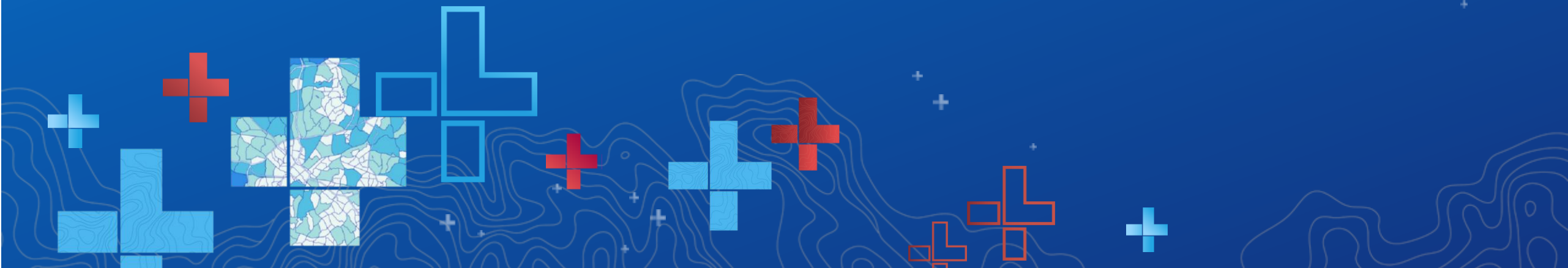
The Forest for the Trees: Making Predictions using Forest-Based Classification and Regression

Lauren Bennett, Flora Vale, Alberto Nieto

2020 ESRI FEDERAL GIS CONFERENCE | WASHINGTON, D.C.

esriurl.com/spatialstats

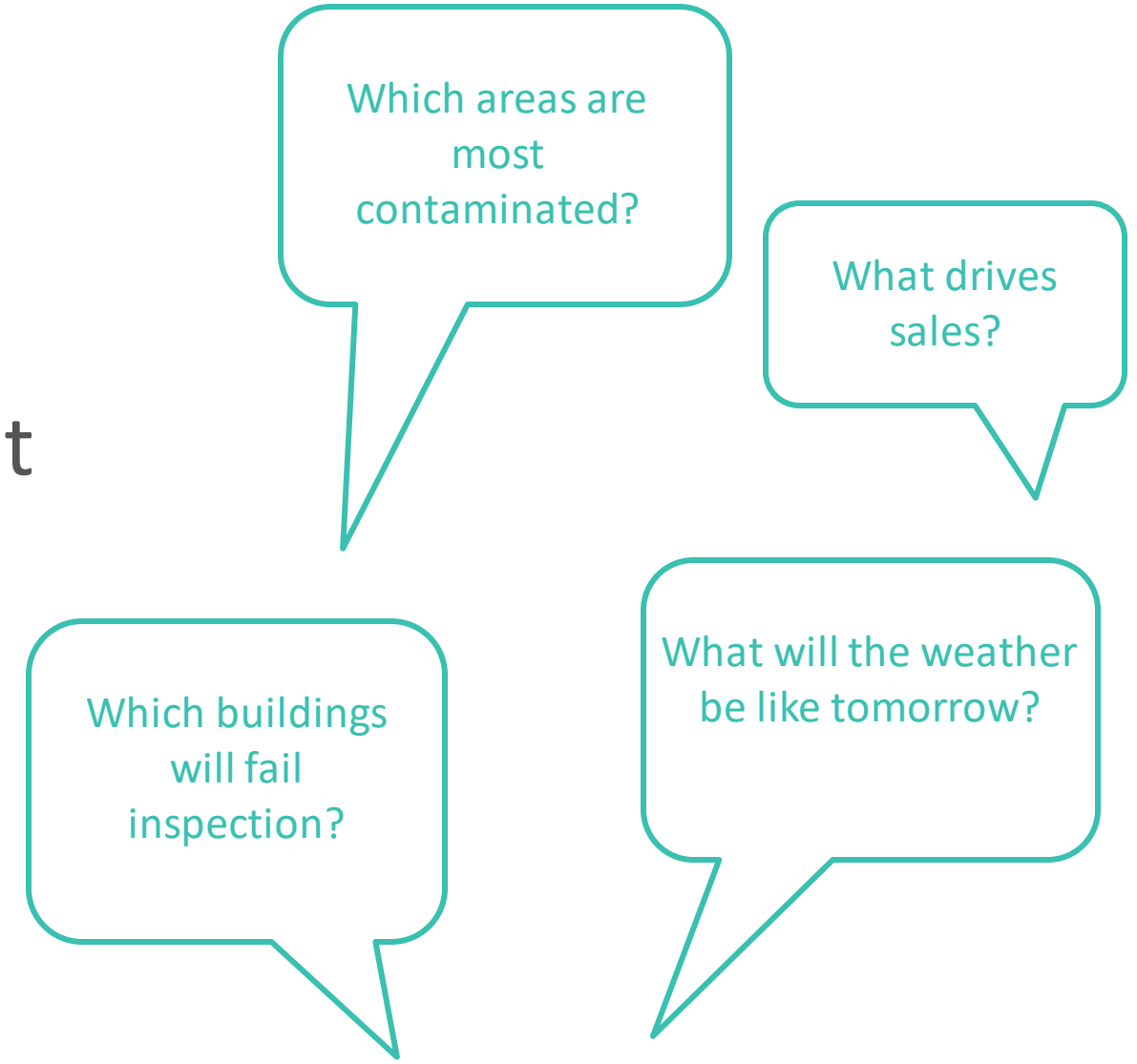


Models

Representative generalizations
used for prediction

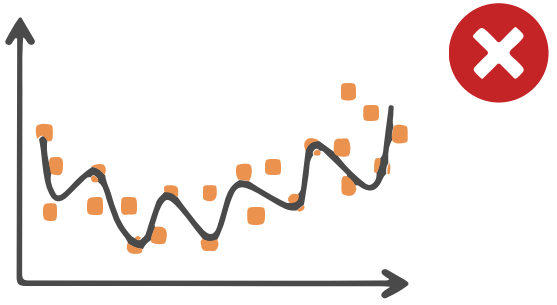
Why model

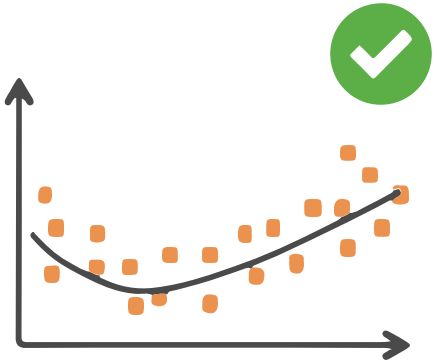
Use information we have to predict information we don't have



When we can’t trust a model

Mimics training dataset and models noise instead of generalizing a trend

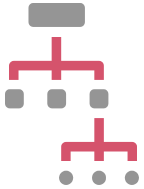




Many many many ways to model

Generalized Linear Regression
Geographically Weighted Regression

Forest-based Classification and Regression



Forest-based Classification & Regression

Predicting using machine learning

Training

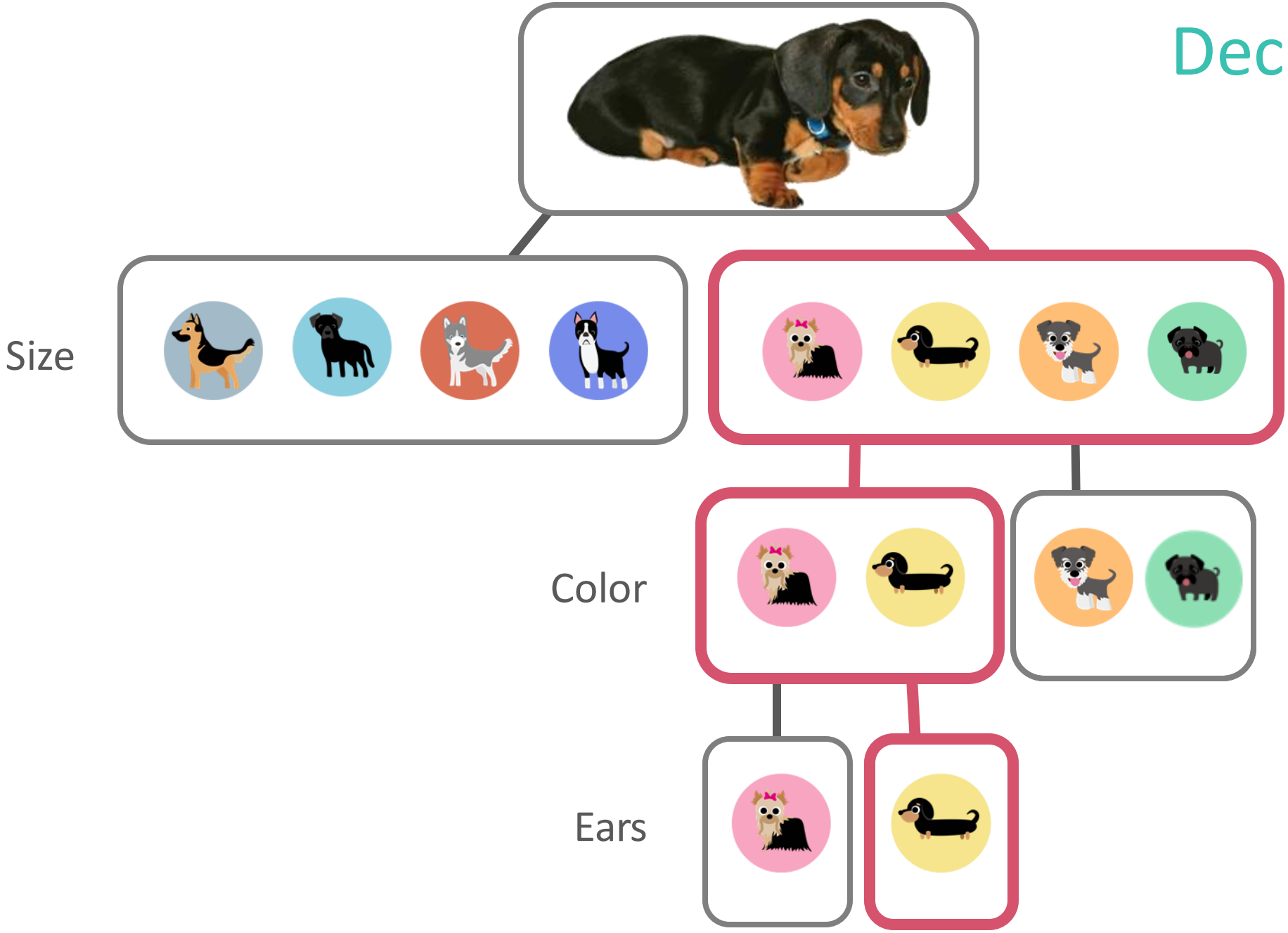
variable to predict

Breed

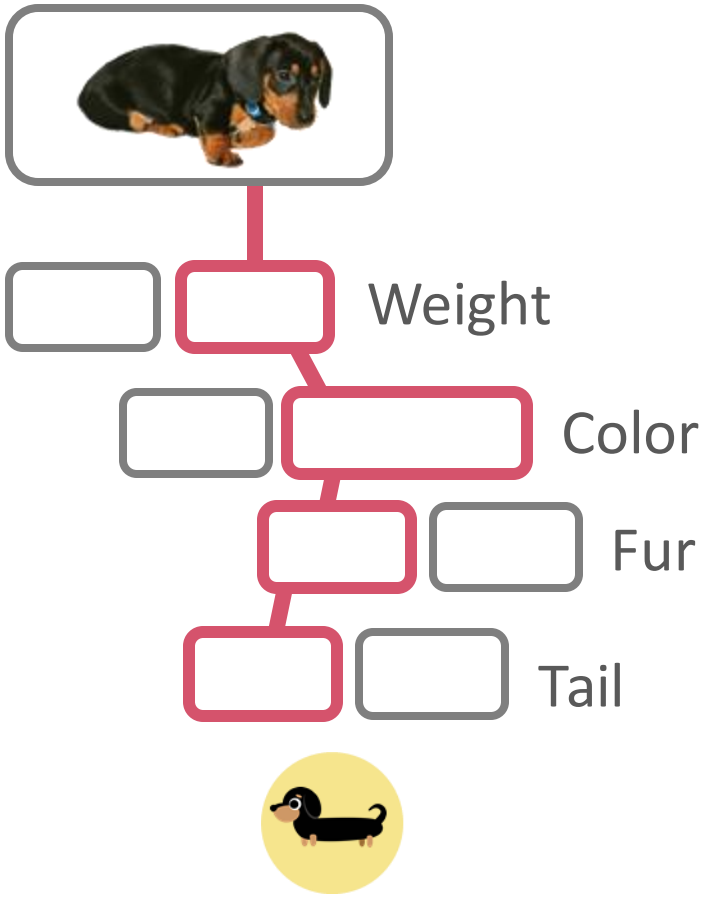
Size
Color
Fur
Ears
Tail
Age
Weight

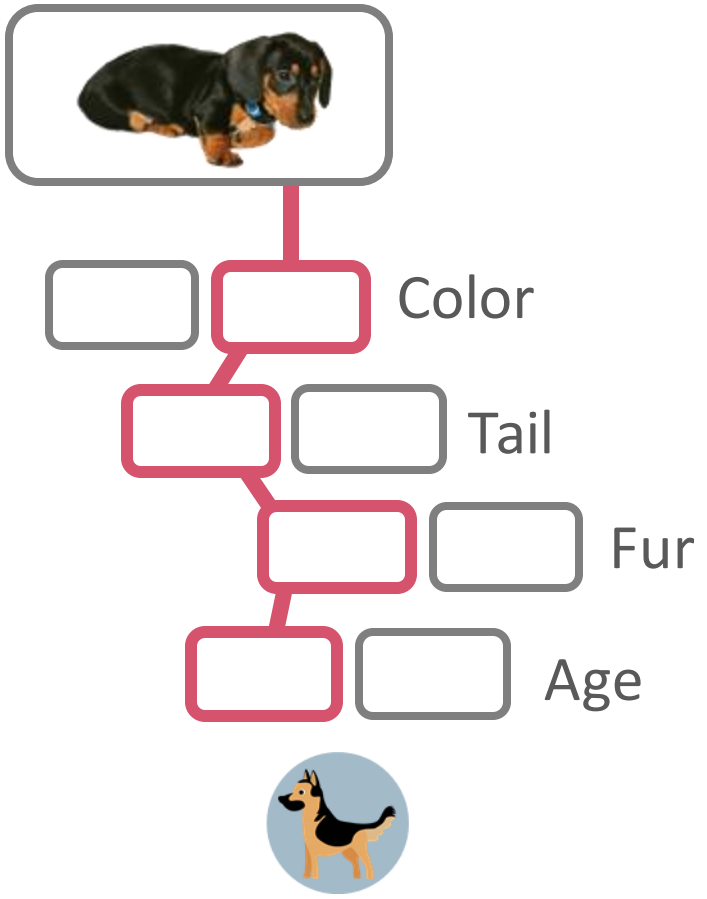
explanatory variables

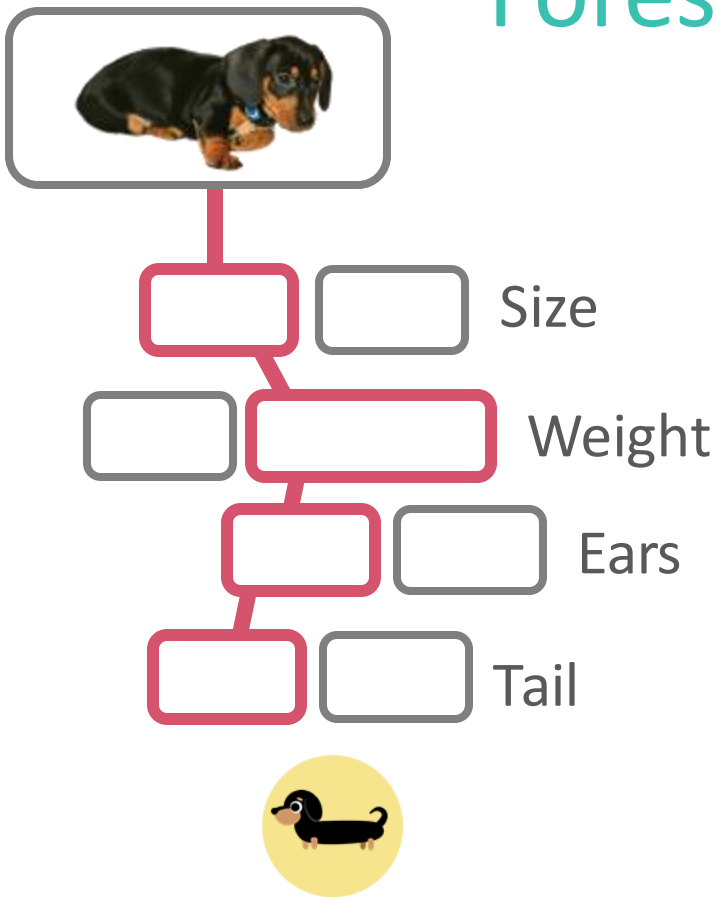
Decision Tree



Forest

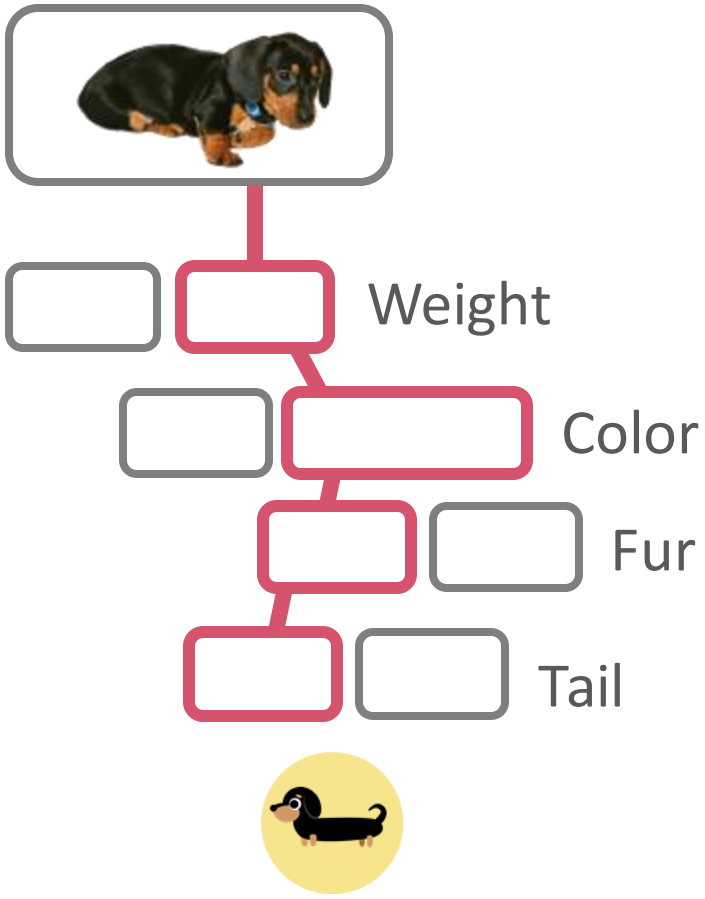


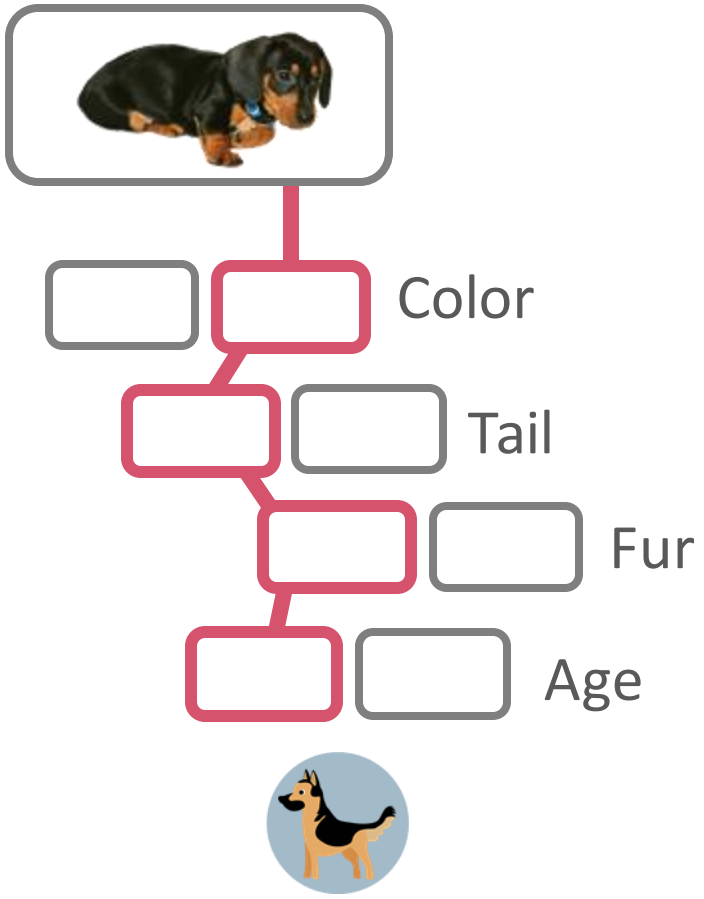


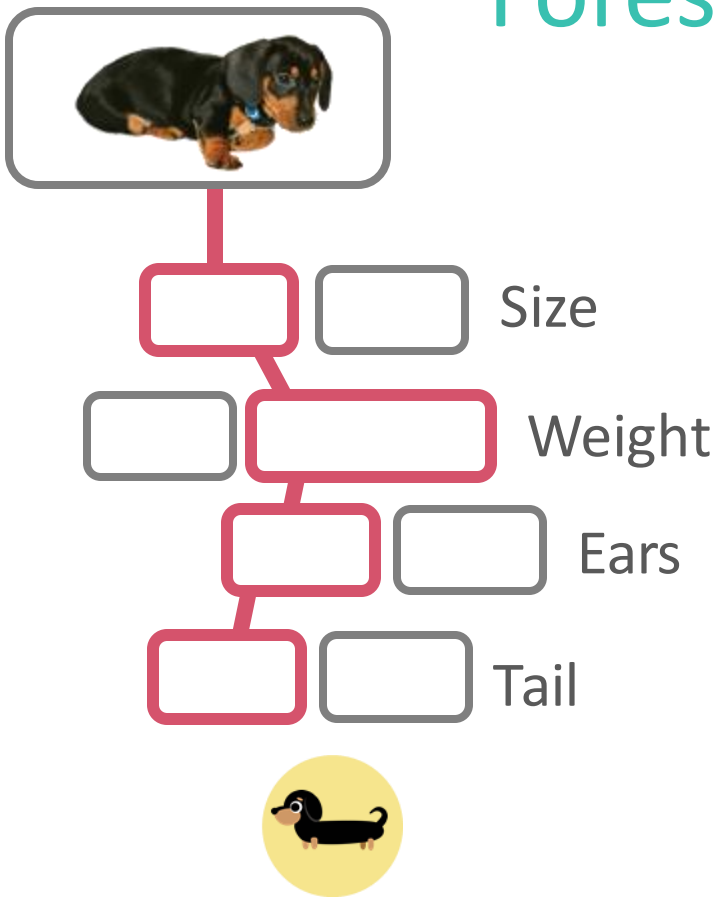


Random subset of data and variables used in each tree

Forest







Majority vote wins

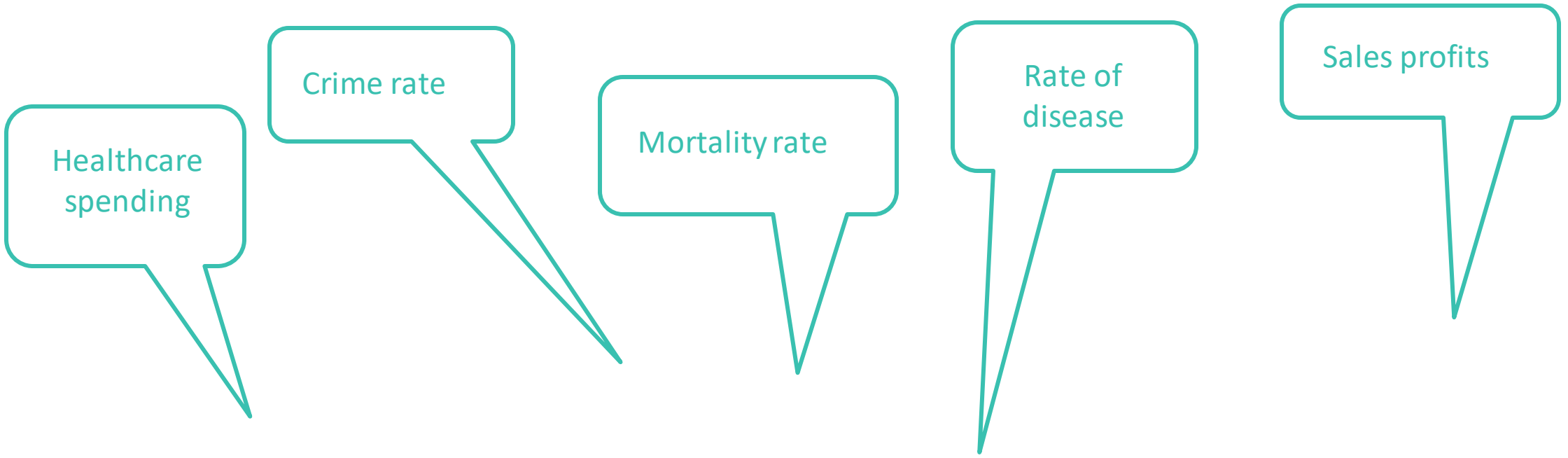
=

Classification

Predict categorical variable

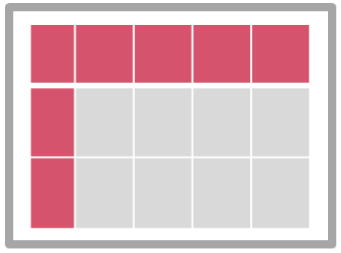
Regression

Predict continuous variable

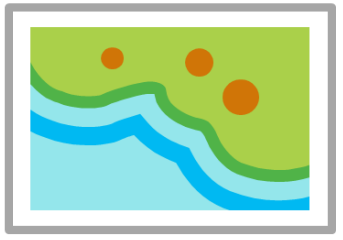


Explanatory Variables

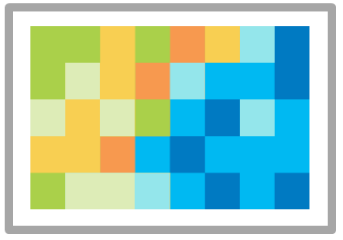
Attributes



Distance features

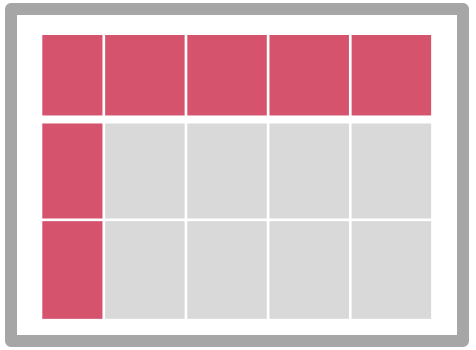


Rasters



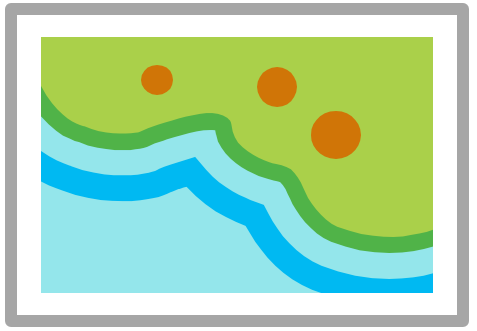
Explanatory Training Variables

Other attributes in the layer containing the Variable to Predict



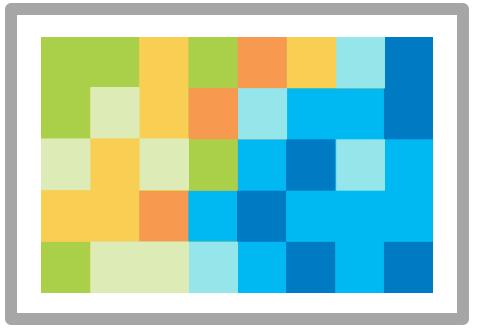
Explanatory Training Distance Features

Features from which distances will be calculated



Explanatory Training Rasters

Rasters from which values will be extracted

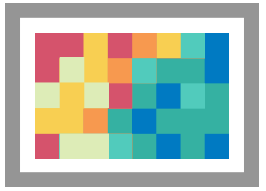


Prediction Type

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

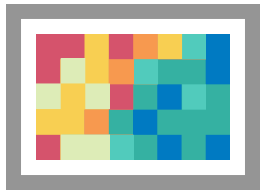
Create a prediction feature class

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

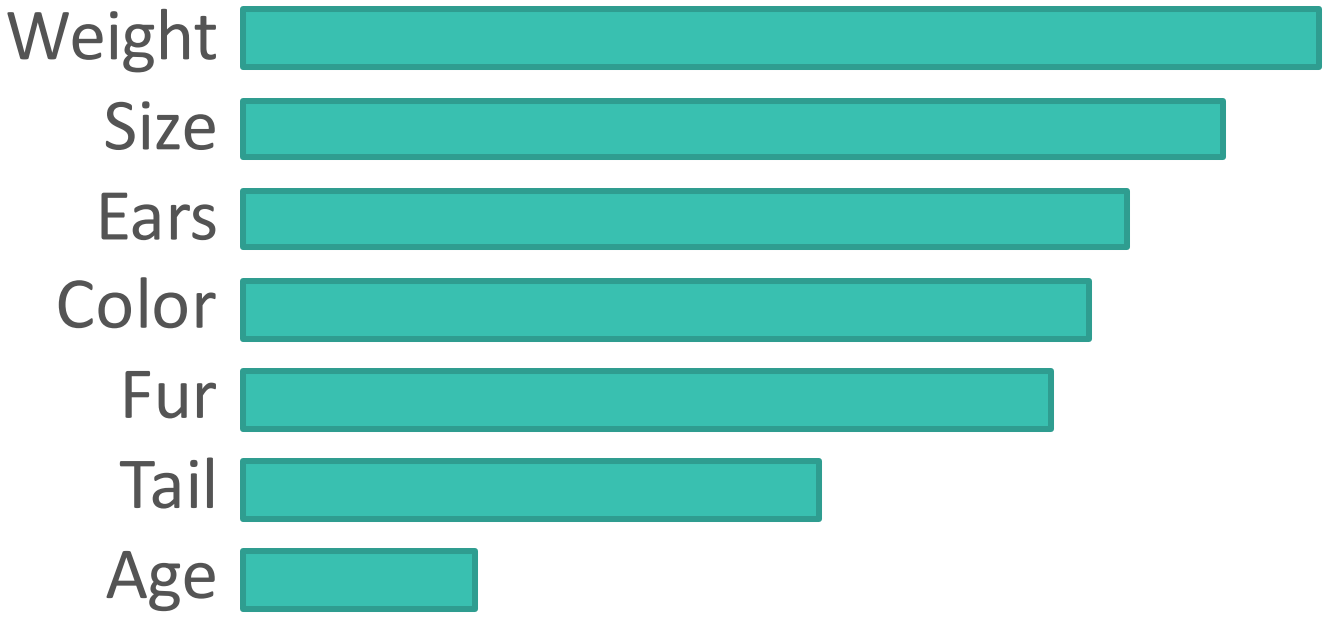
Predict values in a different study area

Predict values in a different time period

Evaluate model performance

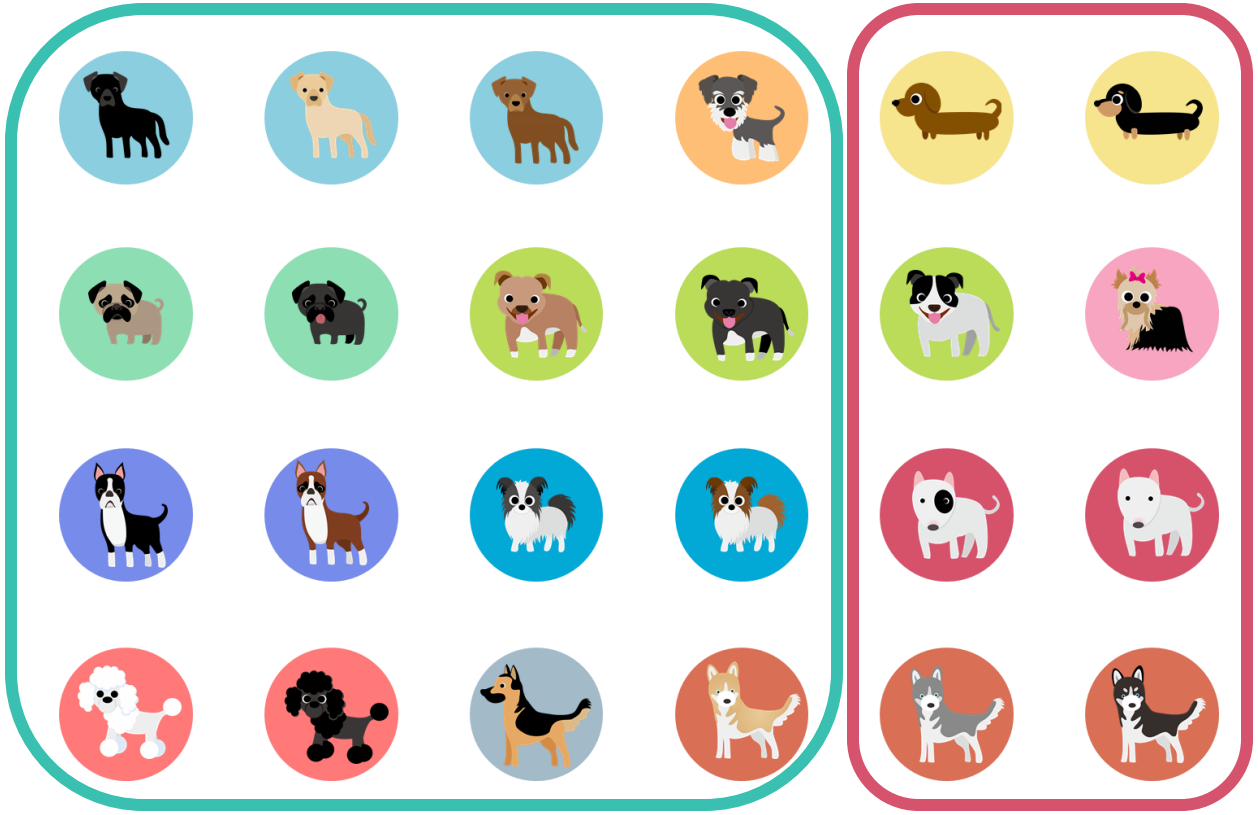
Variable importance

How well does each variable do in splitting the trees?



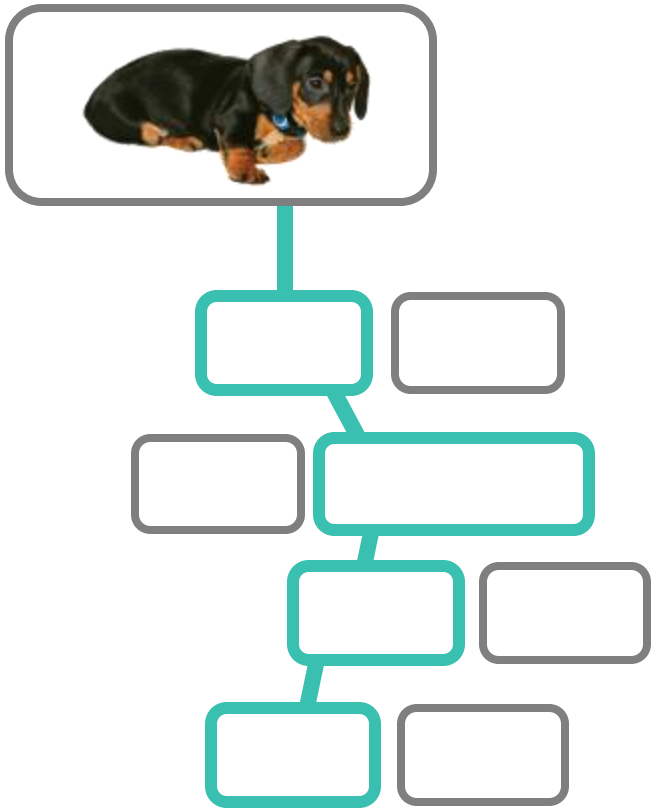
Out Of Bag errors

How well can each tree predict the excluded features?



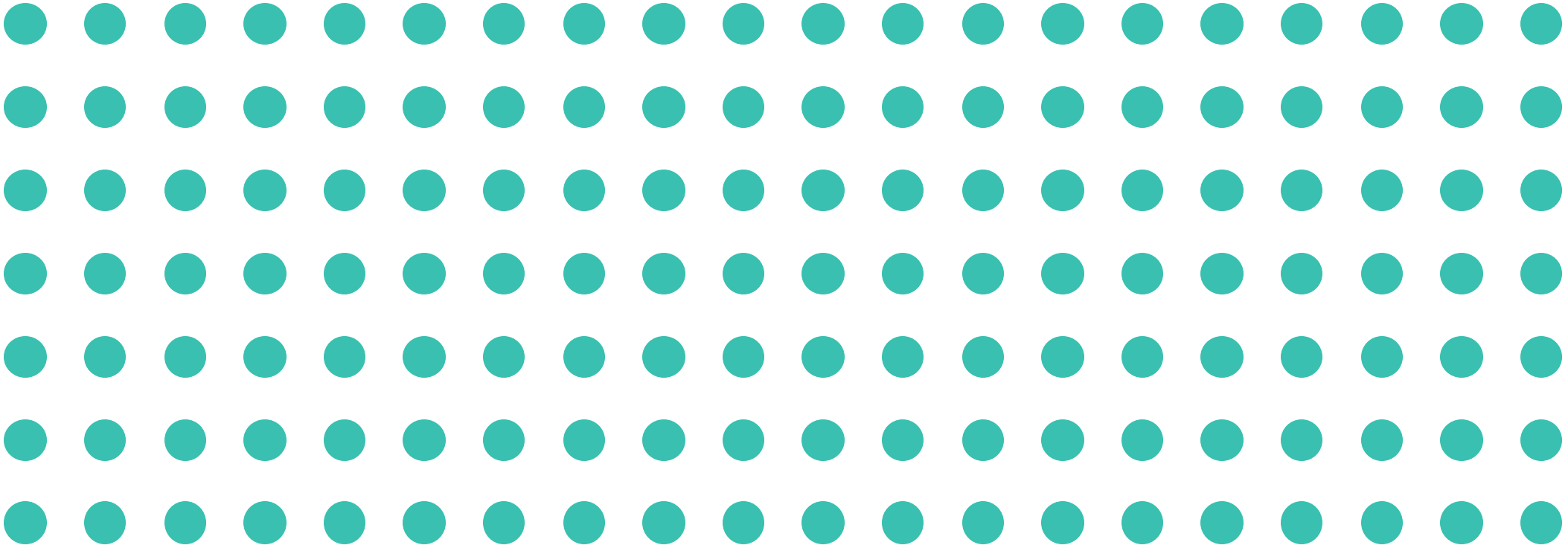
2/3 included (randomly)

1/3 excluded



Model Validation

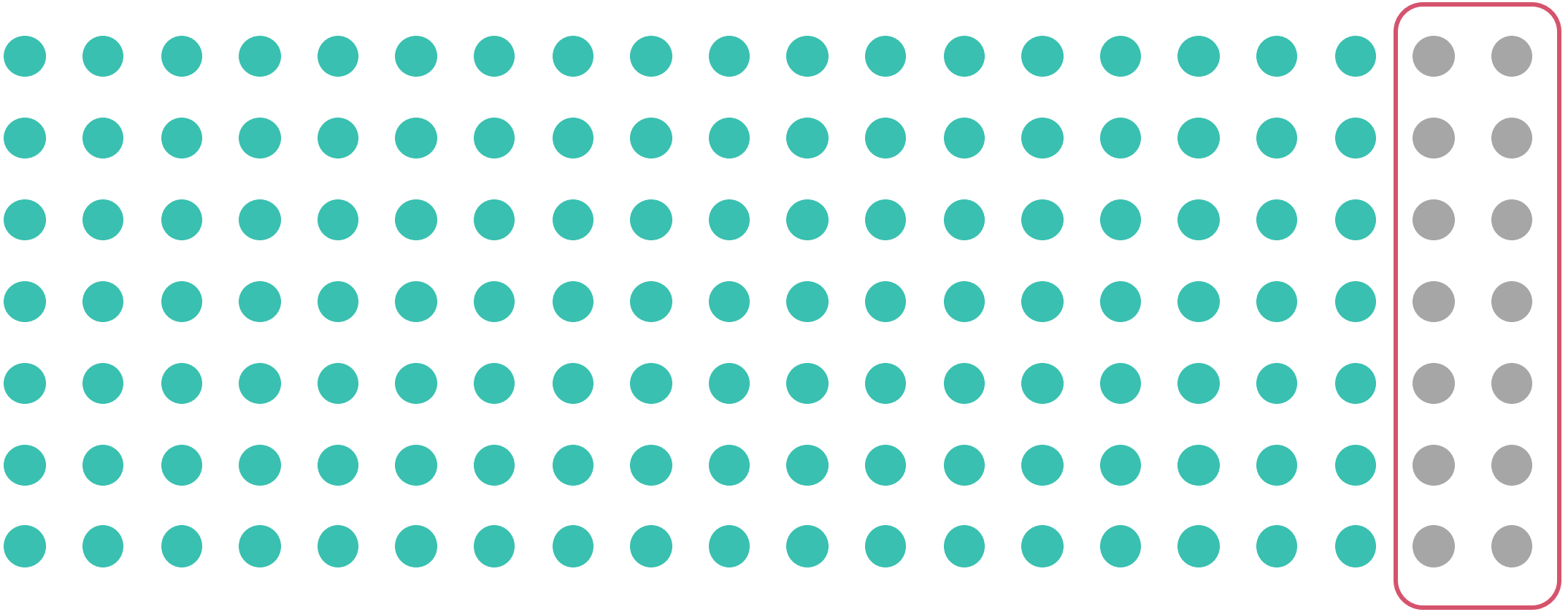
Training features



Model Validation

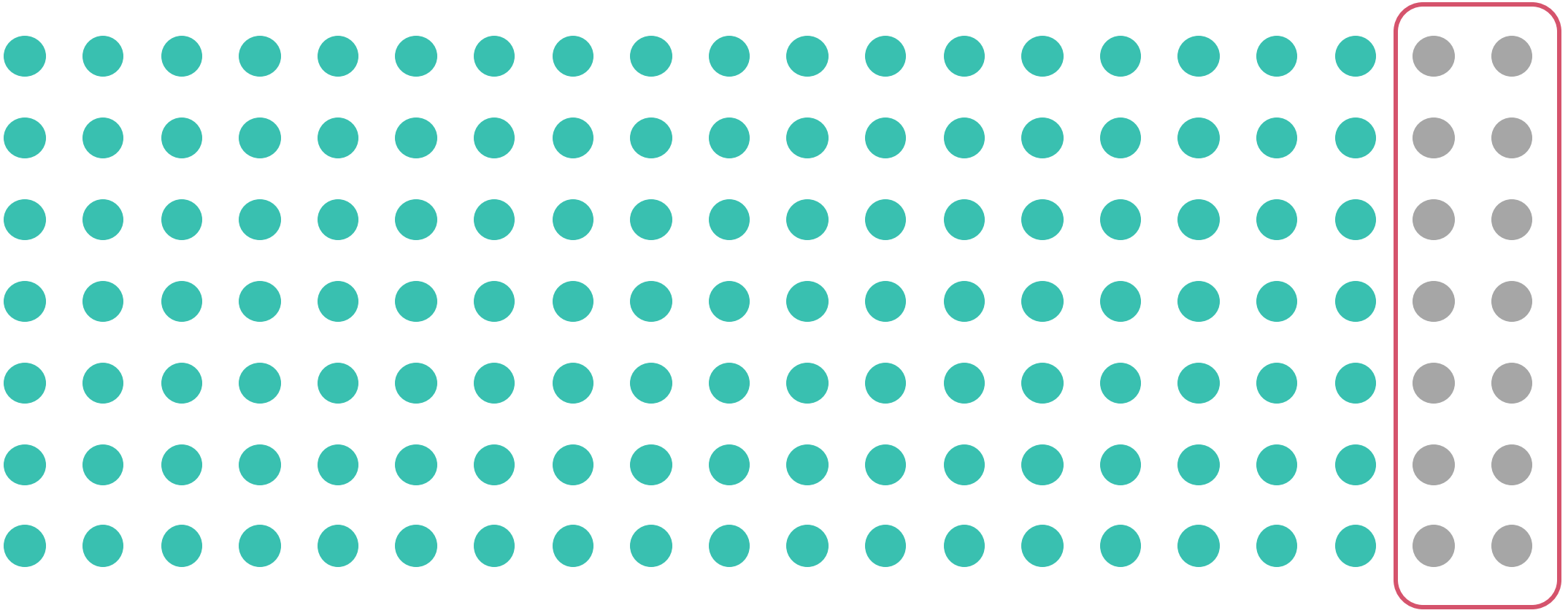
Training features

10% held back

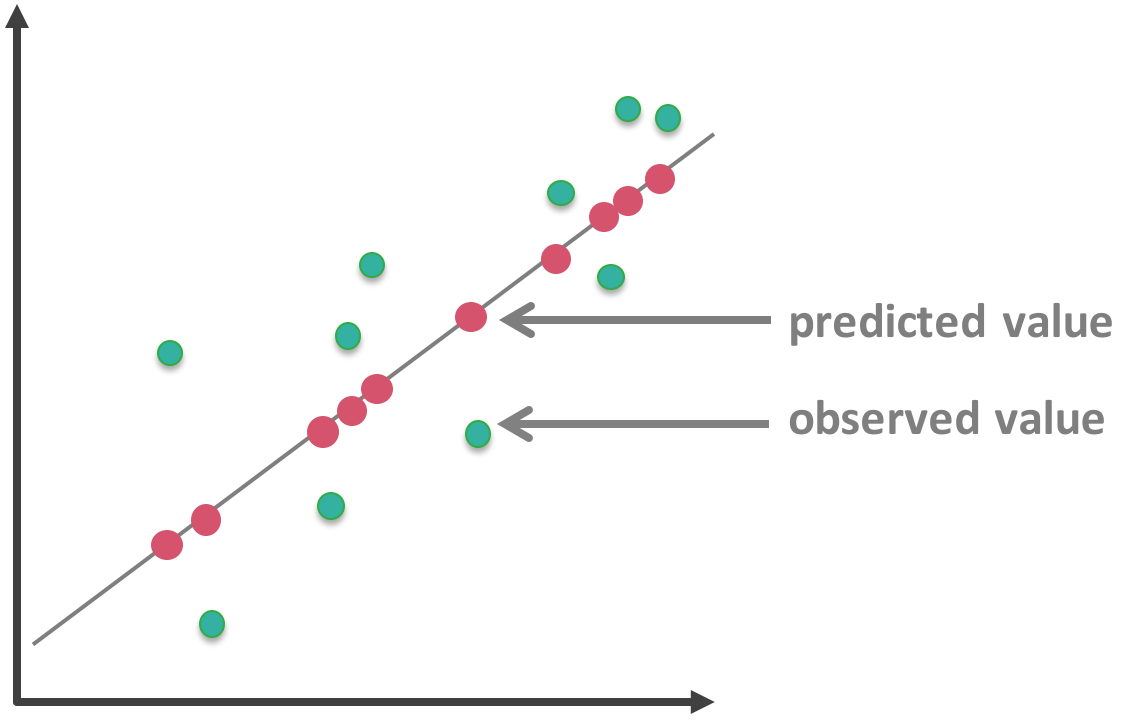


Model Validation

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



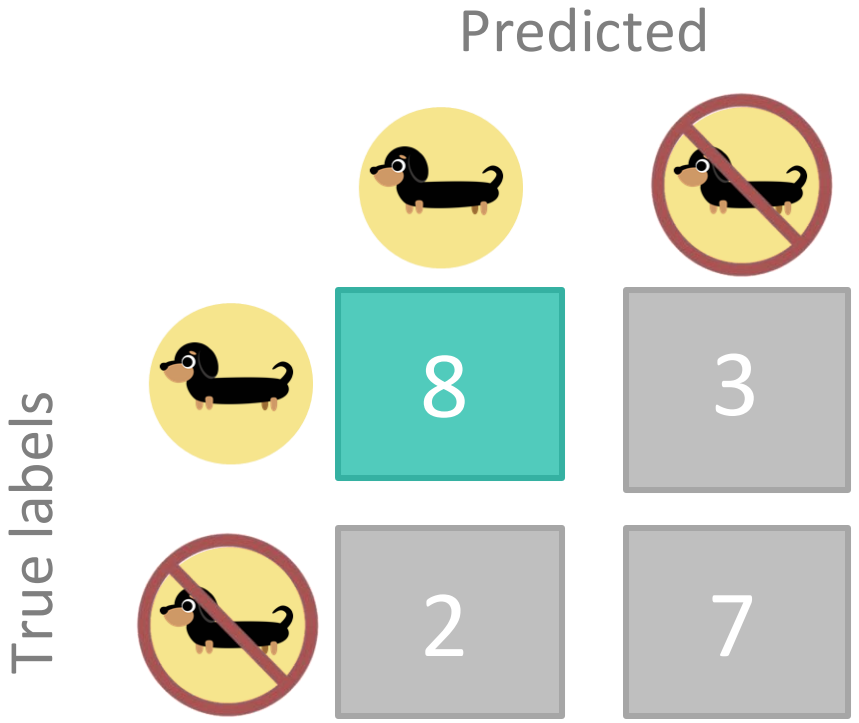
R-squared



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

Confusion matrix

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

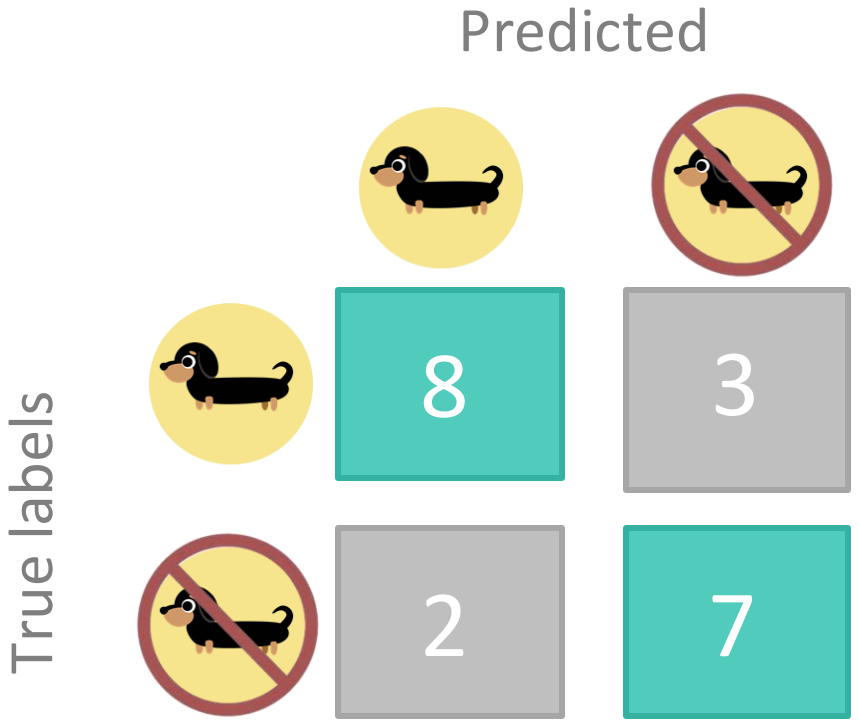


Sensitivity for  80%

8/(8+2)

Confusion matrix

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



Accuracy for  75%

15/20

Modeling workflow

Step 0. Prepare your data

Step 1. Train a model

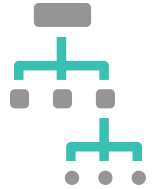
Step 2. Evaluate model performance

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."

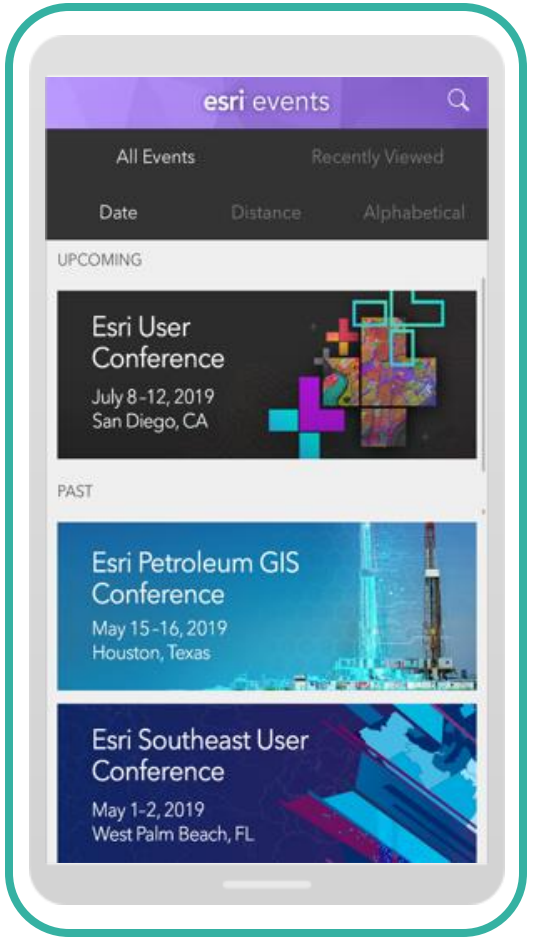
- George E. P. Box

Want to learn more???

esriurl.com/spatialstats

Please fill out a course survey!!!

lbennett@esri.com
anieto@esri.com
fvale@esri.com



TUESDAY

~~1:45p Data Visualization for Spatial Analysis 146C~~

~~3:00p Machine Learning in ArcGIS 146C~~

~~4:15p From Means and Medians to Machine Learning: Spatial Statistics Basics and Innovations 146C~~

WEDNESDAY

~~8:30a Machine Learning in ArcGIS 146C~~

~~11a Data Visualization for Spatial Analysis 146C~~

~~1:30p From Means and Medians to Machine Learning: Spatial Statistics Basics and Innovations 146C~~

~~2:45p Spatial Data Mining: Cluster Analysis and Space Time Analysis 146C~~

~~4:00p Beyond Where: Modeling Spatial Relationships and Making Predictions 146C~~

PARTY!!

~~5:15p The Forest for the Trees: Making Predictions Using Forest-Based Classification and Regression 146C~~