Active Learning with Astronomical Data

Alasdair Tran
u4921817

Supervisors:
Cheng Soon Ong, Christian Wolf, Justin Domke

19 June 2015

Sloan Digital Sky Survey

Photometric measurements of 800 million objects, out of which 3 million objects are spectroscopically labelled.

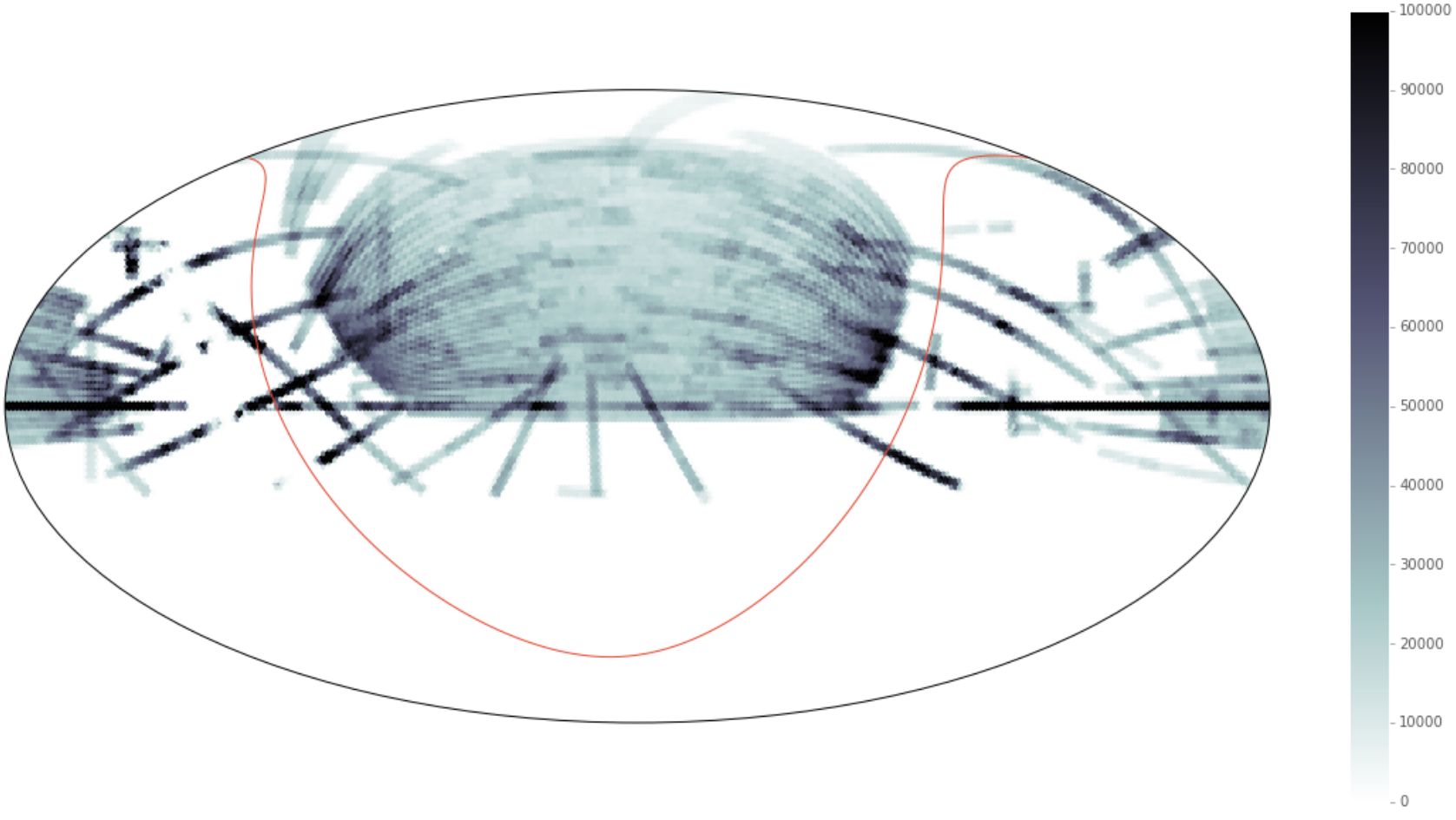

Figure: The coverage of the Sloan survey.

Photometry vs Spectroscopy

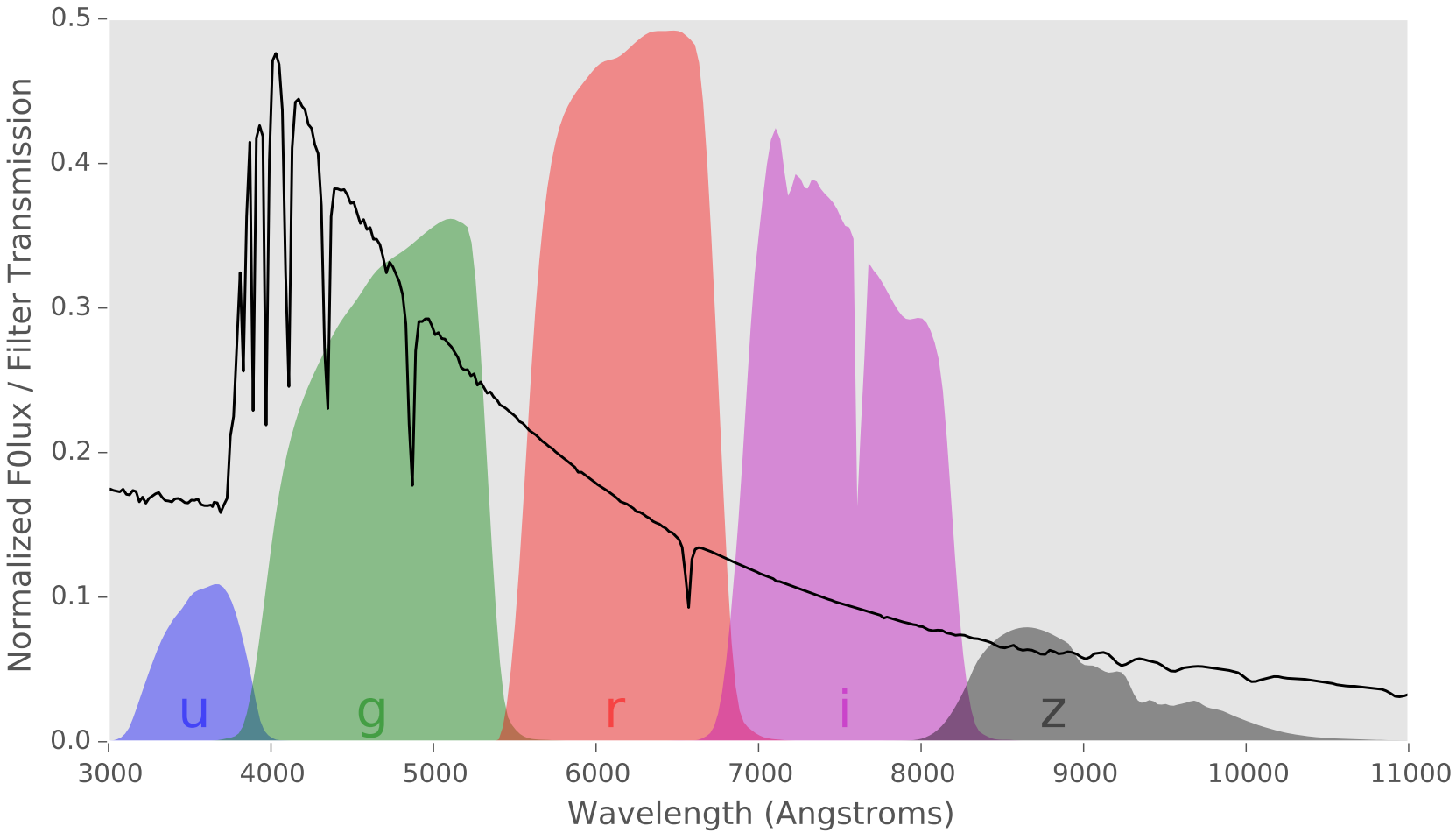

Figure: SDSS Filters and Vega Spectrum

Improving the Features

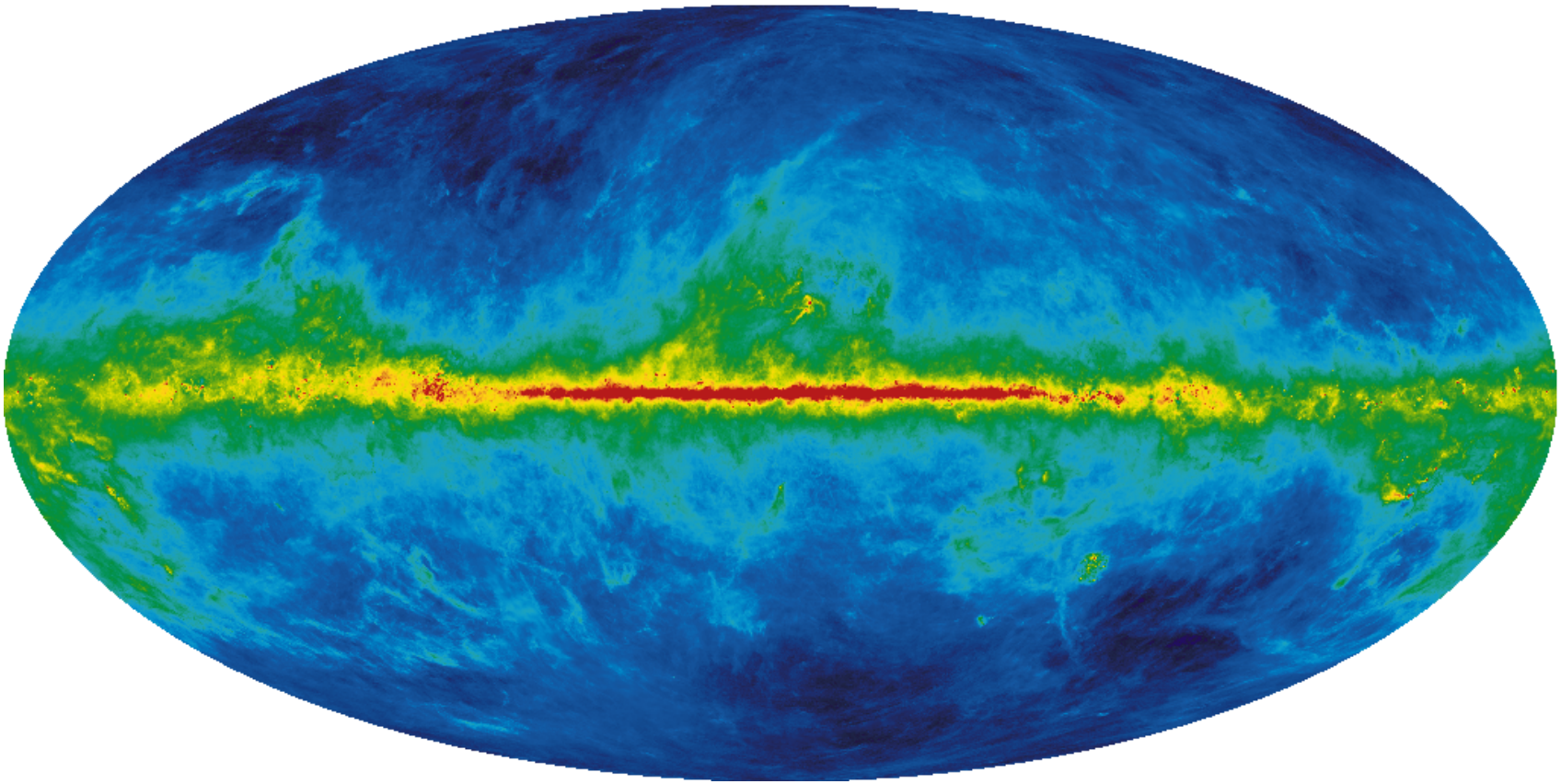

Figure: Map of Galactic Reddening, E(B-V), SFD (1998)[1]

[1] Image from LAMBDA.

Improving the Features

Magnitudes measurements:

- u-band
- g-band
- r-band
- i-band
- z-band

Colour measurements:

- $u - g$ index
- $g - r$ index
- $r - i$ index
- $i - z$ index

Colour indices are independent of distance.

Training Data

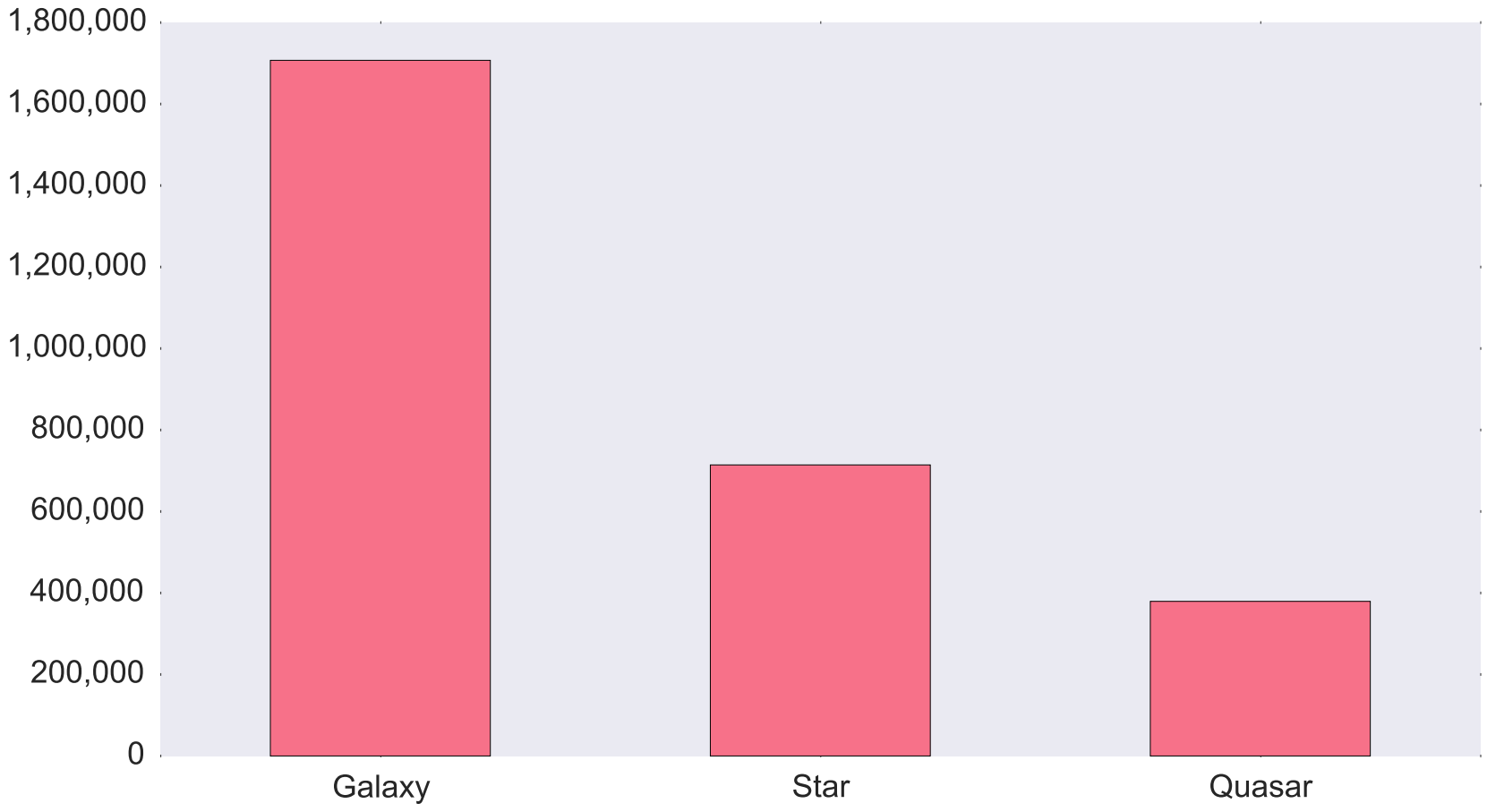

Figure: Distribution of Classes in the Training Set.

Learning Curves

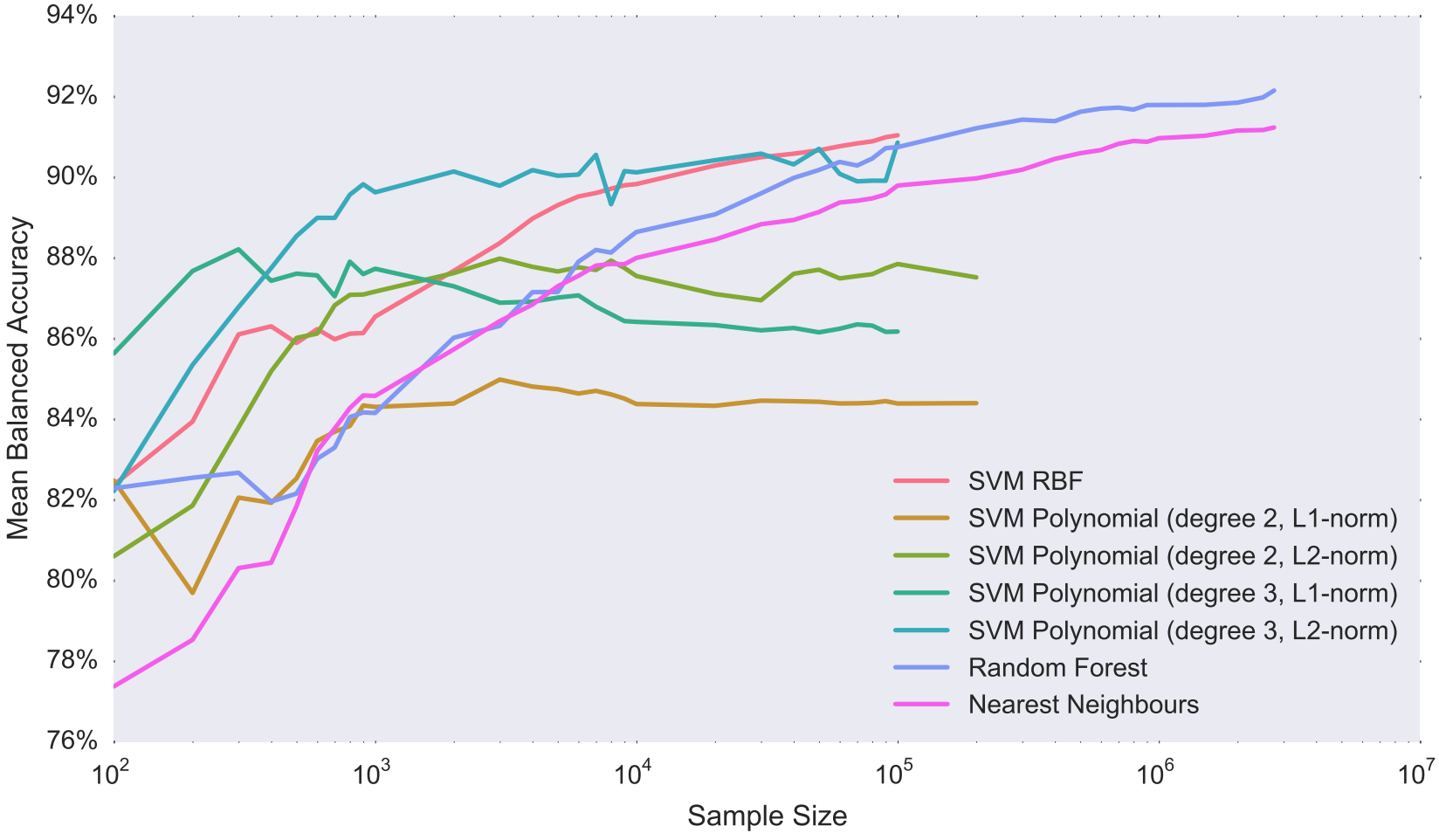

Figure: Learning Curves of Various Classifiers.

Predicting with Test Data

99% of galaxies are correctly classified.

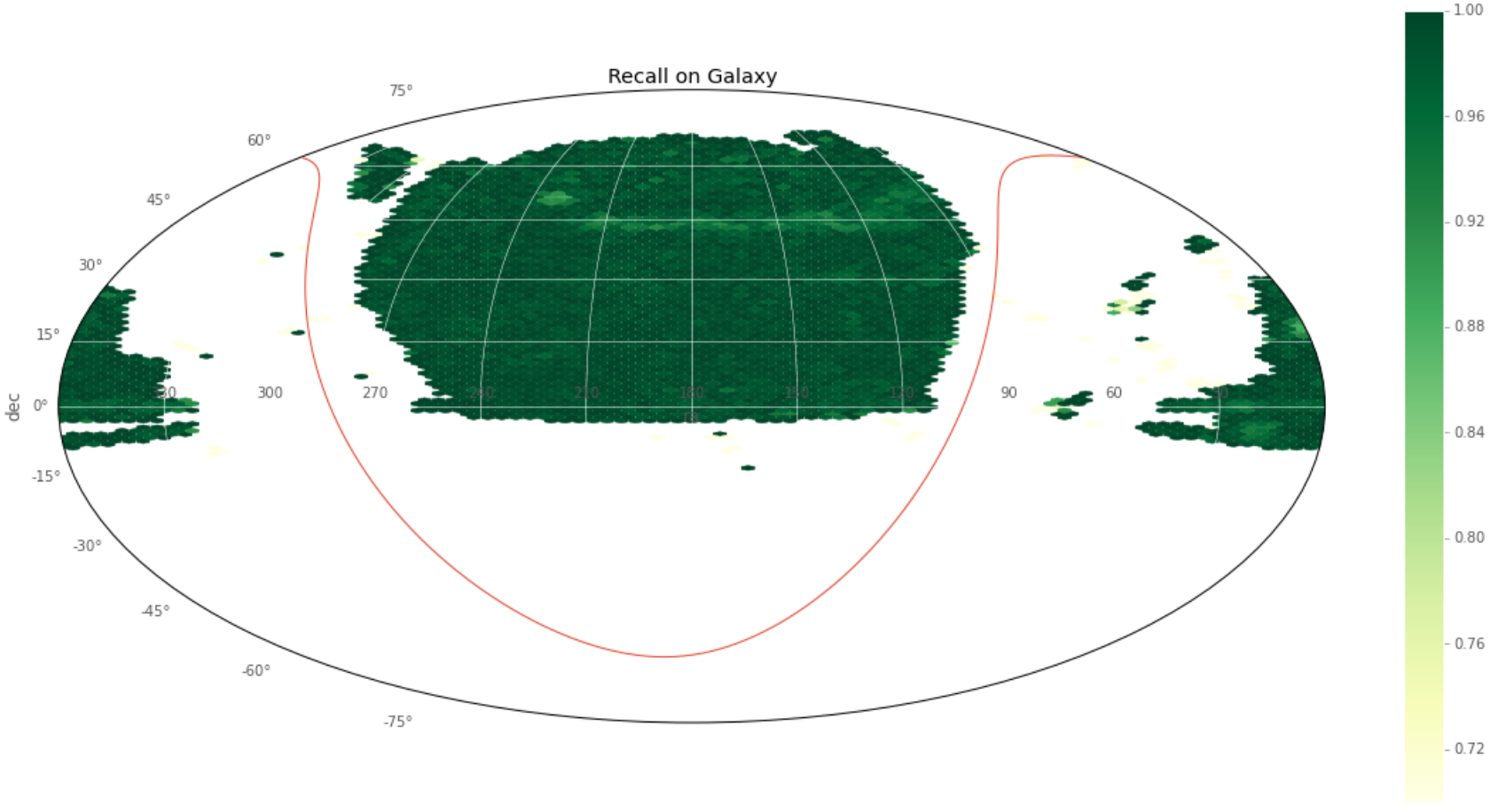

Figure: Random Forest's Recall on Galaxies.

Predicting with Test Data

6% of stars are misclassified as quasars.

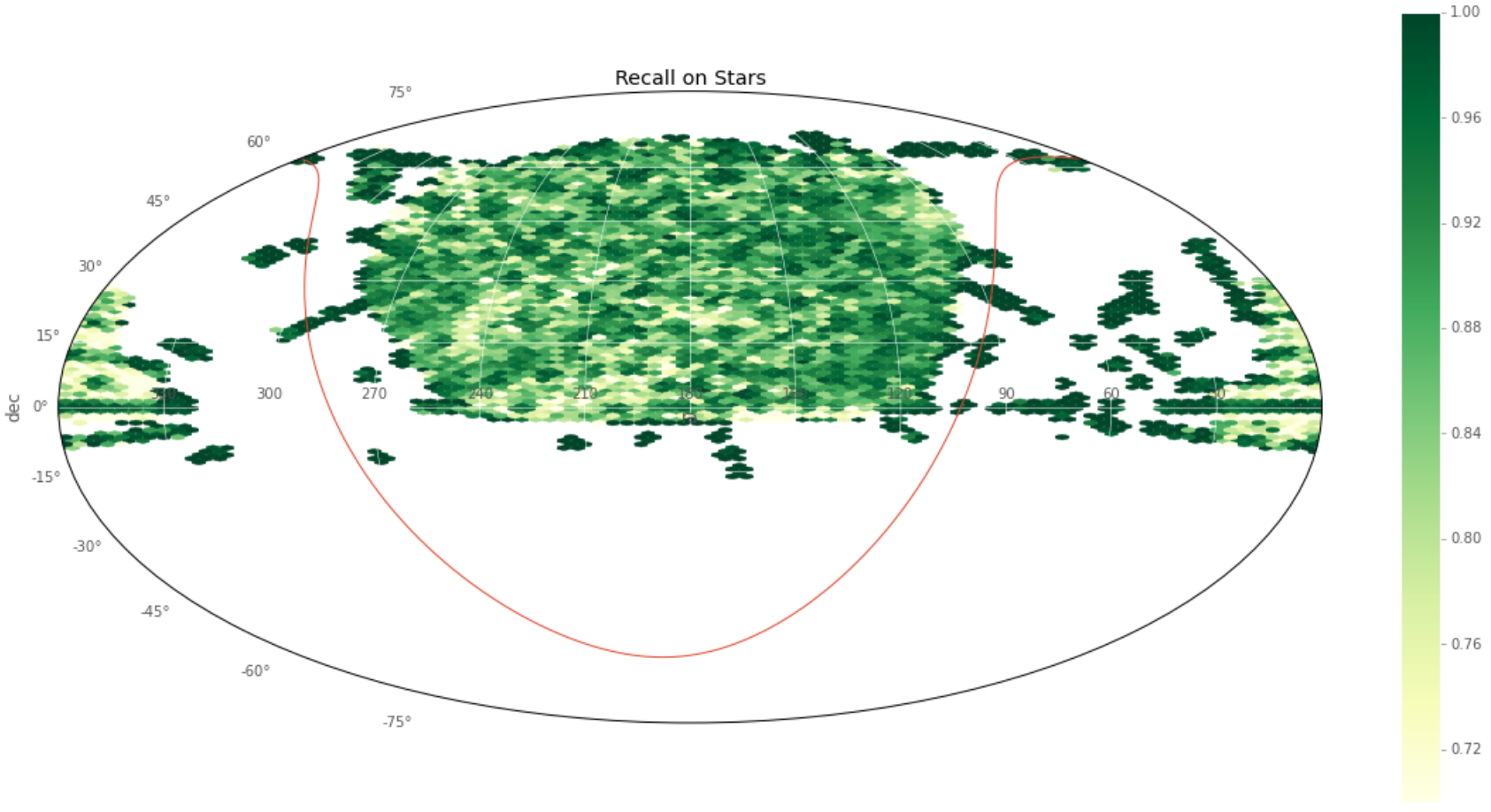

Figure: Random Forest's Recall on Stars.

Predicting with Test Data

8% of quasars are misclassified as stars.

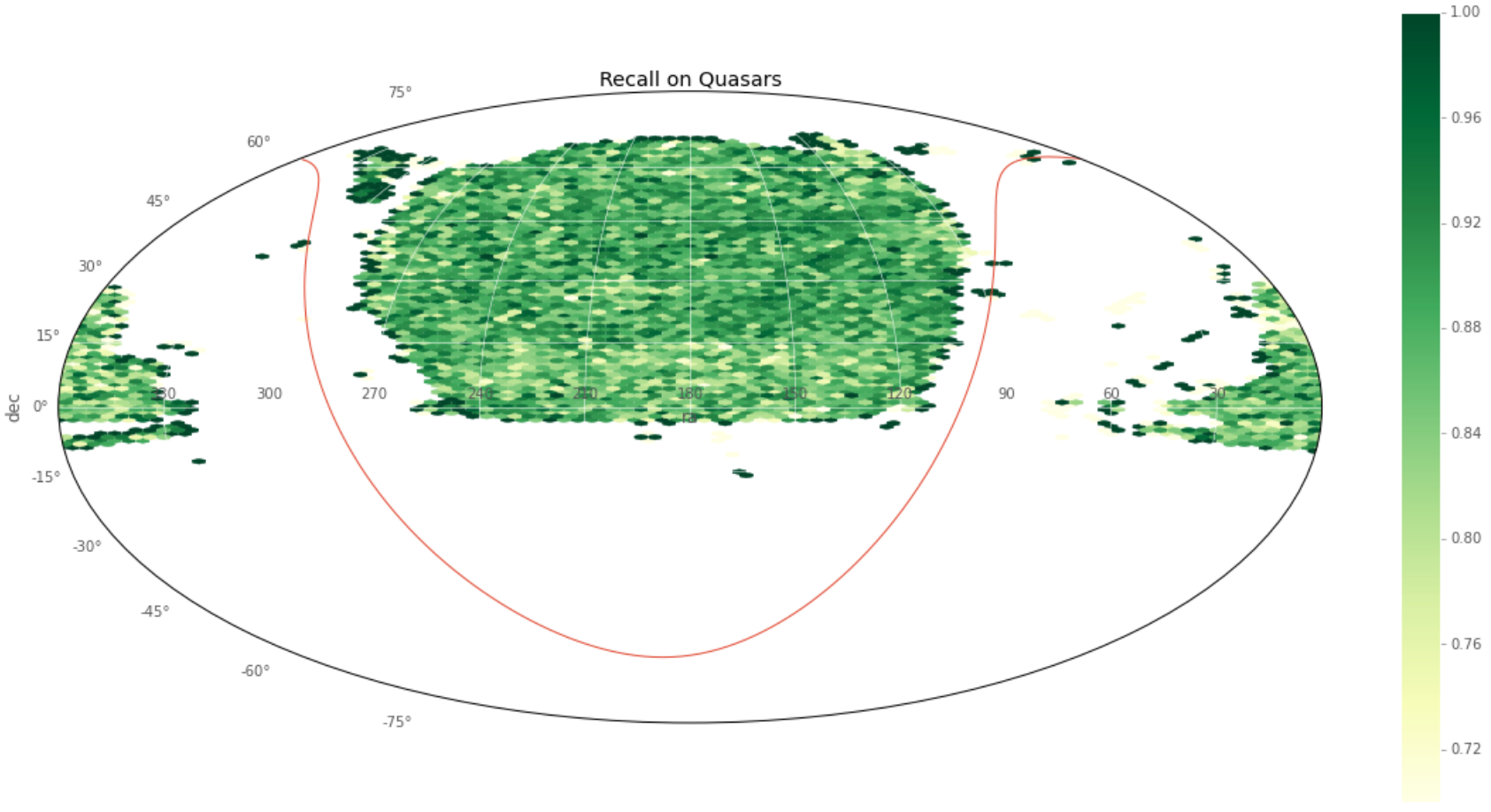

Figure: Random Forest's Recall on Quasars.

Predicting Unknowns

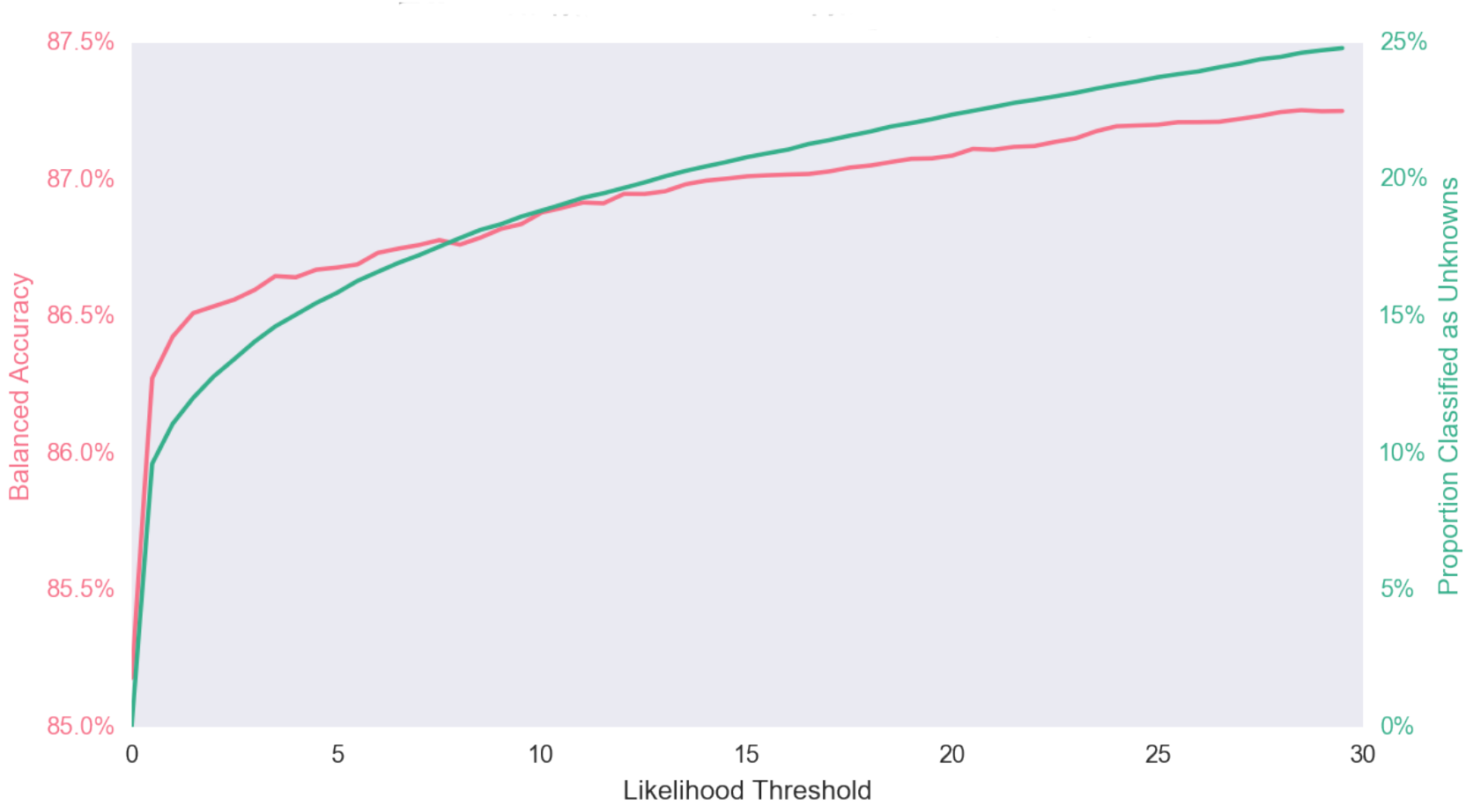

Figure: Effect of the Likelihood Threshold (QDA).

Active Learning: Motivation

- Labelling is expensive.
- One solution is to actively query the expert to obtain the training set.
- The goal is to beat random sampling.

Active Learning: Algorithm

Start with a partial training set and an unlabelled pool.
Repeat the following until we have enough training data:

1. Select T random examples from the pool.
2. Rank these T examples according to an **active learning rule**.
3. Give the expert the highest-ranked example for labelling.
4. Add this new labelled example to the training set.
5. Retrain **the classifier**.

Active Learning: Uncertainty Sampling Heuristics

- Pick the example whose prediction vector p displays the greatest Shannon entropy (information content):

$$H = -\sum_{c} p_c \log p_c$$

- Pick the example with the smallest margin (difference between the two largest values in the prediction vector p):

$$M = |\mathbb{P}(c_1 \mid x) - \mathbb{P}(c_2 \mid x)|$$

Active Learning: Query by Bagging Heuristics

1. Use bagging to train B classifiers $f_1, f_2, ..., f_B$.
2. Rank candidates by disagreement among f_i:
   - Margin-based disagreement: average the prediction of f_i and choose the example with the smallest margin.
   - Choose the example with the highest average Kullback-Leibler divergence from the average:

$$\frac{1}{B}\sum_{b=1}^{B} \mathrm{KL}(f_b || f_{\mathrm{avg}})$$

Active Learning on the Sloan Dataset

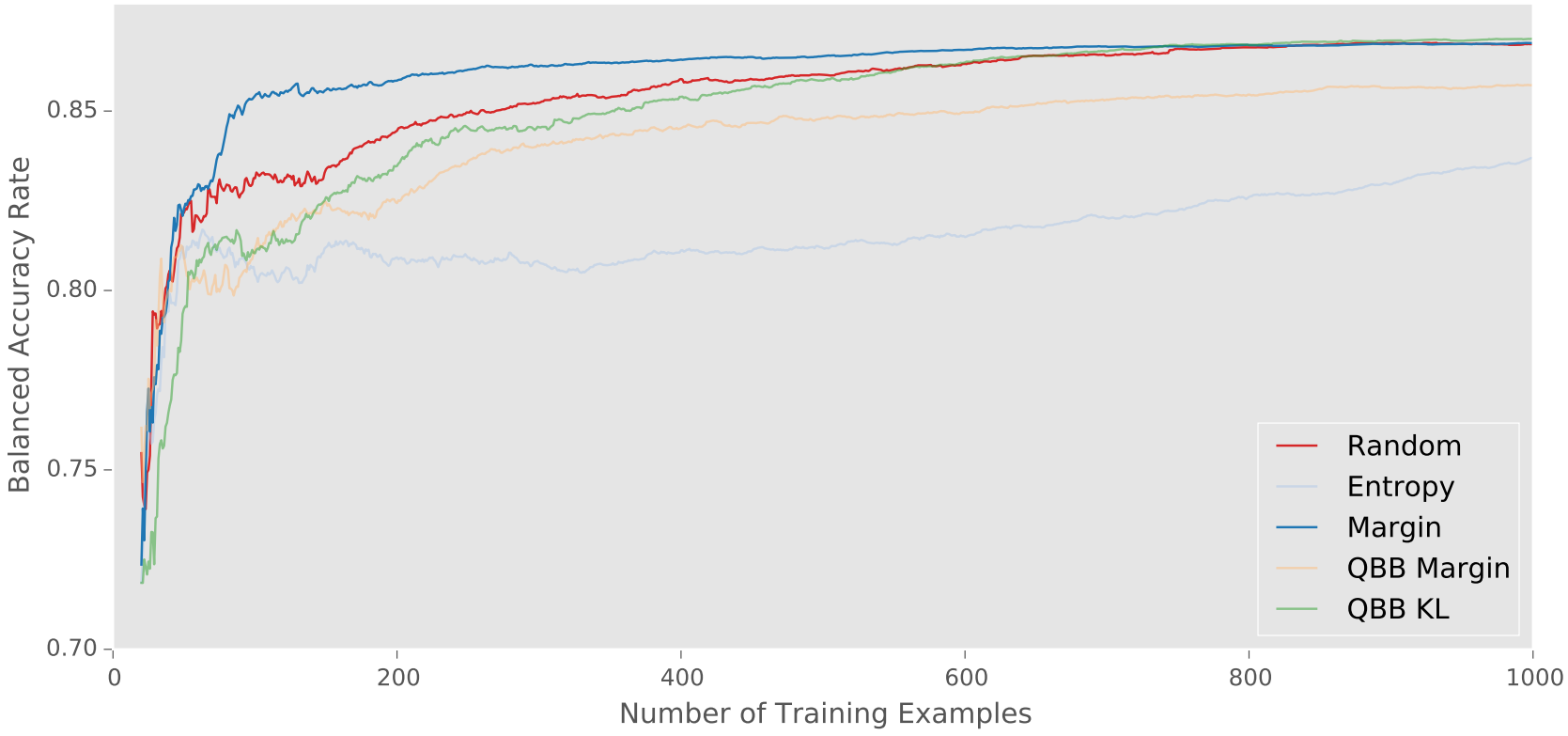

Figure: Logistic Regression Learning Curve (Unmodified Unlabelled Pool).

Active Learning on the Sloan Dataset

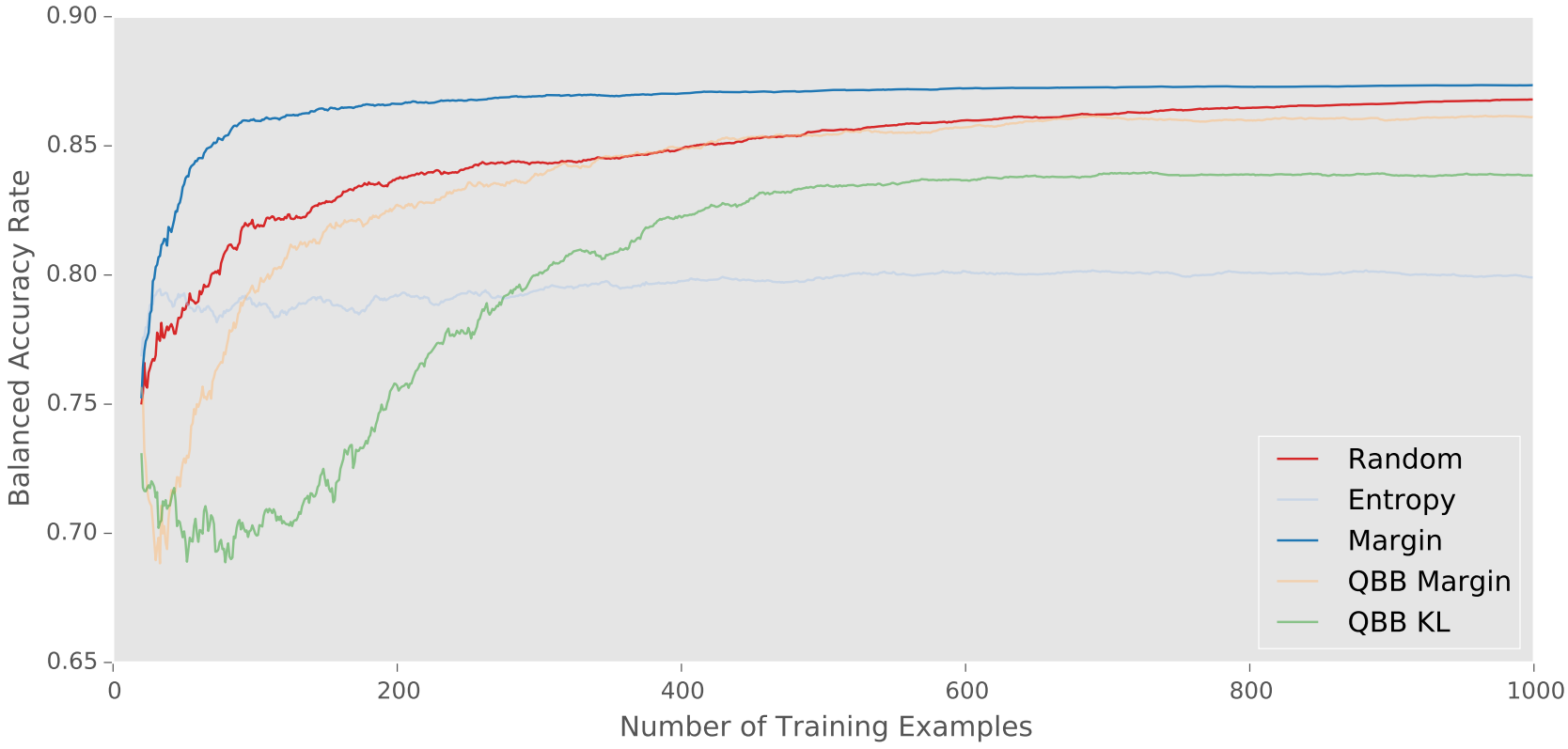

Figure: Logistic Regression Learning Curve (Balanced Unlabelled Pool).

Active Learning with Bandits

- **Arms**: clusters of examples in the unlabelled pool.
- **Context**: mean/variance of distance between individual points in the cluster, proportion of labelled points in the cluster, etc.
- **Reward**: cosine distance between the prediction vectors of the new and old model.
- **Action**: select cluster at each step to maximise the rewards.

Concluding Remarks

- Use robust optimisation to take into account of the uncertainty in measurement errors.
- Theoretical analysis of convergence in multi-class active learning.
- My project is hosted at `github.com/alasdairtran`.