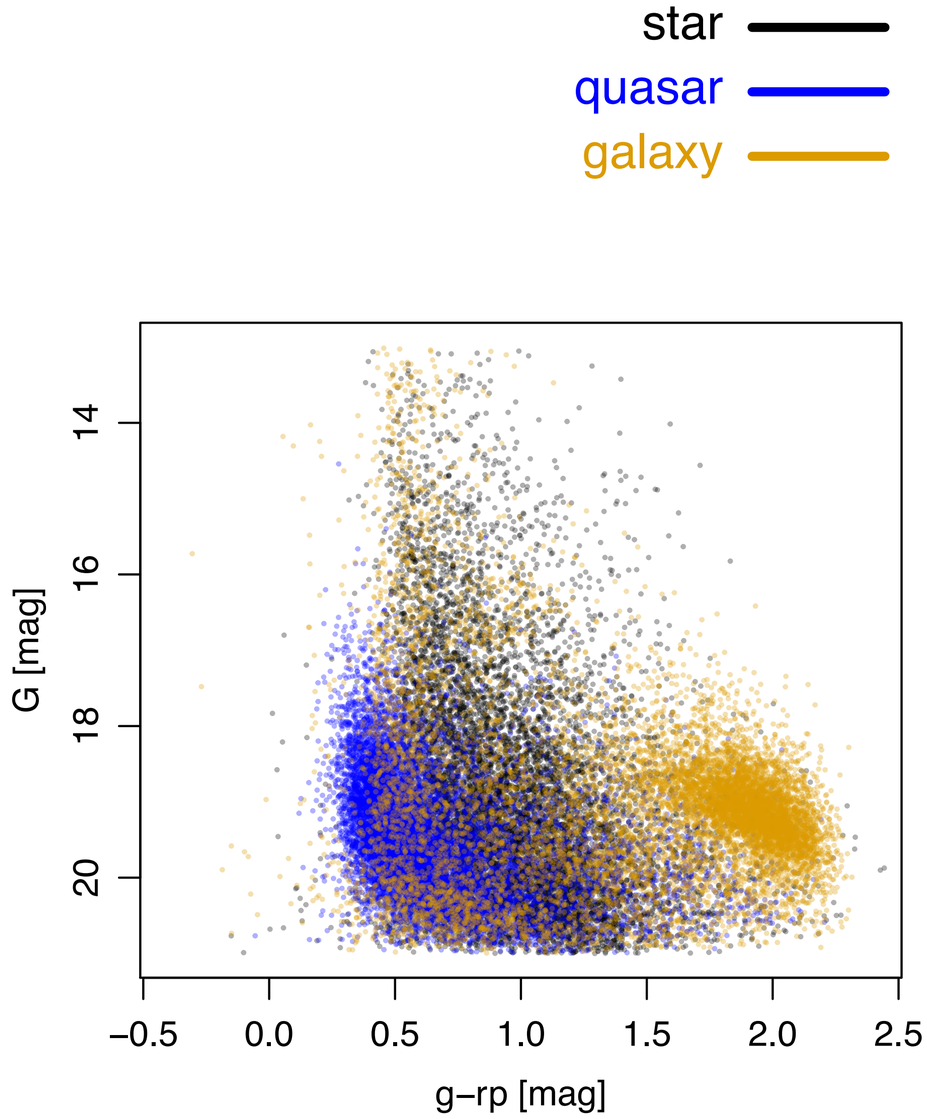
Quasar and galaxy classification in Gaia Data Release 2

Coryn Bailer-Jones
Rene Andrae, Morgan Fouesneau
Max Planck Institute for Astronomy, Heidelberg

ESO AI Conference, 22-26 July 2019



Objective and motivation

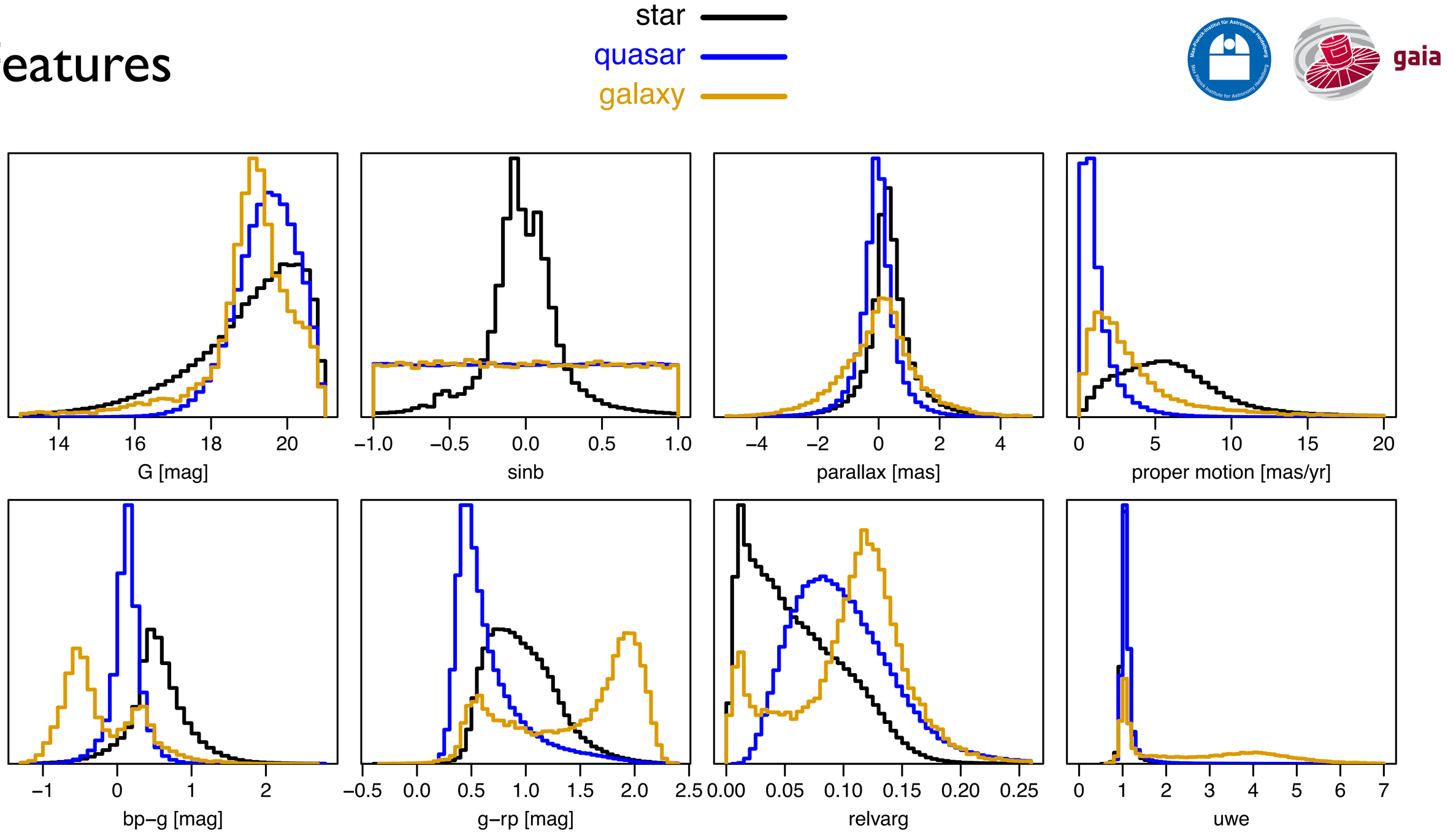
- Gaia DR2: primarily 5-parameter astrometry and 3-band photometry
- Here classify into three classes: `star, quasar, galaxy`
- Use only Gaia DR2 data
  - large, homogeneous data set
  - independent of selection functions from surveys
  - see how well we can do with minimal information (component of Gaia DR3 classifier)
- Probabilistic classifier, empirically trained
  - Gaussian Mixture Model

Coryn Bailer-Jones

Classes are defined by the training set

- Quasars
  - 500 000 from catalogue of Secrest et al. 2015 from All-WISE 2-colour criterion (x-matched)
- Galaxies
  - 25 000 from SDSS-12 with spectroscopic classifications (x-matched)
  - some stellar contamination (although obvious white dwarfs removed)
- Stars
  - a random subset of *all* Gaia sources
  - class is contaminated by quasars and galaxies and should really be called “anonymous”

Features


star
quasar
galaxy
G [mag]
sinb
parallax [mas]
proper motion [mas/yr]
bp−g [mag]
g−rp [mag]
relvarg
uwe

You must accommodate class imbalance!

- Quasars and galaxies are much rarer than stars
- Cannot accommodate this by setting class fractions in training set

1) Adjust classification probabilities to reflect *class prior* $\quad P_k \rightarrow \frac{1}{Z}\pi_k P_k$
2) Class fractions in test set must also reflect class prior
   - can use any class fraction you like and then adjust confusion matrix

We use class prior of $\quad (\pi_{\rm star}, \pi_{\rm qso}, \pi_{\rm gal}) = (3000, 30, 1)$

Coryn Bailer-Jones

Results on test set: confusion matrix

true class		assigned class: star	quasar	galaxy	completeness	equal prior: completeness
	STAR	228265.6	603.7	138.3	0.9968	0.9399
	QUASAR	509.1	1773.6	7.4	0.7745	0.9595
	GALAXY	150.3	50.3	181.1	0.4744	0.6522
	purity	0.9971	0.7306	0.5541		
equal prior:	purity	0.9994	0.1998	0.0469		

classes assigned by maximum probability

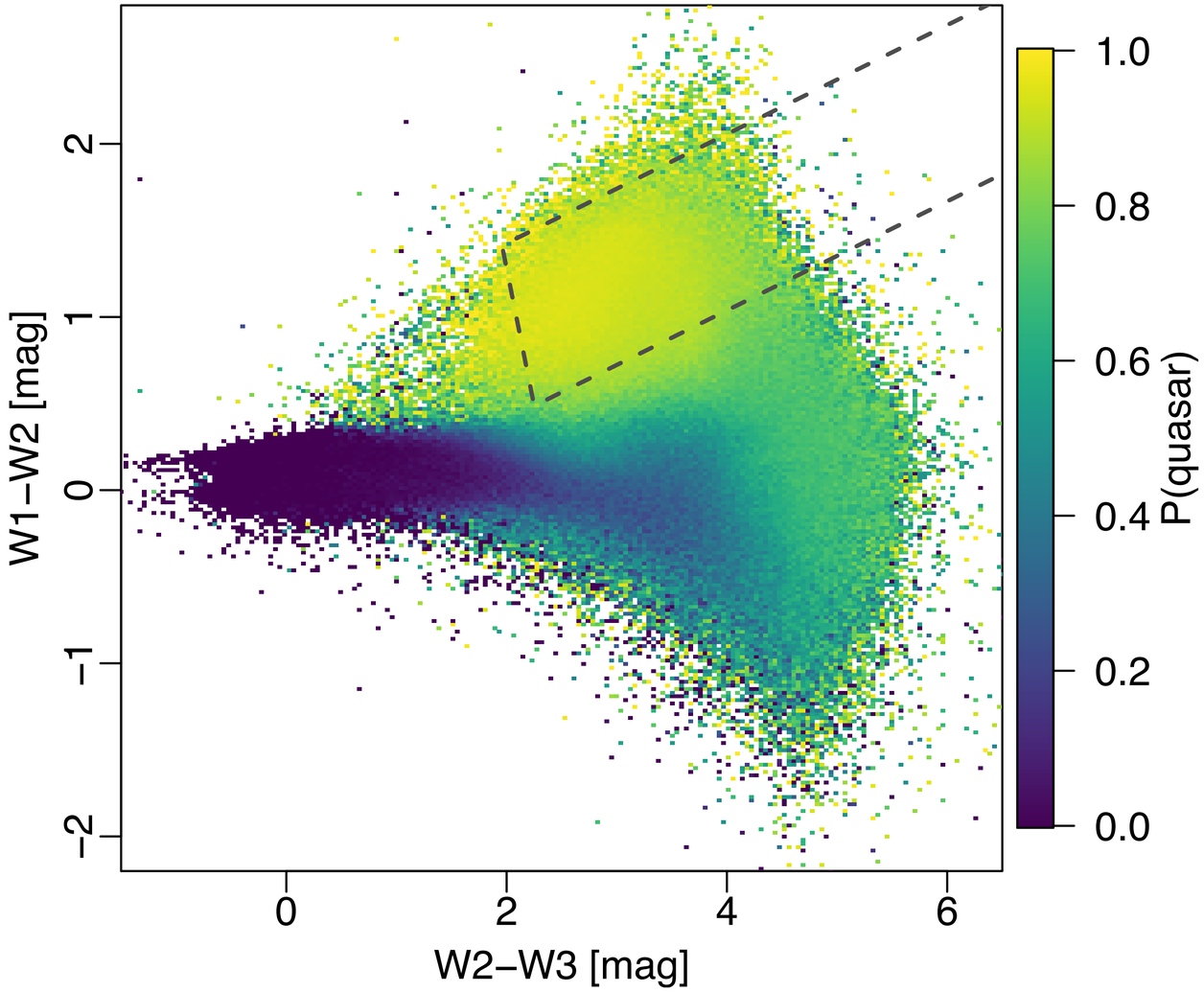
Coryn Bailer-Jones

Application to all of Gaia DR2

- 1.22 billion sources will all 8 features with G>13 mag
- Number of objects found with P>0.5:
  - quasars: 3.6 million
  - galaxies: 0.7 million

Coryn Bailer-Jones

Quasars: distribution in WISE colours



dashed box is 2-colour selection criterion from Mateos et al. 2012

Coryn Bailer-Jones

Summary

- Empirical classification of Gaia-DR2 into 3 classes using only Gaia-DR2 data
- Gaussian Mixture Model using 8 weakly-discriminating features
- Not accommodating class imbalance gives both incorrect class probabilities and optimistic performance predictions
- Performance on test set
  - quasars: completeness and purity around 0.75
  - galaxies: completeness and purity around 0.50
  - not bad considering quasars assumed to be 100x rarer than stars galaxies 3000x rarer
- Catalogue of 3.6 million quasar candidates with P>0.5