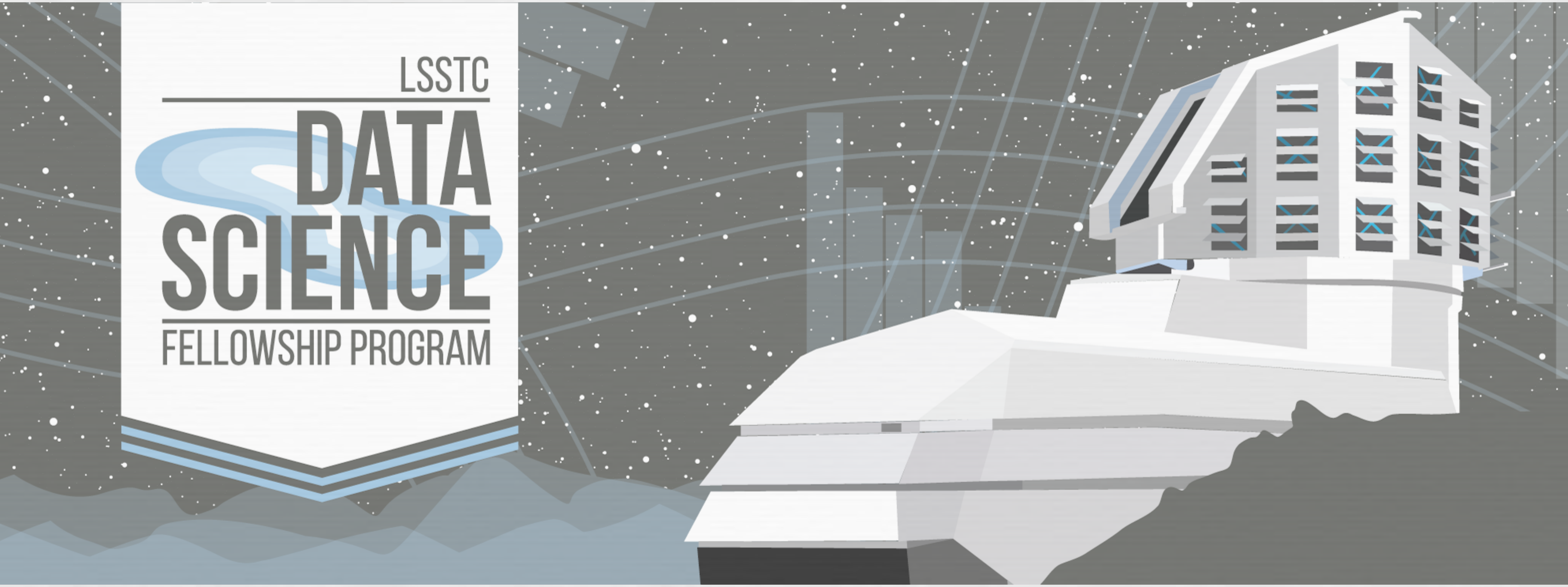
Developing the Machine Learning Workflow



Adam A Miller

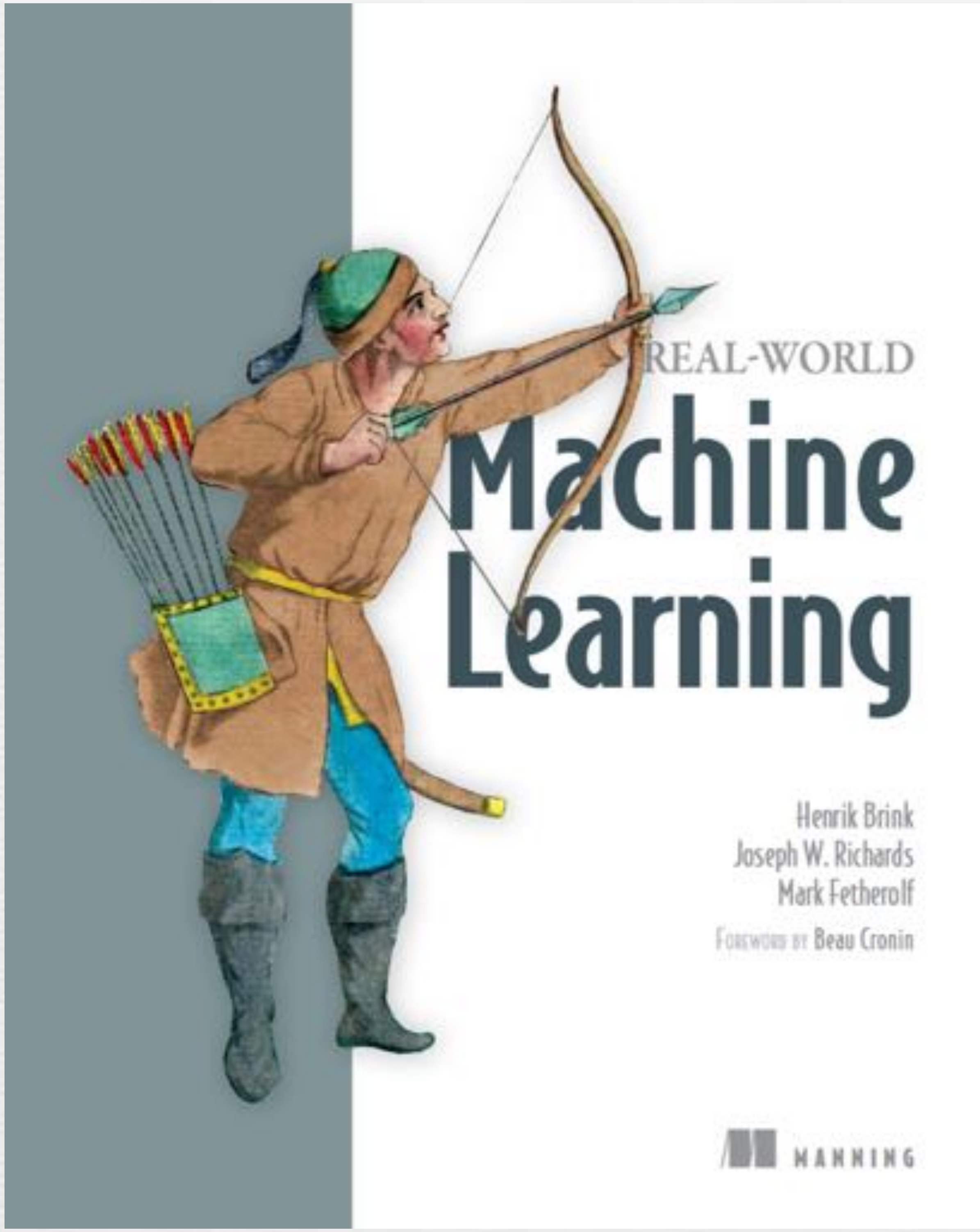
Northwestern/Adler Planetarium

2017 LSSTC DSFP
27 Jan 2016

CIERA
Center for Interdisciplinary Exploration and Research in Astrophysics

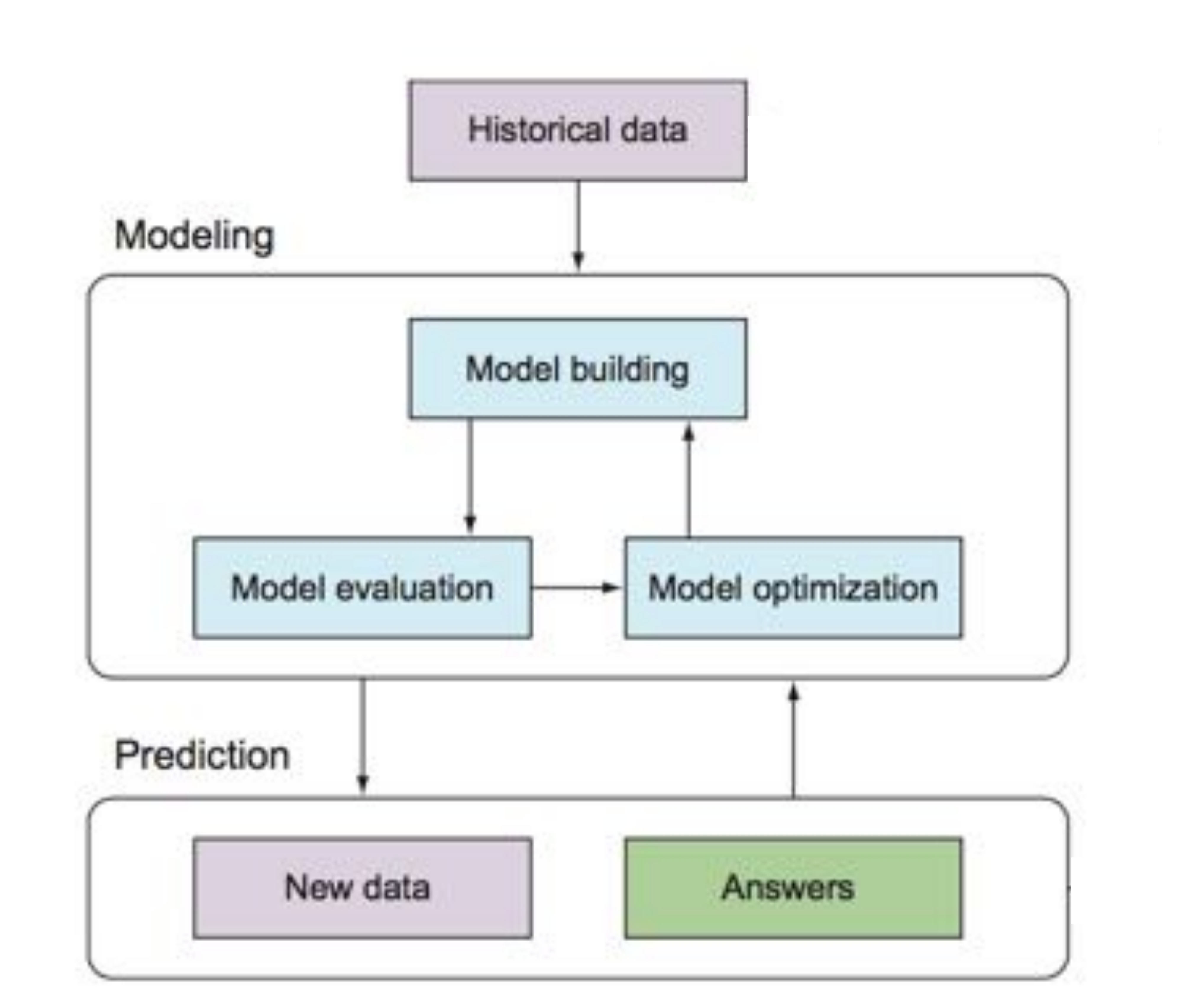
ADLER
PLANETARIUM

Developing the Machine Learning Workflow



Brink, Richards, & Fetherolf 16

The Machine Learning Workflow



BRF16

Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

Worry About The Data

Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

Select features to use in the model

Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

Select features to use in the model

Determine “ground truth” or labels for the training set

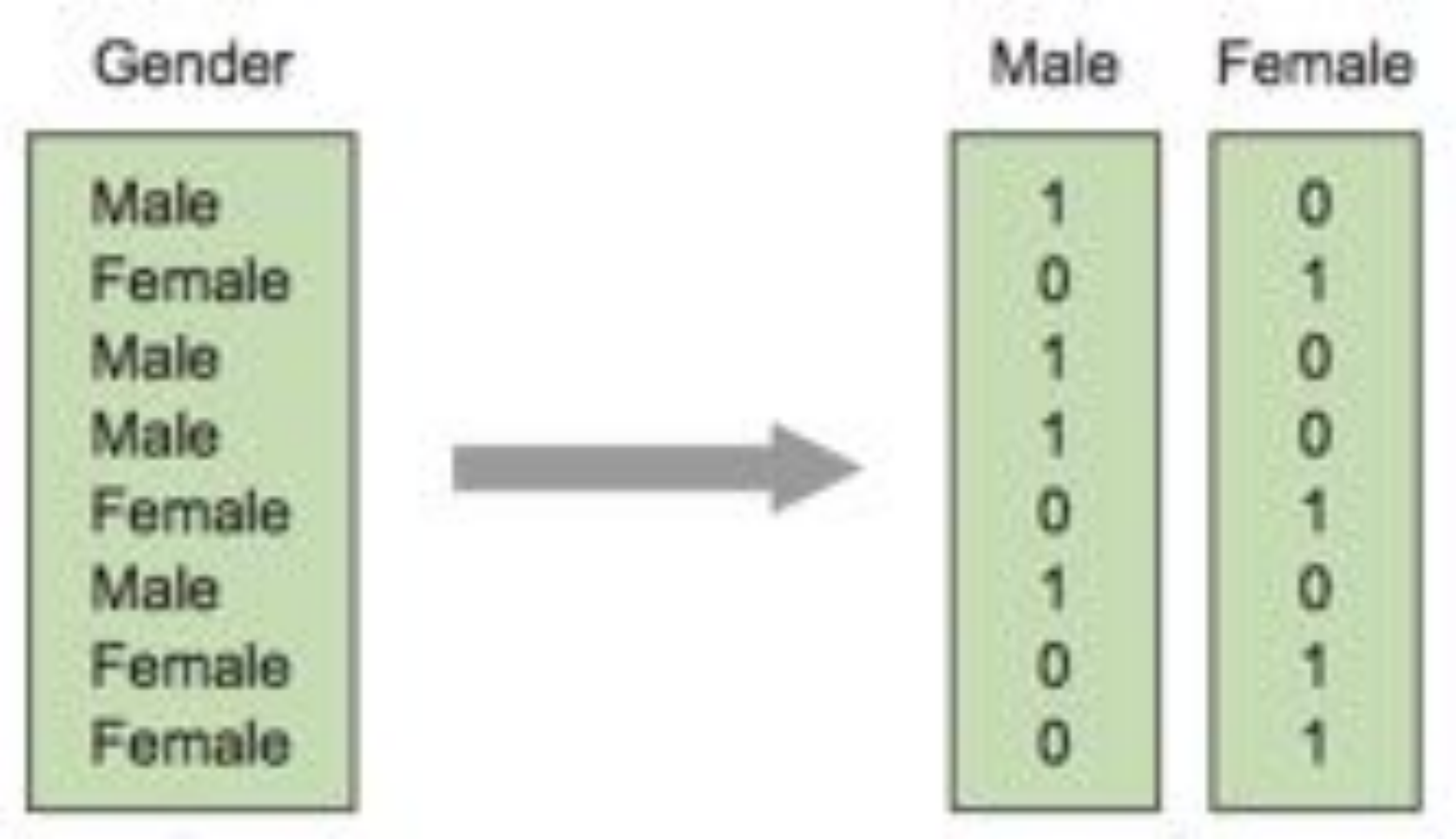
Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

Select features to use in the model

Determine “ground truth” or labels for the training set

Convert categorical features



Gender
Male
Female
Male
Male
Female
Male
Female
Female

Male	Female
1	0
0	1
1	0
1	0
0	1
1	0
0	1
0	1

Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

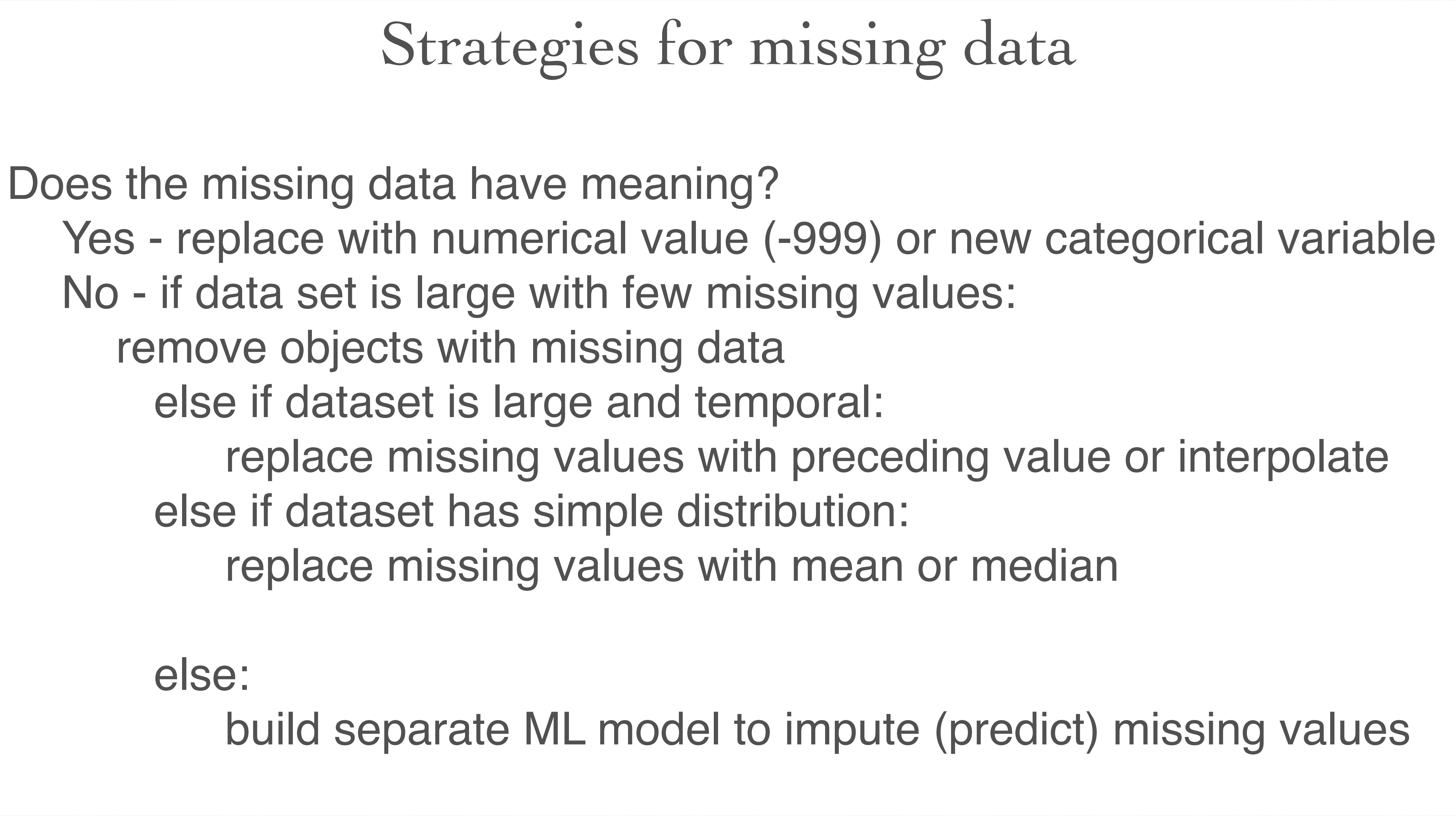
Select features to use in the model

Determine “ground truth” or labels for the training set

Convert categorical features

Impute (or throw out?) missing data

Data Preparation



Strategies for missing data

```
Does the missing data have meaning?
    Yes - replace with numerical value (-999) or new categorical variable
    No - if data set is large with few missing values:
        remove objects with missing data
           else if dataset is large and temporal:
                replace missing values with preceding value or interpolate
           else if dataset has simple distribution:
                replace missing values with mean or median

           else:
                build separate ML model to impute (predict) missing values
```

Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

Select features to use in the model

Determine “ground truth” or labels for the training set

Convert categorical features

Impute (or throw out?) missing data

Normalize the features

Data Preparation

Query, observe, simulate, etc. - collect data that needs to be modeled

Select features to use in the model

Determine “ground truth” or labels for the training set

Convert categorical features

Impute (or throw out?) missing data

Normalize the features

Visualize the data

Worry About The Data

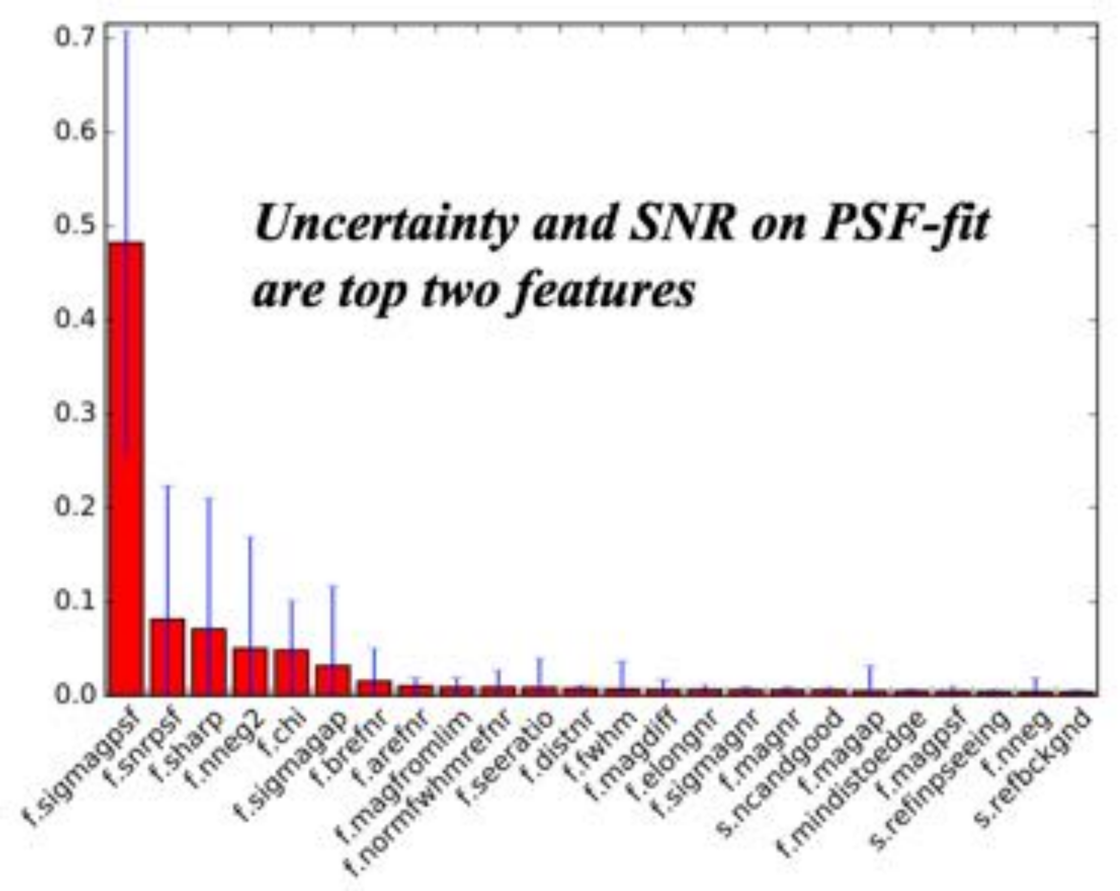
Feature Engineering

Add new features - if necessary

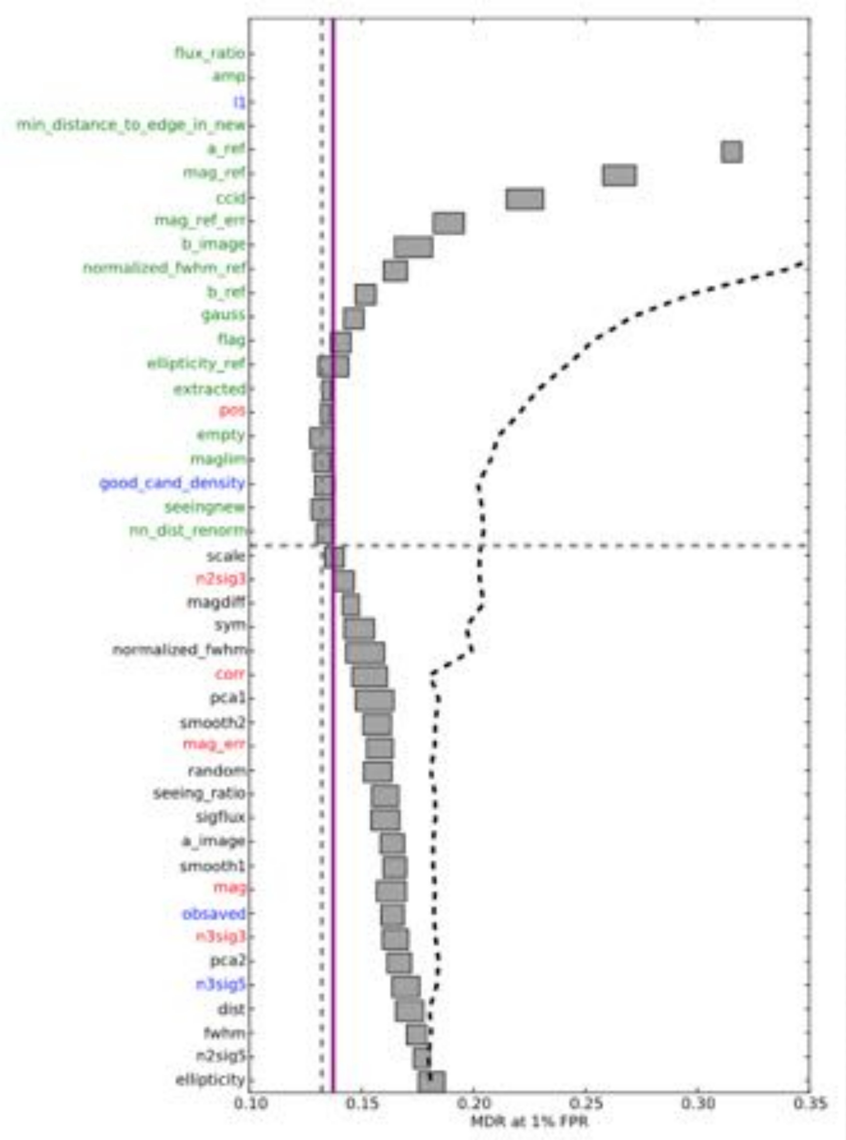
- Utilize domain knowledge to create/compute new features
- Combine features or represent in an alternative fashion

Remove noisy/uniformative features - if necessary

- Determine feature importance (RF)
- Forward/backward selection to iteratively remove features

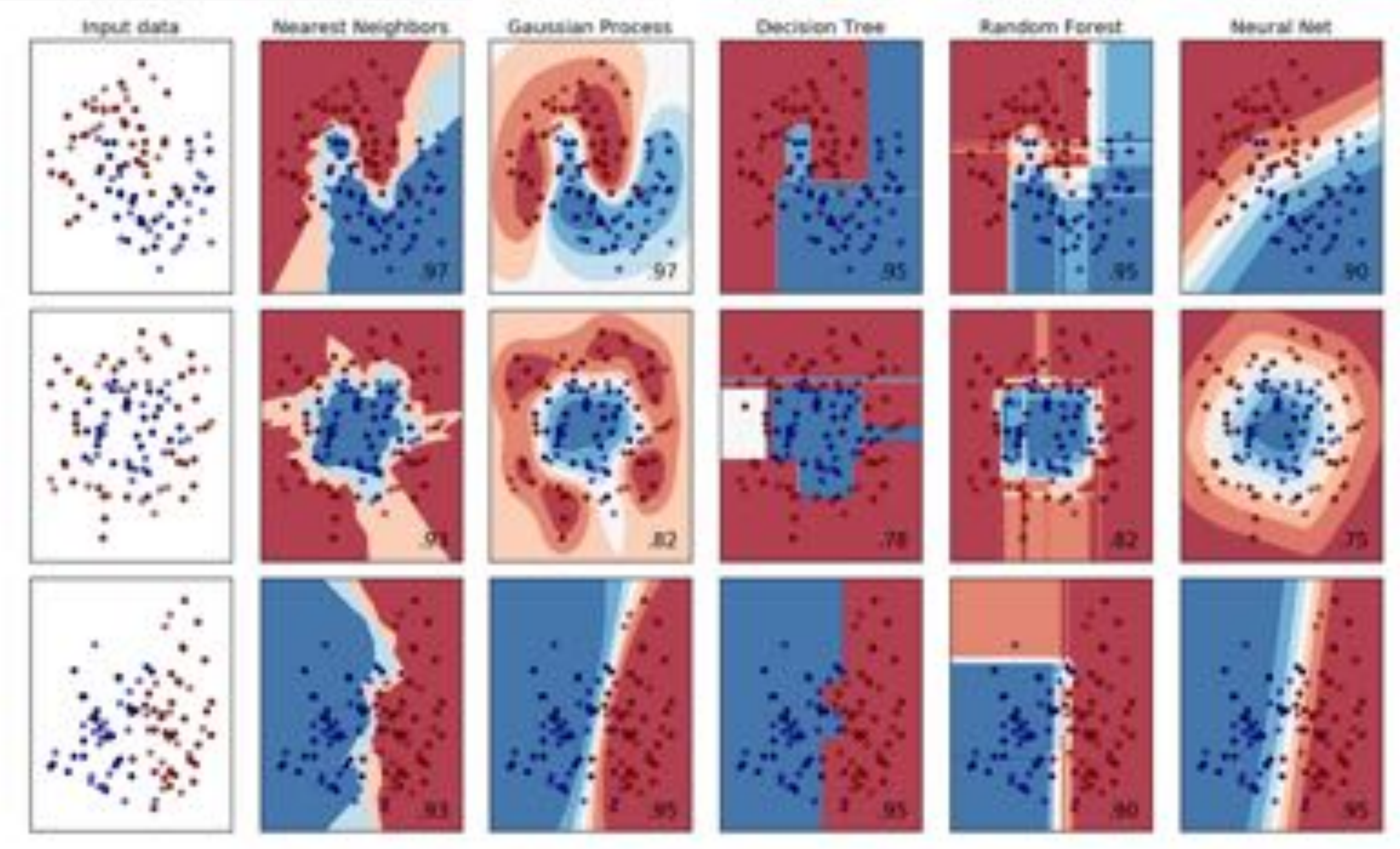


Rebbapragada+16



Brink+13

Model Selection

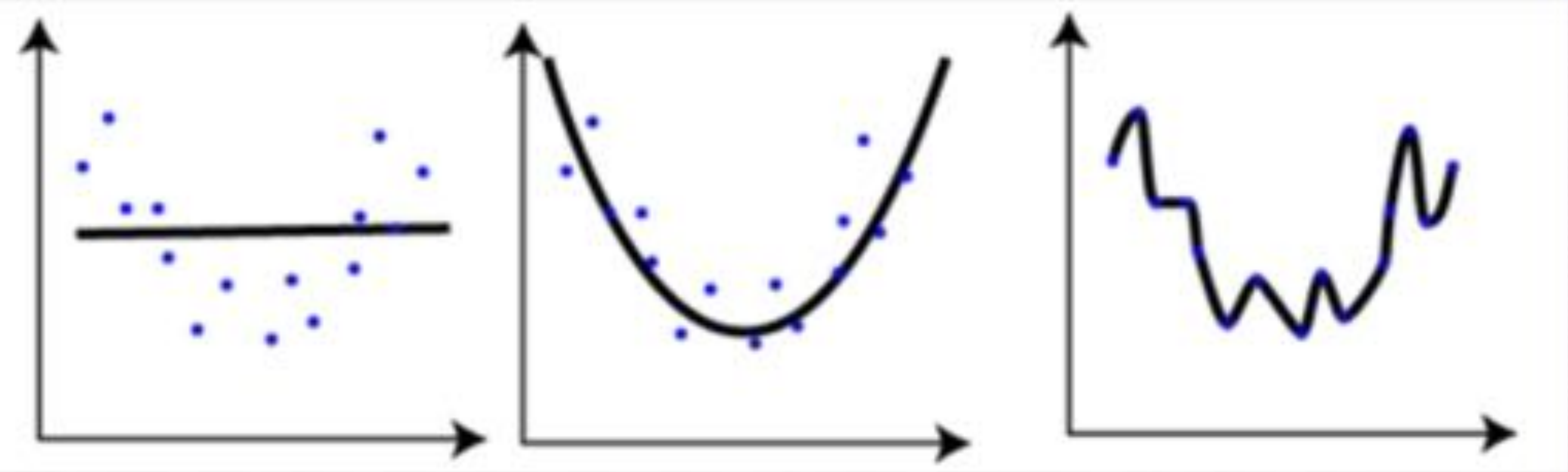


credit:scikit-learn

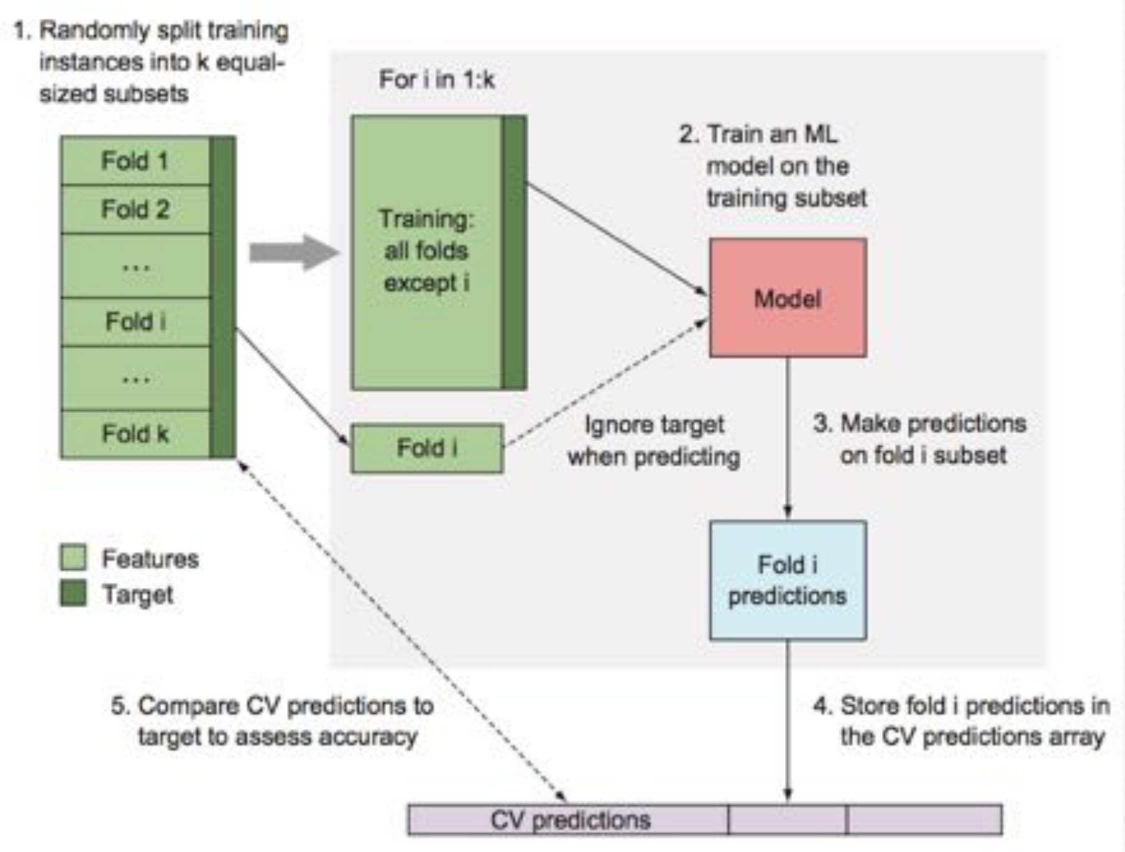
Worry About The Data

Model Evaluation

Avoid under- and over-fitting



Daniela’s lecture (Day 2)



1. Randomly split training instances into k equal-sized subsets

For i in 1:k

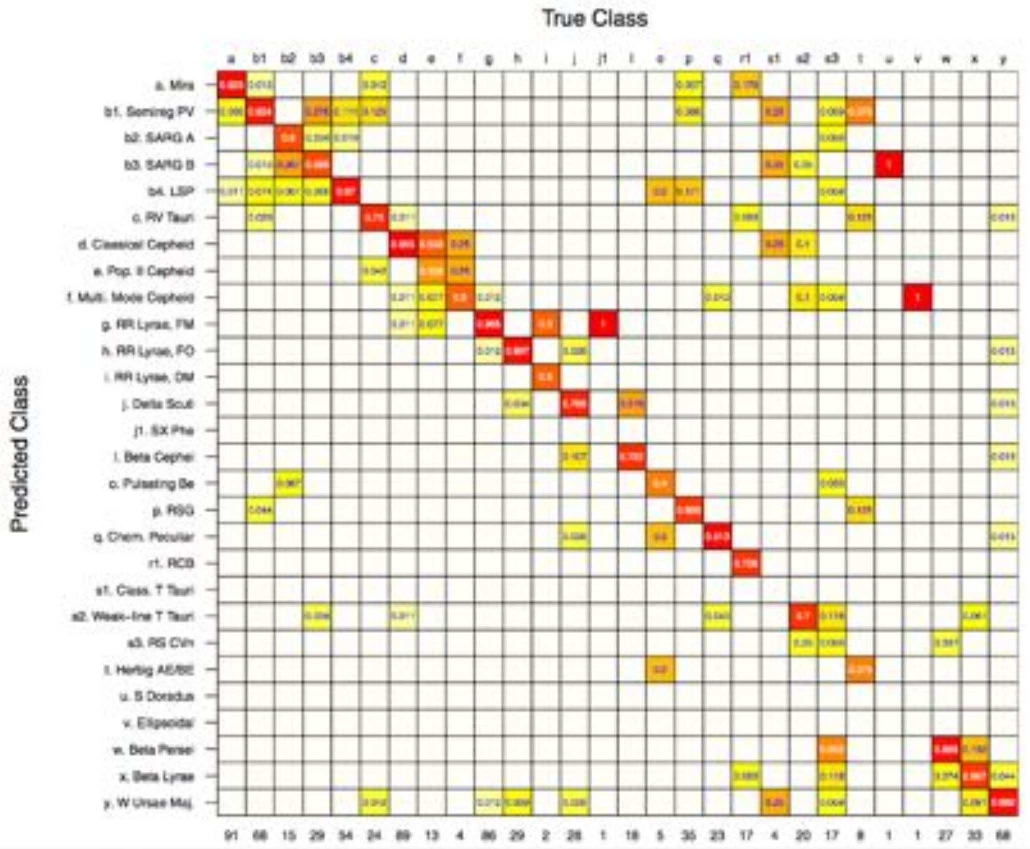
2. Train an ML model on the training subset

3. Make predictions on fold i subset

4. Store fold i predictions in the CV predictions array

5. Compare CV predictions to target to assess accuracy

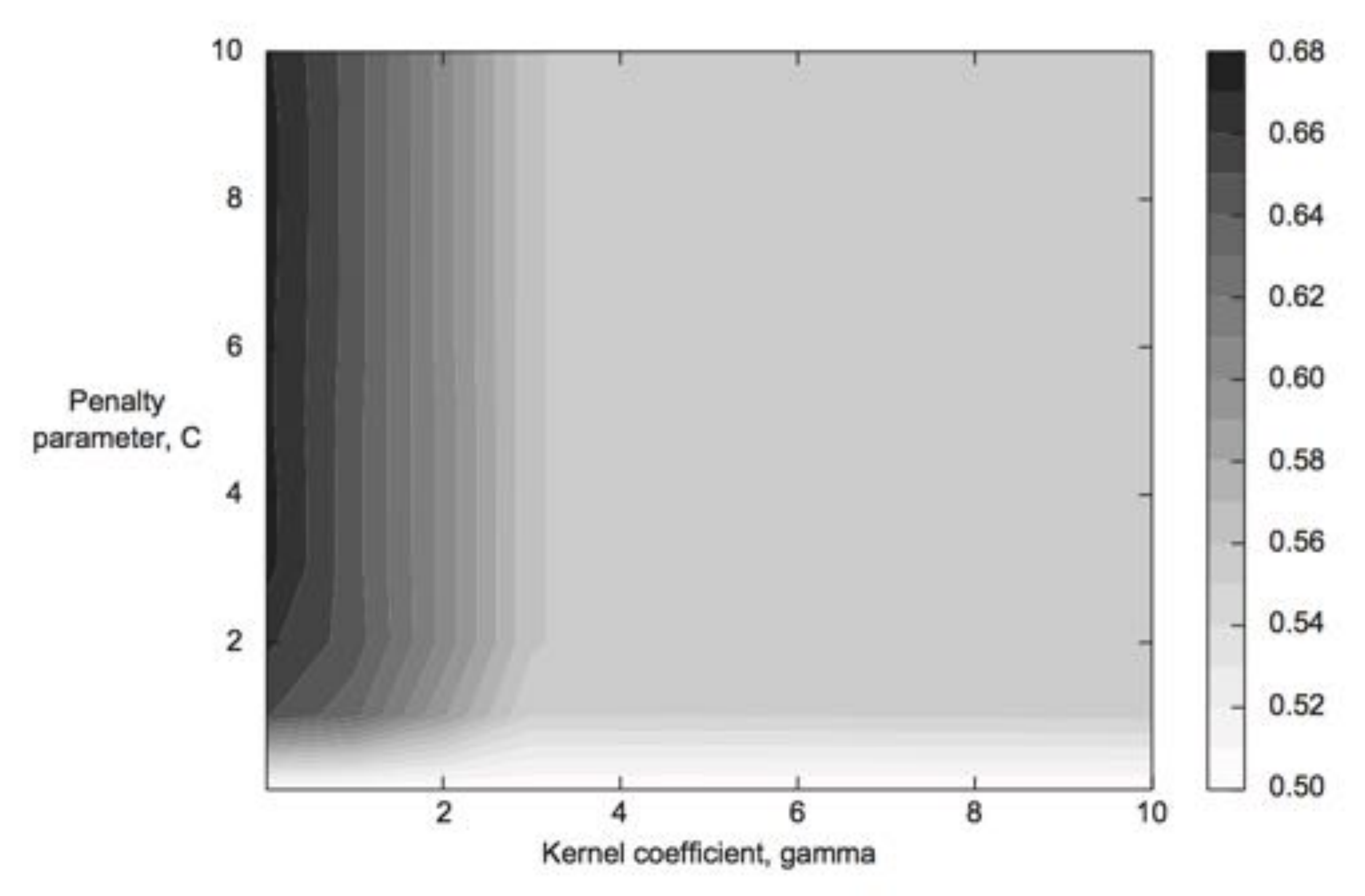
BRF 16



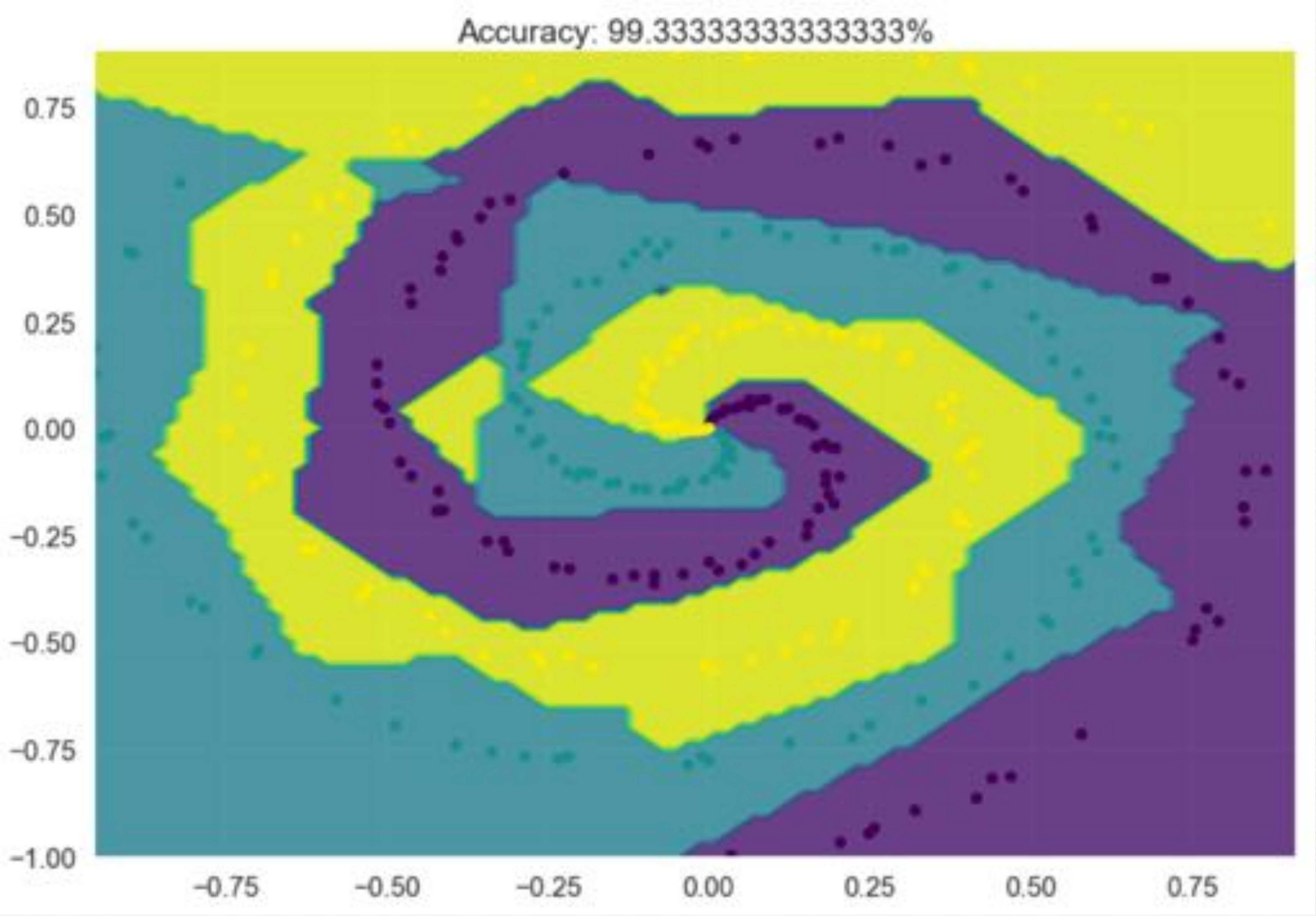
Richards+12

Model Optimization

Identify optimal tuning parameters via grid search

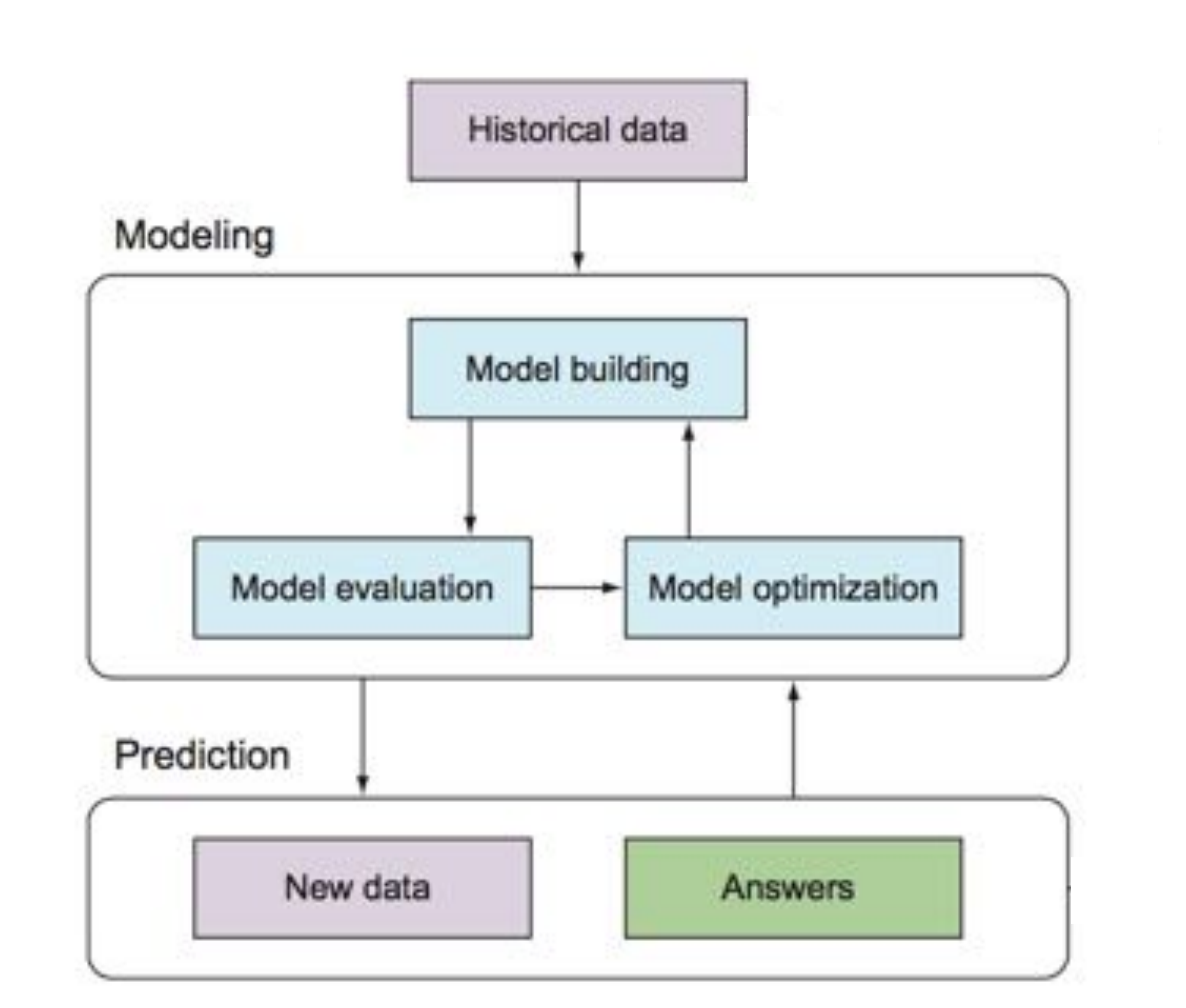


Model Prediction



Worry About The Data

The Machine Learning Workflow



BRF16