate shift*
    - iteratively build training data sets? semi-supervised methods?

Classifier comparison

	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 in %
class assignment by highest probability
K=3 or 4 class problem

Open issues and questions

4. Dimensionality reduction?
5. Is a posterior probability sufficient as goodness-of-fit?
6. Class discovery / novelty detection
   - which unsupervised methods?
   - how to feed back into supervised classifiers?
7. How to define the training data distribution for *regression* problems?

Summary and Conclusions

- Assign probabilities; use thresholds to build ad hoc samples
- Class fractions in training data can bias classifier
- Take into account priors on target population
  - failure to do so gives inferior results
  - train model once on equal class fractions then adjust probabilities
- 62% quasar sample completeness with zero contamination
- see arXiv:0809.3373 (MNRAS, in press)