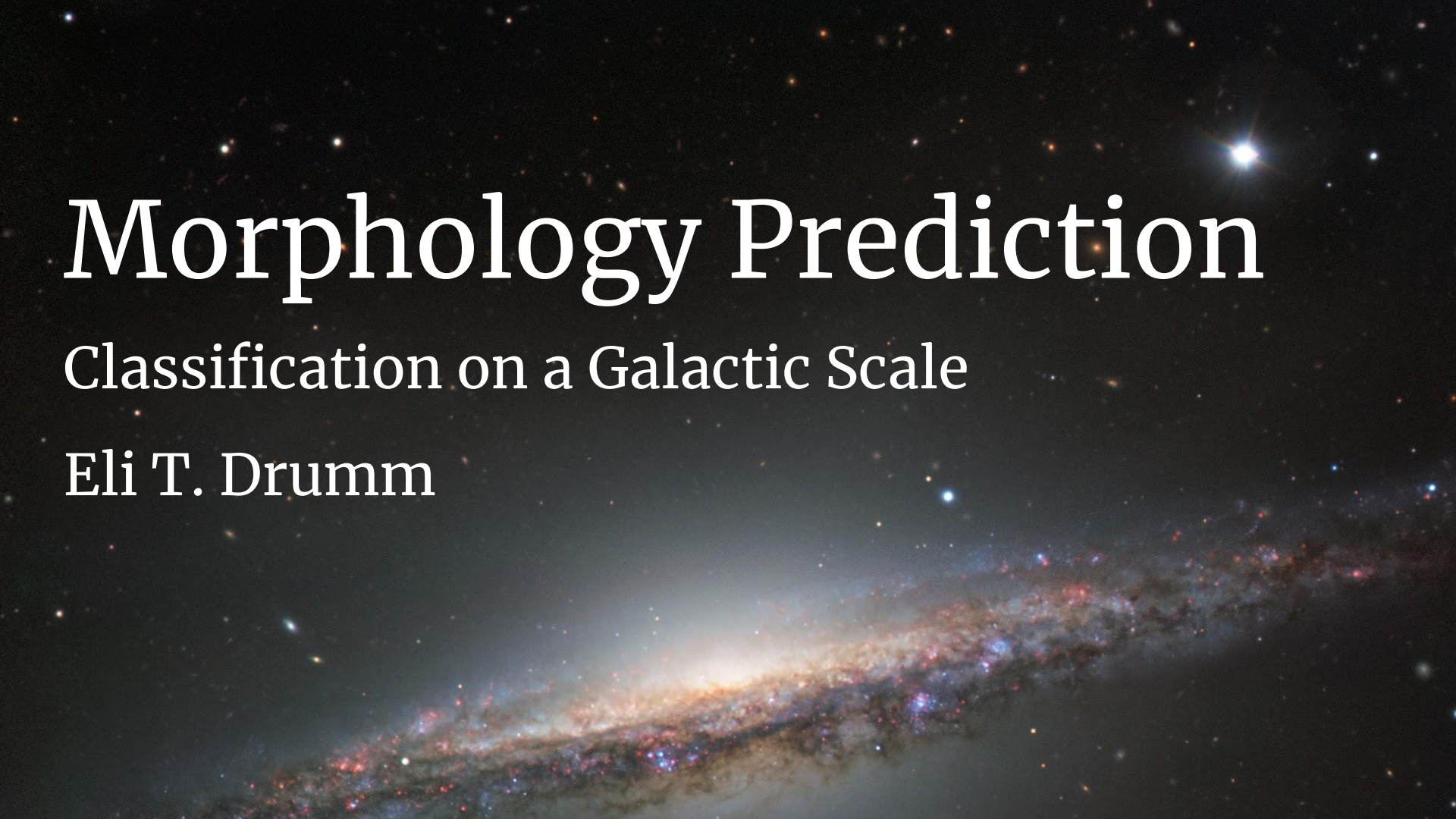


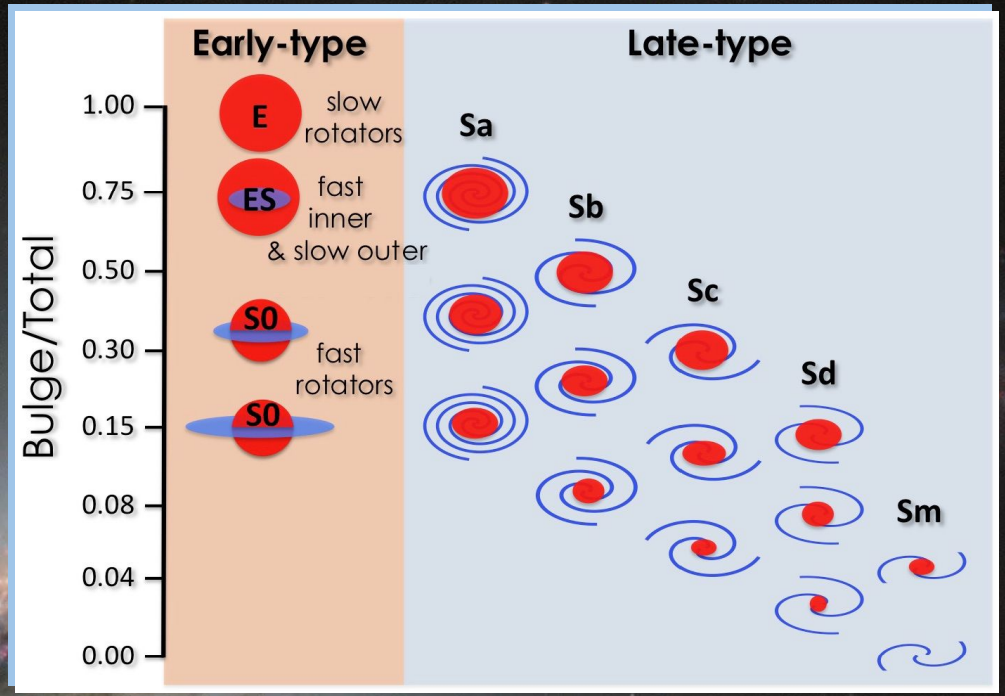
# Morphology Prediction

## Classification on a Galactic Scale

Eli T. Drumm

# Background

- Galaxy Zoo project
- Paper (and dataset) by Schawinski et al. (2009)
- Early- vs. late-type




Wikipedia

# Feature selection

- Spectroscopy
  - *ugriz*
- $M_{\odot}$
- Redshift
- Luminosity
- etc.

# Linear model

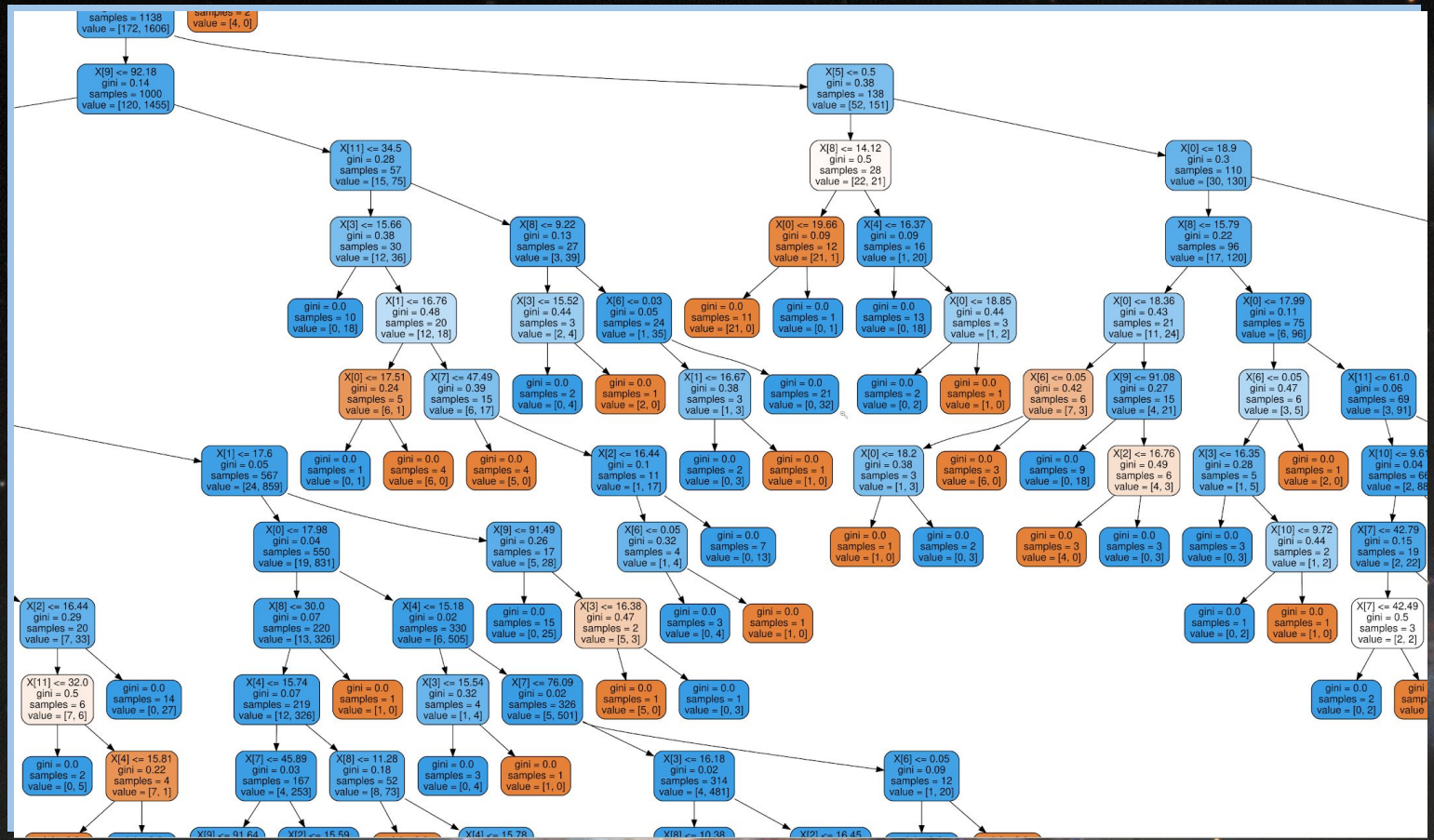
- Logistic regression, 5-fold CV
  - aaaand...

- Accuracy is 87%
- ROC AUC is 0.923

| | precision | recall | f1 |
|---|---|---|---|
| **non-late** | 0.85 | 0.78 | 0.81 |
| **late-type** | 0.88 | 0.92 | 0.90 |

# Trees and forests

- Can we do better?
- Random forest
  - Again, with 5-fold CV

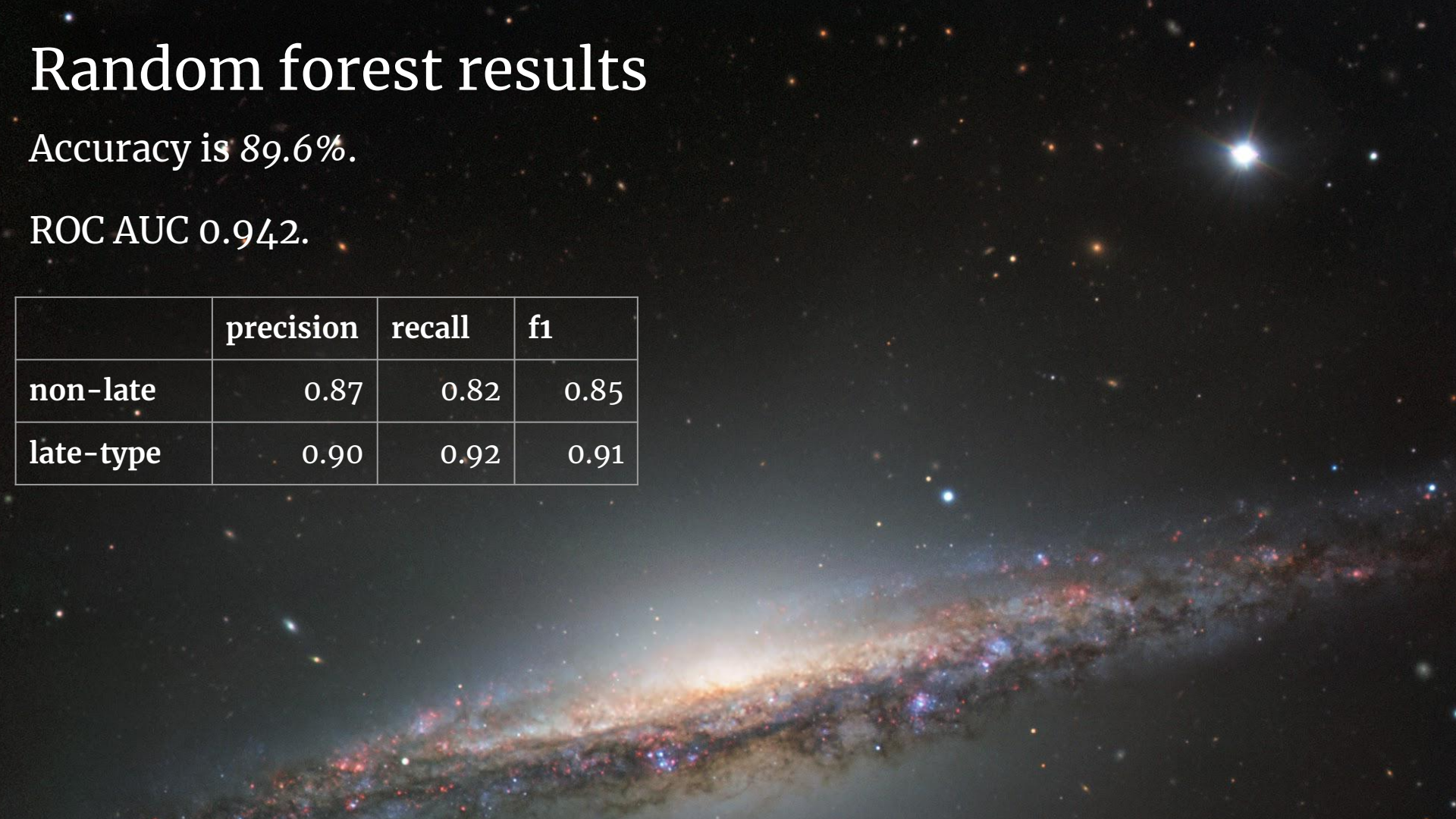

samples = 1138
value = [172, 1606]
samples = 2
value = [4, 0]
X[9] <= 92.18
gini = 0.14
samples = 1000
value = [120, 1455]
X[5] <= 0.5
gini = 0.38
samples = 138
value = [52, 151]
X[11] <= 34.5
gini = 0.28
samples = 57
value = [15, 75]
X[8] <= 14.12
gini = 0.5
samples = 28
value = [22, 21]
X[0] <= 18.9
gini = 0.3
samples = 110
value = [30, 130]
X[3] <= 15.66
gini = 0.38
samples = 30
value = [12, 36]
X[8] <= 9.22
gini = 0.13
samples = 27
value = [3, 39]
X[0] <= 19.66
gini = 0.09
samples = 12
value = [21, 1]
X[4] <= 16.37
gini = 0.09
samples = 16
value = [1, 20]
X[8] <= 15.79
gini = 0.22
samples = 96
value = [17, 120]
gini = 0.0
samples = 10
value = [0, 18]
X[1] <= 16.76
gini = 0.48
samples = 20
value = [12, 18]
X[3] <= 15.52
gini = 0.44
samples = 3
value = [2, 4]
X[6] <= 0.03
gini = 0.05
samples = 24
value = [1, 35]
gini = 0.0
samples = 11
value = [21, 0]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 13
value = [0, 18]
X[0] <= 18.85
gini = 0.44
samples = 3
value = [1, 2]
X[0] <= 18.36
gini = 0.43
samples = 21
value = [11, 24]
X[0] <= 17.99
gini = 0.11
samples = 75
value = [6, 96]
X[0] <= 17.51
gini = 0.24
samples = 5
value = [6, 1]
X[7] <= 47.49
gini = 0.39
samples = 15
value = [6, 17]
gini = 0.0
samples = 2
value = [0, 4]
gini = 0.0
samples = 1
value = [2, 0]
X[1] <= 16.67
gini = 0.38
samples = 3
value = [1, 3]
gini = 0.0
samples = 21
value = [0, 32]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.0
samples = 1
value = [1, 0]
X[6] <= 0.05
gini = 0.42
samples = 6
value = [7, 3]
X[9] <= 91.08
gini = 0.27
samples = 15
value = [4, 21]
X[6] <= 0.05
gini = 0.47
samples = 6
value = [3, 5]
X[11] <= 61.0
gini = 0.06
samples = 69
value = [3, 91]
X[1] <= 17.6
gini = 0.05
samples = 567
value = [24, 859]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 4
value = [6, 0]
gini = 0.0
samples = 4
value = [5, 0]
X[2] <= 16.44
gini = 0.1
samples = 11
value = [1, 17]
gini = 0.0
samples = 2
value = [0, 3]
gini = 0.0
samples = 1
value = [1, 0]
X[0] <= 18.2
gini = 0.38
samples = 3
value = [1, 3]
gini = 0.0
samples = 3
value = [6, 0]
gini = 0.0
samples = 9
value = [0, 18]
X[2] <= 16.76
gini = 0.49
samples = 6
value = [4, 3]
X[3] <= 16.35
gini = 0.28
samples = 5
value = [1, 5]
gini = 0.0
samples = 1
value = [2, 0]
X[10] <= 9.61
gini = 0.04
samples = 66
value = [2, 88]
X[0] <= 17.98
gini = 0.04
samples = 550
value = [19, 831]
X[9] <= 91.49
gini = 0.26
samples = 17
value = [5, 28]
X[6] <= 0.05
gini = 0.32
samples = 4
value = [1, 4]
gini = 0.0
samples = 7
value = [0, 13]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 2
value = [0, 3]
gini = 0.0
samples = 3
value = [4, 0]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.0
samples = 3
value = [0, 3]
X[10] <= 9.72
gini = 0.44
samples = 2
value = [1, 2]
X[7] <= 42.79
gini = 0.15
samples = 19
value = [2, 22]
X[2] <= 16.44
gini = 0.29
samples = 20
value = [7, 33]
X[8] <= 30.0
gini = 0.07
samples = 220
value = [13, 326]
X[4] <= 15.18
gini = 0.02
samples = 330
value = [6, 505]
gini = 0.0
samples = 15
value = [0, 25]
X[3] <= 16.38
gini = 0.47
samples = 2
value = [5, 3]
gini = 0.0
samples = 1
value = [0, 2]
gini = 0.0
samples = 1
value = [1, 0]
X[7] <= 42.49
gini = 0.5
samples = 3
value = [2, 2]
X[11] <= 32.0
gini = 0.5
samples = 6
value = [7, 6]
gini = 0.0
samples = 14
value = [0, 27]
X[4] <= 15.74
gini = 0.07
samples = 219
value = [12, 326]
gini = 0.0
samples = 1
value = [1, 0]
X[3] <= 15.54
gini = 0.32
samples = 4
value = [1, 4]
X[7] <= 76.09
gini = 0.02
samples = 326
value = [5, 501]
gini = 0.0
samples = 1
value = [5, 0]
gini = 0.0
samples = 1
value = [0, 3]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.0
samples = 2
value = [0, 5]
X[4] <= 15.81
gini = 0.22
samples = 4
value = [7, 1]
X[7] <= 45.89
gini = 0.03
samples = 167
value = [4, 253]
X[8] <= 11.28
gini = 0.18
samples = 52
value = [8, 73]
gini = 0.0
samples = 3
value = [0, 4]
gini = 0.0
samples = 1
value = [1, 0]
X[3] <= 16.18
gini = 0.02
samples = 314
value = [4, 481]
X[6] <= 0.05
gini = 0.09
samples = 12
value = [1, 20]

# Random forest results

Accuracy is *89.6%*.

ROC AUC 0.942.

| | precision | recall | f1 |
|---|---|---|---|
| **non-late** | 0.87 | 0.82 | 0.85 |
| **late-type** | 0.90 | 0.92 | 0.91 |

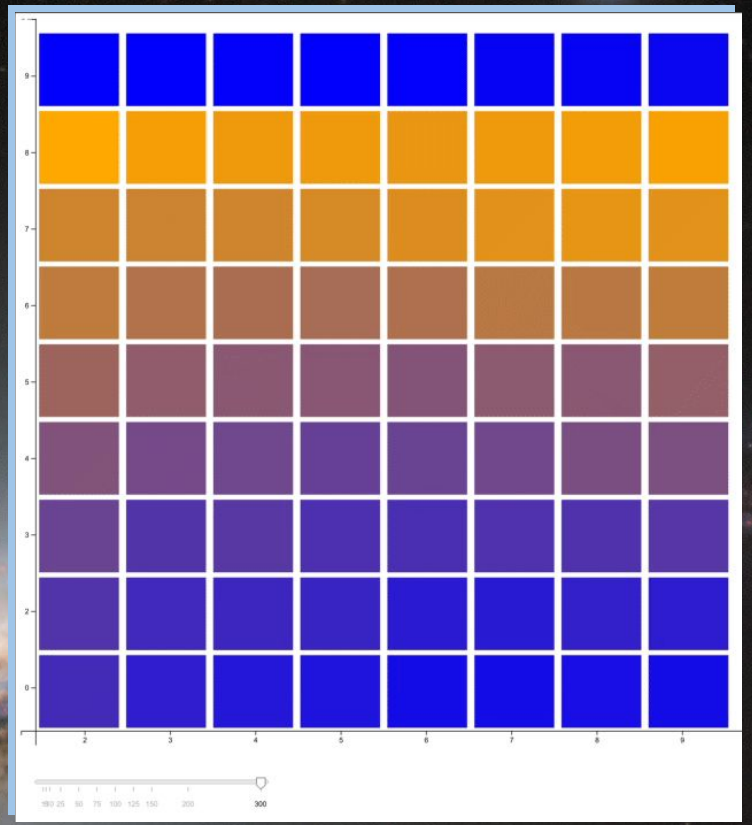


# Hiking the random forests

We can do a grid search to find the best parameters.

General lessons:

- Plant more trees
- ...but expect diminishing returns
- Unlimited max depth

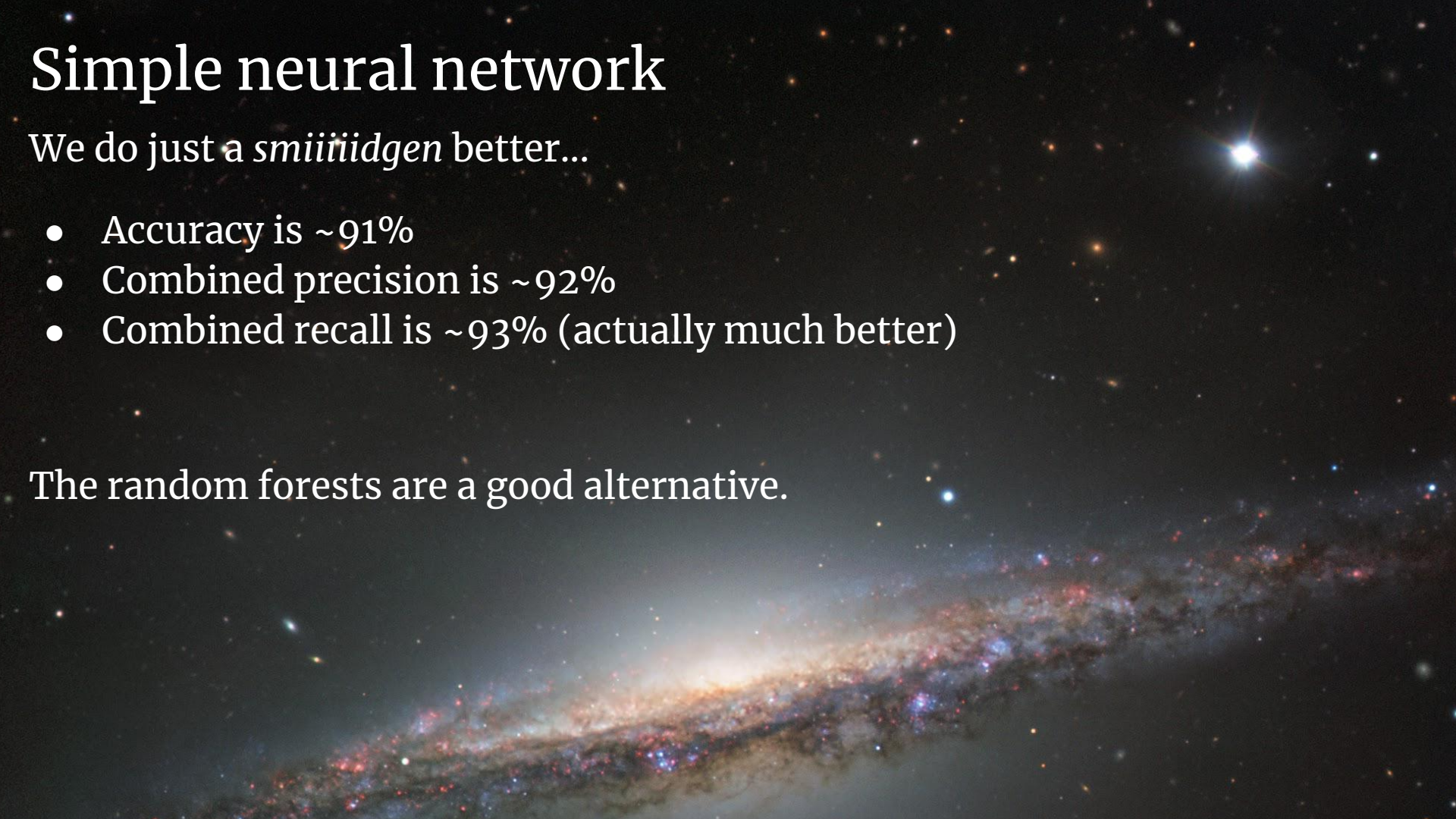


# Simple neural network

We do just a *smiiiiidgen* better...

- Accuracy is ~91%
- Combined precision is ~92%
- Combined recall is ~93% (actually much better)

The random forests are a good alternative.



# Features revisited

- Feature importance metrics from forests
  - Most important: velocity dispersion and its error term

- Reengineering features
  - $u-r$, e.g.

- How do we know what features to engineer?
  - Who's this "we" anyway?

# Takeaways

- Other data: Galaxy Zoo 2, Galaxy Zoo Hubble
- Without computed data, prediction is difficult
    - *“At the best galaxy redshift, the stellar velocity dispersion is also determined. This is done by computing a PCA basis of 24 eigenspectra from the ELODIE stellar library (Prugniel & Soubiran 2001), convolved and binned to match the instrumental resolution and constant-velocity pixel scale of the reduced SDSS spectra, and broadened by Gaussian kernels of successively larger velocity width ranging from 100 to 850 km/s in steps of 25 km/s. The broadened stellar template sets are redshifted to the best-fit galaxy redshift, and the spectrum is modeled as a least-squares linear combination of the basis at each trial broadening, masking pixels at the position of common emission lines in the galaxy-redshift rest frame. The best-fit velocity dispersion is determined by fitting locally for the position of the minimum of chi-squared versus trial velocity dispersion in the neighborhood of the lowest gridded chi-squared value. Velocity-dispersion error estimates are determined from the curvature of the chi-squared curve at the global minimum, and are set to a negative value if the best value occurs at the high-velocity end of the fitting range. Reported best-fit velocity-dispersion values less than about 100 km/s are below the resolution limit of the SDSS spectrograph and are to be regarded with caution.”*
- Morale: data science can’t be isolated
    - Work with astronomers or whomever!