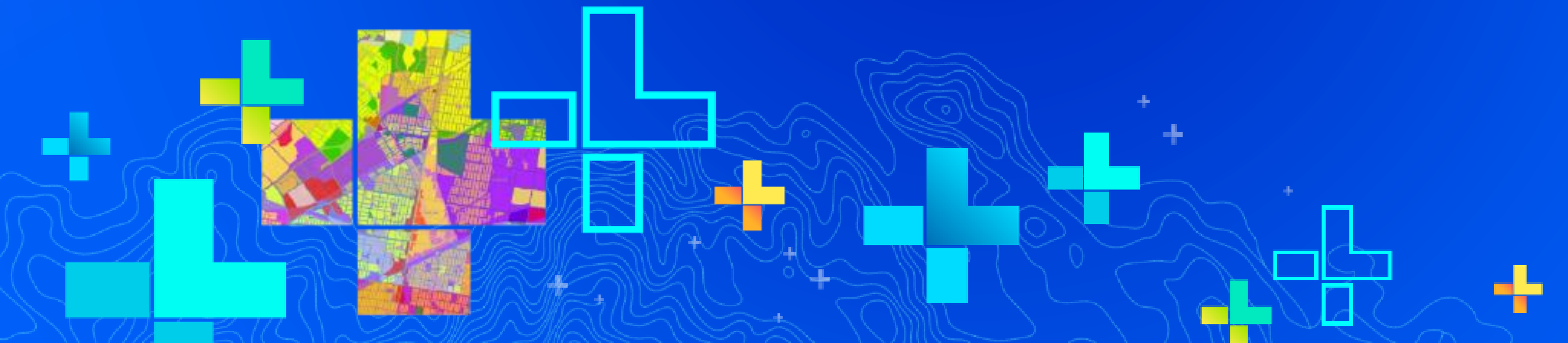
The Forest for the Trees: Making Predictions using Forest-Based Classification and Regression

Lauren Bennett

Alberto Nieto

Flora Vale

esriurl.com/spatialstats



SEE
WHAT
OTHERS
CAN'T

Models

Representative generalizations used for prediction

Why model

Use information we have to predict information we don’t have

Which areas are most contaminated?

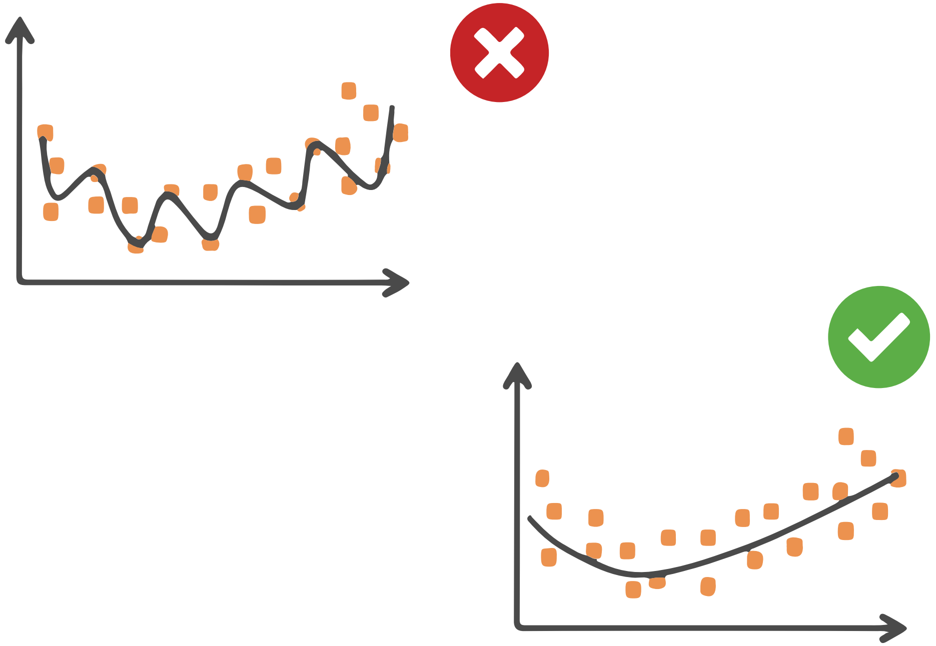
What drives sales?

Which buildings will fail inspection?

What will the weather be like tomorrow?

When we can’t trust a model

Mimics training dataset and models noise instead of generalizing a trend



Many many ways to model

Generalized Linear Regression
Geographically Weighted Regression

Forest-based Classification and Regression

Forest-based Classification & Regression

Predicting using machine learning

Training

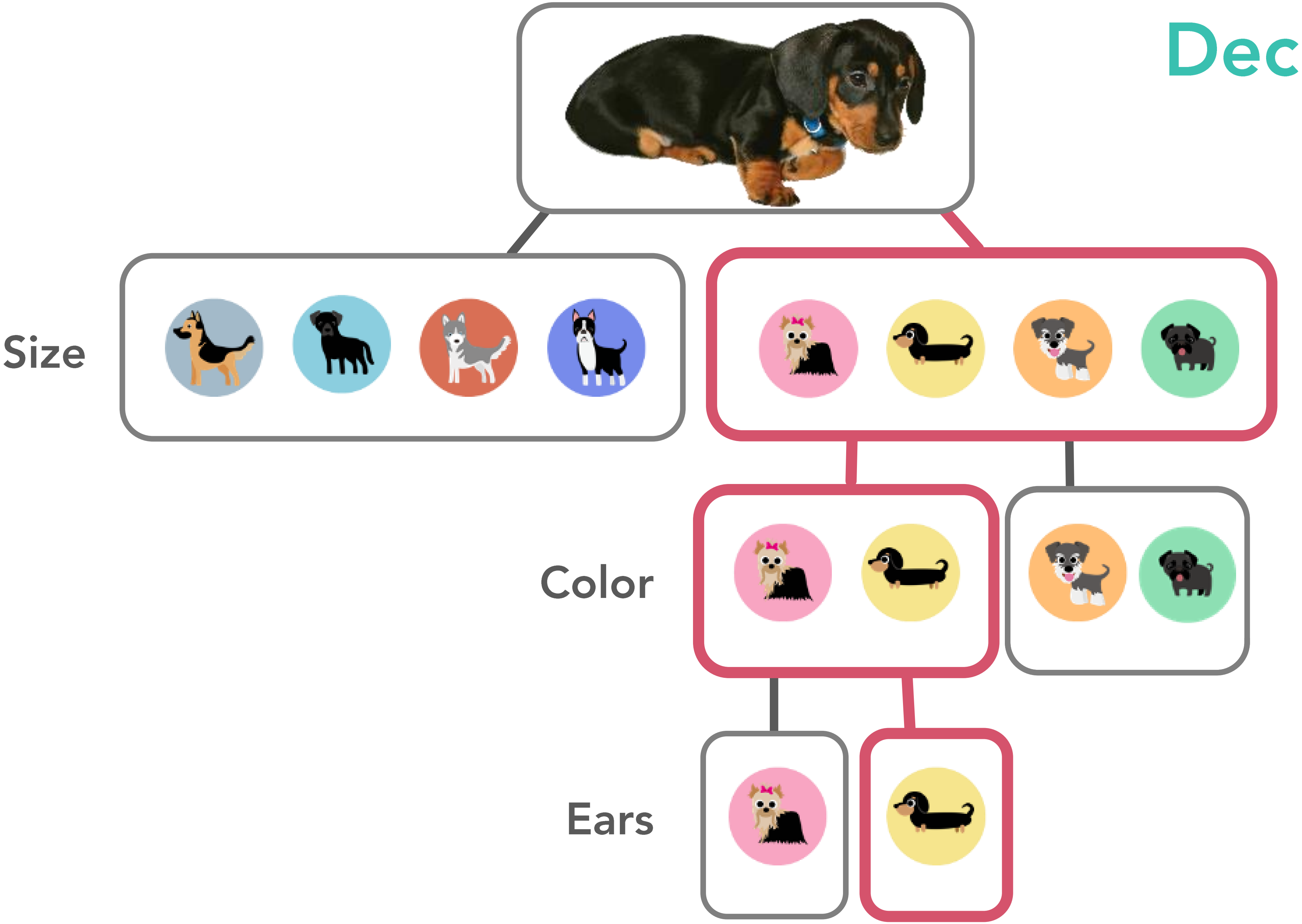
variable to predict

Breed

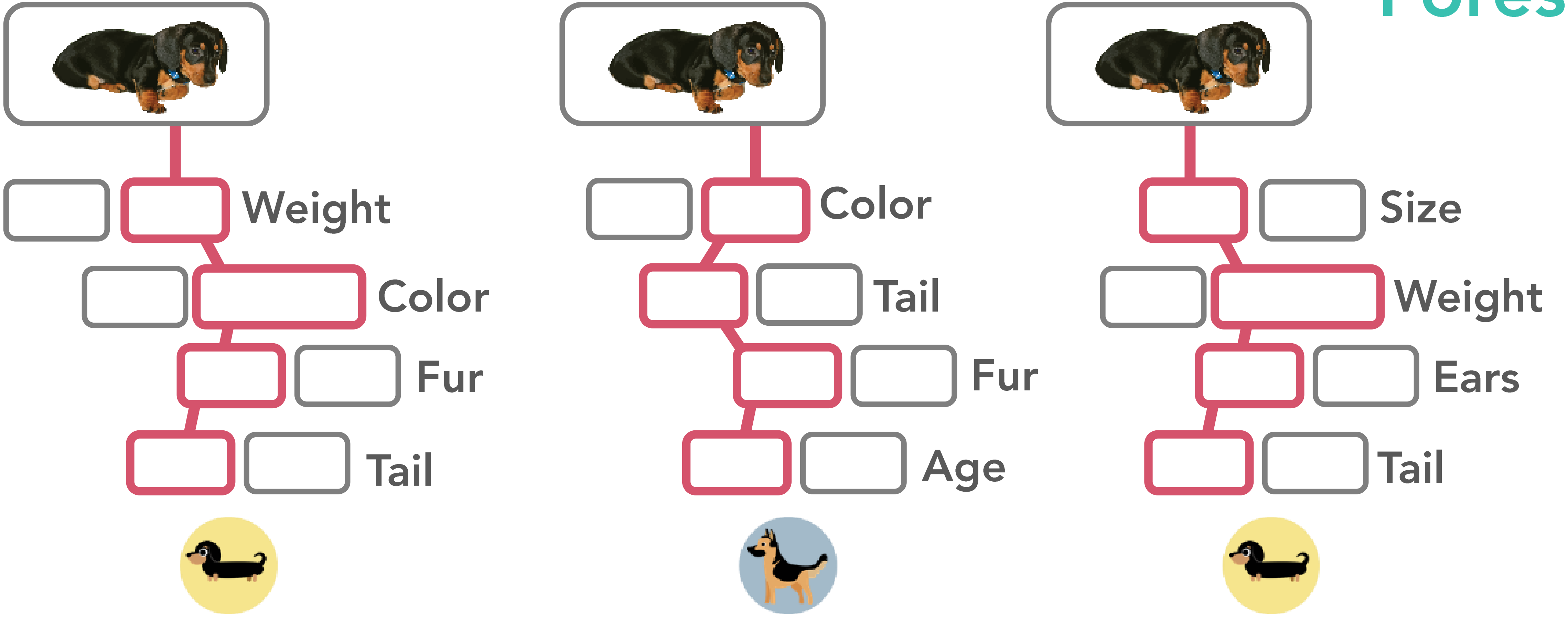
Size
Color
Fur
Ears
Tail
Age
Weight

explanatory variables

Decision Tree

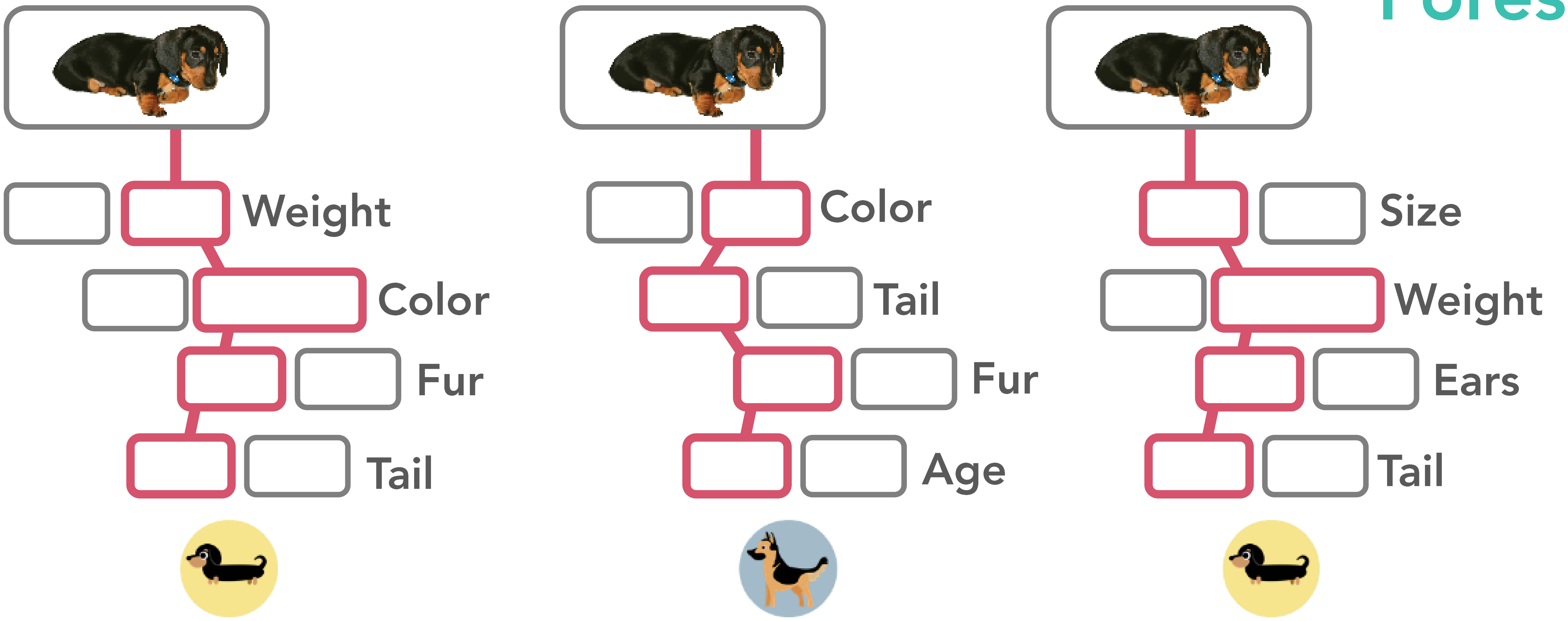


Forest



Random subset of data and variables used in each tree

Forest



Majority vote wins

Classification

Predict categorical variable

- Presence of disease
- Crime type
- Causes of forest fires
- Species distribution
- Dog breed

Regression

Predict continuous variable

Healthcare spending

Crime rate

Mortality rate

Rate of disease

Sales profits

Explanatory Variables

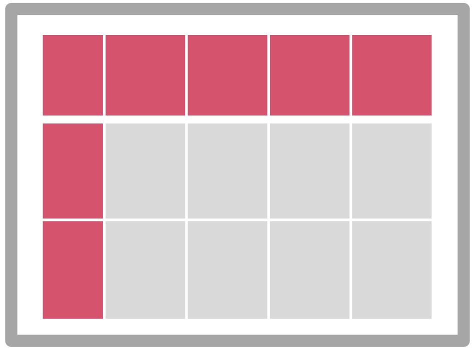
Attributes

Distance features

Rasters

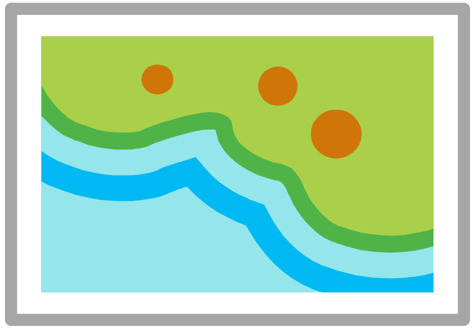
Explanatory Training Variables

Other attributes in the layer containing the Variable to Predict



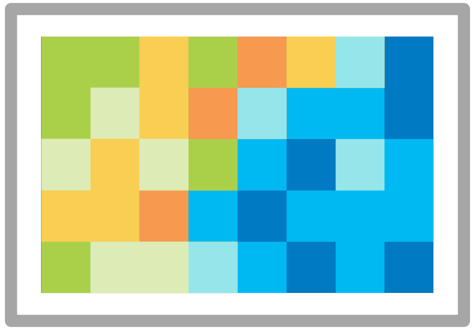
Explanatory Training Distance Features

Features from which distances will be calculated



Explanatory Training Rasters

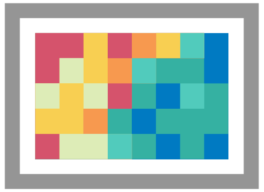
Rasters from which values will be extracted



Prediction Type

- Train only
- Predict to features

- Predict to rasters



Train only

Assess model performance

How accurate is the model?

Which variables were most important for prediction?

Predict to features

Create a prediction feature class

Predict missing values in study area

Predict values in a different study area

Predict values in a different time period

Predict to raster

Create a prediction surface

All explanatory variables must be rasters

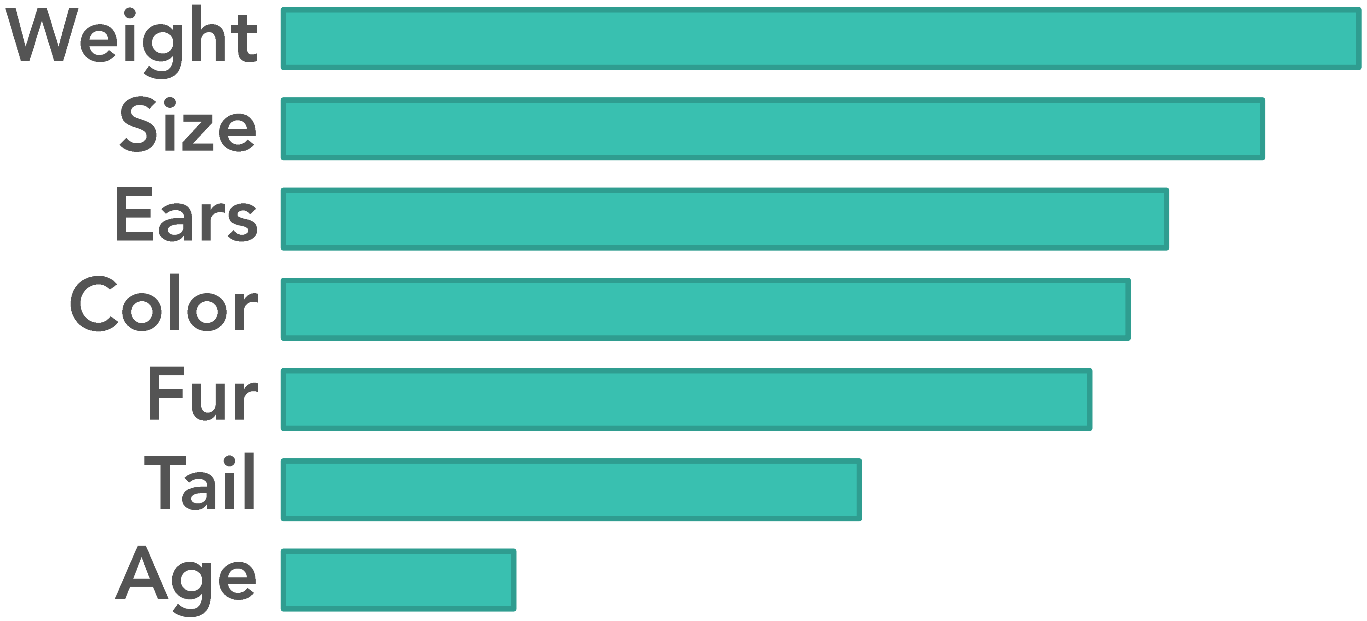
Predict values in a different study area

Predict values in a different time period

Evaluate model performance

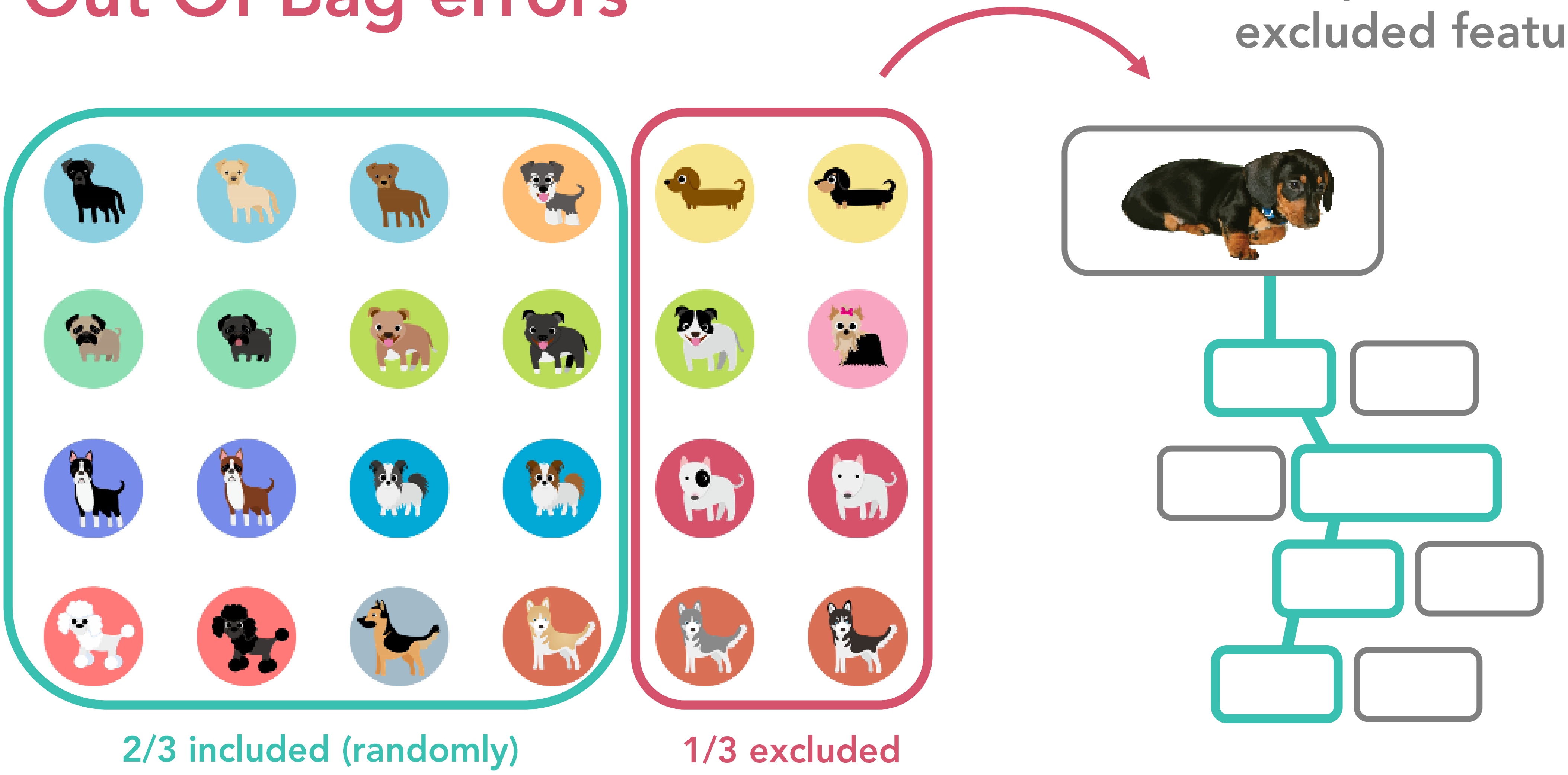
Variable importance

How well does each variable do in splitting the trees?



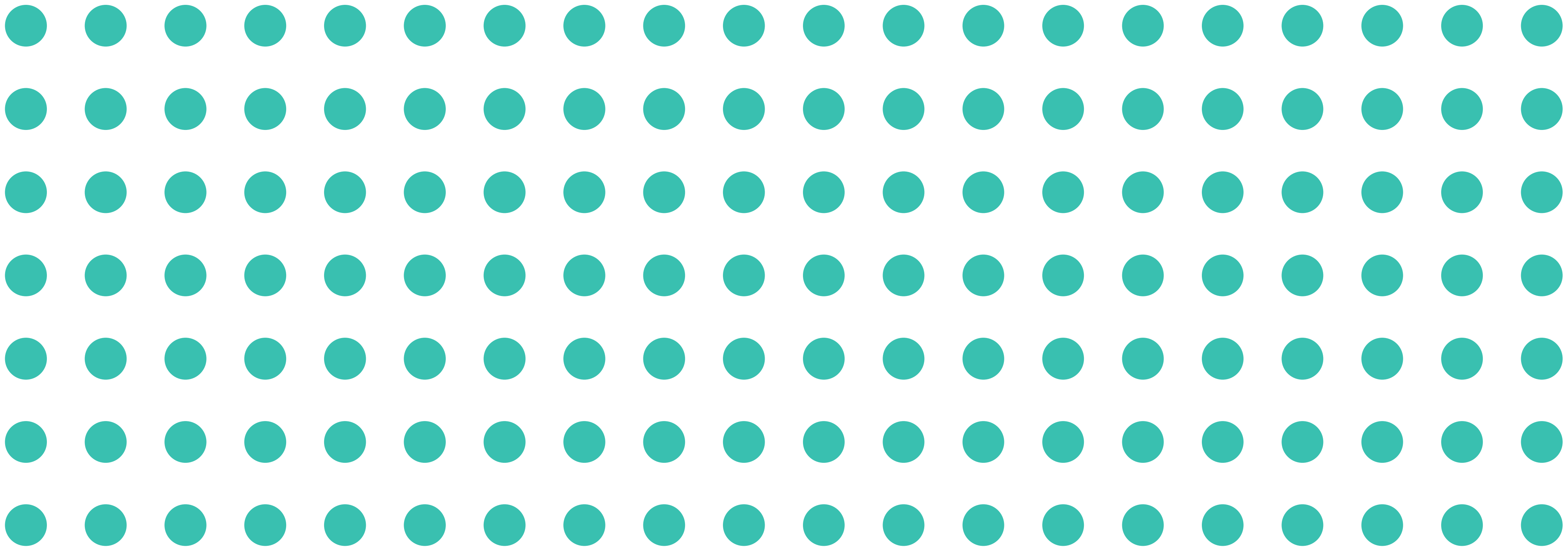
Out Of Bag errors

How well can each tree predict the excluded features?

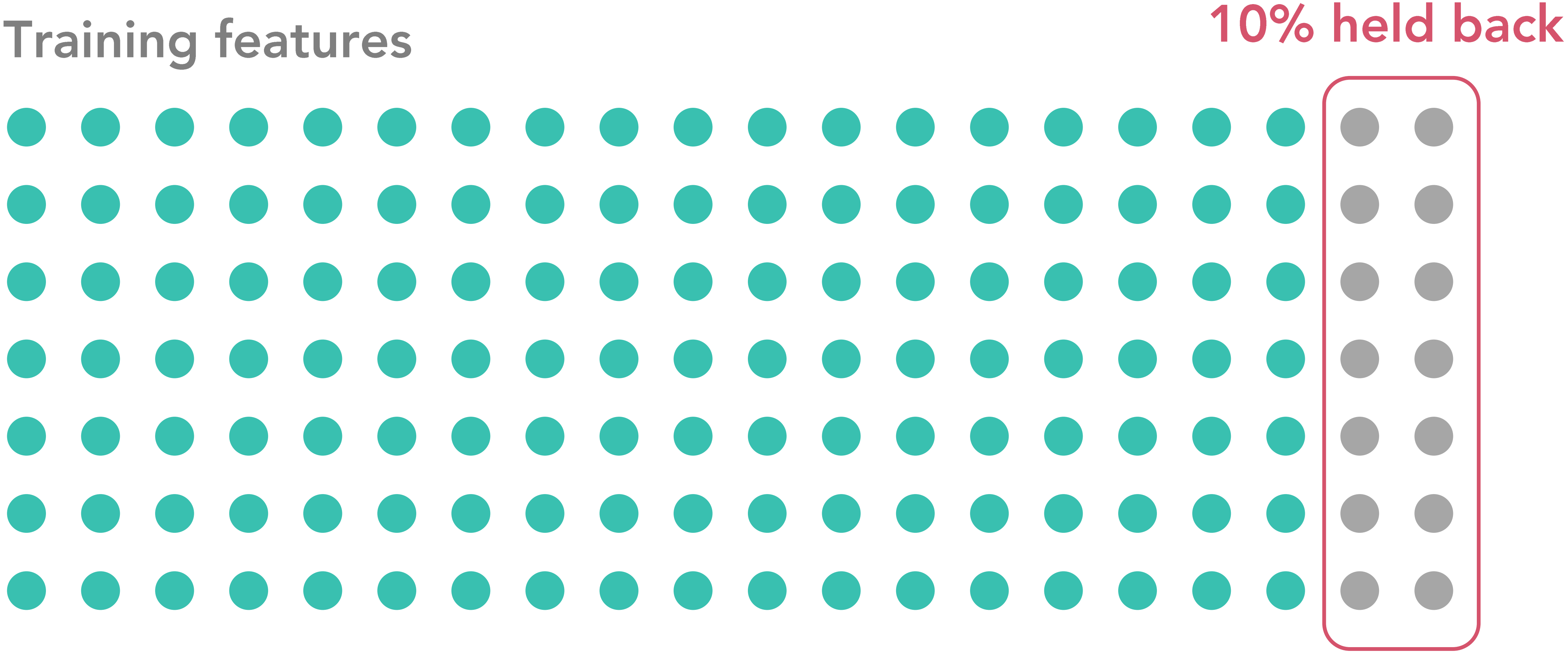


Model Validation

Training features

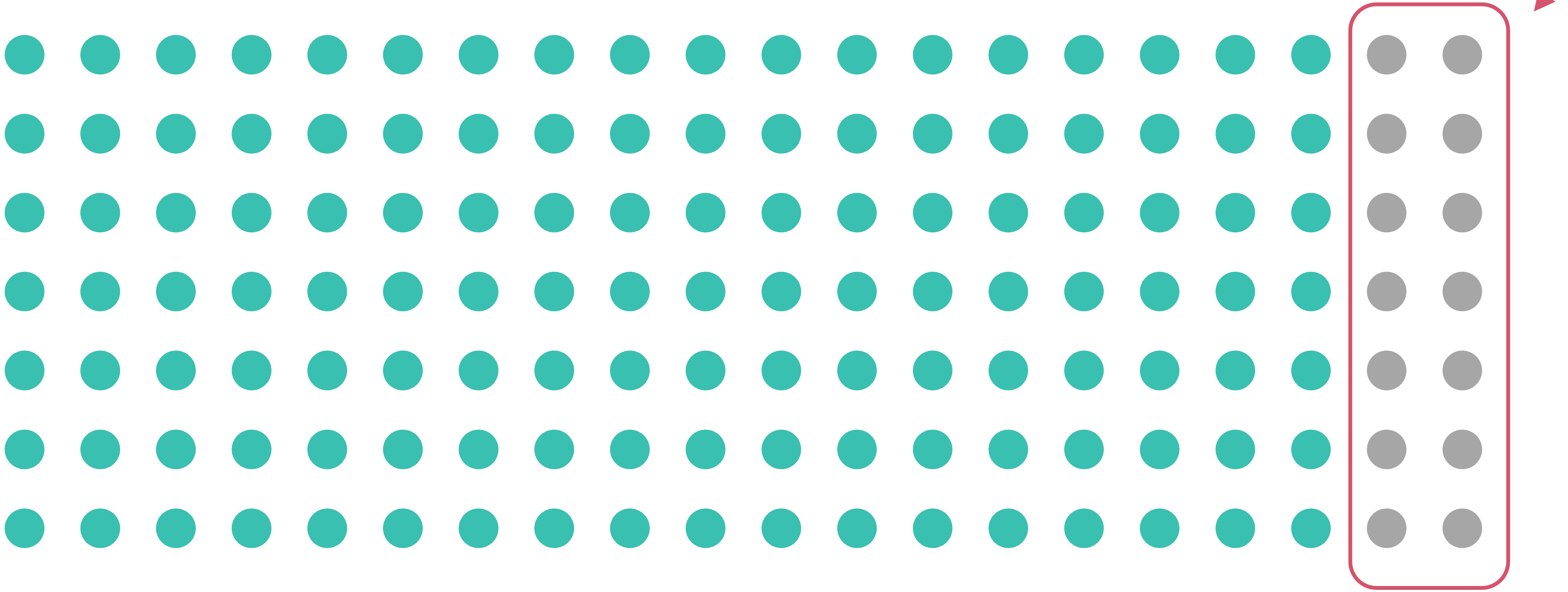


Model Validation



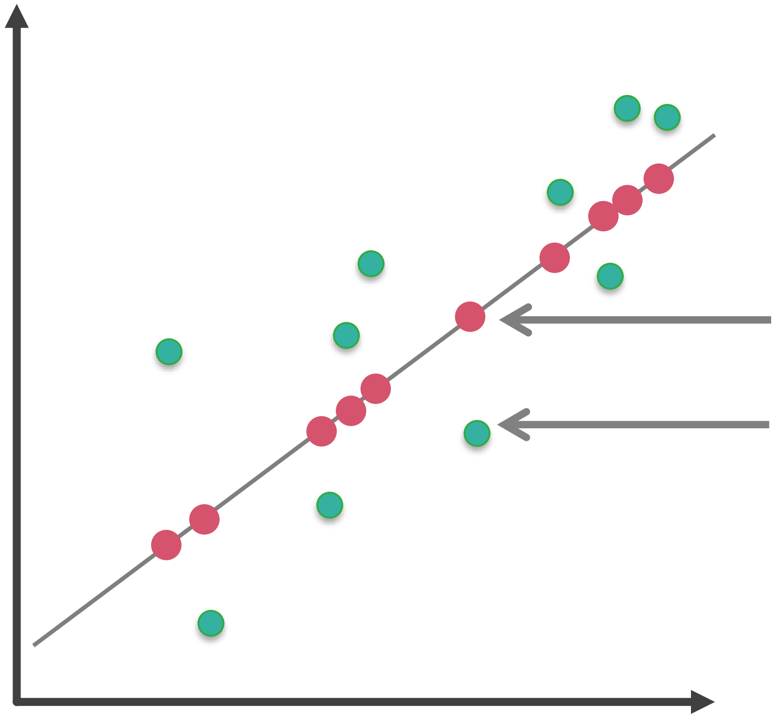
Model Validation

How well can the forest predict the features not used in training?



R-squared

How well can the forest predict (regression) the features not used in training?

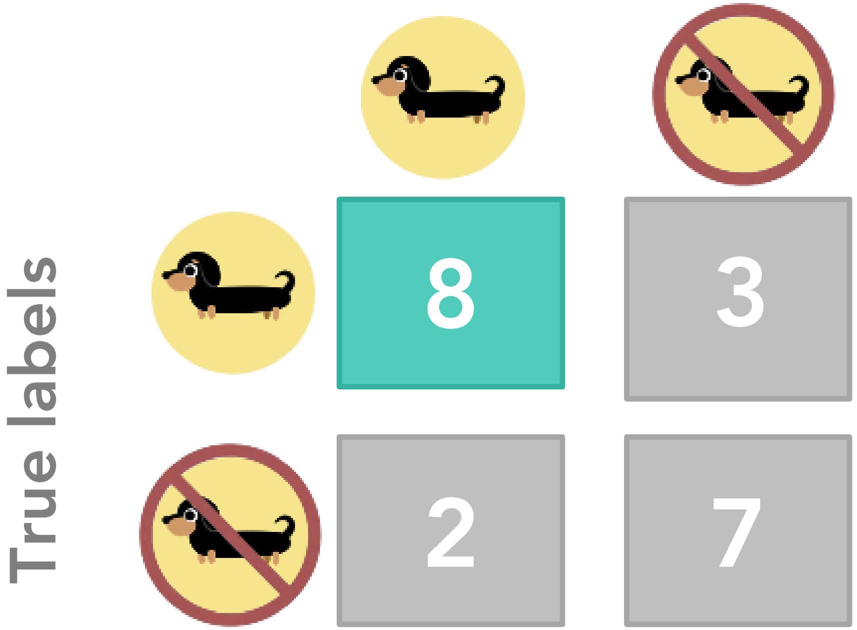


predicted value

observed value

Confusion matrix

How well can the forest predict (classification) the features not used in training?



True labels \ Predicted	🐕	🚫🐕
🐕	8	3
🚫🐕	2	7

Sensitivity for 🐕 80%

8/(8+2)

Confusion matrix

How well can the forest predict (classification) the features not used in training?

Predicted

True labels

	🐕	🚫🐕
🐕	8	3
🚫🐕	2	7

Accuracy for 🐕 75%

15/20

Modeling workflow

Step 0. Prepare your data

Step 1. Train a model

Step 2. Evaluate model performance

Step 3. Train again with different parameters

Step 4. Compare models

Step 5. Repeat...∞

Step 6. Use best model to predict unknown values

Demo

"Essentially, all models are wrong, but some are useful."

- George E. P. Box

esri | THE SCIENCE OF WHERE

esriurl.com/spatialstats