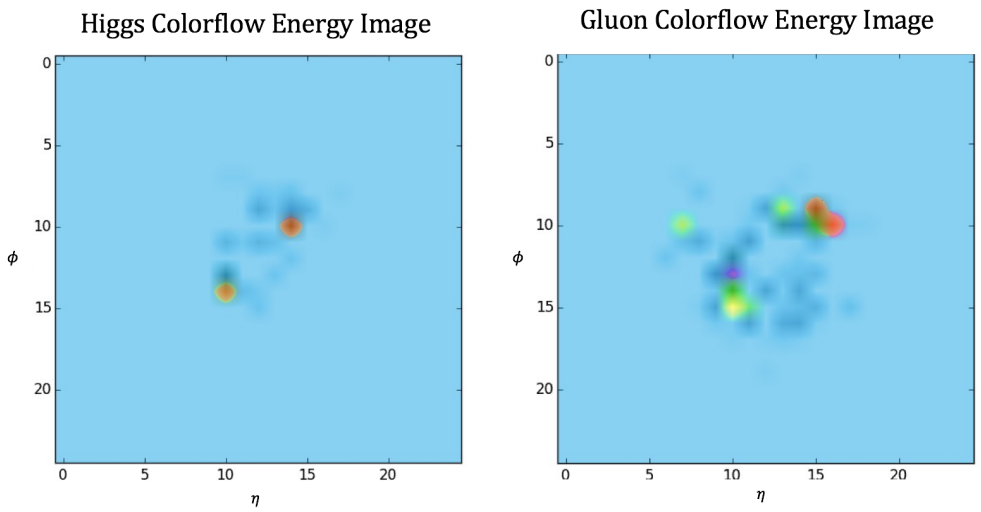
Measuring Colour Flow and Machine Learning

Guy Amdur, Anton Apostolatos, Leonard Bronner

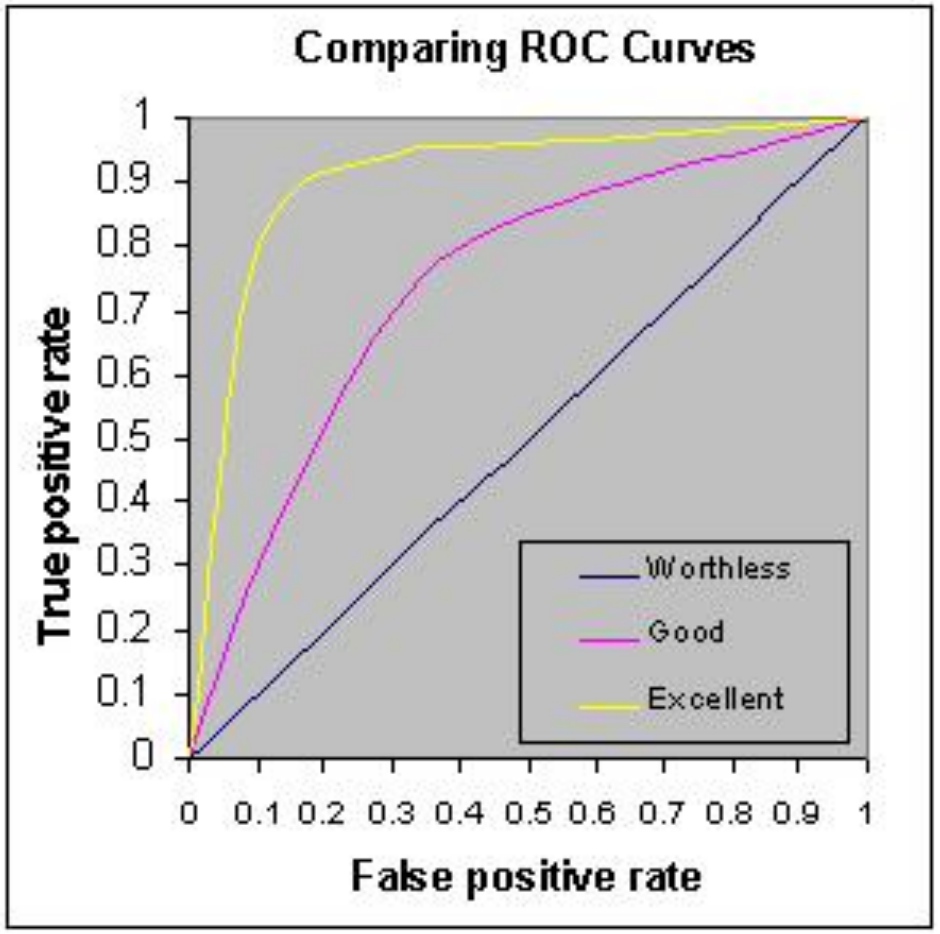
The Problem and the Data

- Distinguishing Higgs boson decay from background data using colorflow energy images
  - Images in cylindrical coordinates η and ϕ with constant resonance, preprocessed so that the jet center is at the center of the image
  - We're given 10k samples in 25x25 images and 50k samples in 100x100 images



Our methods

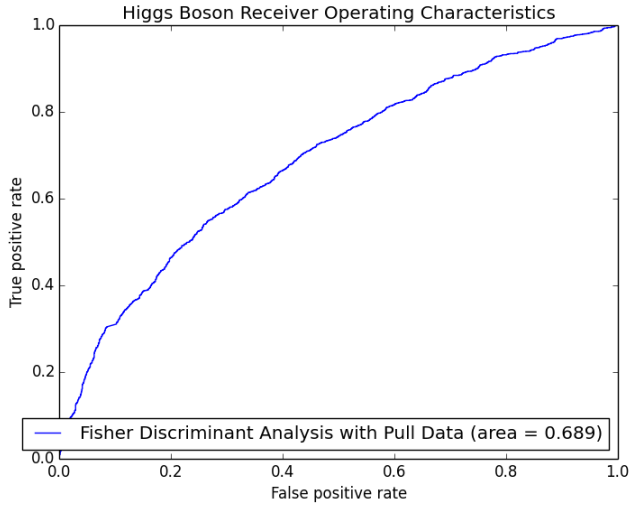
- We split the data into a development set and a test set
- Learn on the development and test on the unseen test set
- Measure and compare ROC curves (plotting TPR and FPR at various thresholds) and the area under this curve

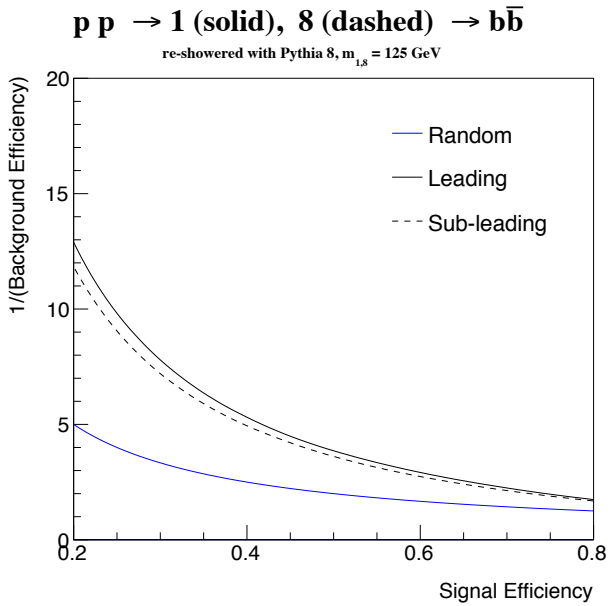


Development (75%)	Test (25%)

Pull data

- Current state of the art classification methods uses jet pull data, which is a superstructure feature of an event which can help determine if jets came from decay of a color-singlet object
- The classification method used is Fisher Discriminant Analysis

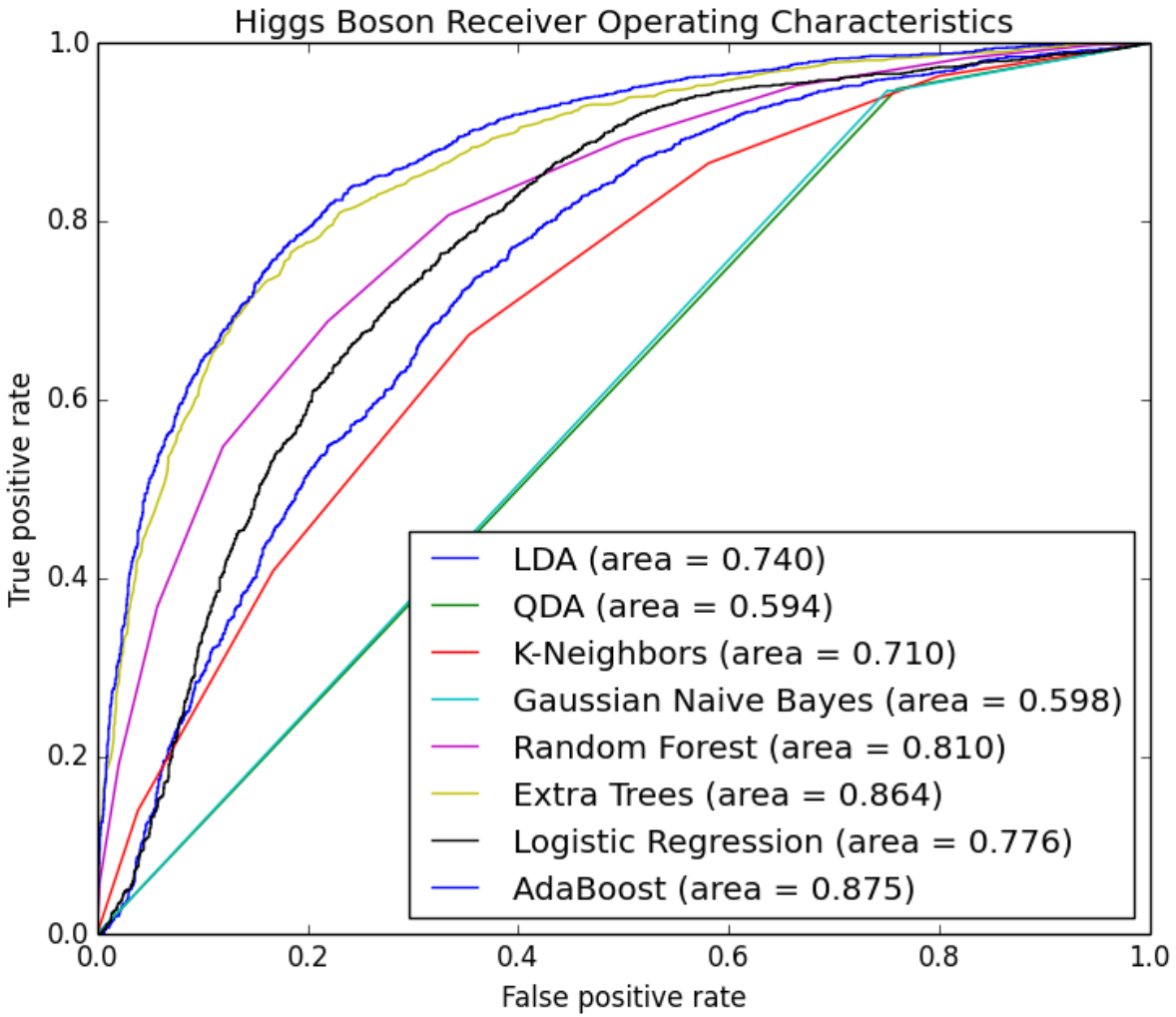




Initial tests

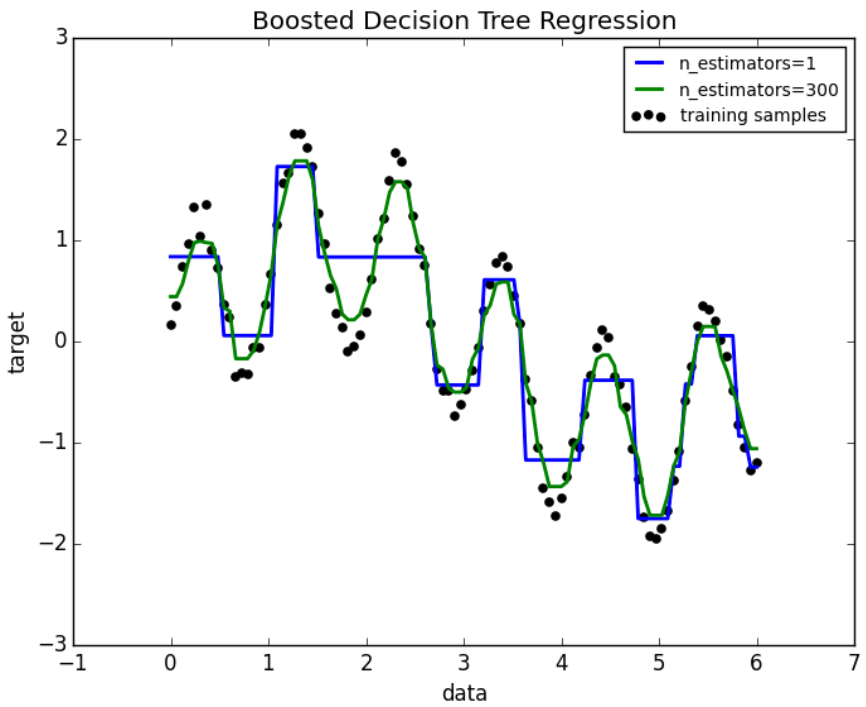
- Vectorize the images from $\mathbb{R}^{25\times25} \rightarrow \mathbb{R}^{625}$ for coarse data and $\mathbb{R}^{100\times100} \rightarrow \mathbb{R}^{10000}$ for the finer granularity data.
- Pass these through various classifiers to compare quality of results
- Motivation: see if learning algorithms are able to pick up on the most important pixels and aspects of an image without human intervention

Results on coarse data

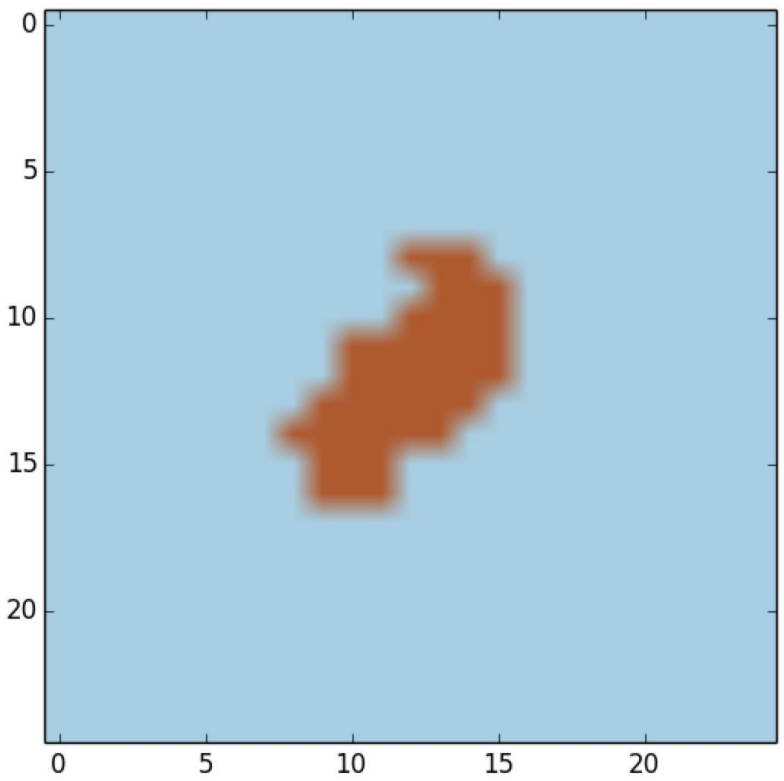


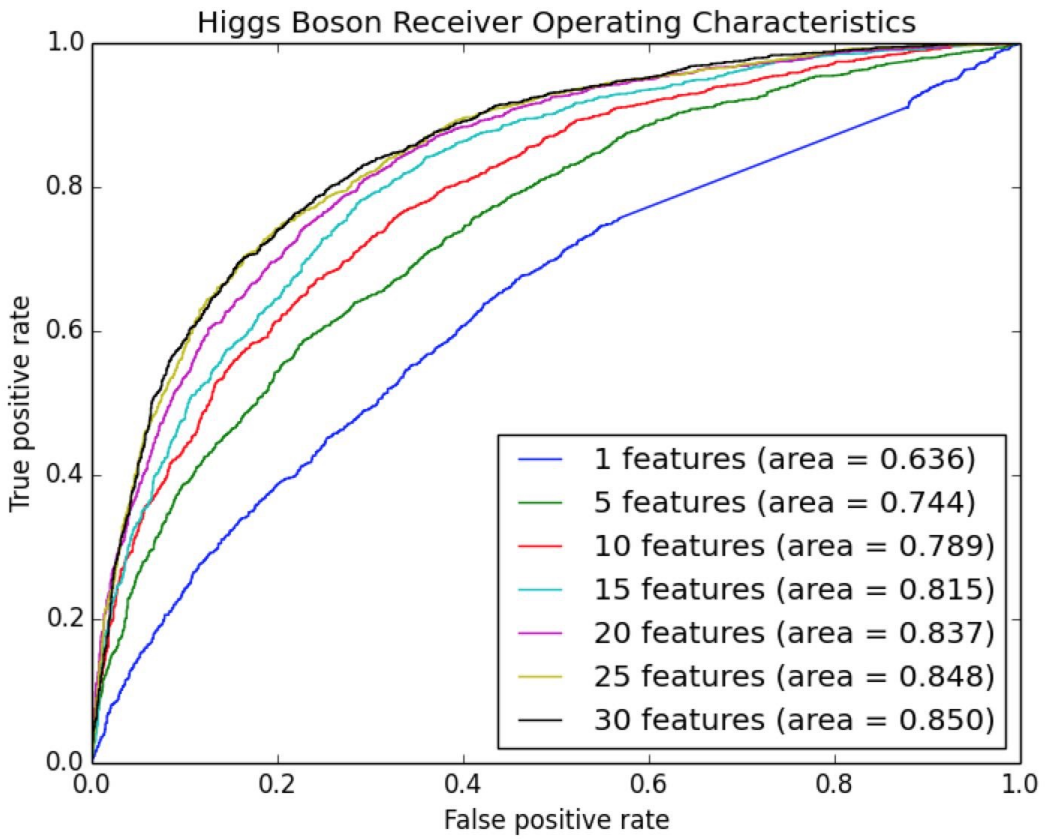
Tree classifiers

- Decision tree learning uses a decision tree as a predictive model, where leaves represent labels (in this case, higgs and not-higgs) and branches represent conjunctions of features from the data
- AdaBoost is a meta-estimator that fits many decision-trees on the dataset, combining them into a weighted sum representing the meta classifier

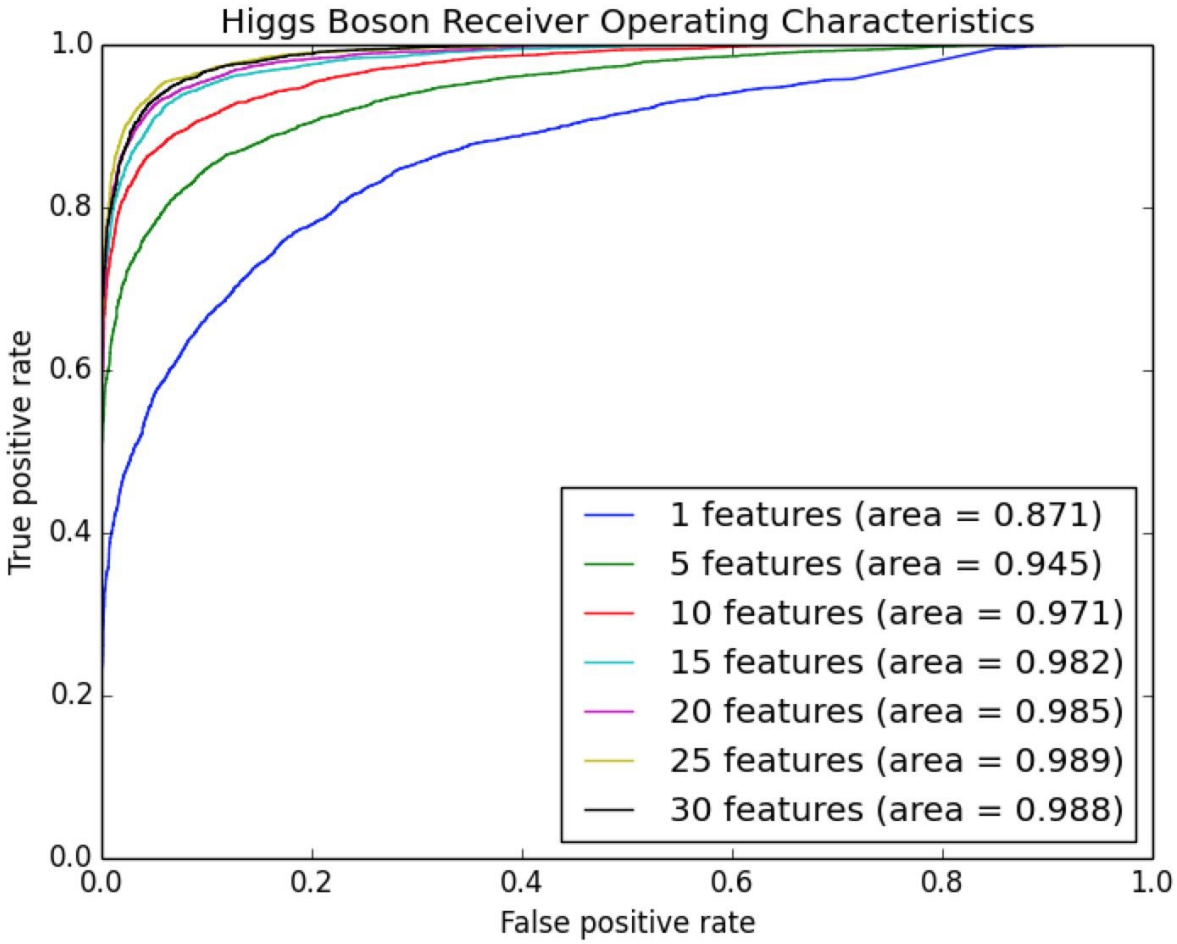


Feature selection





Overfitting on training data



Coarse versus Fine Images

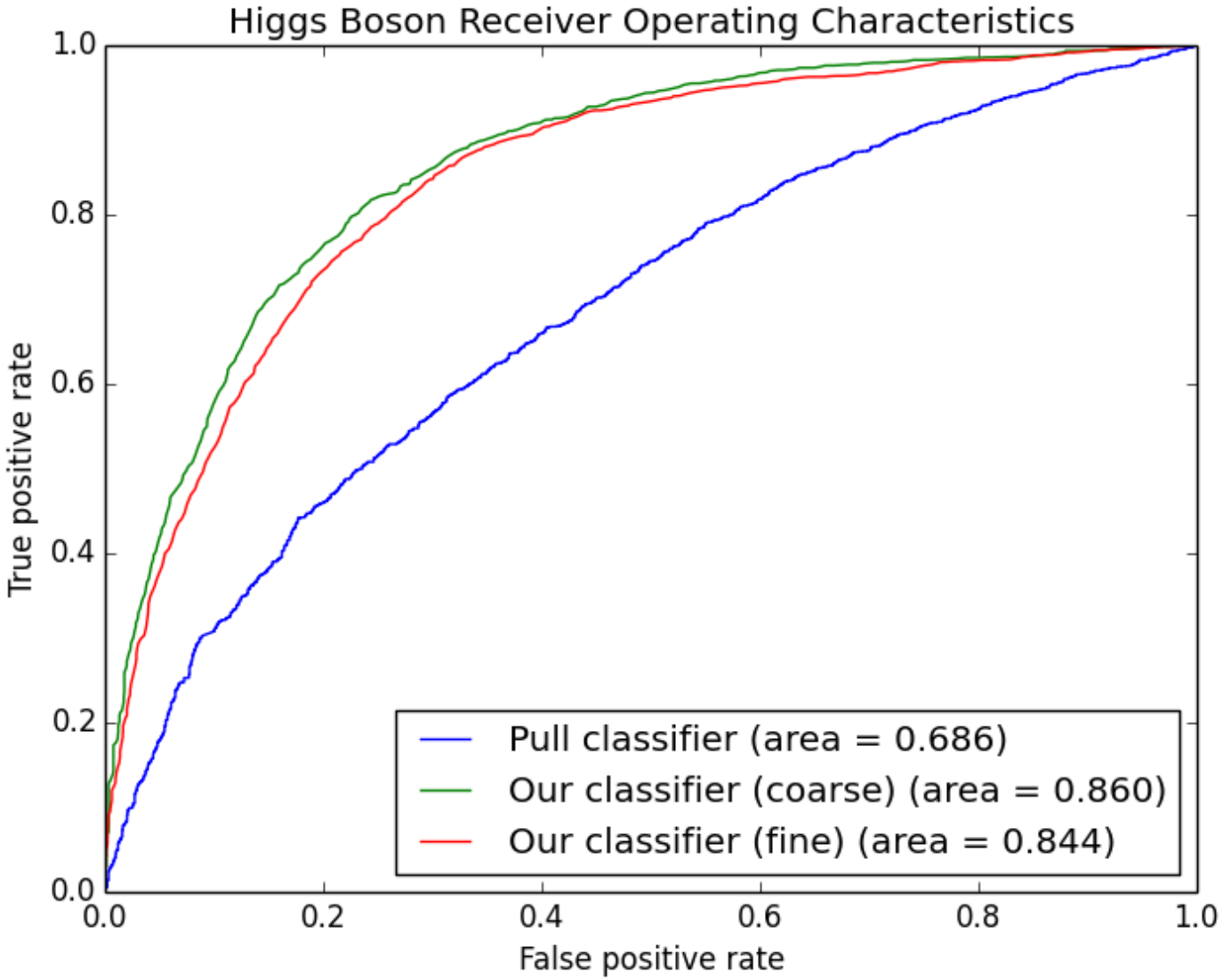
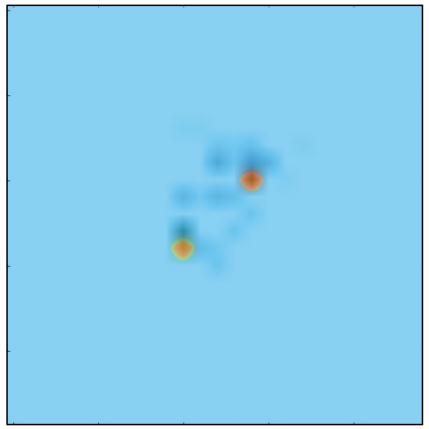


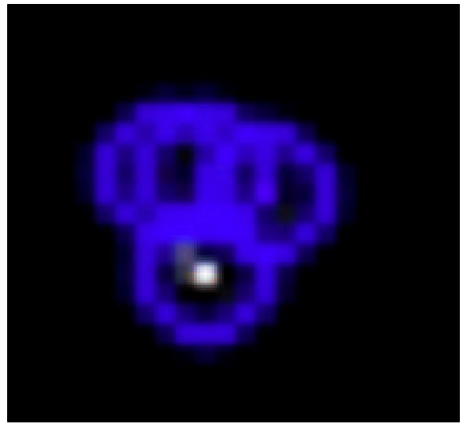
Image recognition

- We experimented with various image descriptors, including ORB, BRIEF, and SIFT for the higher granularity images, to see if we could build on top of vectorized pixel features.
- The results were unsatisfactory, achieving an AUC of only 0.550.
- These descriptors weren't able to capture many of the nuances of the sparse pixel matrices.
- Image histograms of pixel density also proved unsuccessful, resulting in an AUC score of 0.660.

Colorflow Image

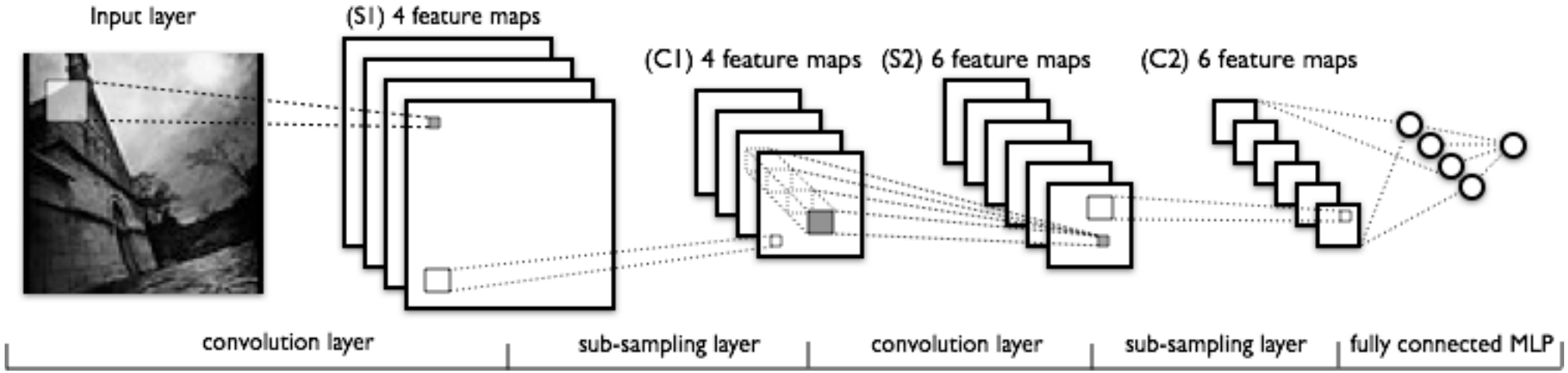


FAST Keypoints



Future work

- Test on images with less consistent conditions (no preprocessing, or non-constant resonance)
- Use a convolutional neural network as a classification method



References

- De Oliveira, Luke, Michael A. Kagan, Lester Mackey, Benjamin Nachman, and Ariel Schwartzman. *Deep Convolutional Architectures for Jet-Images at the Large Hadron Collider*. N.d. Stanford.
- "Deep Learning Tutorials¶." *Deep Learning Tutorials — DeepLearning 0.1 Documentation*. LISA Lab, n.d. Web. 07 Dec. 2015.
- Drucker, Harris. "Improving Regressors Using Boosting Techniques." Proceedings of the 14th International Conference on Machine Learning, Aug. 1997. Web.
- Genesereth, Michael, and Michael Thielscher. "Propositional Nets." *General Game Playing*. Stanford Logic Group, n.d. Web.

Special thanks to Ariel Schwartzman, Ben Nachman and Michael Kagan for their support and guidance during this project