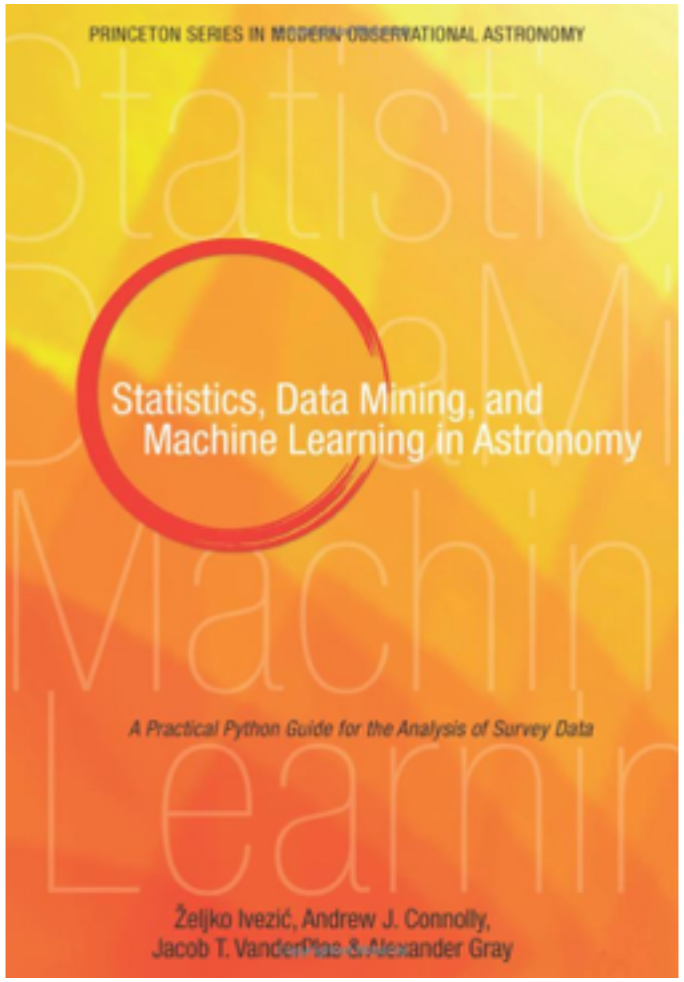
Astro ML course

Lecture 1

Course plan

15 discussion/hands-on sessions
One person leads the discussion
Everyone reads the book

Content and format are flexible!



Project

Choose a method from the book and solve a real problem

1-3 people per project

Peer-review

Deadline in May

- What is machine learning?
- Some terminology
- The bias-variance trade-off
- Real-world example:
  predicting galaxy escape fractions with
  Lasso regression
- Practical stuff: how to access the data sets
  used in the book

What is machine learning?

A computer program is said to learn from an experience E with respect to some task T and some performance measure P if its performance on T as measured by P improves with experience E.

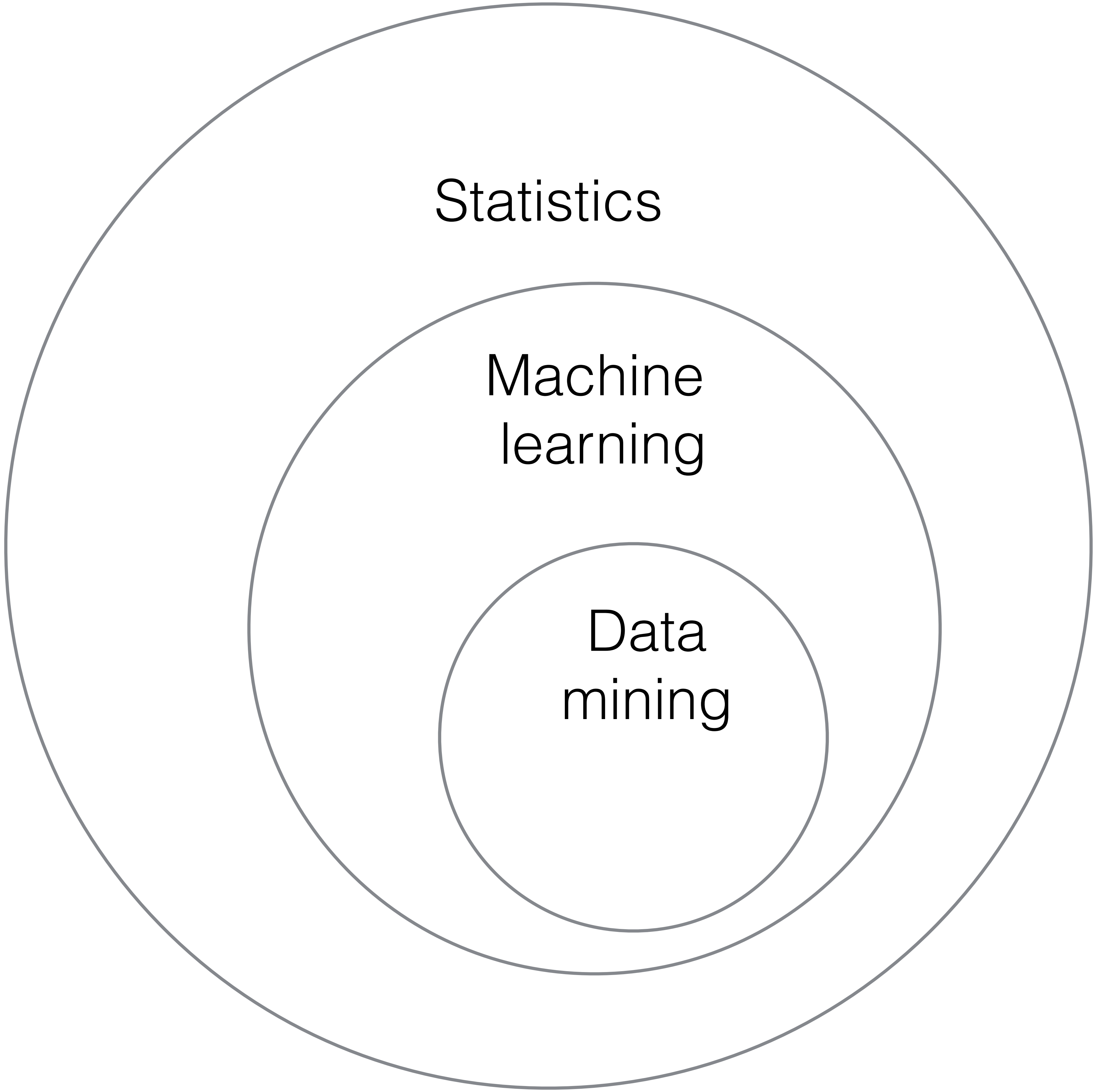
-Andrew Ng

”…analysis and interpretation of data, often involving large quantities of data, and often resorting to numerical methods…”

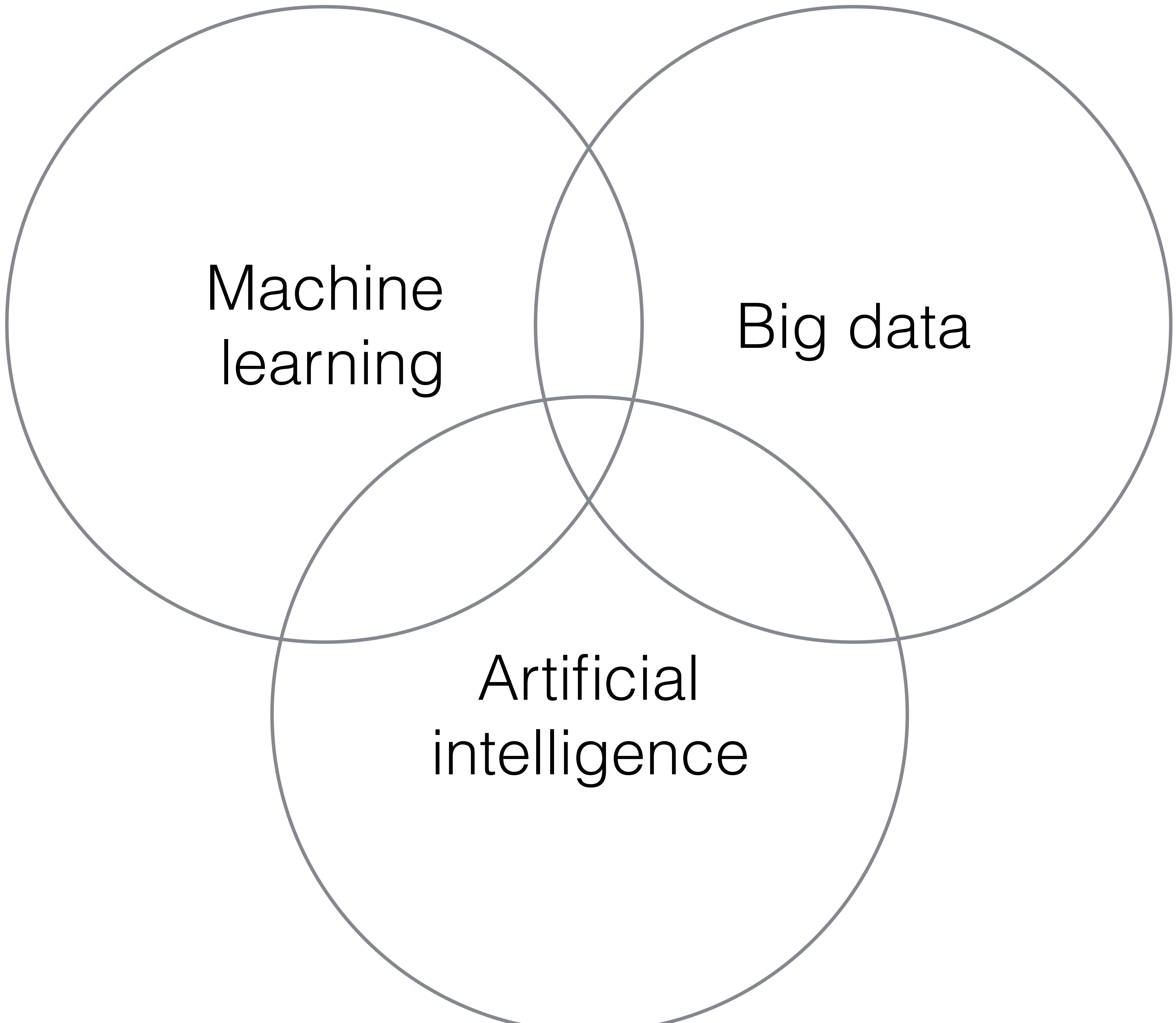
-The course book

”Machine learning is what computer scientists say when they mean ’statistics’ ”

-Some person on the internet


Statistics
Machine
learning
Data
mining

Buzzwords



Machine learning/statistical learning

Supervised learning (6,7)

Regression

Linear regression,
Lasso regression,
Ridge regression,
Neural networks
…

Classification

Logistic regression,
Support Vector
Machines,
K Nearest Neighbor,
Decision trees
…

Unsupervised learning, (8,9)
(Data mining)

Clustering, dimensionality reduction, density estimation

K-means
Principal Components
Stochastic Neighbor
embedding
…

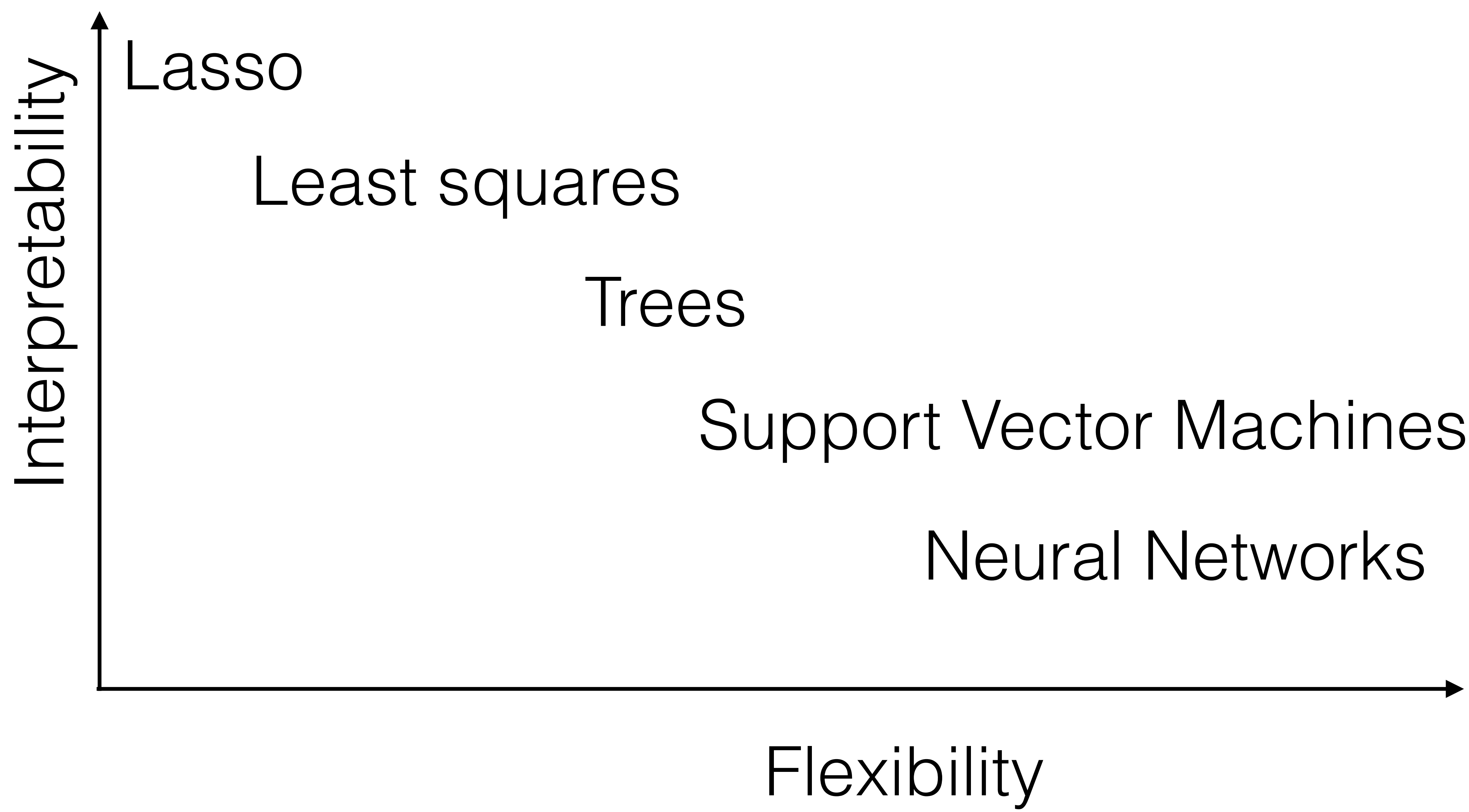
Anomaly detection

Examples

Regression:
Given galaxy magnitudes in a number of filters, predict the redshift of the galaxy

Classification:
Given an image of an object, determine whether the object is a star or a galaxy

Clustering:
Given a set of stellar spectra, find out if there are natural groups of stars


Interpretability
Lasso
Least squares
Trees
Support Vector Machines
Neural Networks
Flexibility

Terminology, Supervised learning

Set of training examples, each with a **label** *y* and *n* **features**

(response, dependent variable)

(predictor, independent variable)

Fit a model that predicts *y*, given an unseen set of features

The model should minimize the **cost function**

(loss function, objective function, fitness function …)

Example: house prices

Supervised learning (regression)

Problem: Predict the price of a house, given its size in square meters

Features: the size in square meters (only one)

Label: the house price

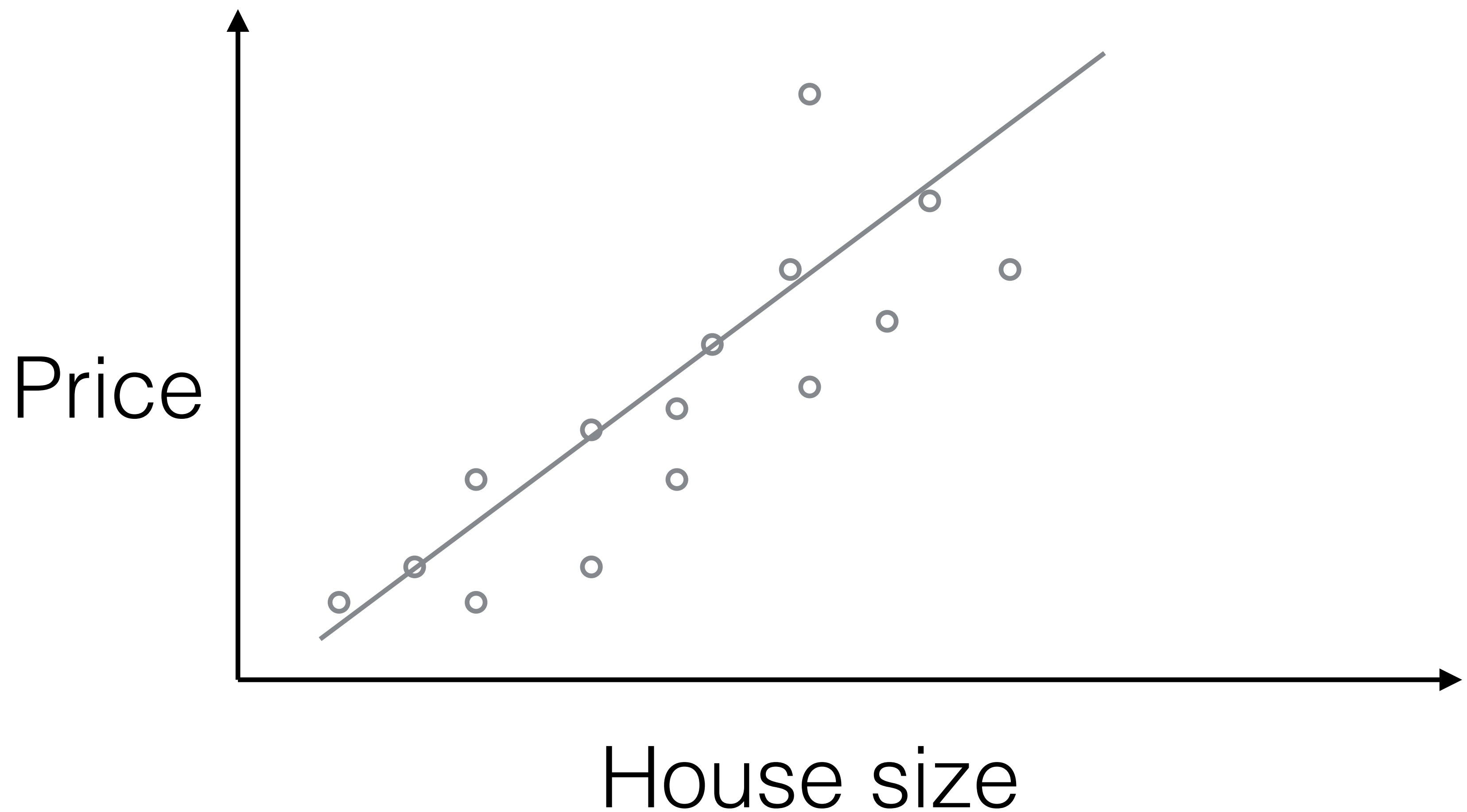
Training set: historical records of house prices and sizes

Model: price = $\theta_0 + \theta_1$*size (linear regression)

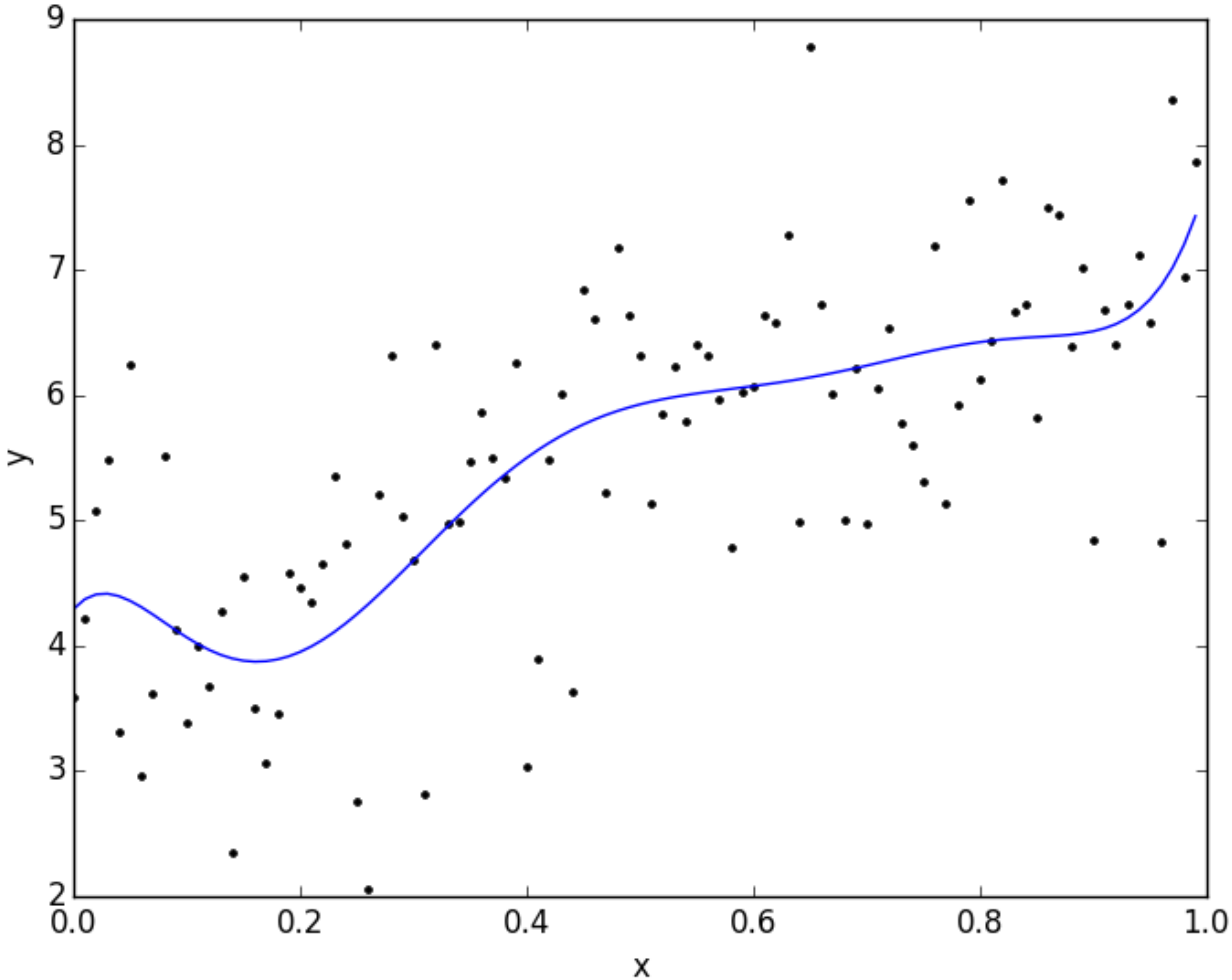
Cost function: $$L(\theta_0, \theta_1) = 1/m \sum_{\mathrm{trainingset}} (\theta_0 + \theta_1 \mathrm{size}_i - \mathrm{price}_i)^2$$

Example: house prices

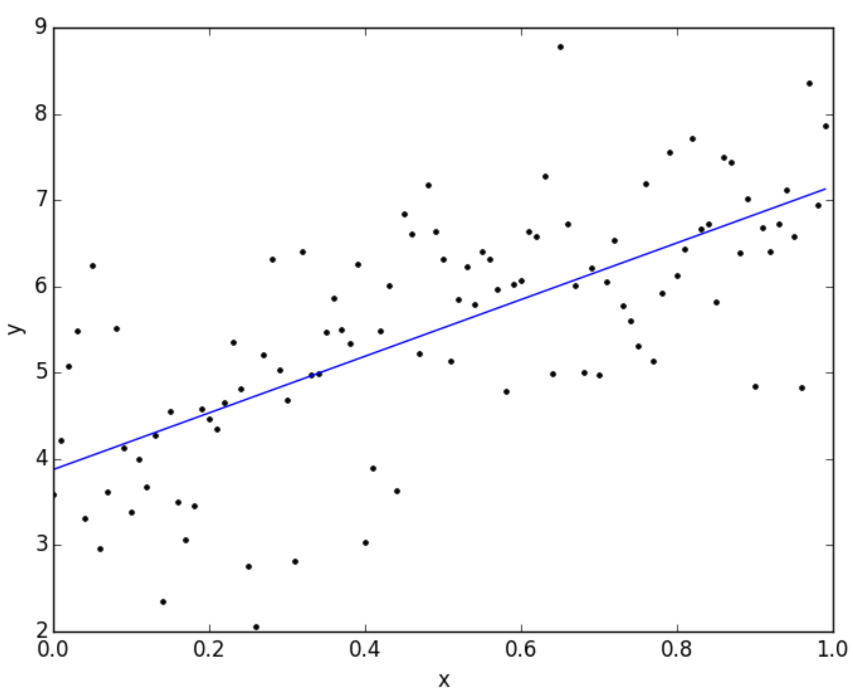
Minimizing the cost function

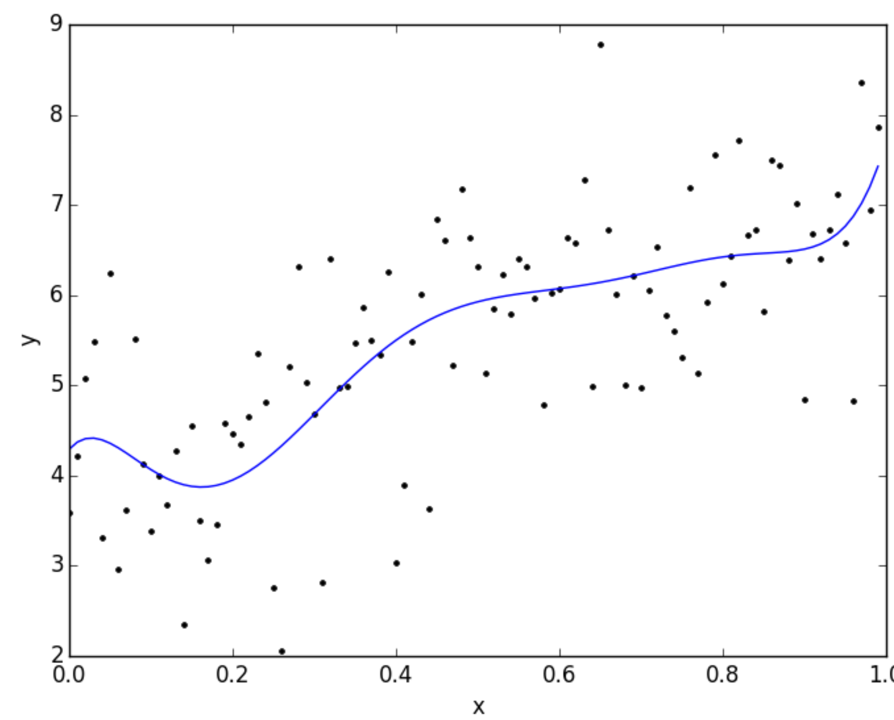


Bias-variance trade-off



Bias-variance trade-off





High-bias model

Underfits data

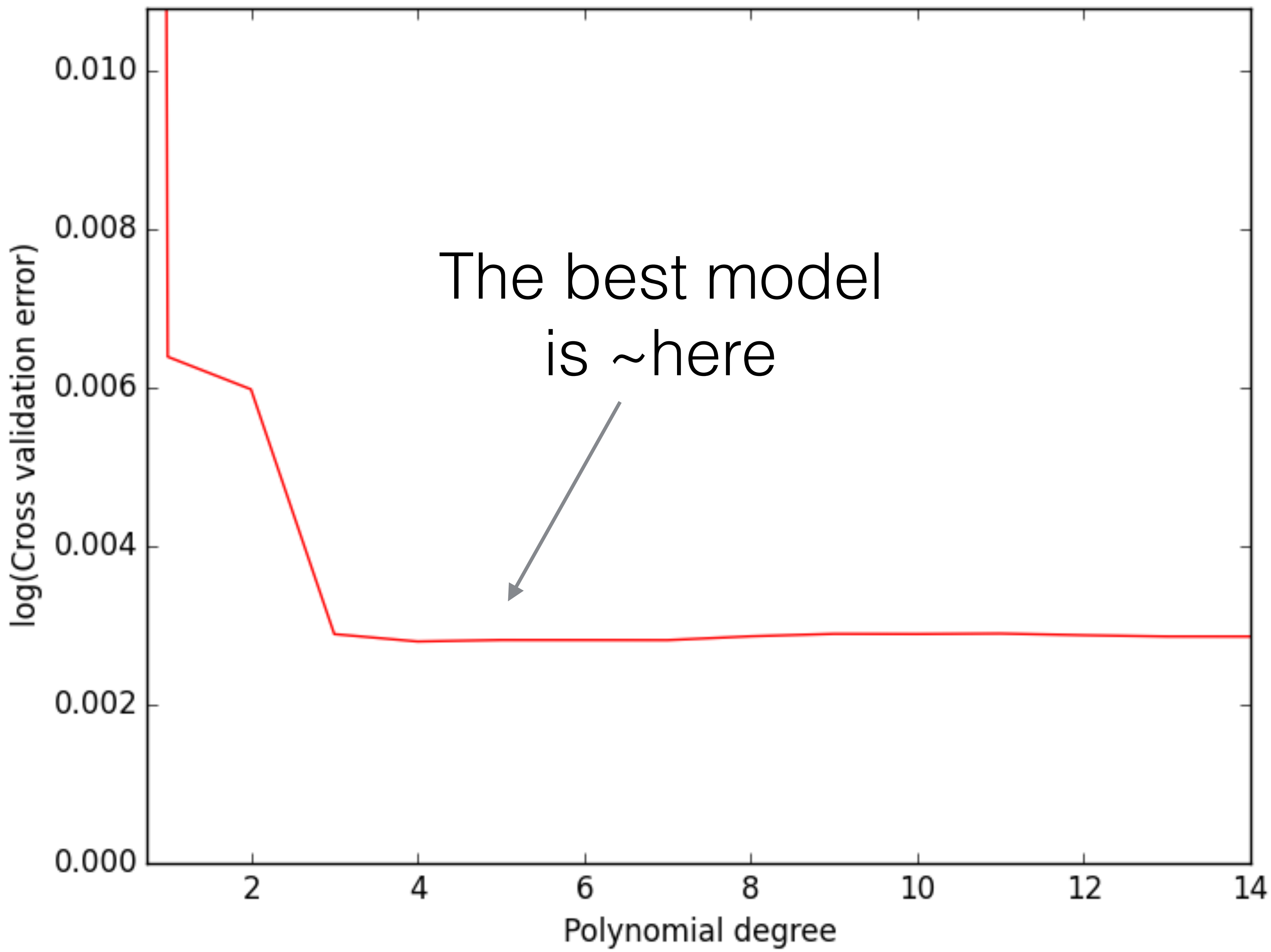
High-variance model

Overfits data

Cross-validation

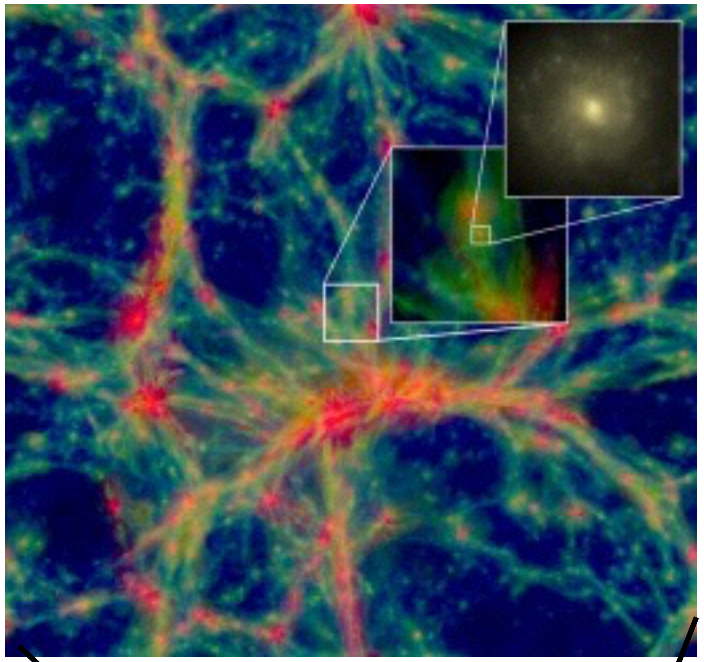
- The best model is one that gives the smallest error for new data points
- Separate data into training and cross-validation set
- Fit parameters on training set for different tuning parameters, evaluate on cross-validation set

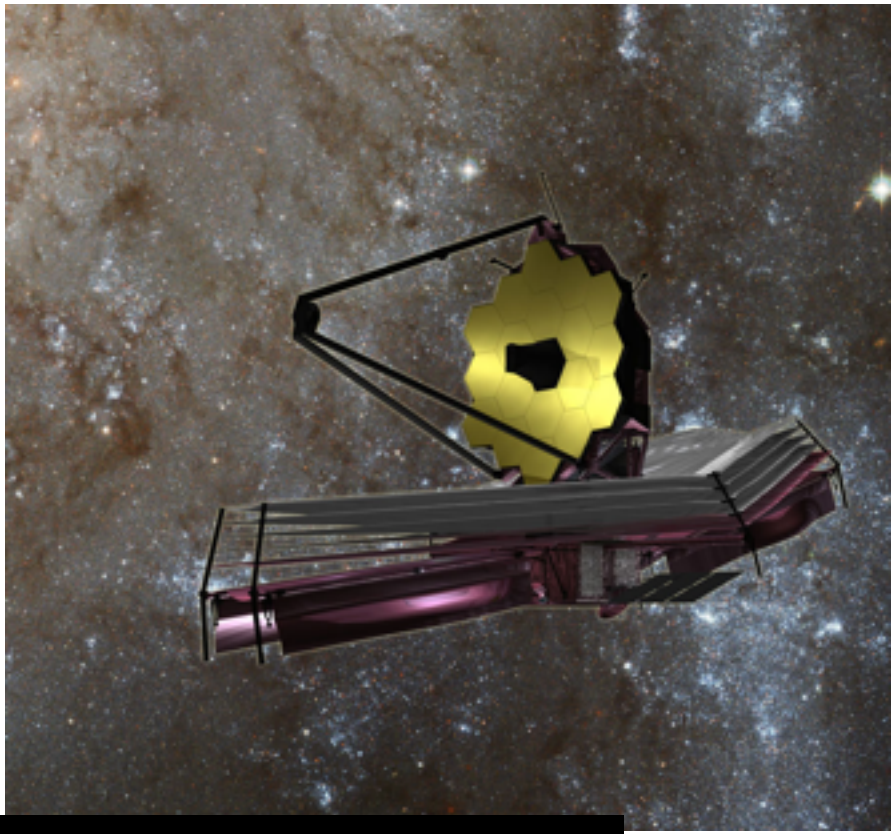
Cross-validation

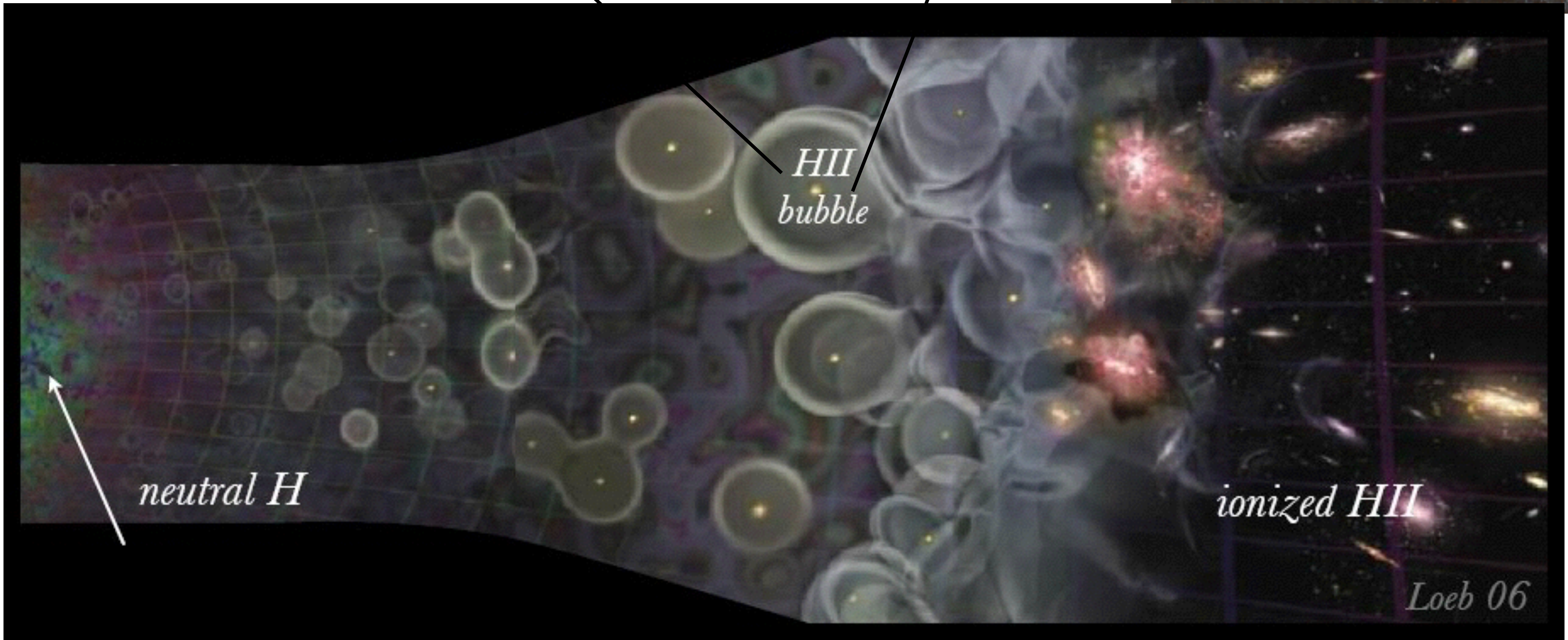


A real-world example

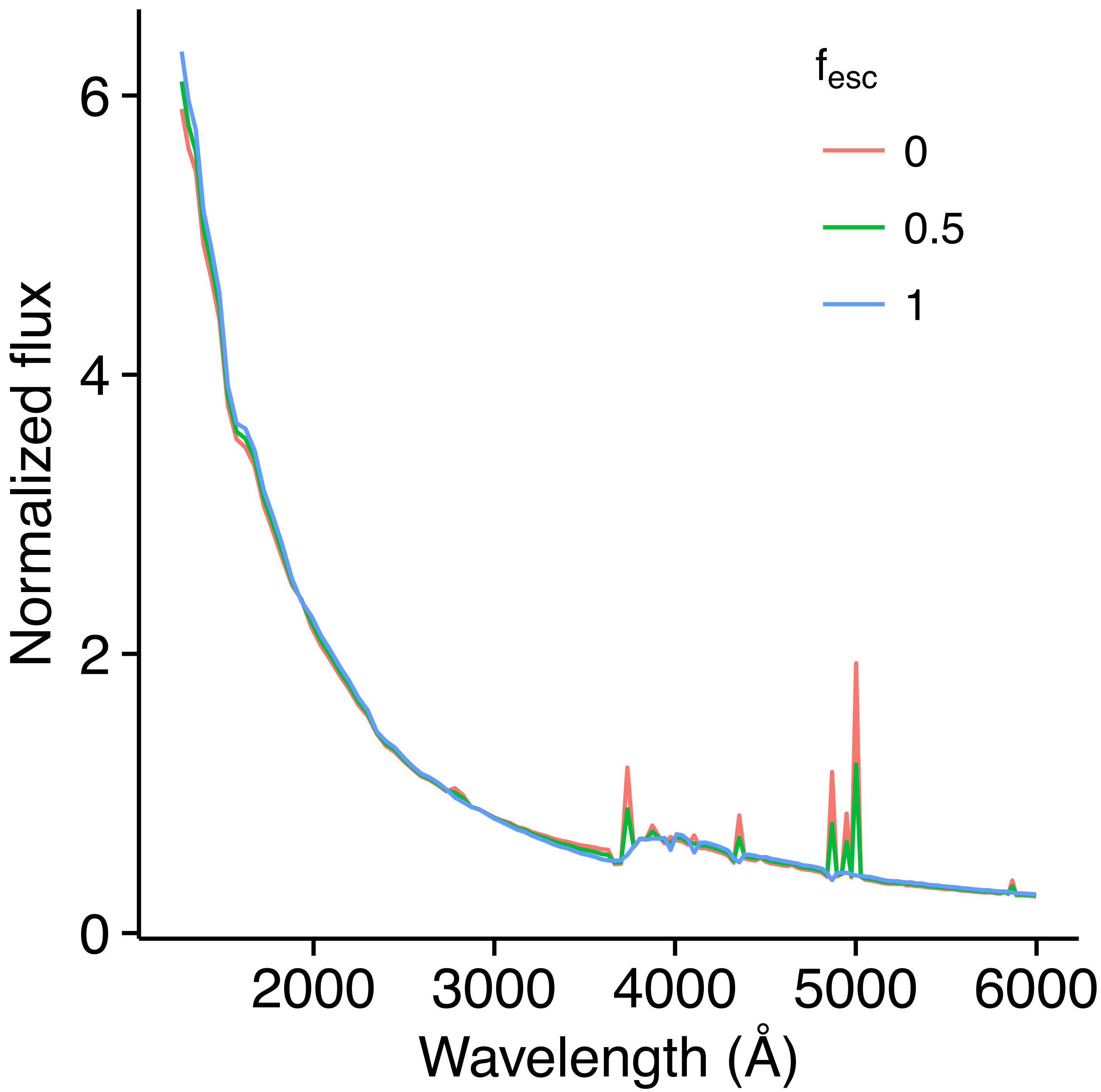
Predicting the escape fraction from galaxy spectra using Lasso regression escape fraction = fraction of ionizing photons that get out of a galaxy



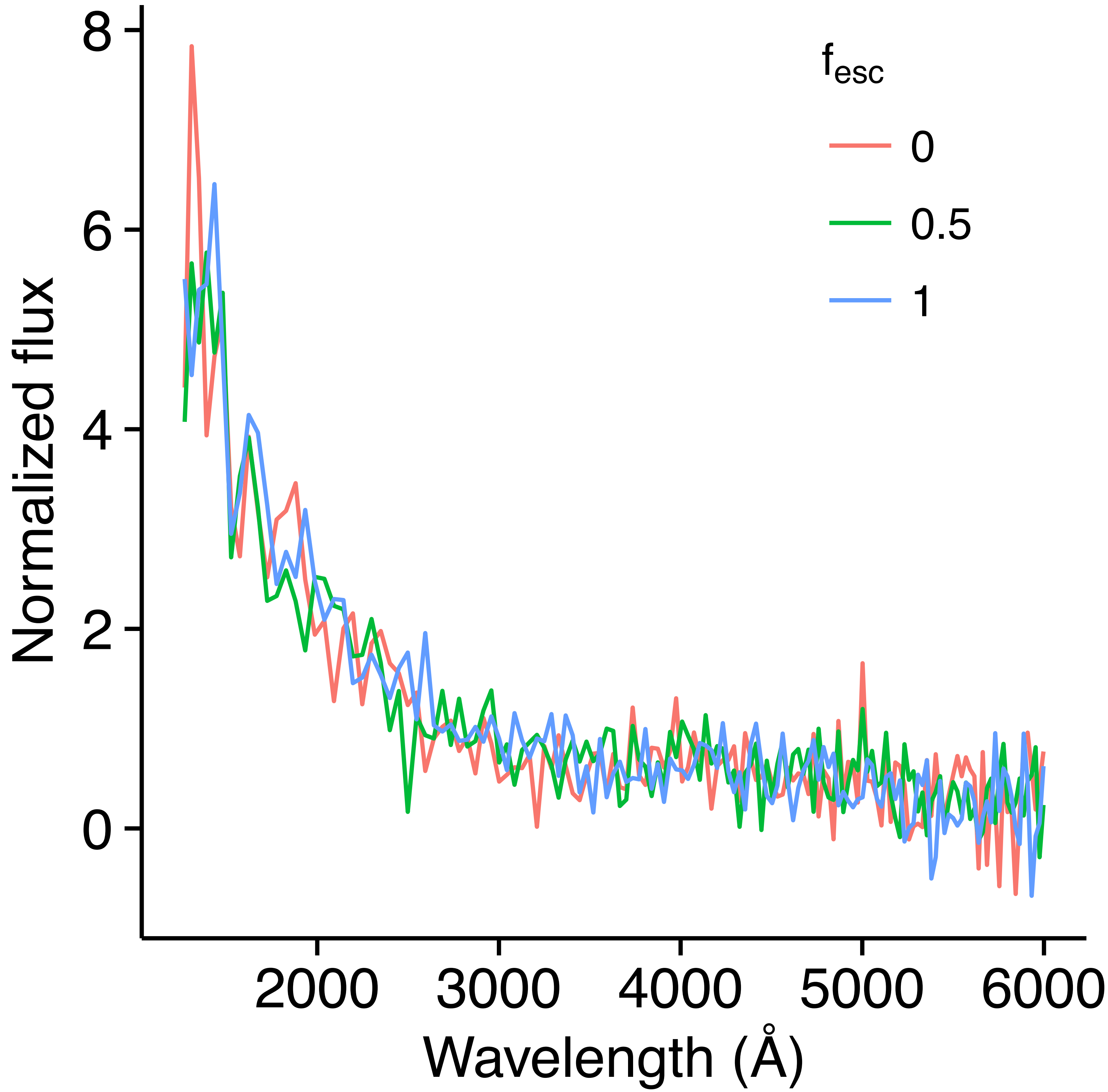




Example spectra



With simulated noise

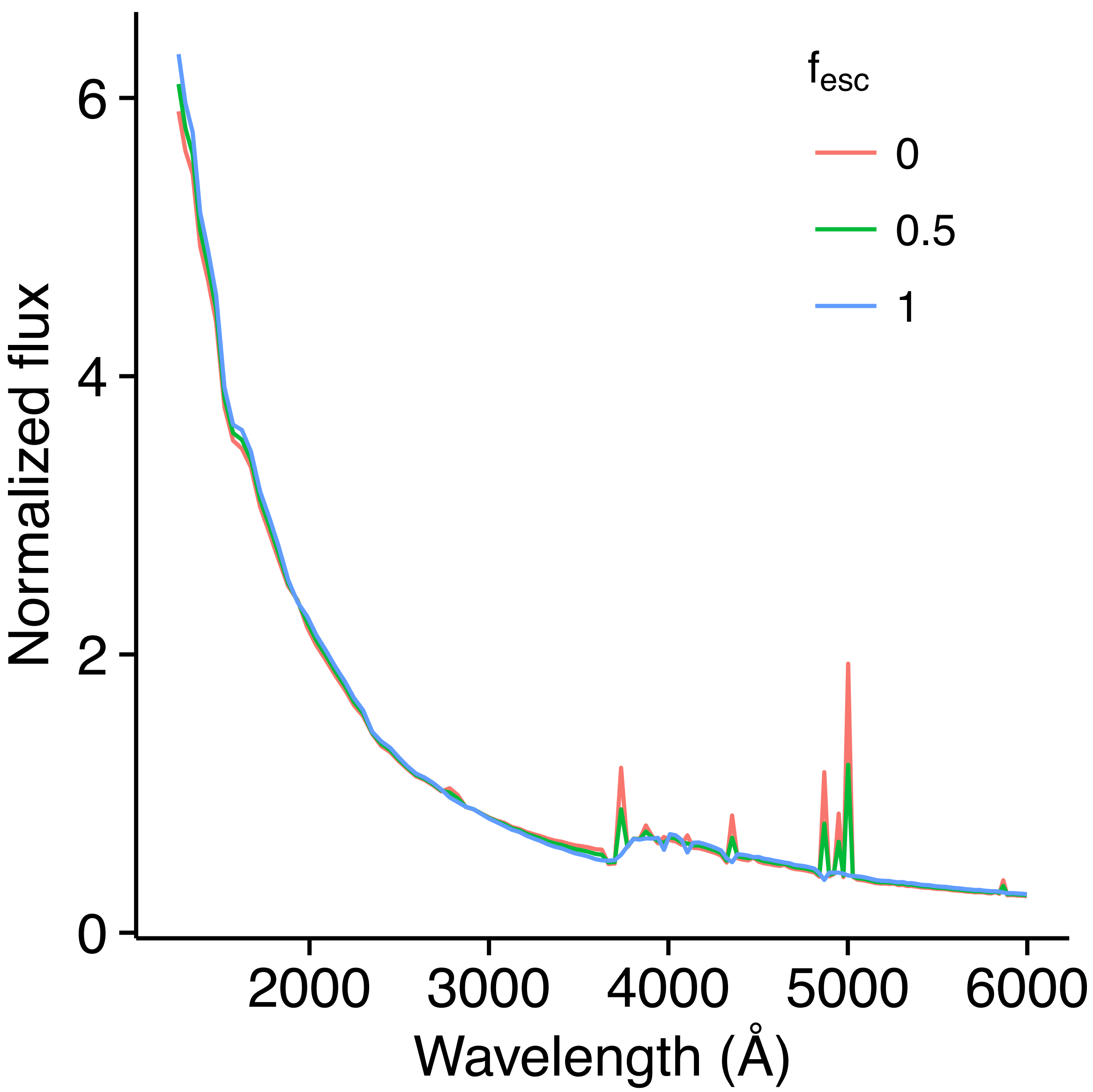


Approach to solve problem

Regression problem (supervised learning)

1. Identify features and labels
2. Pre-process data
3. Assume a model
4. Define a cost function
5. Use cross-validation to pick best model

1. Identify features and labels



Labels: escape fractions

Features: e.g. flux in each bin

2. Preprocess data

Make sure all spectra are binned in the same way

Normalize to have mean=1

Normalize each feature to have same variance

3. Assume a model

Flux in bin i

$$\hat{f}_{\mathrm{esc}} = \theta_0 + \sum_{i=1}^{N} \theta_i f_{\lambda,i}$$

Keep it simple,
not necessarily make the optimal model

4. Define a cost function

Lasso regression:

$$L(\theta) = \sum_{i=1}^{m} (\hat{f}_{\rm esc} - f_{\rm esc})^2 + \lambda \sum_{j=1}^{N} |\theta_j|$$

$$\hat{f}_{\rm esc} = \theta_0 + \sum_{i=1}^{N} \theta_i f_{\lambda,i}$$

Regularization term

Equivalent version of Lasso

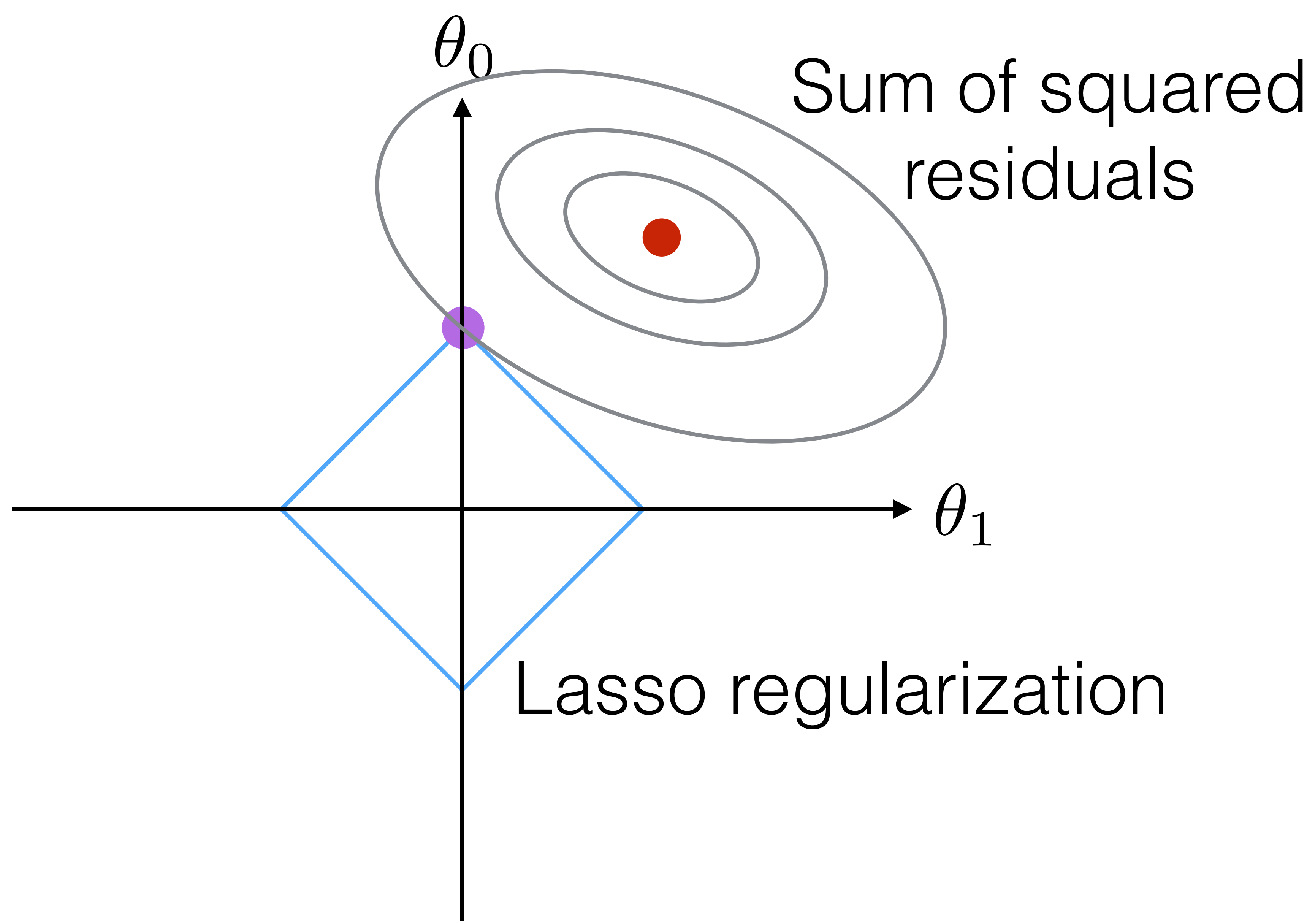
$$\underset{\theta}{\operatorname{argmin}} \sum (\hat{f}_{\mathrm{esc}} - f_{\mathrm{esc}})^2$$

s.t.

$$\sum_i |\theta_i| \leq s$$

Trivial example: fitting a line

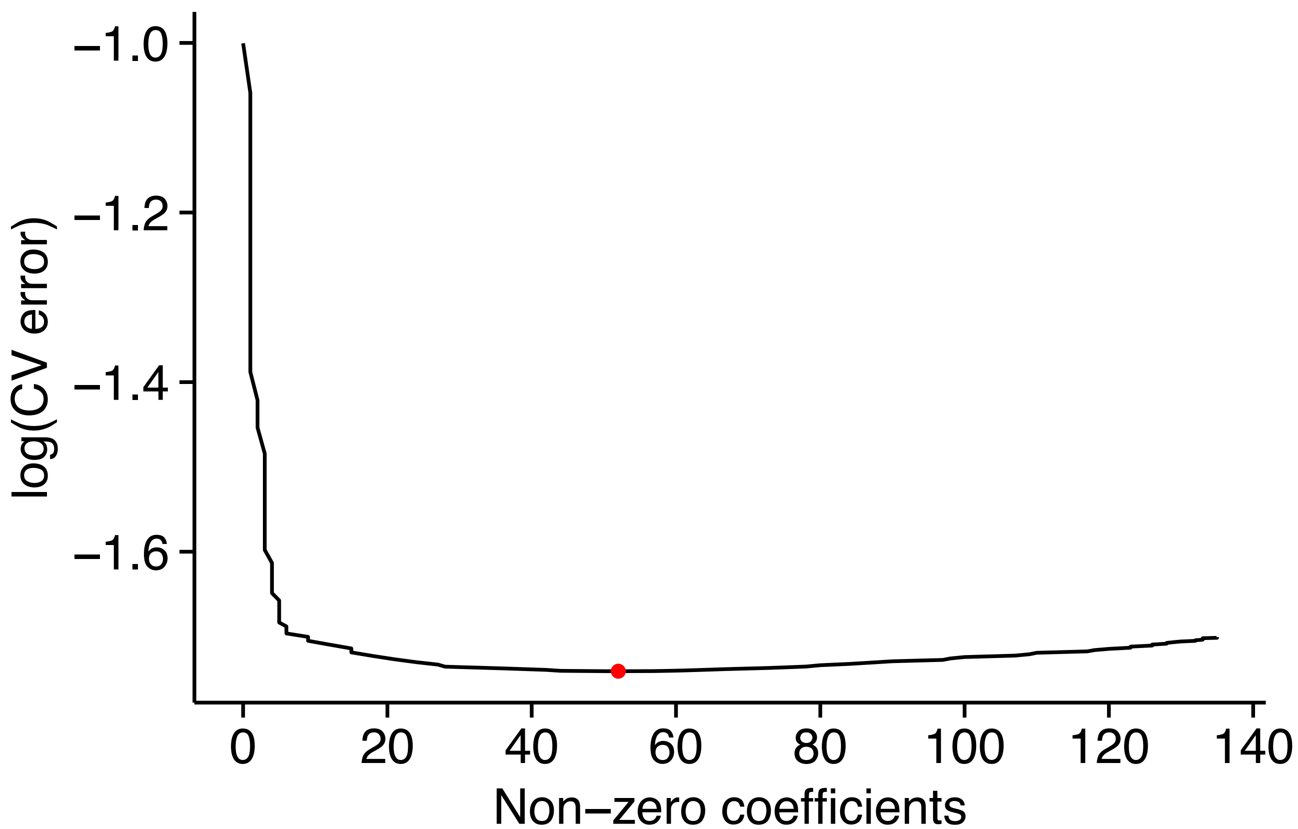
$$\hat{y} = \theta_0 + \theta_1 x$$

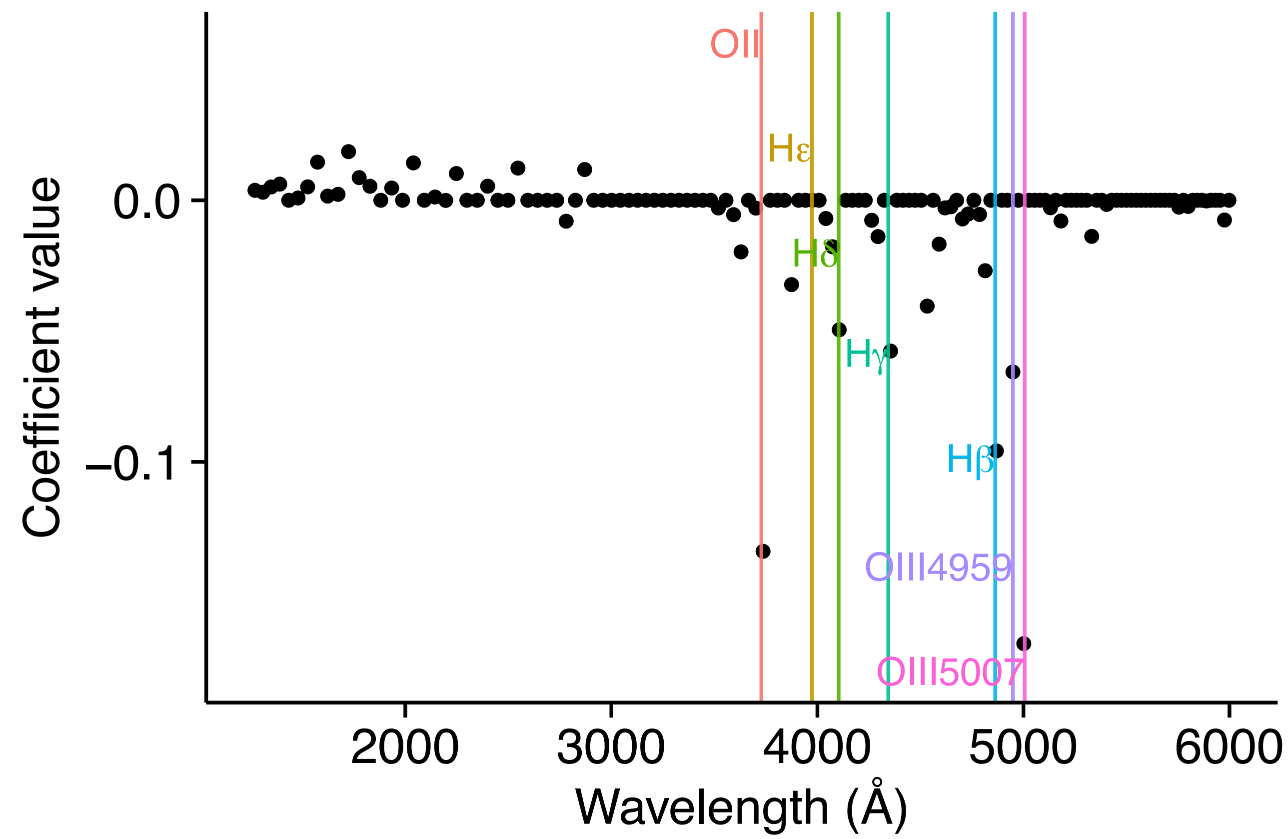


5. Fit model using cross-validation

Fitting model = minimizing cost function

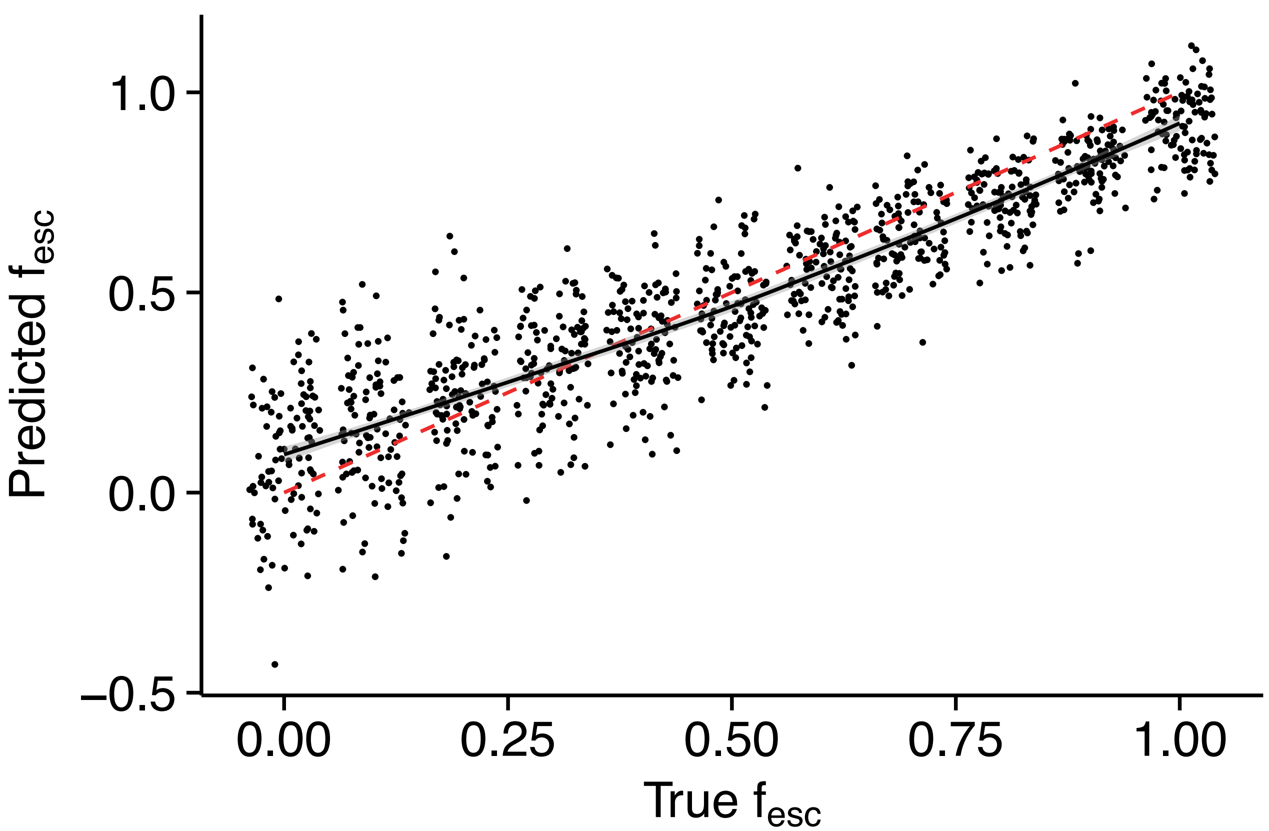
We want to find the best value for the regularization parameter



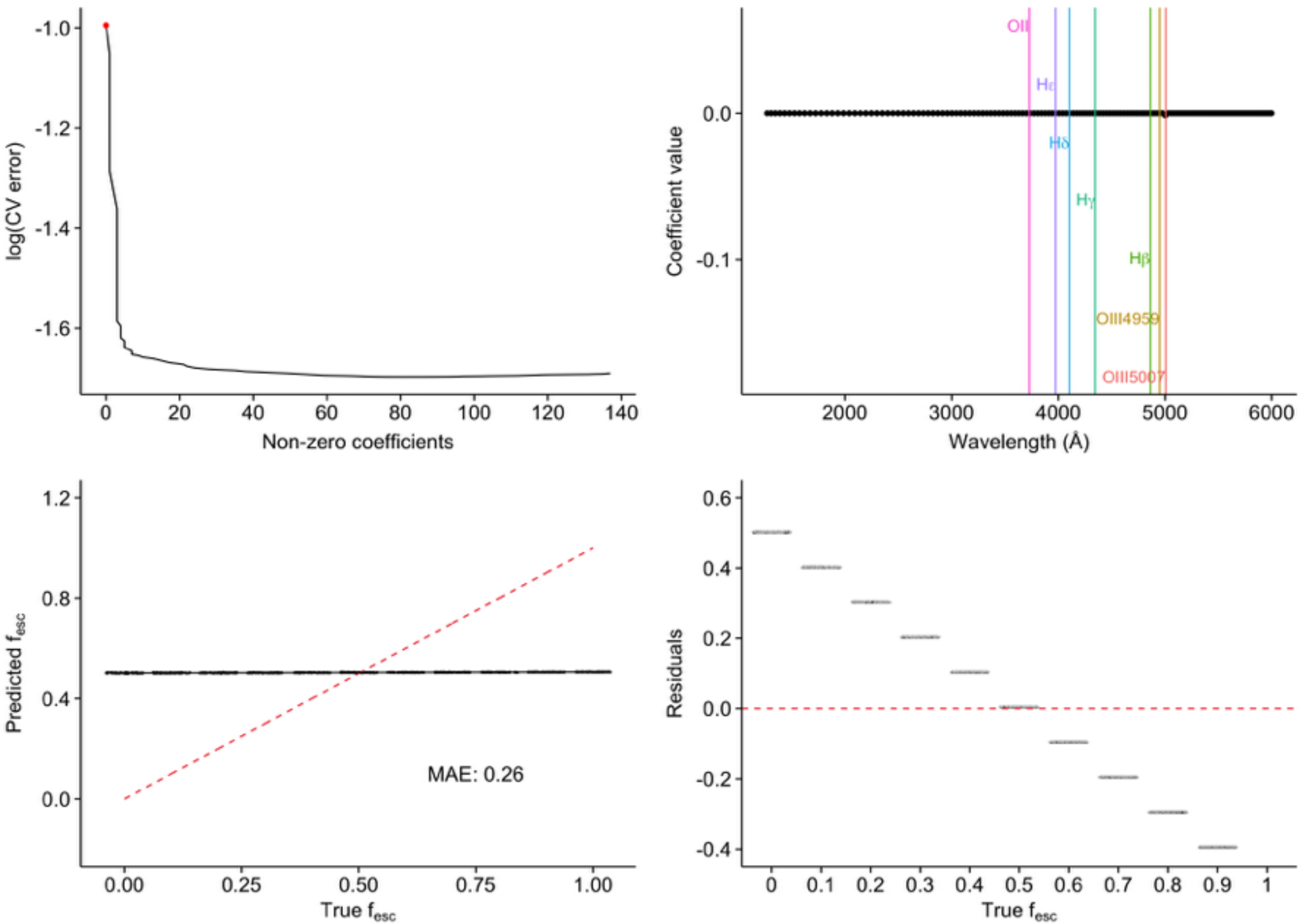


← Regularization parameter

Results on test set



MAE=0.10

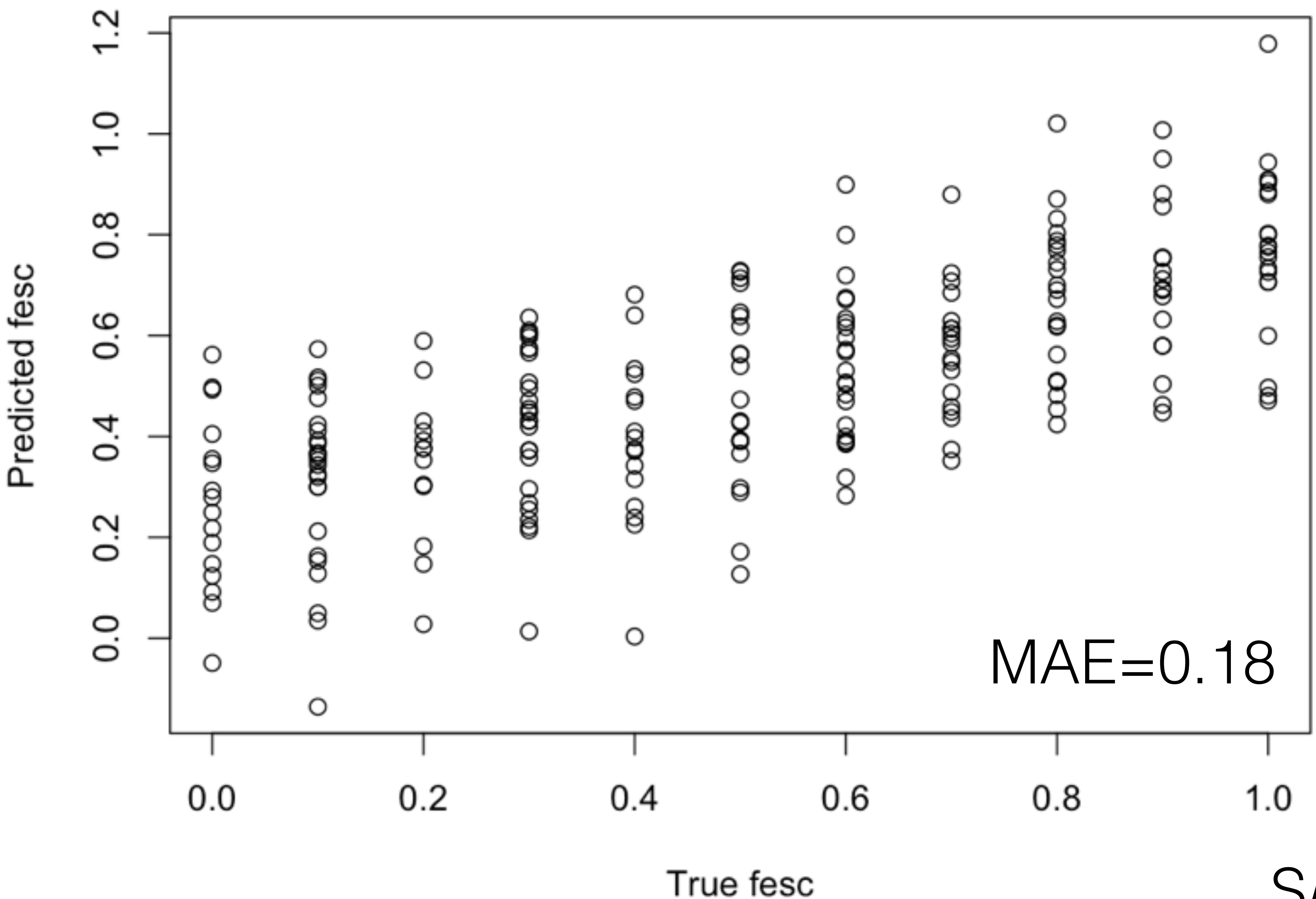


log(CV error)
-1.0
-1.2
-1.4
-1.6
0
20
40
60
80
100
120
140
Non-zero coefficients
Coefficient value
0.0
-0.1
OII
Hε
Hδ
Hγ
Hβ
OIII4959
OIII5007
2000
3000
4000
5000
6000
Wavelength (Å)
Predicted f_esc
1.2
0.8
0.4
0.0
MAE: 0.26
0.00
0.25
0.50
0.75
1.00
True f_esc
Residuals
0.6
0.4
0.2
0.0
-0.2
-0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
True f_esc

Previous approach

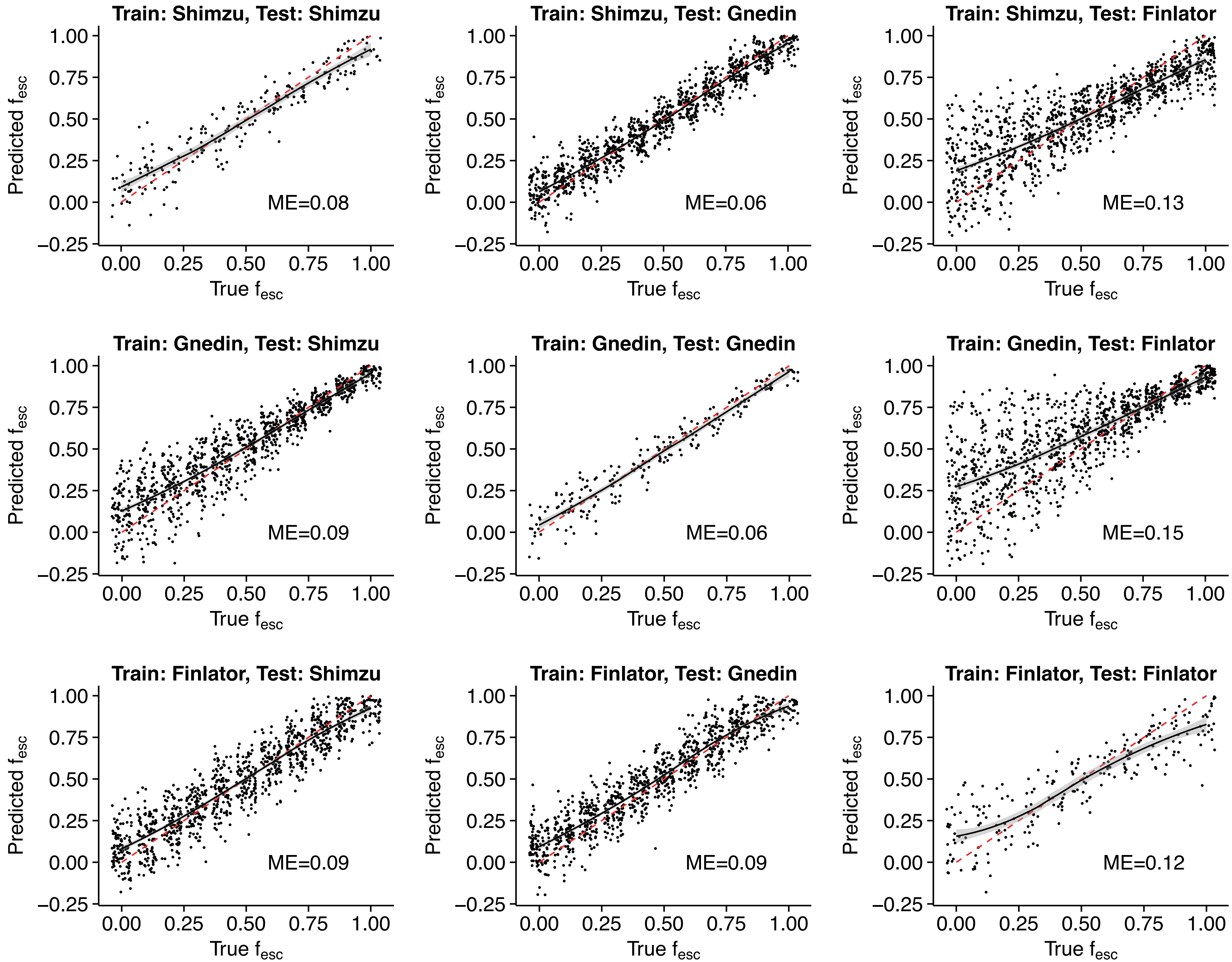
Zackrisson, Inoue, Jensen 2013

Slope of spectrum + strength of Hβ line



S/N=5

Still simulation dependent…



Data sets in the book

- SDSS - photometry and spectra of millions of objects
- 2MASS - photometry for stars from SDSS
- LINEAR - variable stars
- LIGO - simulated gravitational wave data
- Asteroid data from various sources