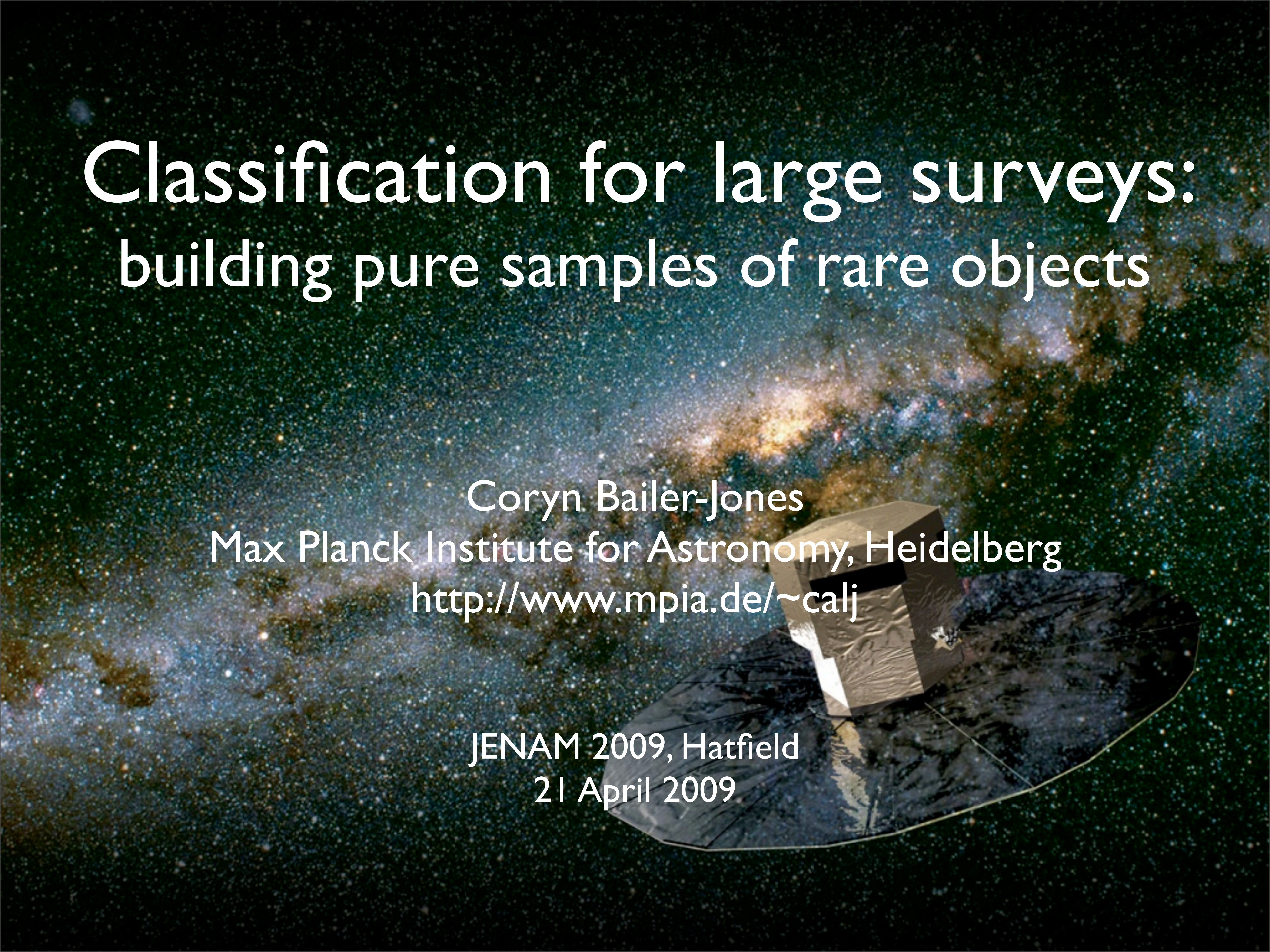


Classification for large surveys:

building pure samples of rare objects

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

JENAM 2009, Hatfield
21 April 2009

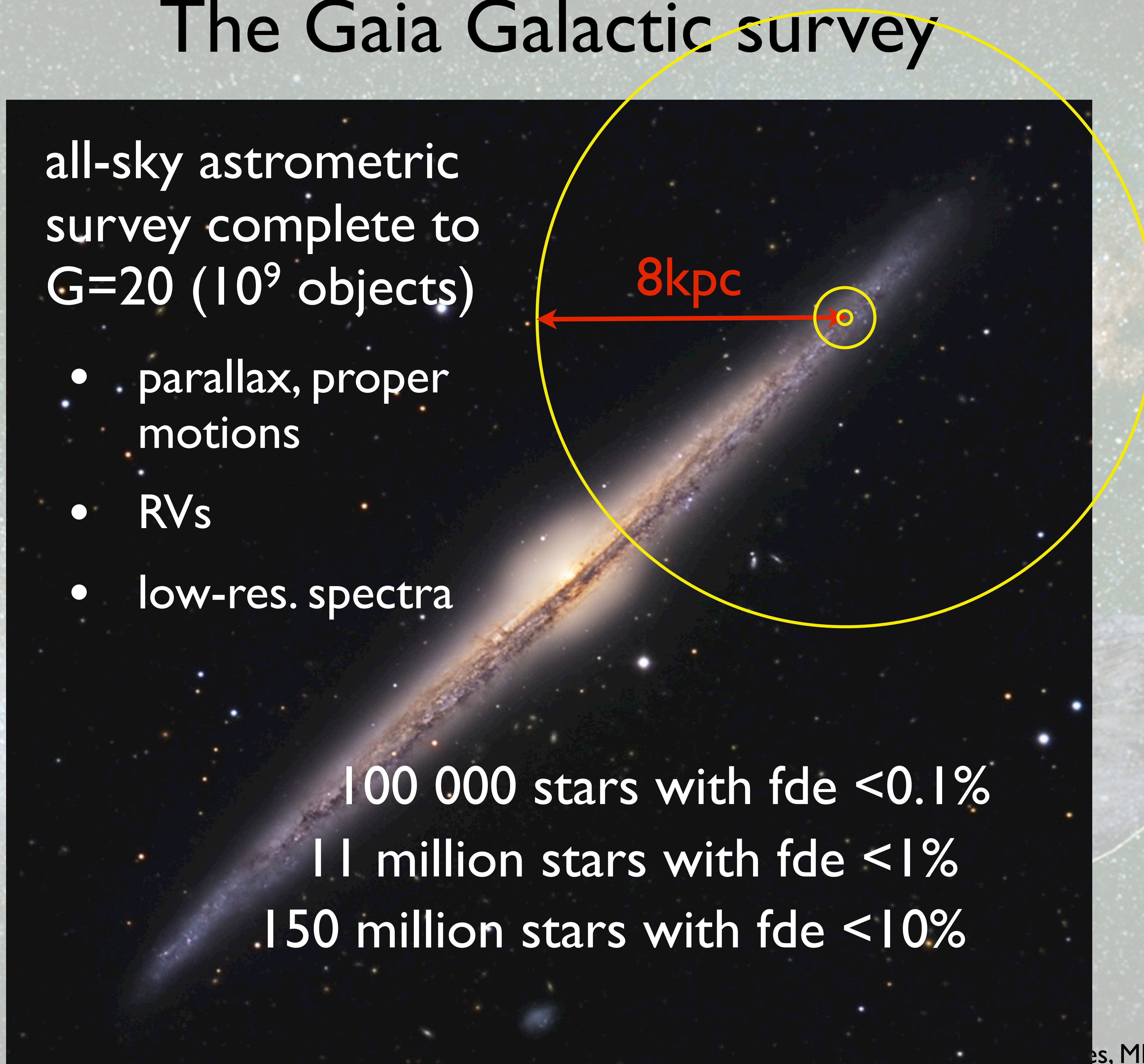
Large surveys

- Goals
  - object classification
  - identification of specific, maybe rare, objects
  - discovery of new types of objects
- Characteristics
  - blind, but we have prior information
  - can usually build models of known objects $\Rightarrow$ supervised learning

The Gaia Galactic survey

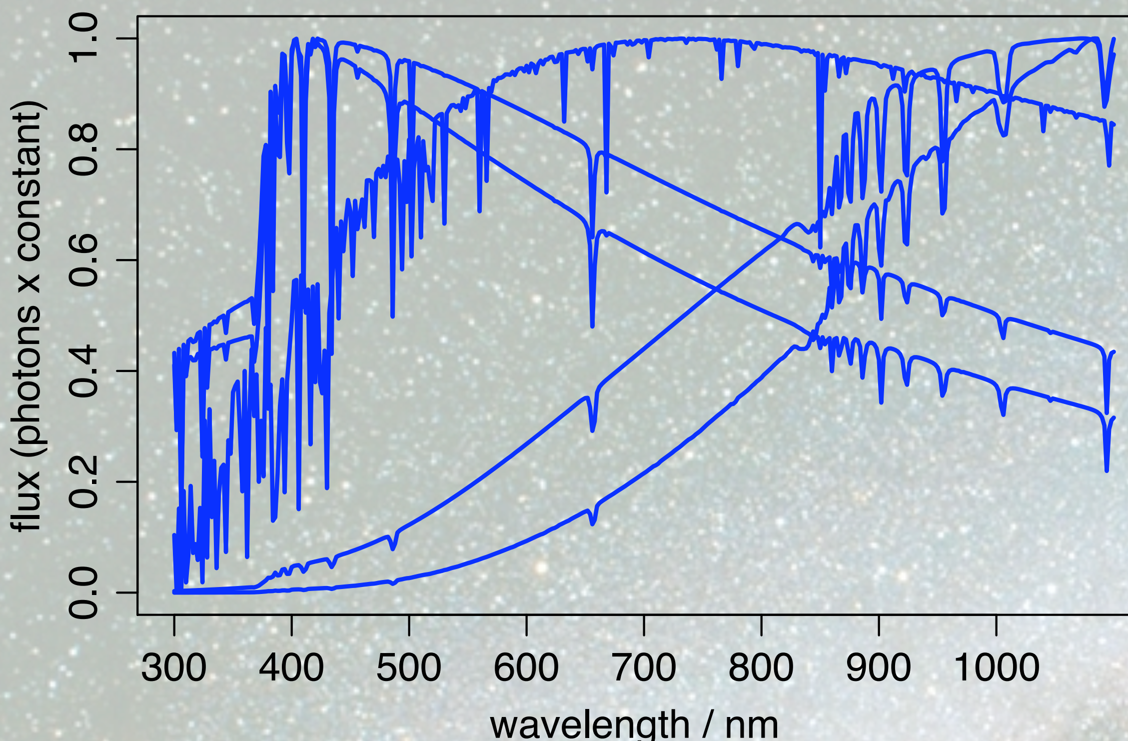
all-sky astrometric survey complete to G=20 (10^9 objects)

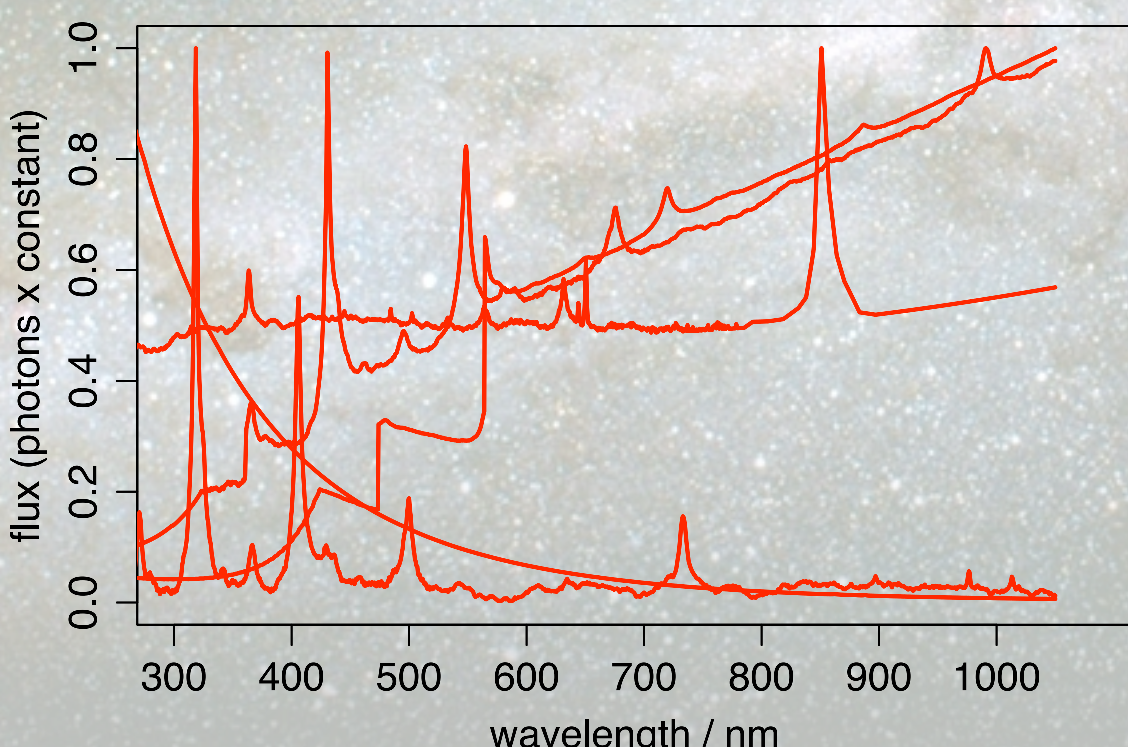
- parallax, proper motions
- RVs
- low-res. spectra



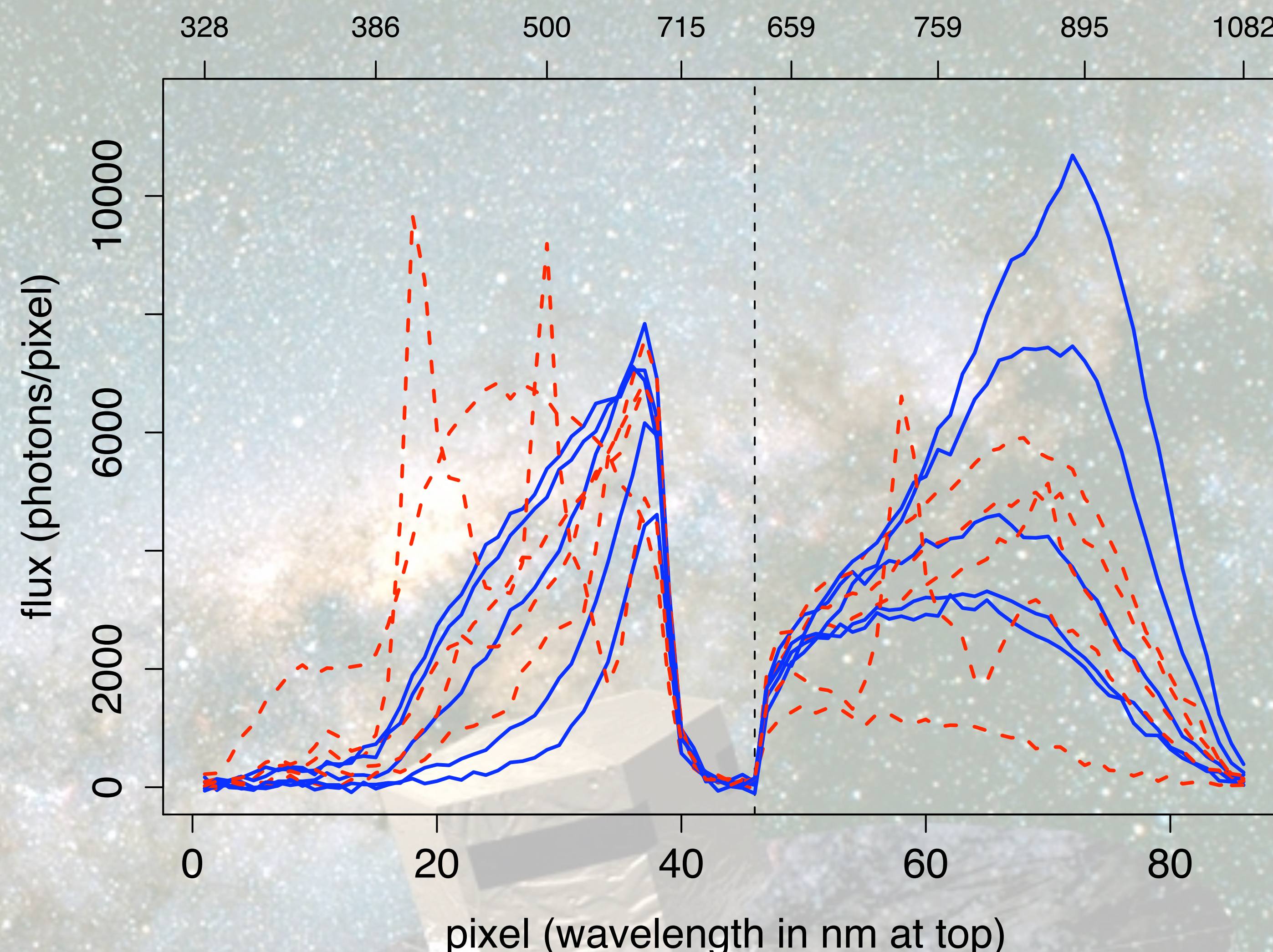
100 000 stars with fde <0.1%
11 million stars with fde <1%
150 million stars with fde <10%

Input spectra





Gaia spectra



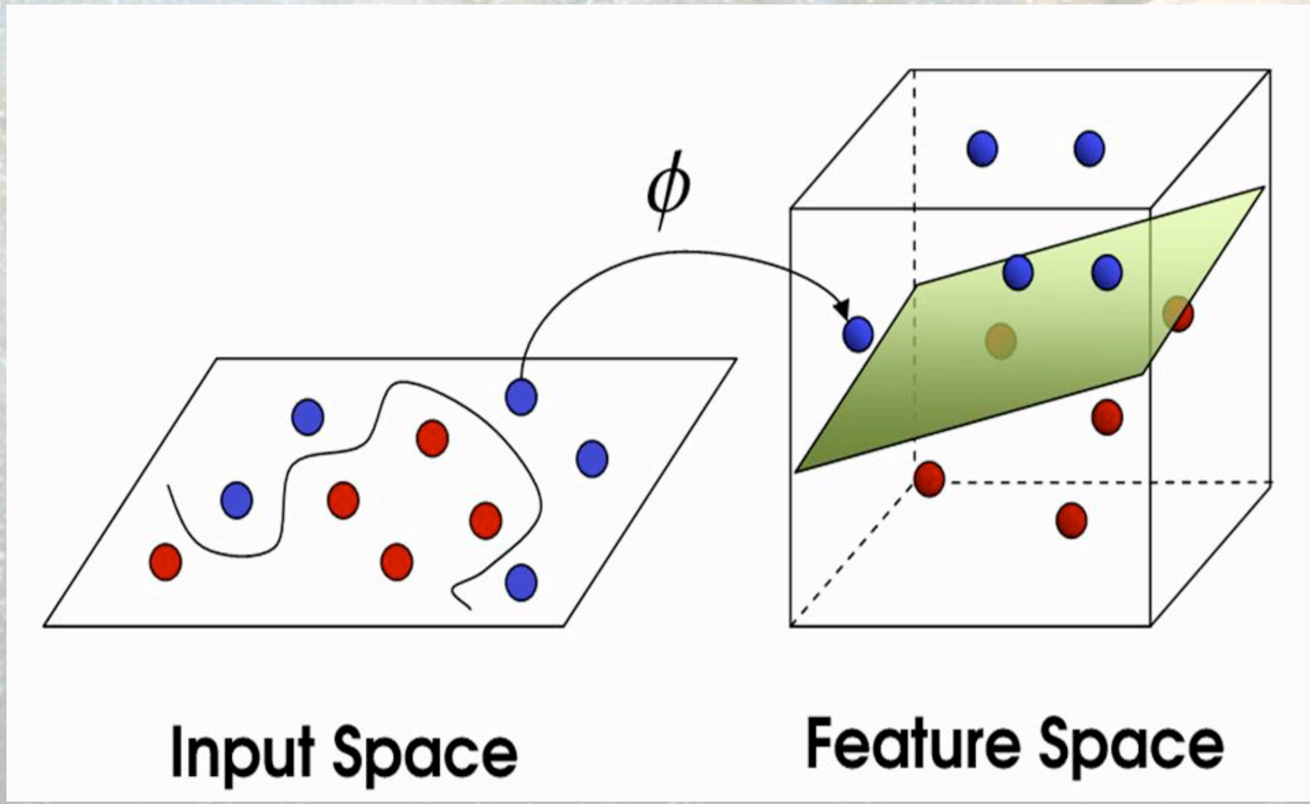
blue = stars

red / dashed = quasars

Classification engine: SVM

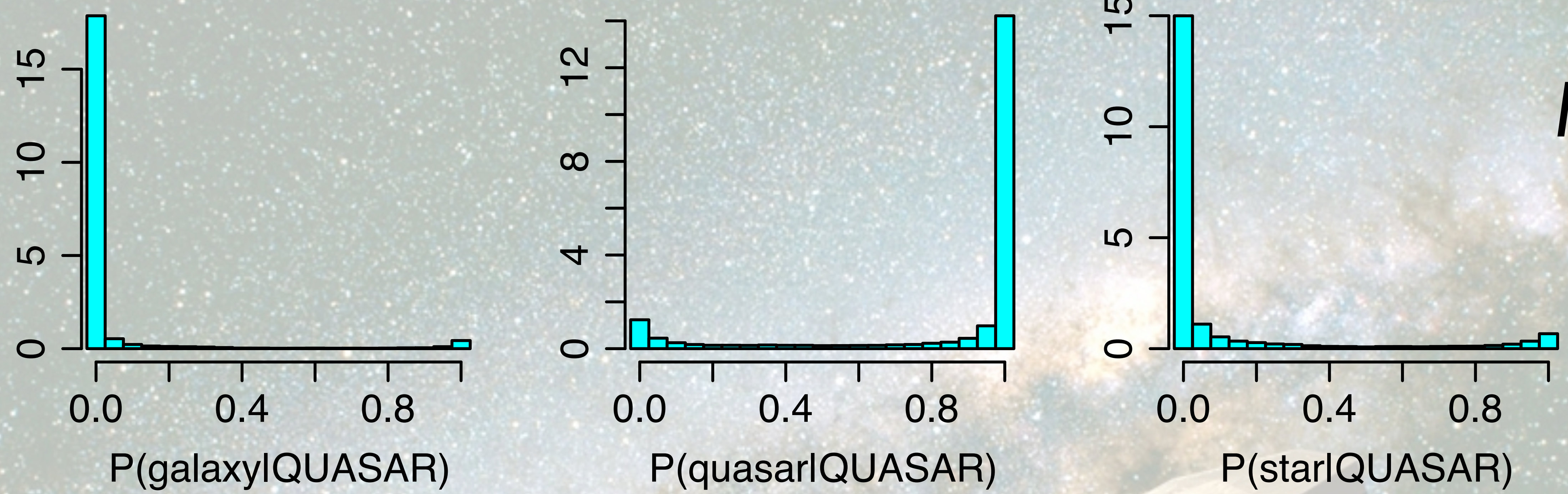
- 3 class problem (star-galaxy-quasar)
- outputs are class probabilities
- train: 5000 of each class test: 60 000 of each class

image credit: www.imres.tech.in



libSVM (Java)

Output probabilities

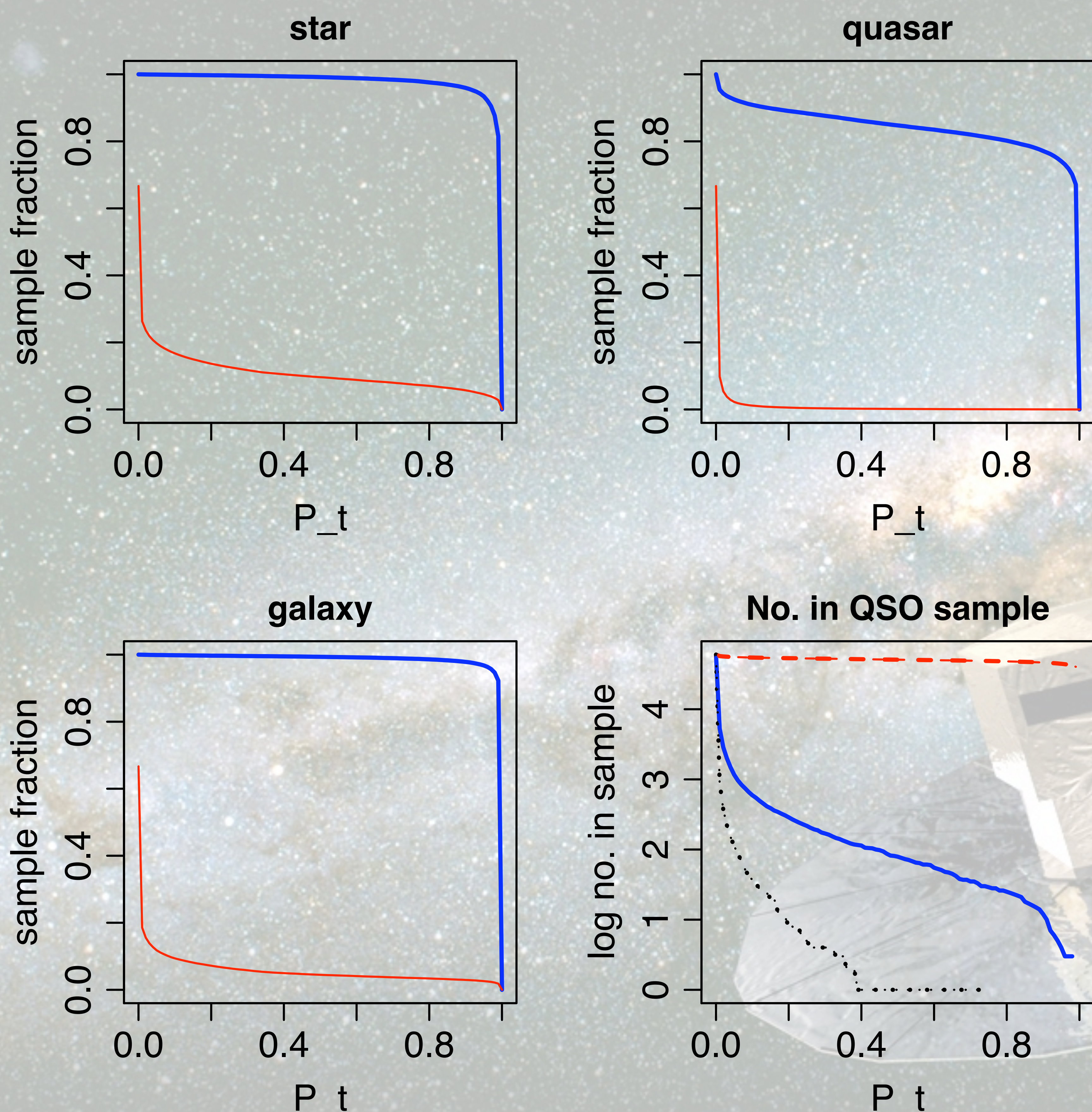


Nominal model

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

Here: assign objects to class with largest probability

Sample building



Nominal model
(equal numbers of objects per class)

blue line is completeness

red line is contamination

Bayesian learning

posterior likelihood prior

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

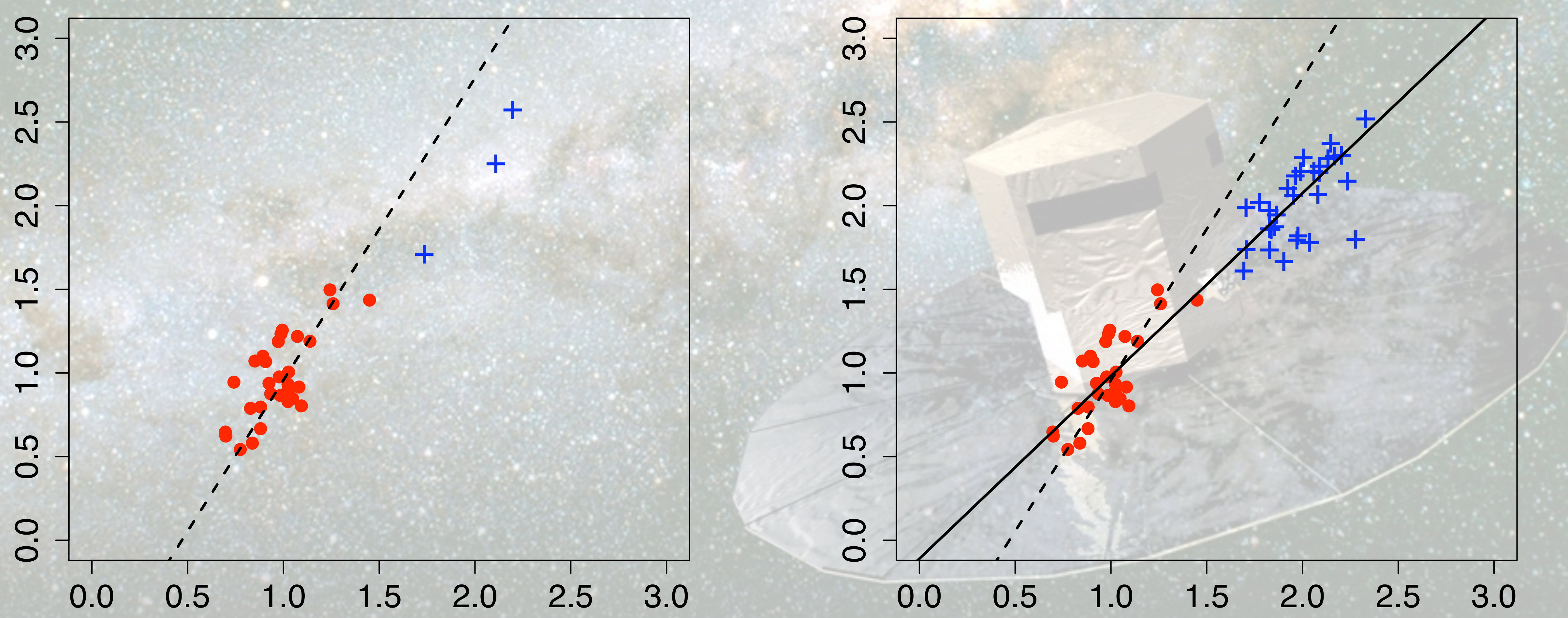
class
e.g. star, galaxy

data
(spectrum)

model

Class imbalance problem

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



Coryn Bailer-Jones, MPIA Heidelberg

The modified model

- Adjust priors to reflect “class fractions” in target population
- Here, quasars are 1000 times rarer (our prior)

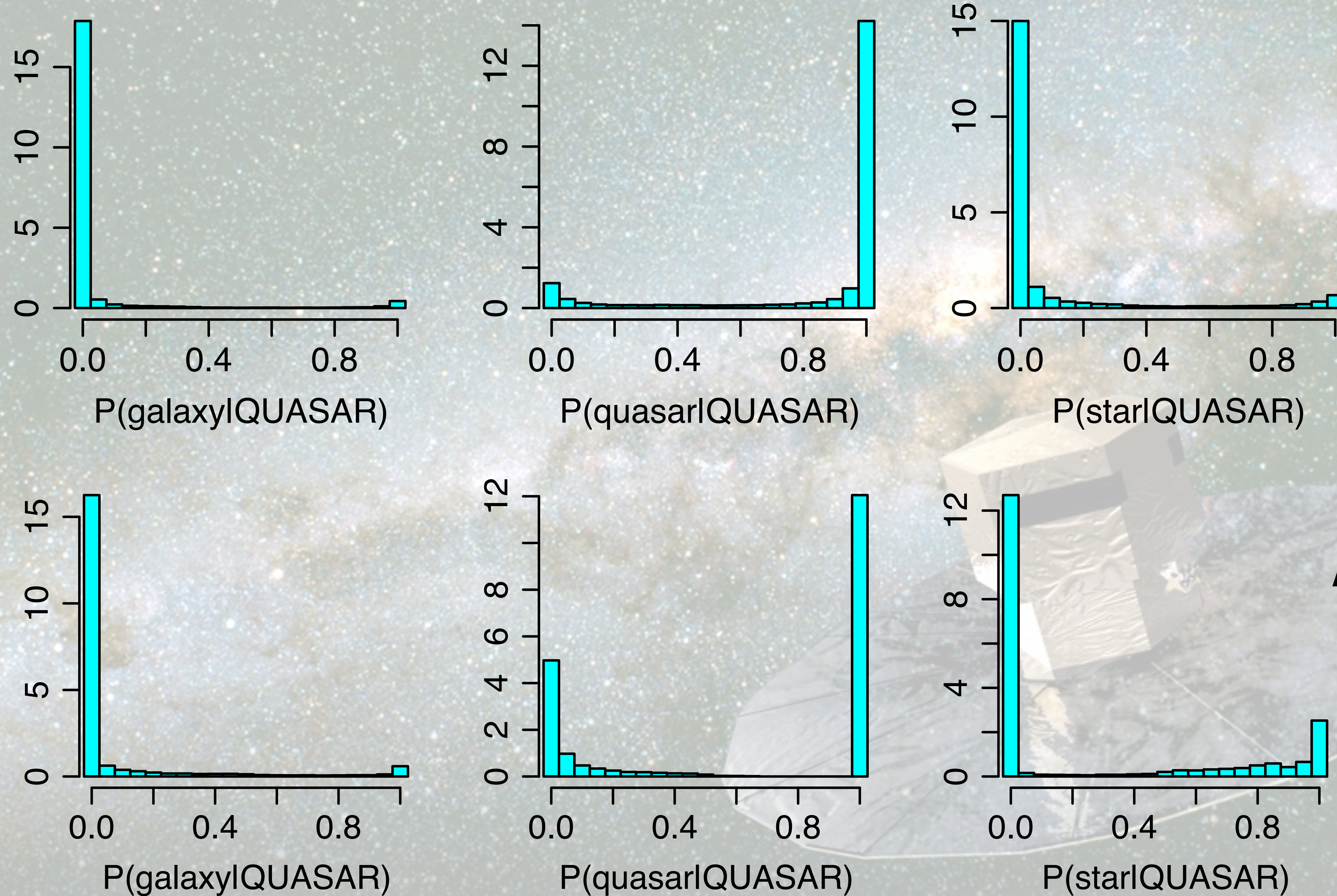
$$\mathbf{f} = (f_{\mathrm{galaxy}}, f_{\mathrm{quasar}}, f_{\mathrm{star}}) \qquad \mathbf{f}^{target} = (1, 0.001, 1)$$

modified posterior nominal posterior priors ratio

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

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

Output probabilities

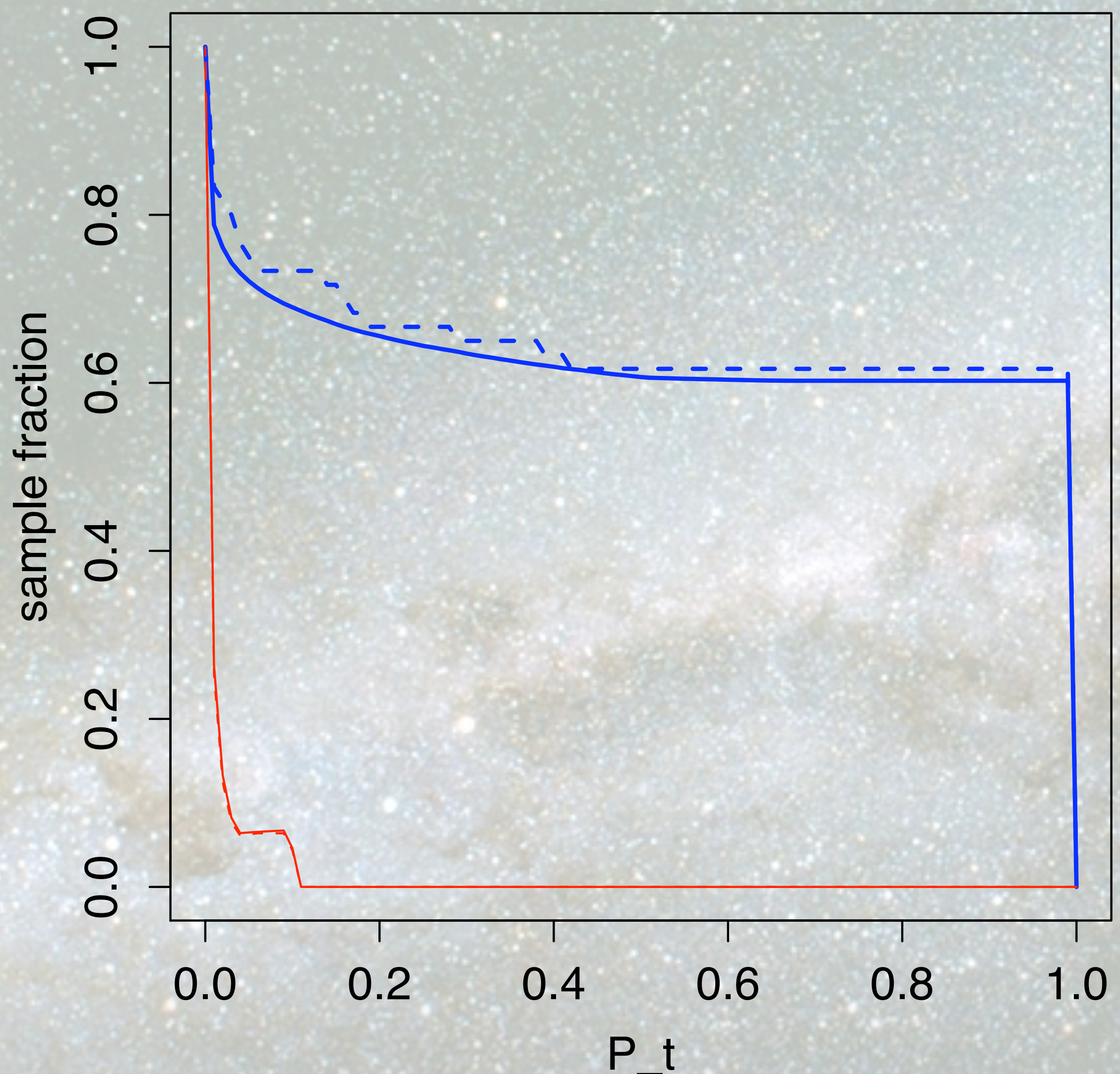


Nominal model

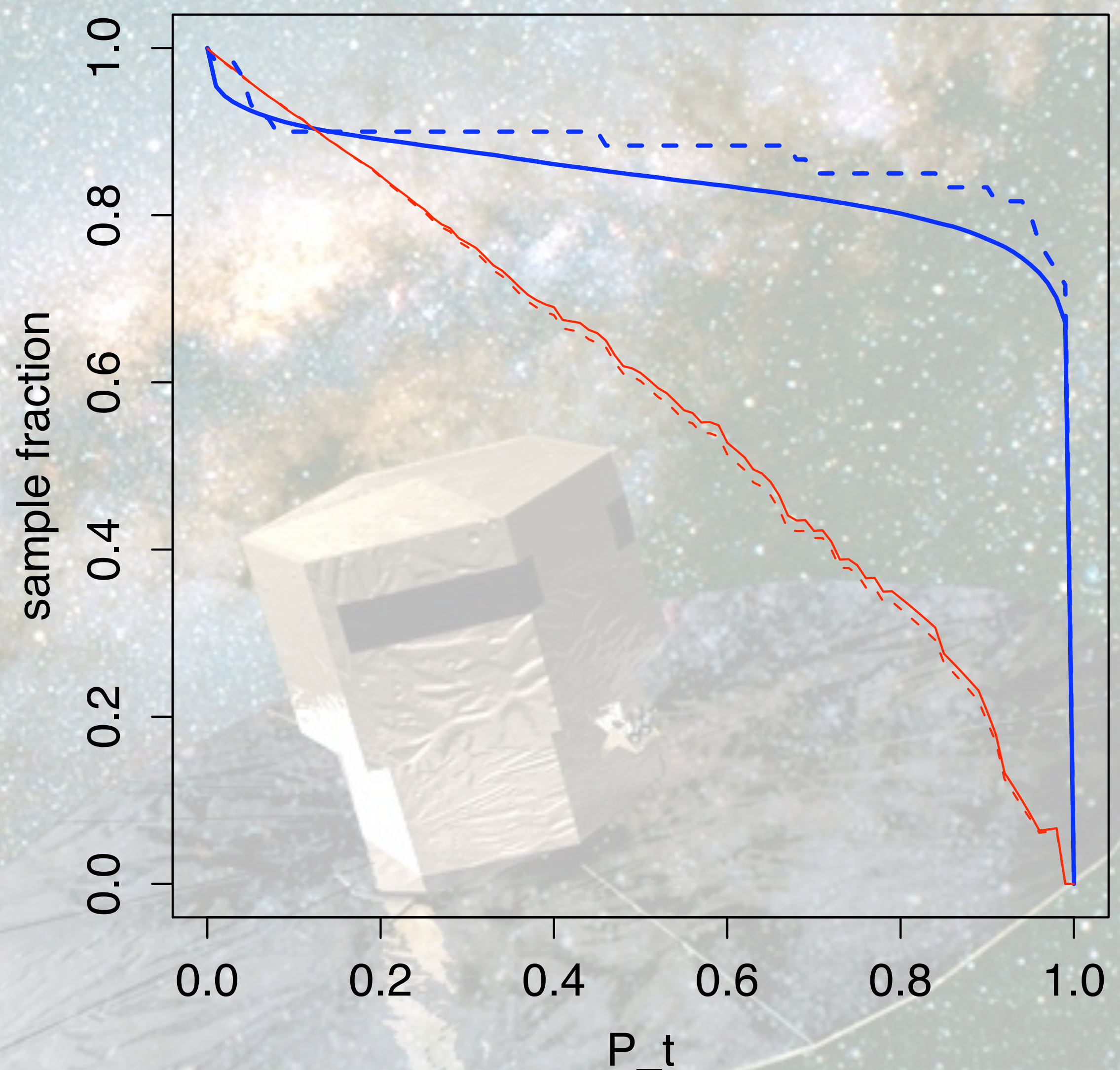
Modified model

The effect on quasar classification

modified model on test
data set with f = (1, 0.001, 1)



nominal model on test
data set with f = (1, 0.001, 1)



blue = quasar completeness red = quasar contamination
solid = predicted dashed = measured

Coryn Bailer-Jones, MPIA Heidelberg

The advantages of the modified model

- With quasars 1000 times rare than stars and galaxies
  - 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

Model-based priors

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

$$P(C_j|\theta) = \sum_{n=1}^{n=N_{test}} \underbrace{P(C_j|x_n,\theta)}_{\text{posterior (model outputs)}} \underbrace{P(x_n|\theta)}_{= \frac{1}{N_{test}}}$$

These calculated priors agree closely with class fractions in target sample

Summary and Conclusions

- Assign probabilities; use thresholds to build ad hoc samples
- Class fractions in training data can bias classifier
  - all models have a prior (may be implicit)!
- Take into account priors on target population
  - train model once on equal class fractions then adjust probabilities
- 62% quasar sample completeness with zero contamination
  - <13 contaminants in sample of 250 000 quasars with Gaia
- Bailer-Jones et al. 2008, MNRAS 391, 1838