Beyond Where: Modeling Spatial Relationships and Making Predictions

Lauren Bennett, PhD

Jenora D'Acosta

Flora Vale

esriurl.com/spatialstats

Models

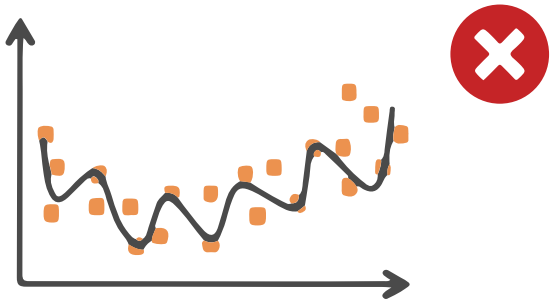
Representative generalizations used for prediction

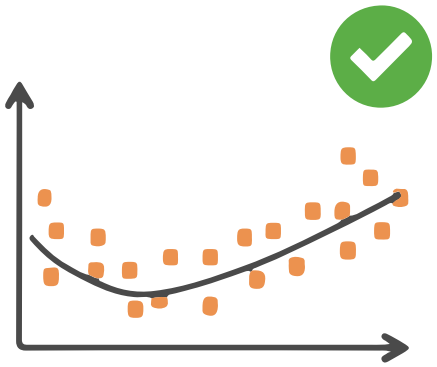
Why model

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

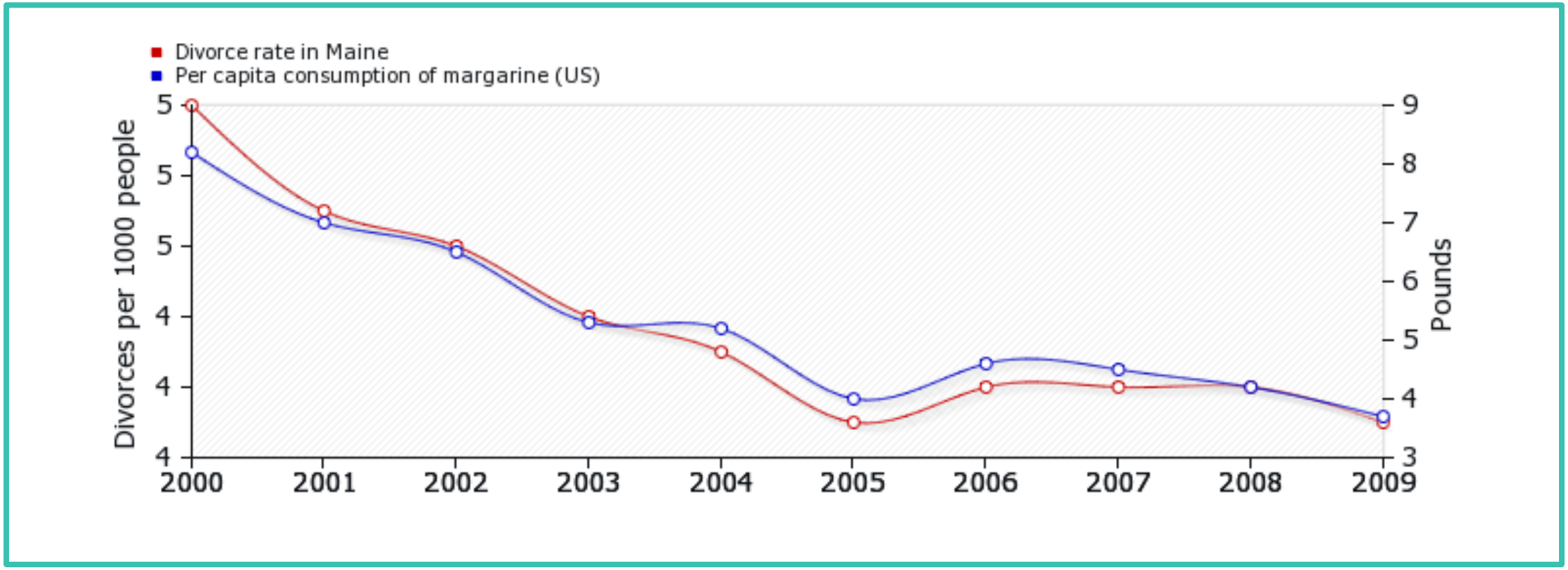
When we can't trust a model

Mimics training dataset and models noise instead of generalizing a trend





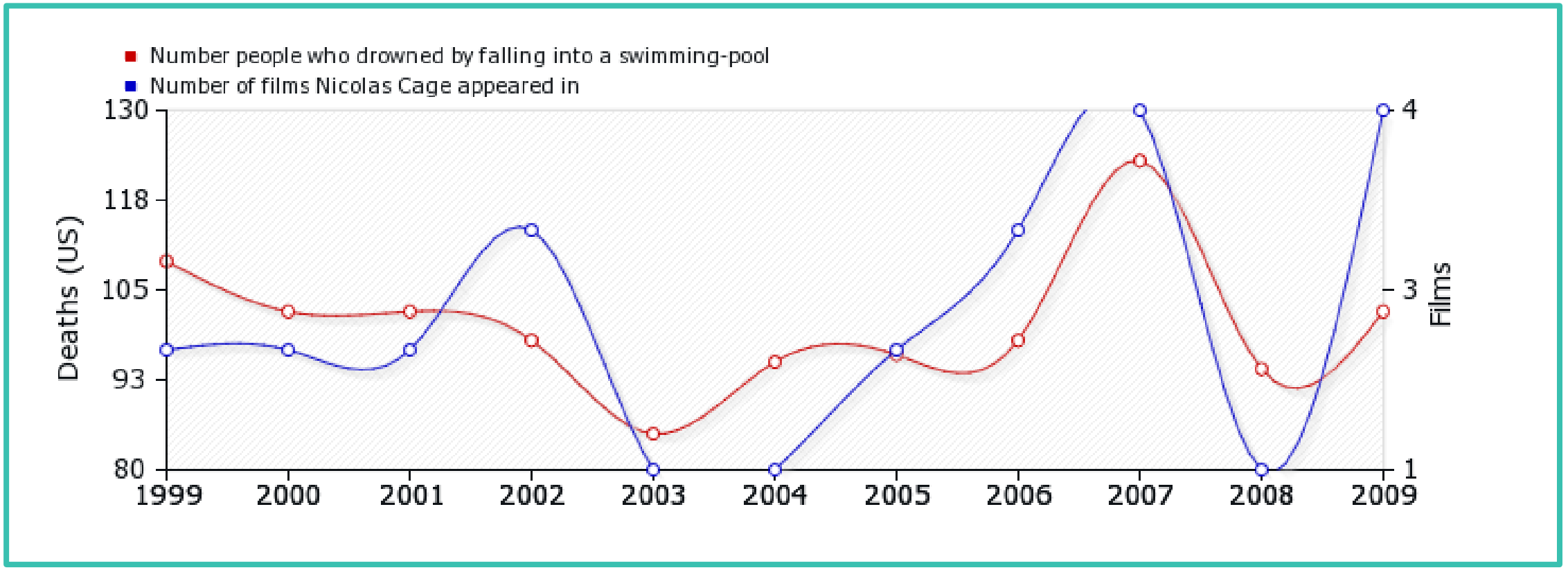
Divorce Rate in Maine vs Per Capita Consumption of Margarine



tylervigen.com

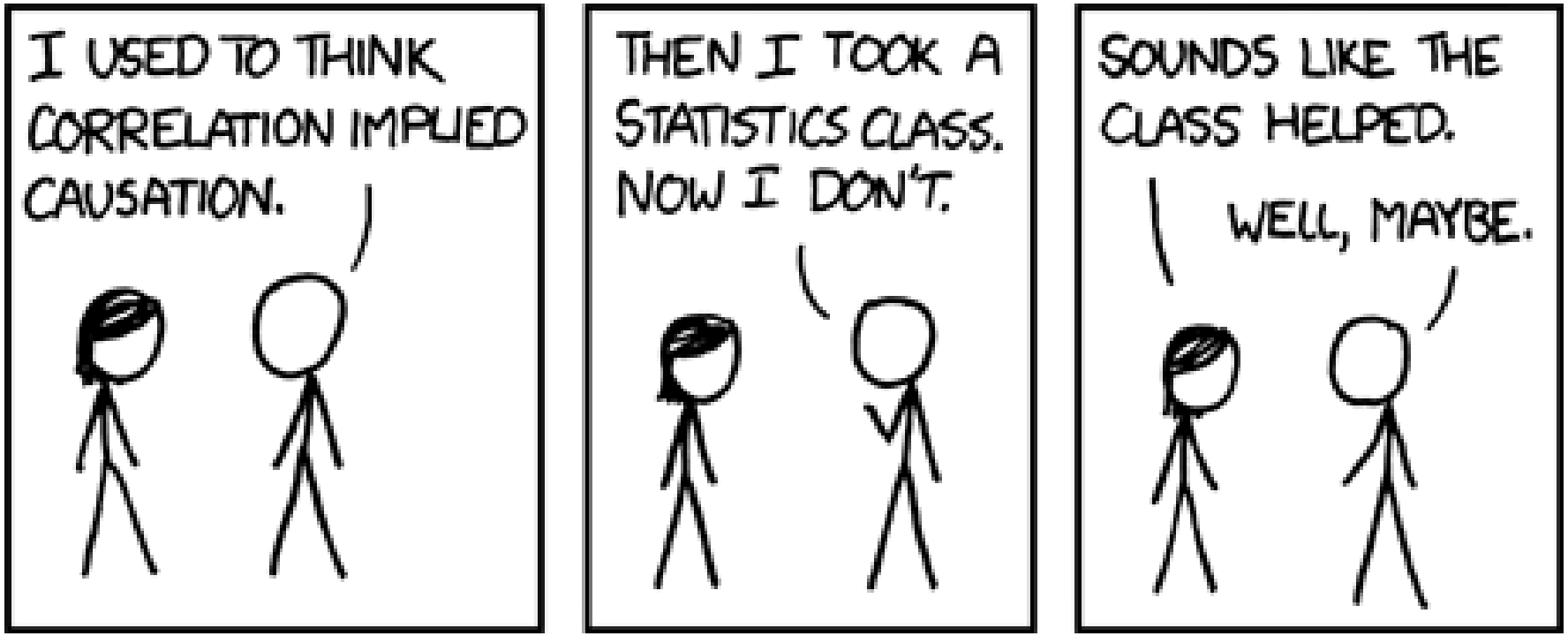
Correlation: 0.992558

Number People Who Drowned by Falling into a Swimming-Pool vs Number of Nicolas Cage Films



Correlation: 0.666004

tylervigen.com

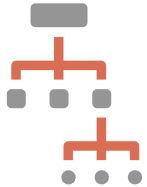


http://xkcd.com/552/

Many ways to model

Ordinary Least Squares
Geographically Weighted Regression

Forest-based Classification and Regression

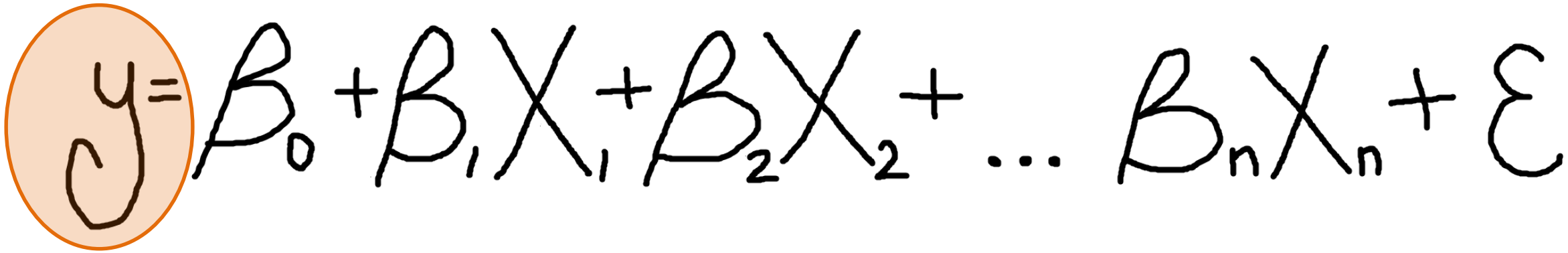


Ordinary Least Squares

Modeling linear relationships

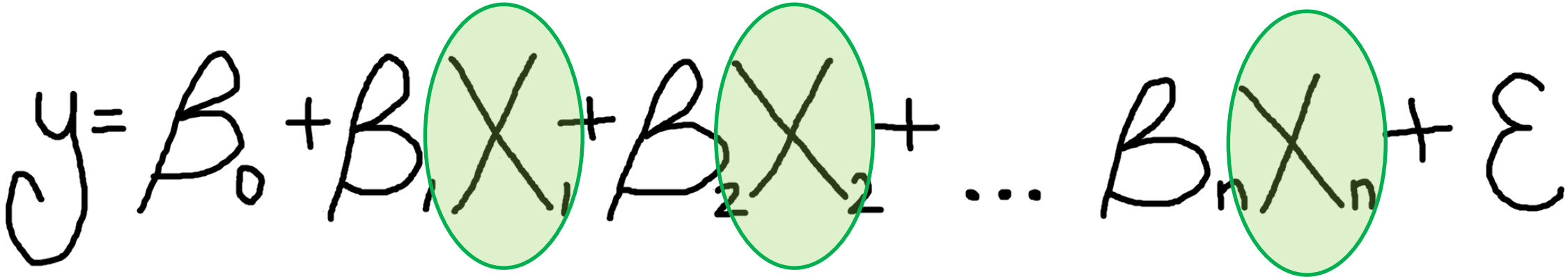
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$

Dependent Variable



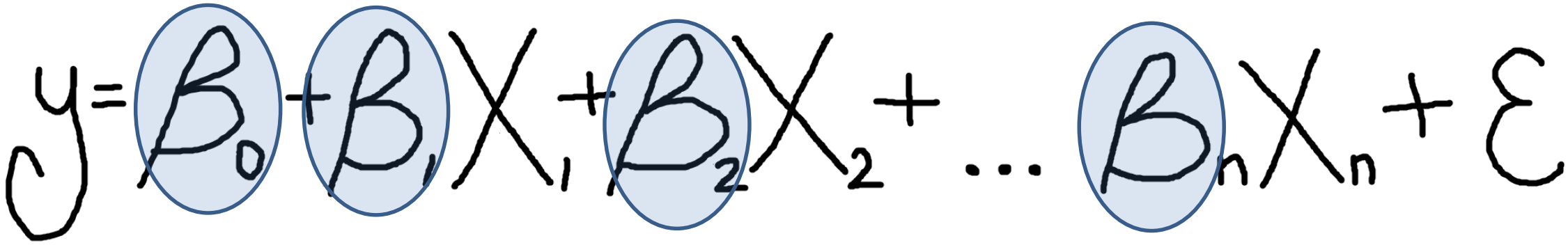
What are you trying to predict or understand?

Explanatory Variables



Variables you believe to cause or explain the dependent variable

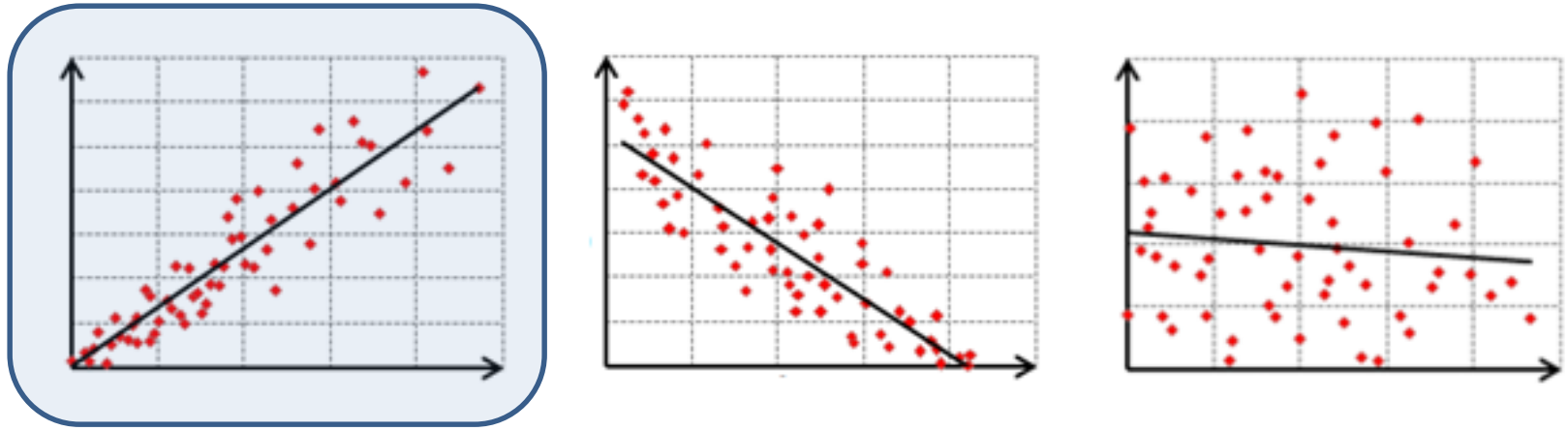
Coefficients



Represent the strength and type of relationship that X has to y

Coefficients

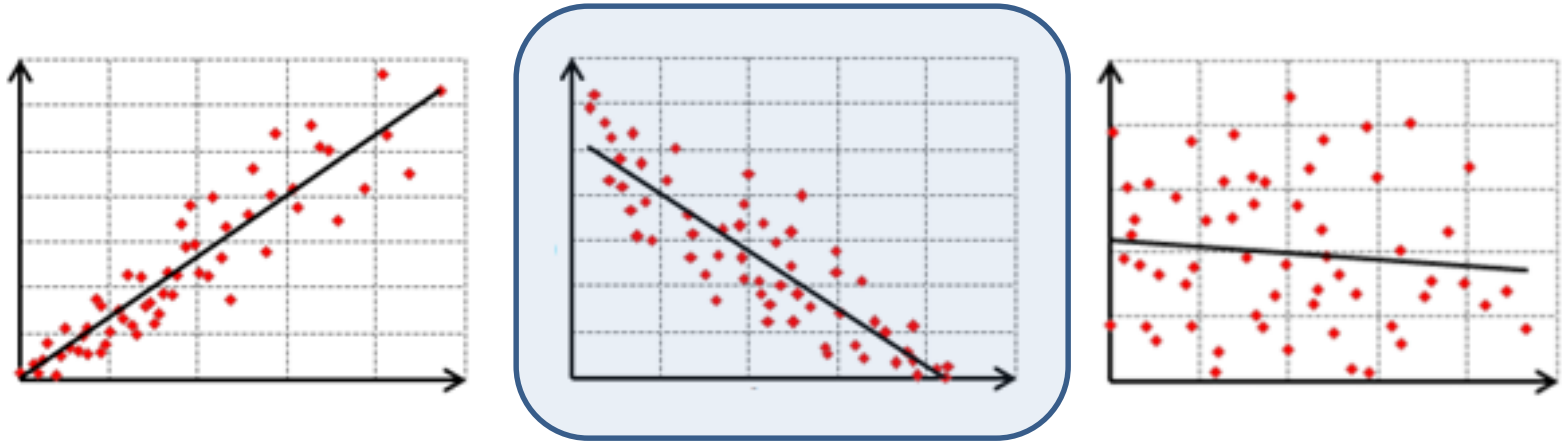
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$



Positive relationship- as obesity rates rise, diabetes rates also rise

Coefficients

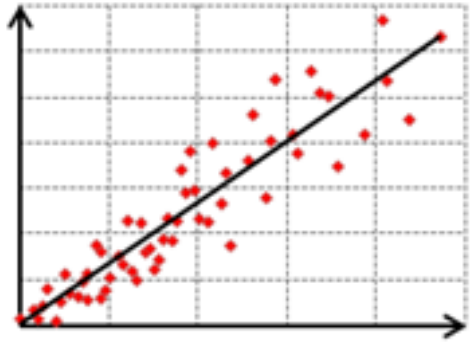
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$

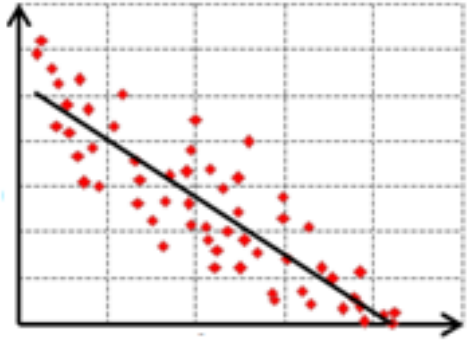


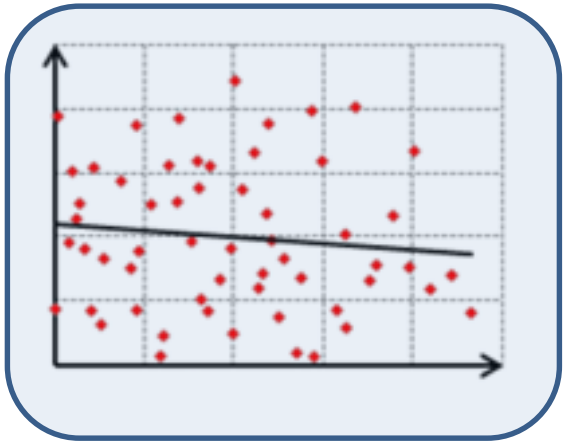
Negative relationship- as foreclosure rates rise, home values drop

Coefficients

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$

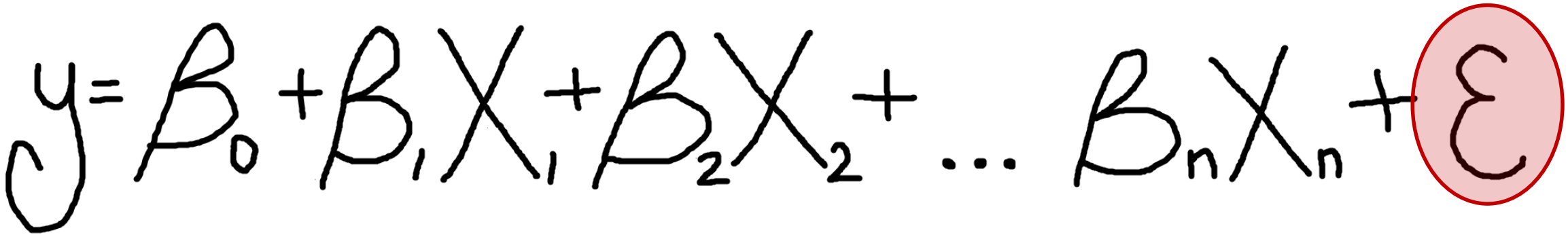






No relationship- the value for X is not correlated with the value for y

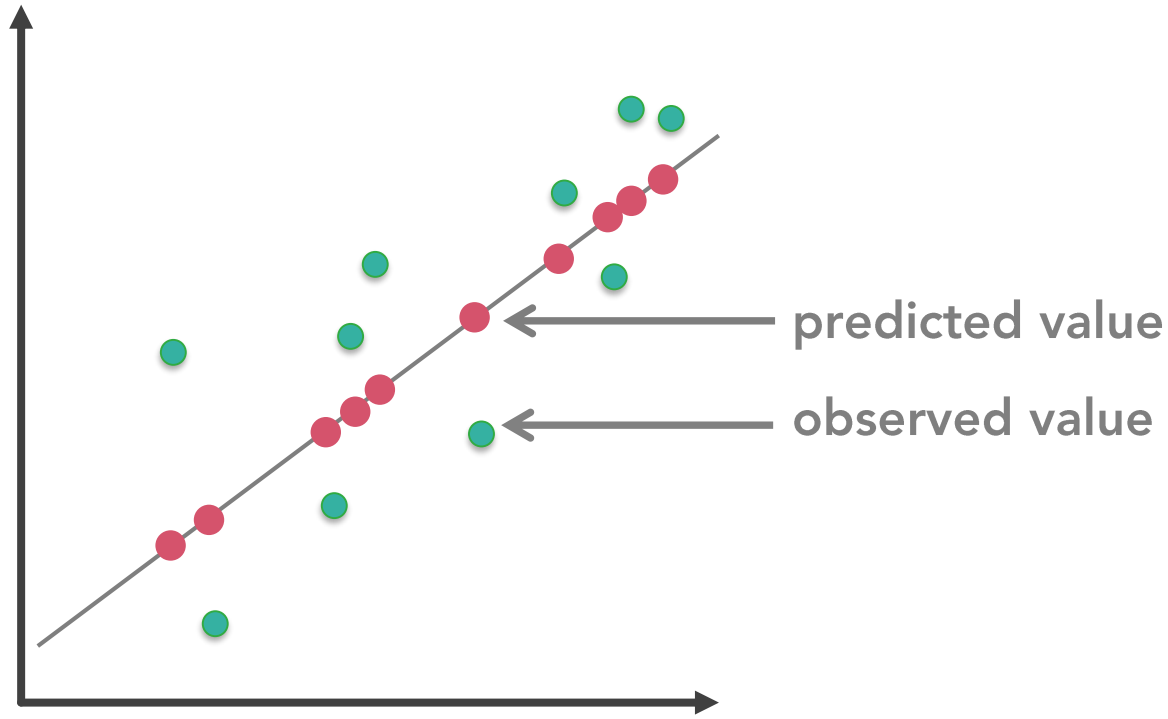
Residual



Model over and under predictions

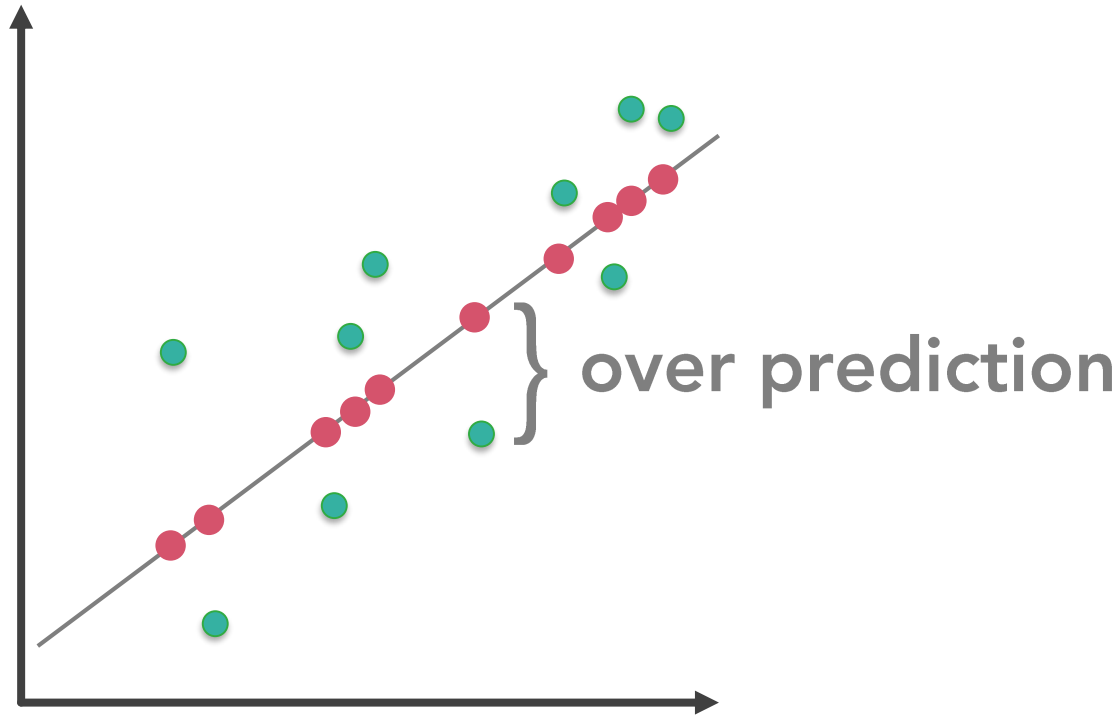
Residual

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$



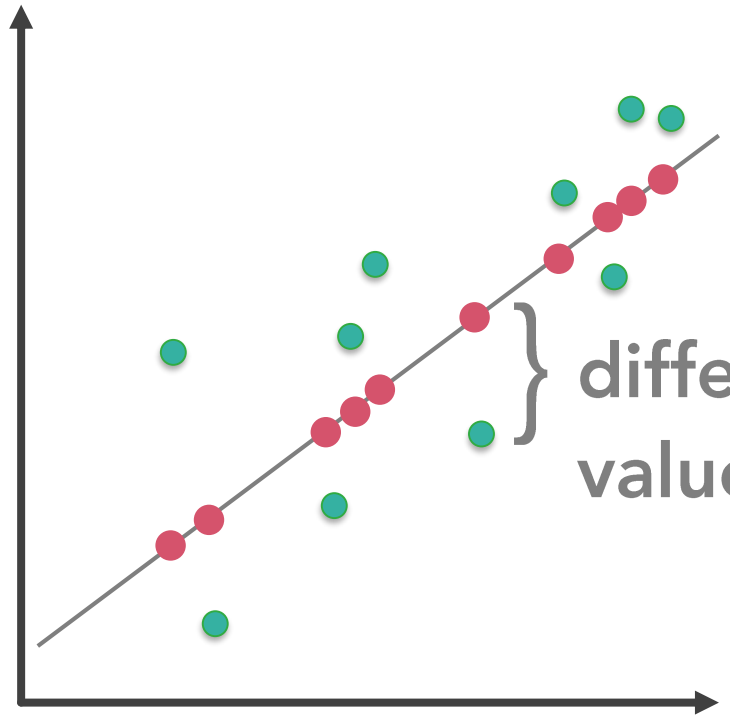
Residual

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$



Residual

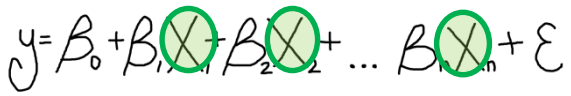
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$



difference between the observed value and the predicted value = ε

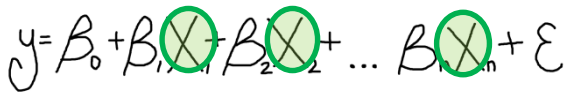
Finding a model we can trust

Every variable should be statistically significant *



Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	9947.038596	44.461950	223.720250	0.000000*	43.232352	230.083219	0.000000*	--------
HOSPBEDSD	453.838027	94.410182	4.807088	0.000006*	95.160978	4.769161	0.000007*	1.789789
EVANDMAND	0.399674	0.036395	10.981628	0.000000*	0.041993	9.517526	0.000000*	3.476159
IMAGINGD	0.359023	0.165592	2.168123	0.032424*	0.161761	2.219469	0.028620*	3.044556
HOUSTOND	-0.001145	0.000115	-9.999029	0.000000*	0.000104	-10.976894	0.000000*	1.329077
PQI10D	1.041130	0.220645	4.718573	0.000009*	0.213298	4.881106	0.000005*	1.528101

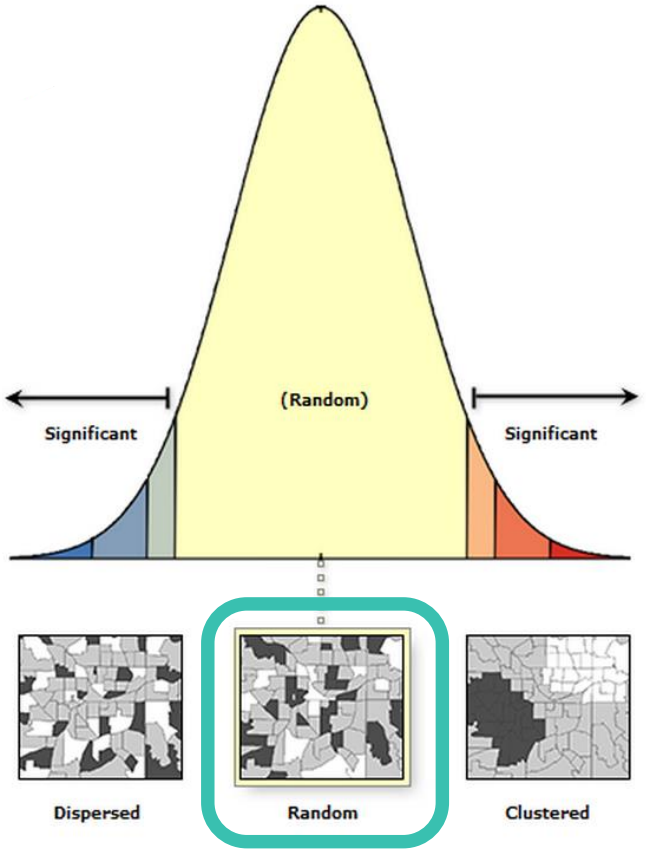
Each variable should tell a different part of the story



Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	9947.038596	44.461950	223.720250	0.000000*	43.232352	230.083219	0.000000*	--------
HOSPBEDSD	453.838027	94.410182	4.807088	0.000006*	95.160978	4.769161	0.000007*	1.789789
EVANDMAND	0.399674	0.036395	10.981628	0.000000*	0.041993	9.517526	0.000000*	3.476159
IMAGINGD	0.359023	0.165592	2.168123	0.032424*	0.161761	2.219469	0.028620*	3.044556
HOUSTOND	-0.001145	0.000115	-9.999029	0.000000*	0.000104	-10.976894	0.000000*	1.329077
PQI10D	1.041130	0.220645	4.718573	0.000009*	0.213298	4.881106	0.000005*	1.528101

Residuals should not be clustered in location or in value

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$

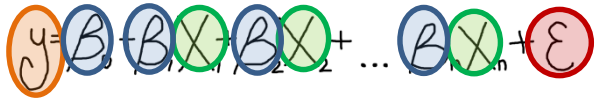


Residuals should not be clustered in location or in value

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \varepsilon$$

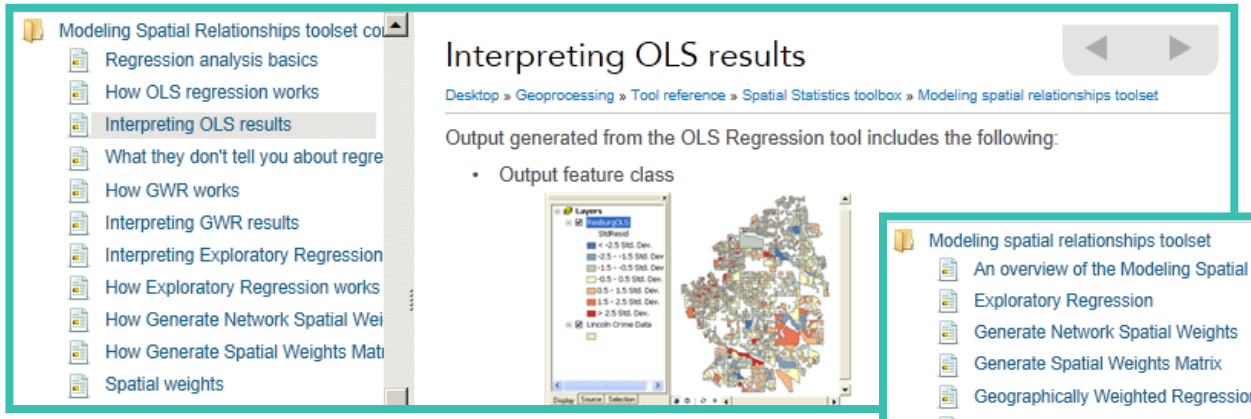
Input Features:	Hot Spot Study Area	Dependent Variable:	TOTAL_COSTS_2010
Number of Observations:	110	Akaike's Information Criterion (AICc) [d]:	1672.966945
Multiple R-Squared [d]:	0.870841	Adjusted R-Squared [d]:	0.864631
Joint F-Statistic [e]:	140.241872	Prob(>F), (5,104) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	556.069919	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	27.470483	Prob(>chi-squared), (5) degrees of freedom:	0.000046*
Jarque-Bera Statistic [g]:	1.591597	Prob(>chi-squared), (2) degrees of freedom:	0.451221

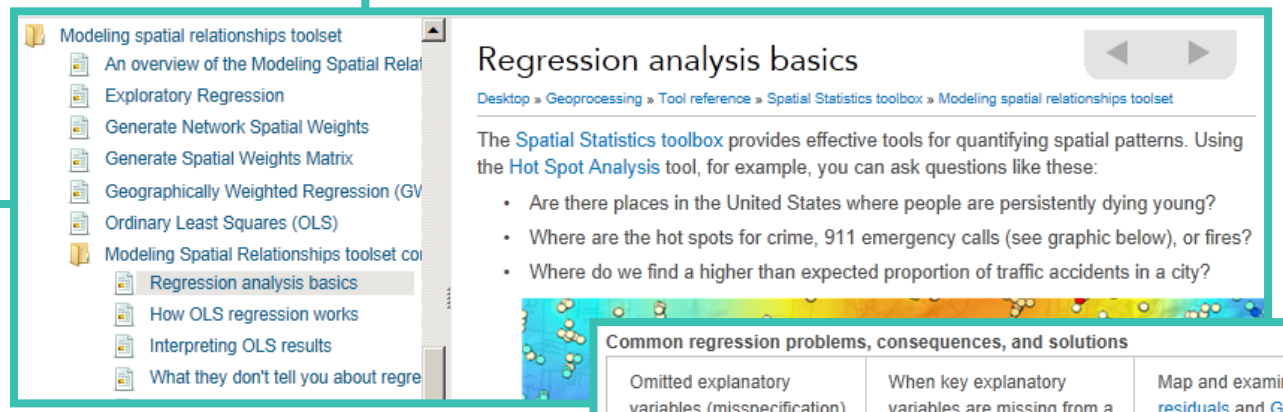
Model should have a strong R-Squared

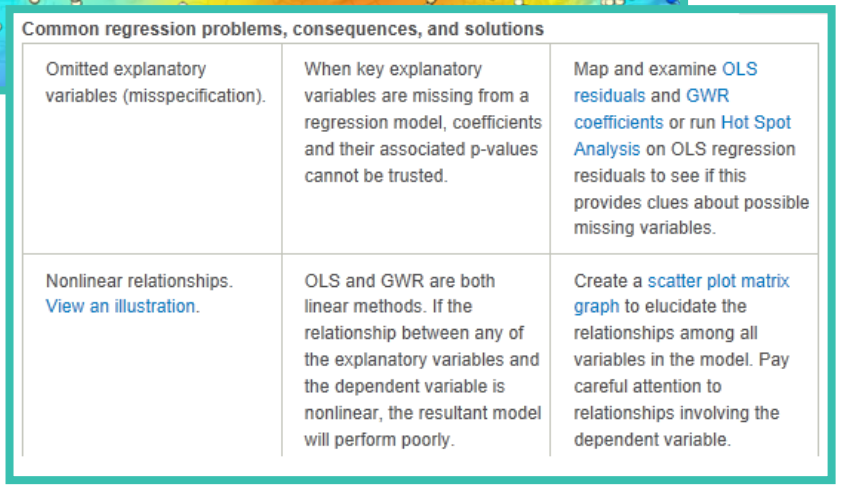


Input Features:	Hot Spot Study Area	Dependent Variable:	TOTAL_COSTS_2010
Number of Observations:	110	Akaike's Information Criterion (AICc) [d]:	1672.966945
Multiple R-Squared [d]:	0.870841	Adjusted R-Squared [d]:	0.864631
Joint F-Statistic [e]:	140.241872	Prob(>F), (5,104) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	556.069919	Prob(>chi-squared), (5) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	27.470483	Prob(>chi-squared), (5) degrees of freedom:	0.000046*
Jarque-Bera Statistic [g]:	1.591597	Prob(>chi-squared), (2) degrees of freedom:	0.451221

Online help is … helpful!







Things to check

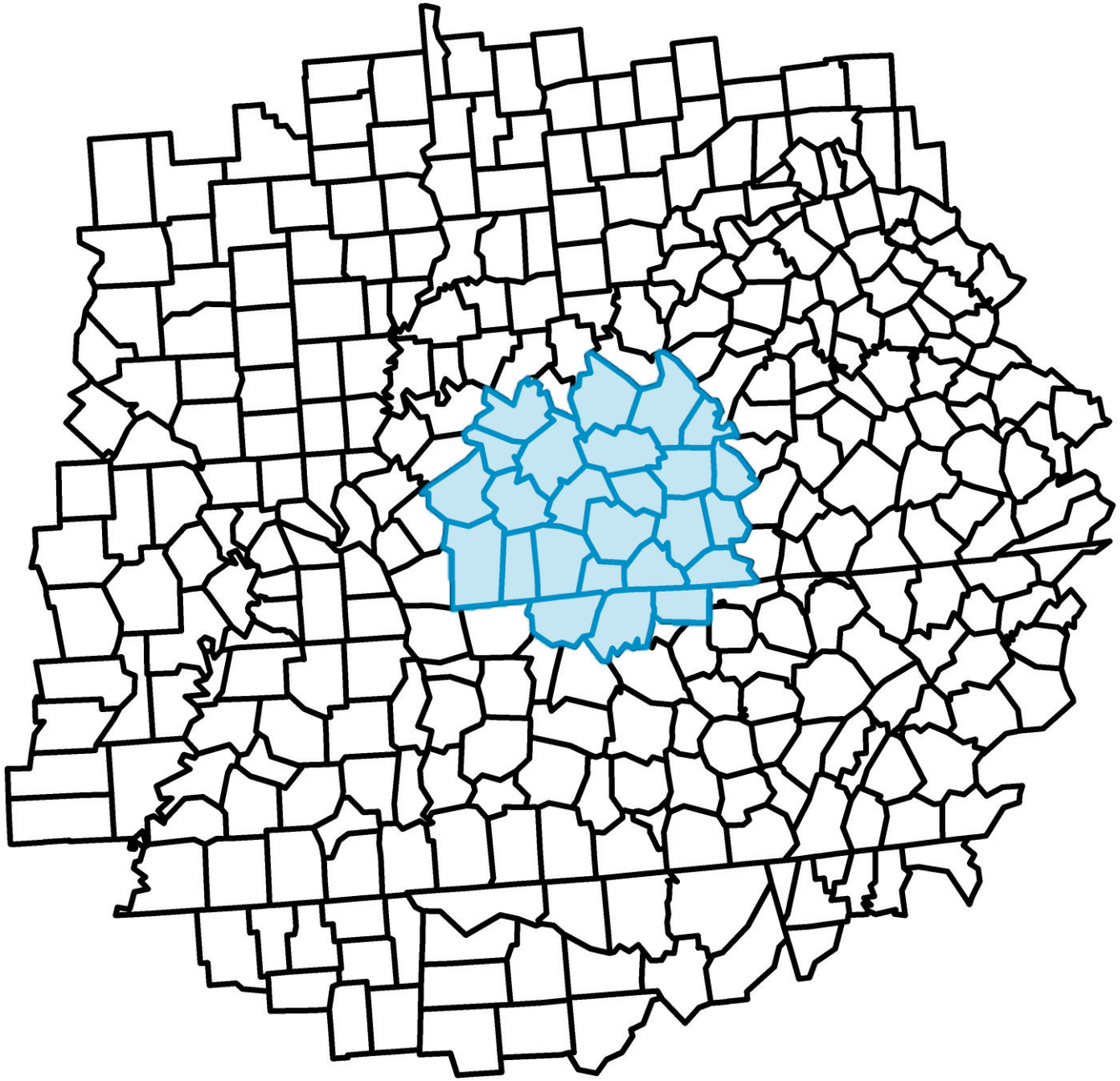
- ✓ Coefficients are statistically significant
- ✓ No redundancy among explanatory variables
- ✓ Residuals are normally distributed
- ✓ Residuals are not spatially autocorrelated
- ✓ Strong Adjusted R2 (good model performance)

Exploratory Regression

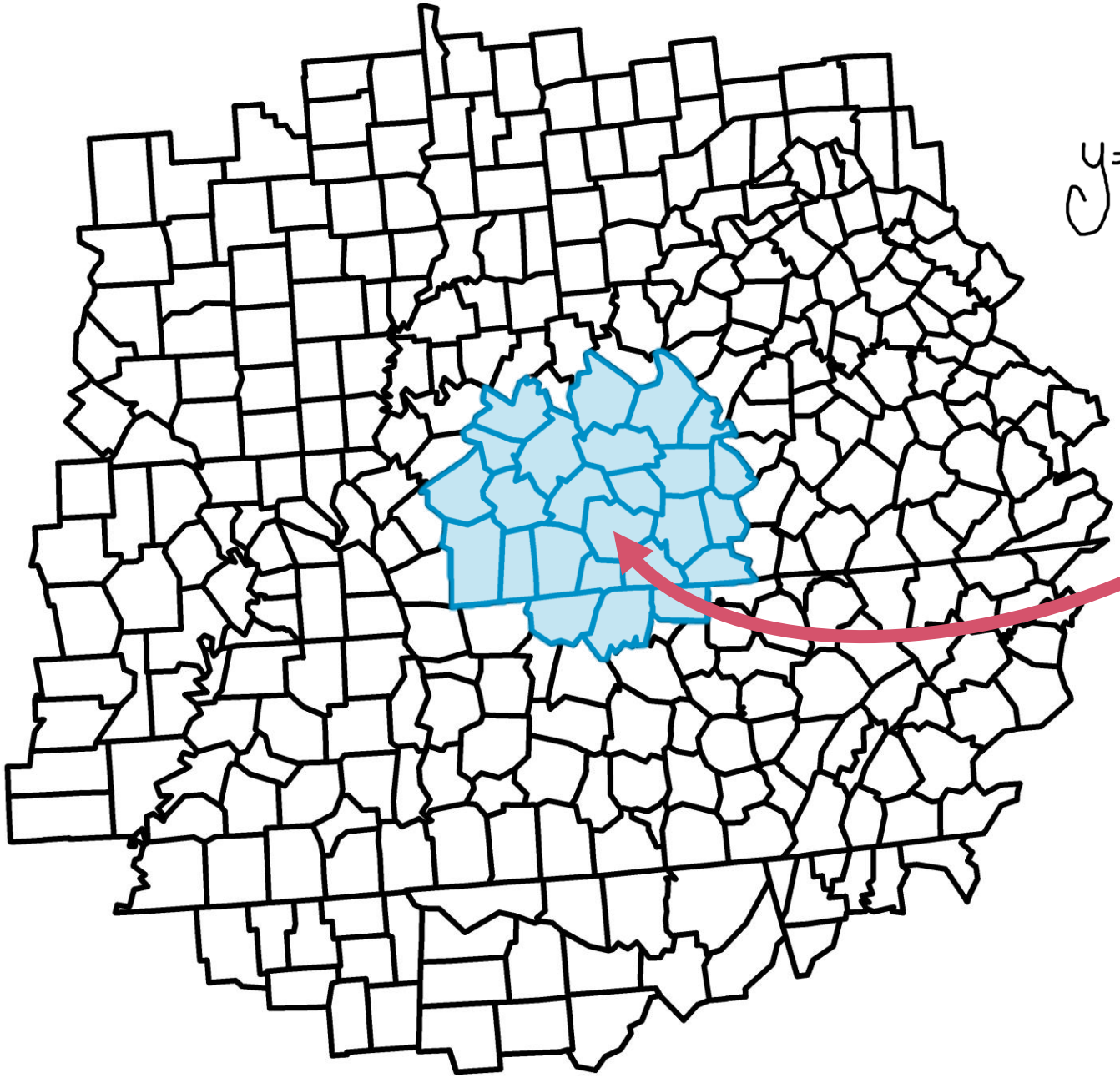
Demo

Geographically Weighted Regression

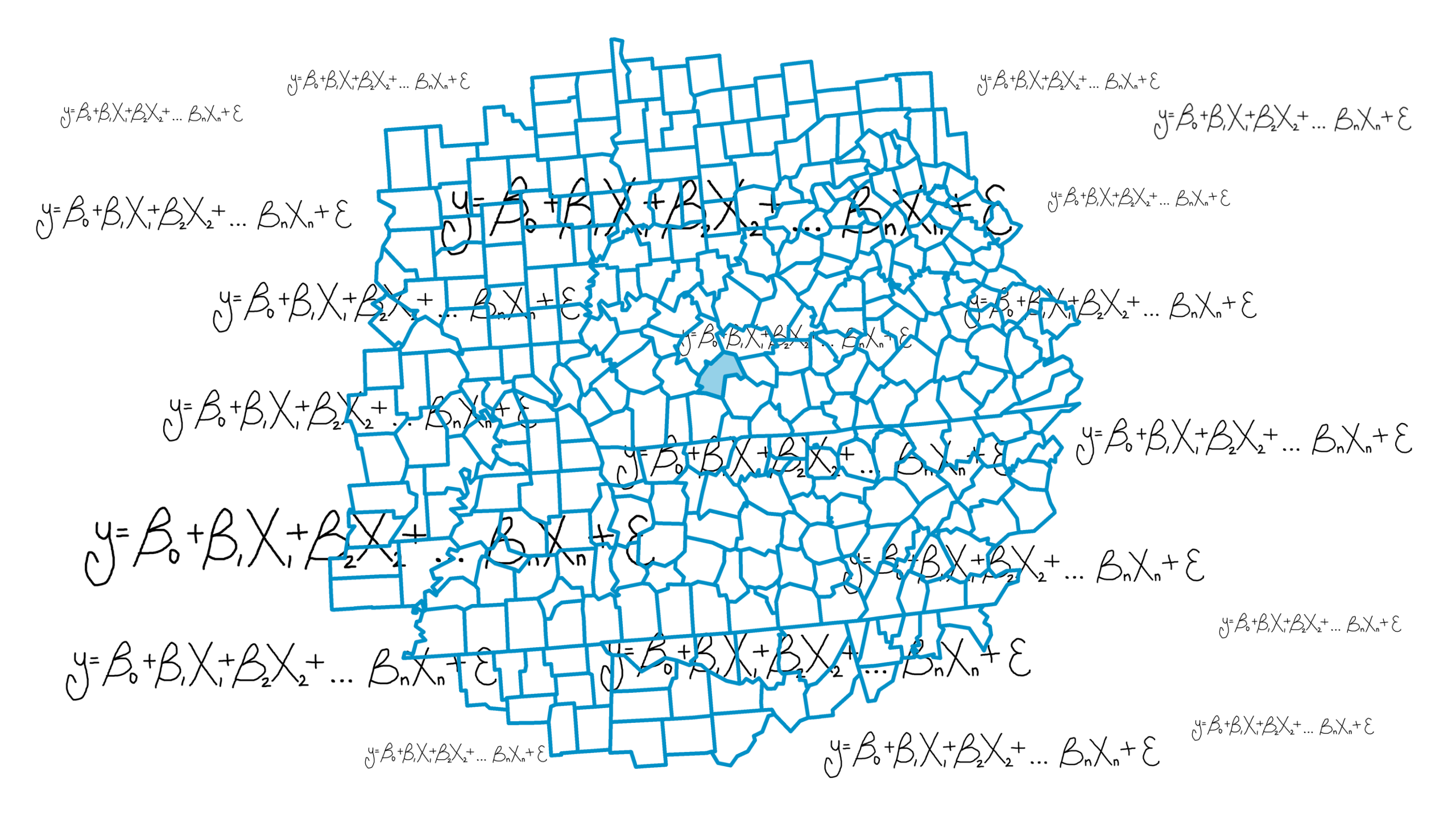
Exploring spatial variation



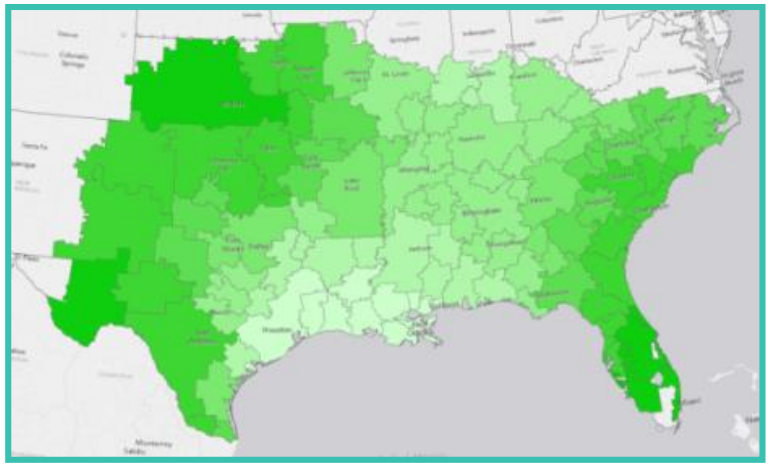
each feature gets a separate equation



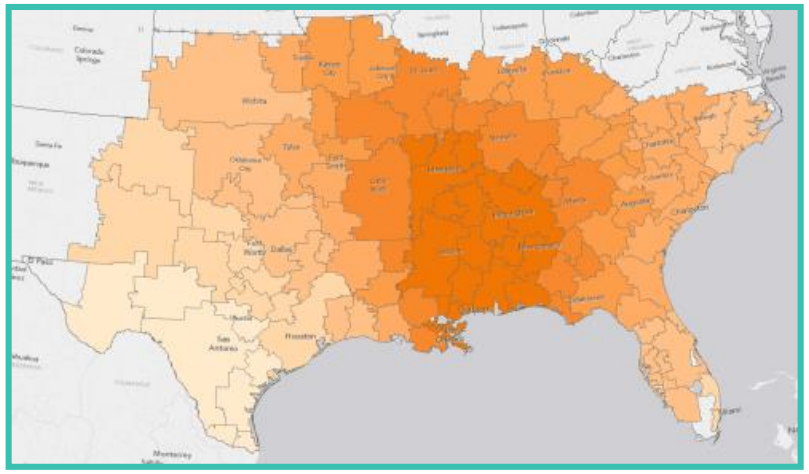
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$


$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$

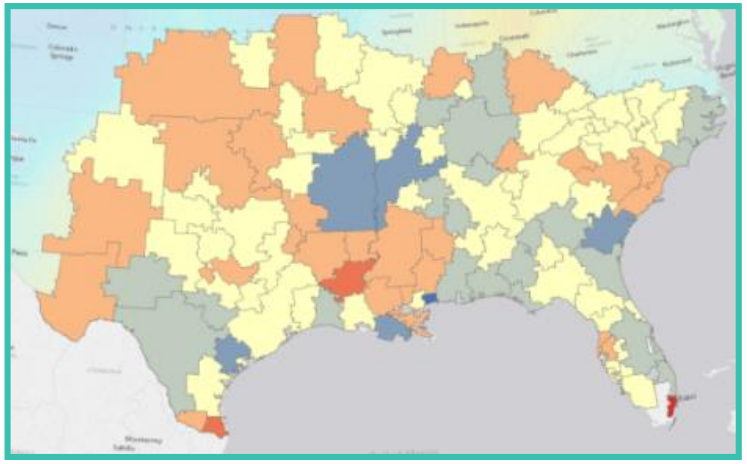
Demo



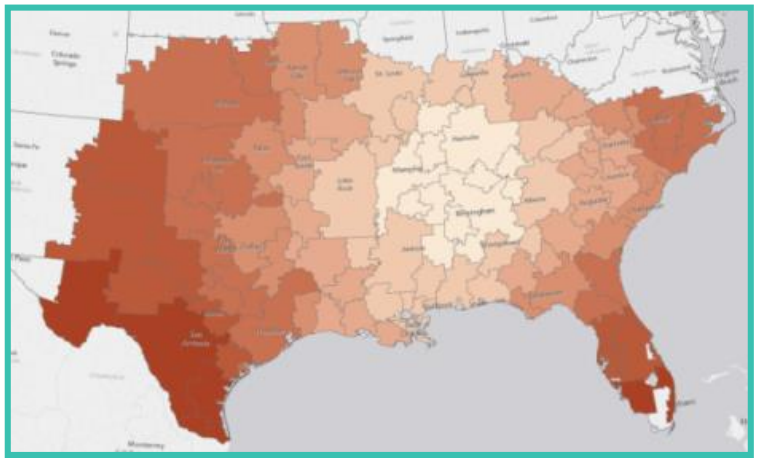
Local R-Squared



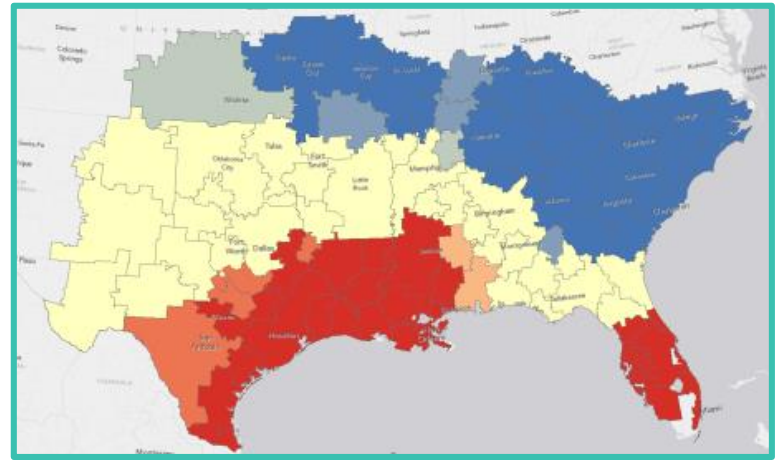
Coefficients



Residuals



Condition Number



Predictions

Forest-based Classification & Regression

Predicting using machine learning

Training

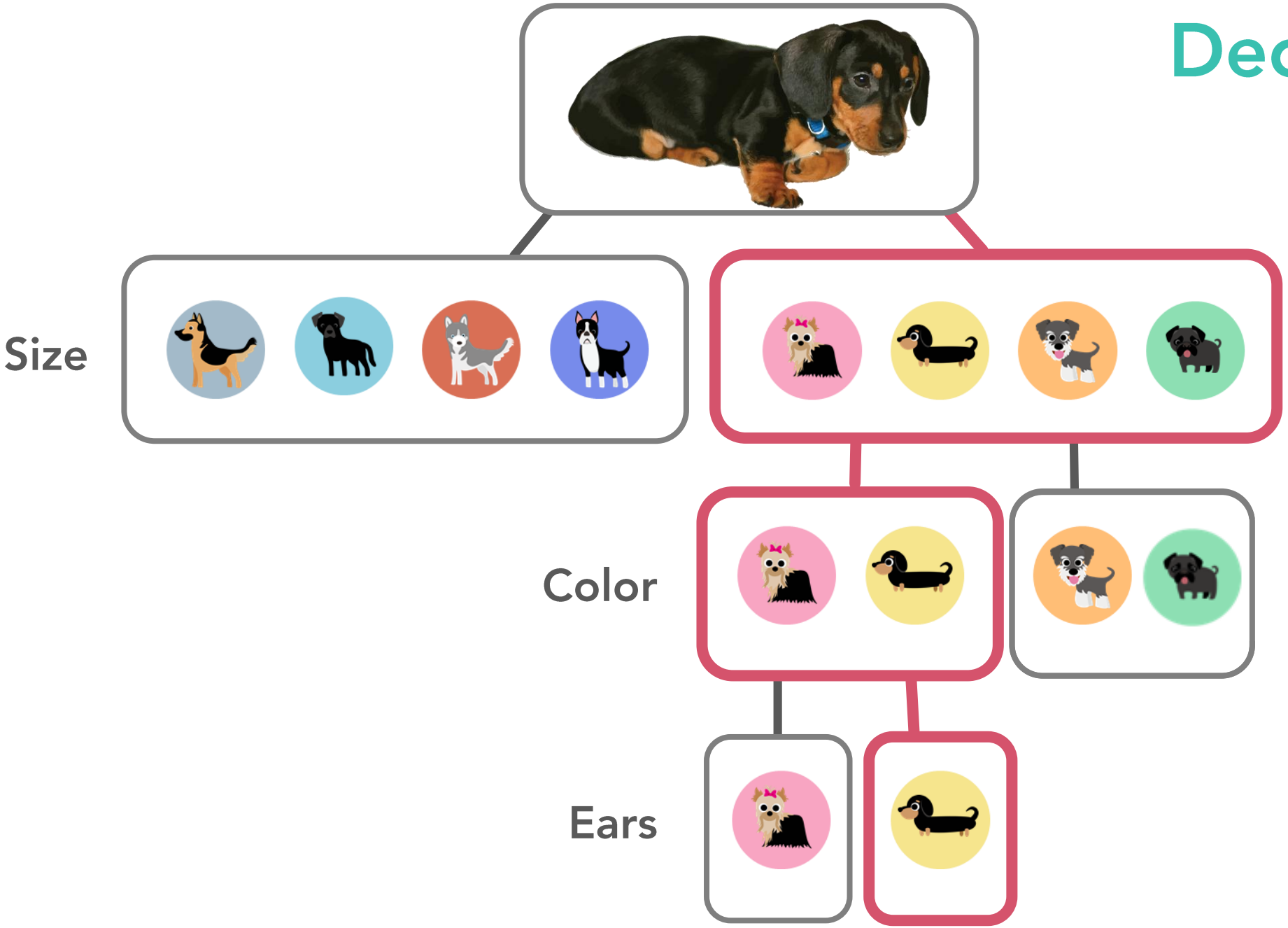
variable to predict

Breed

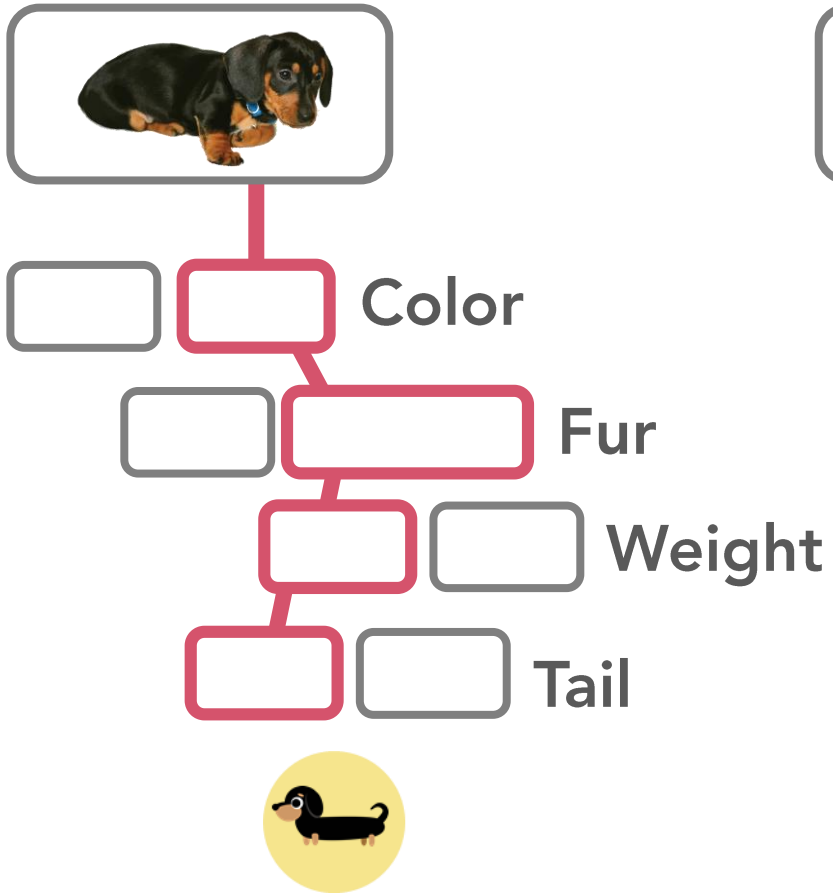
explanatory variables

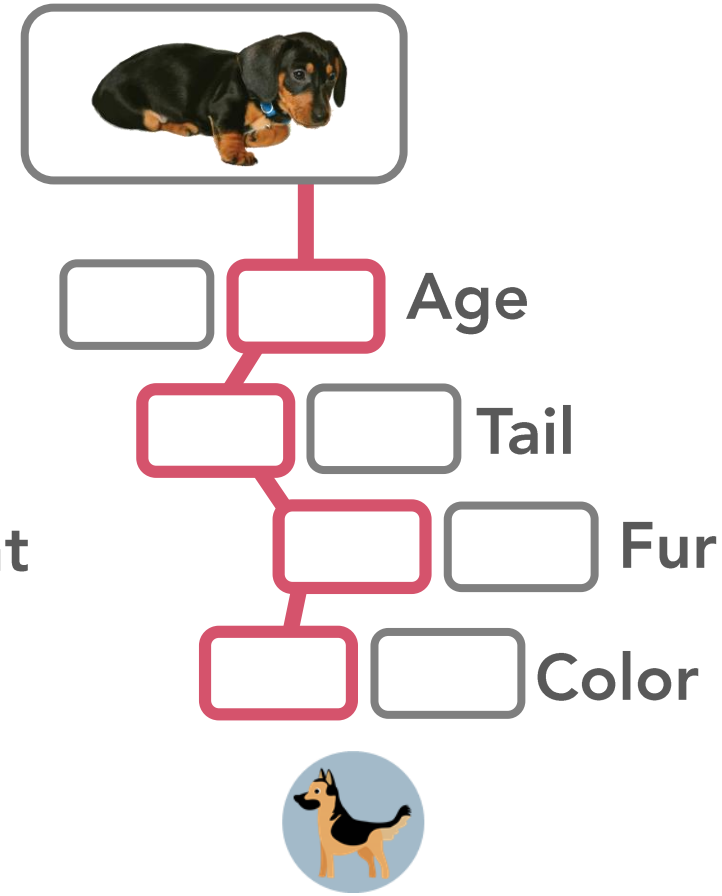
- Size
- Color
- Fur
- Ears
- Tail
- Age
- Weight

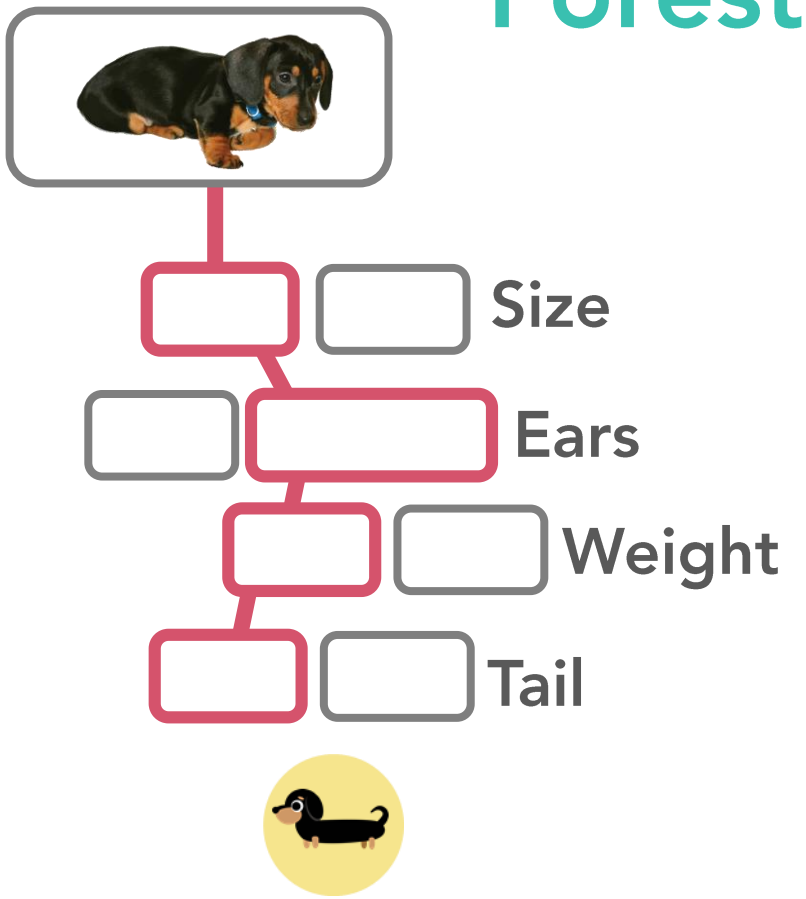
Decision Tree



Forest

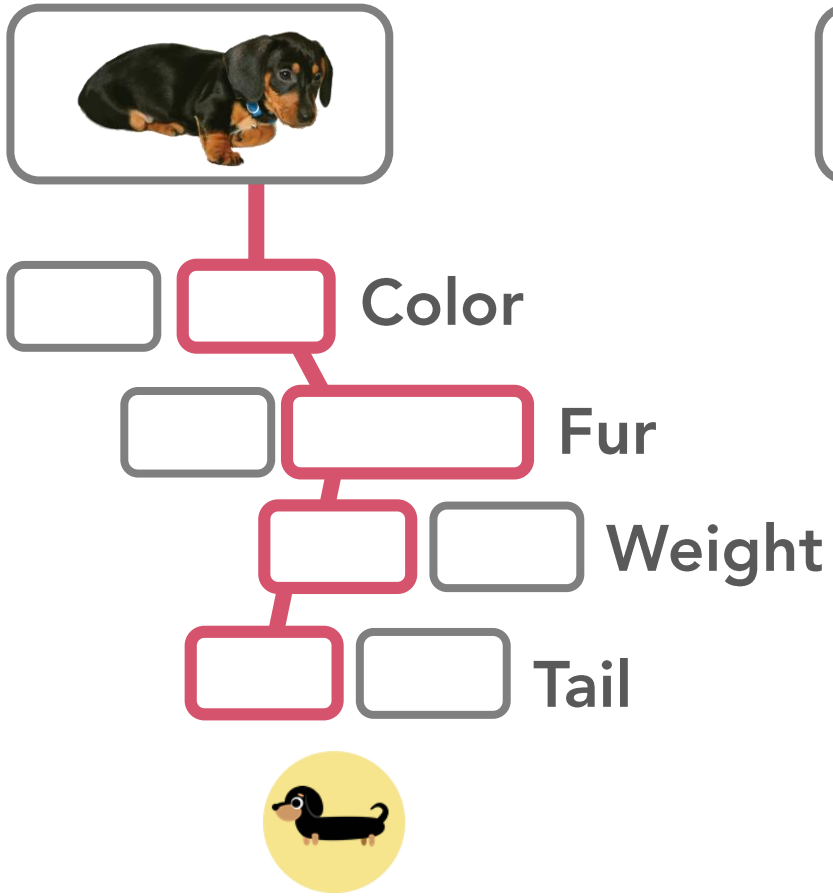


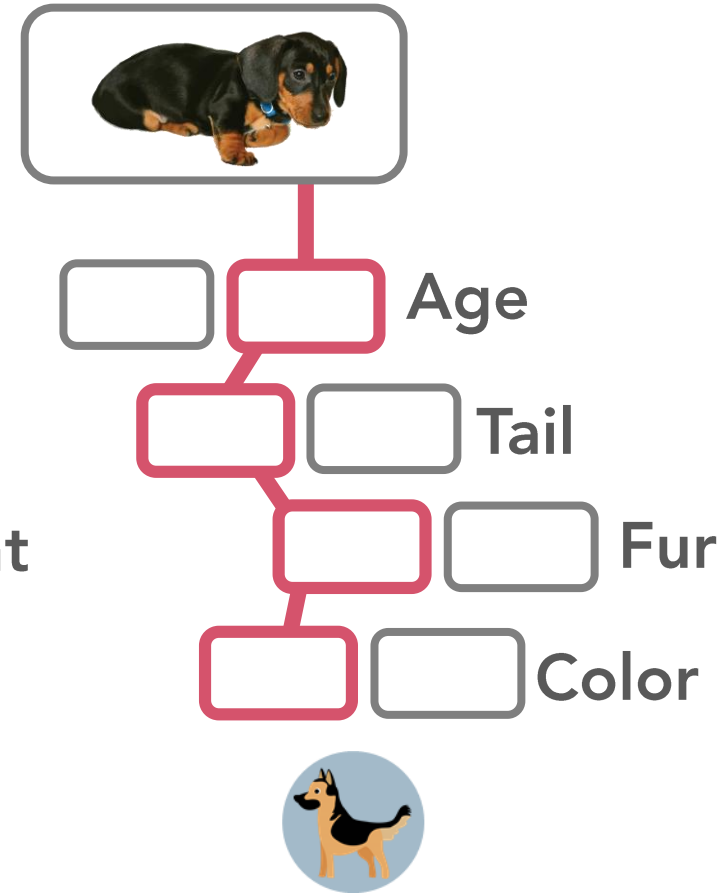


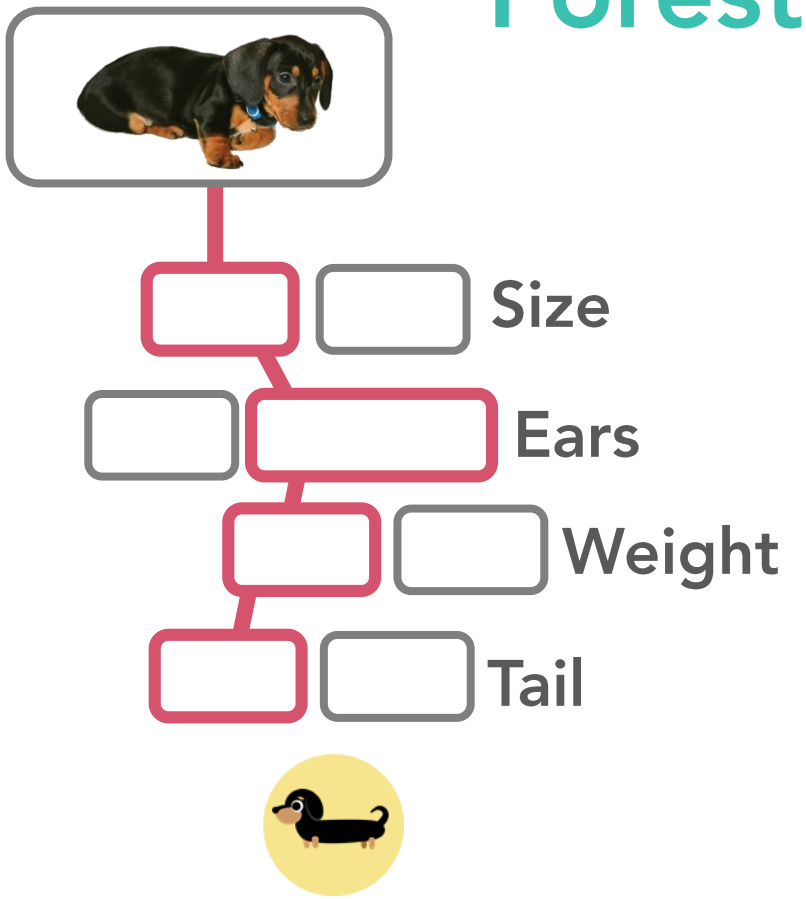


Random subset of data and variables used in each tree

Forest

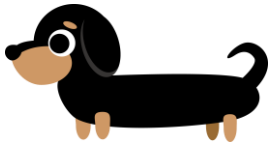






Majority vote wins

=

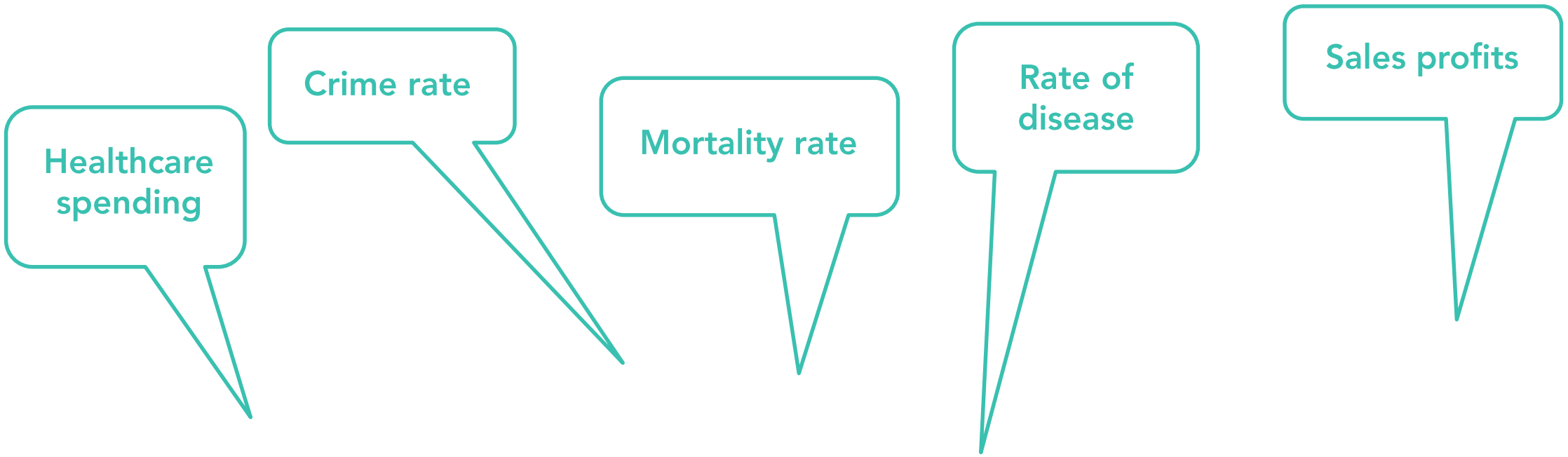


Classification

Predict categorical variable

Regression

Predict continuous variable

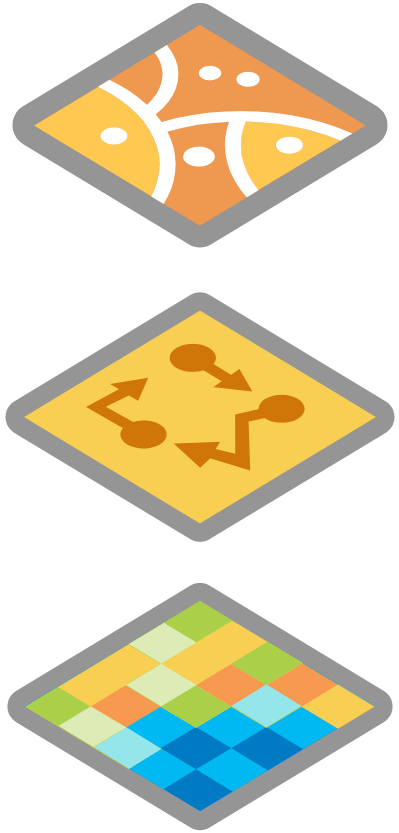


Explanatory Variables

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

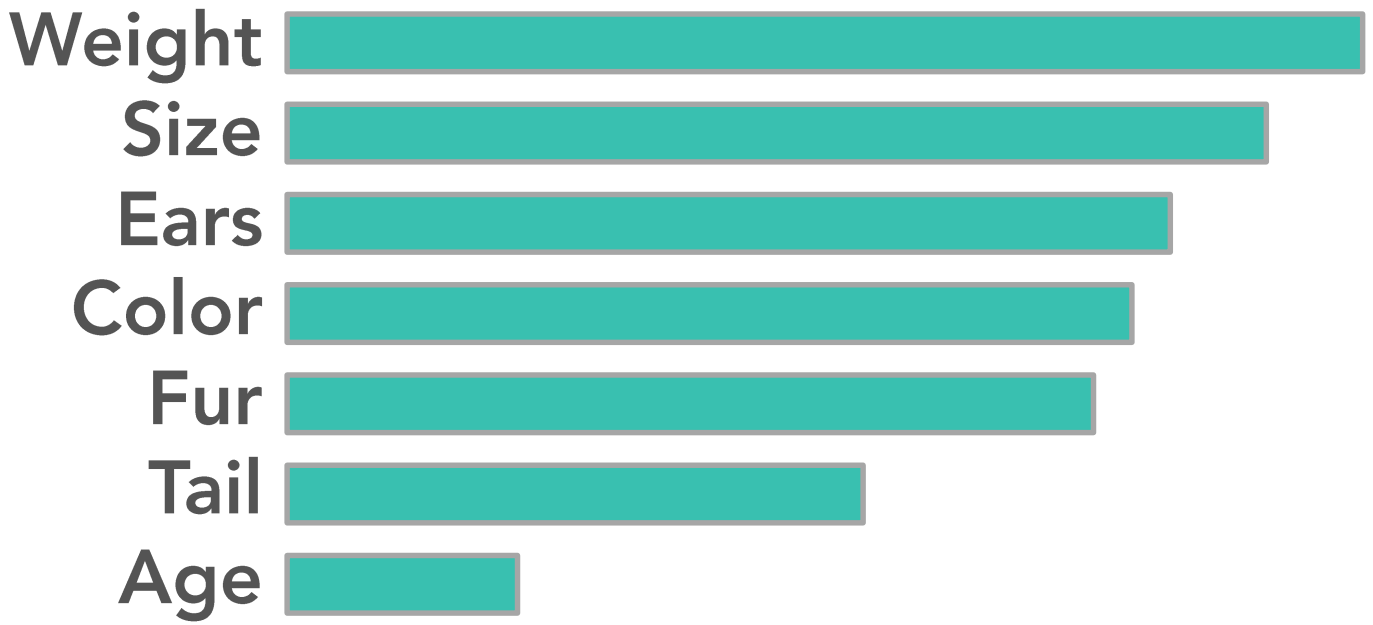
Predict values in a different study area

Predict values in a different time period

Finding a model we can trust

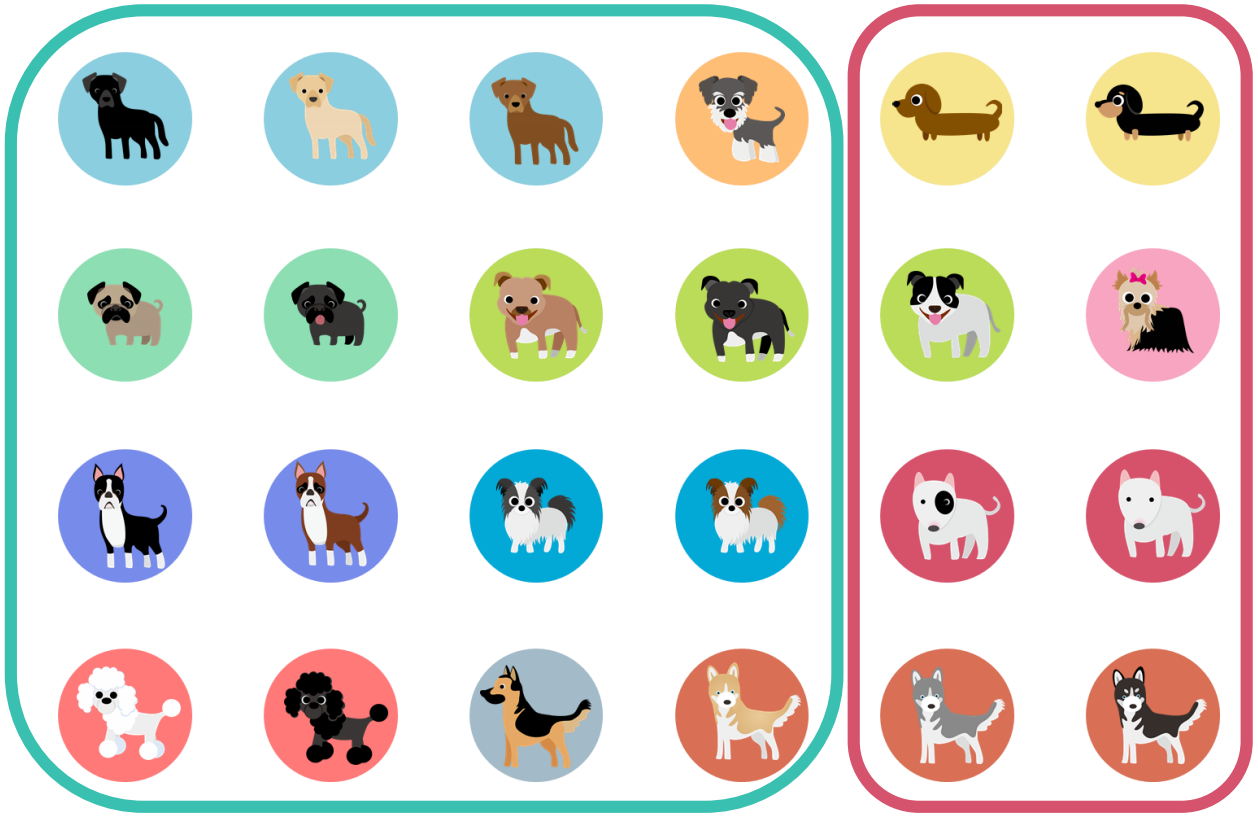
Variable importance

How well does each variable do in splitting the trees?



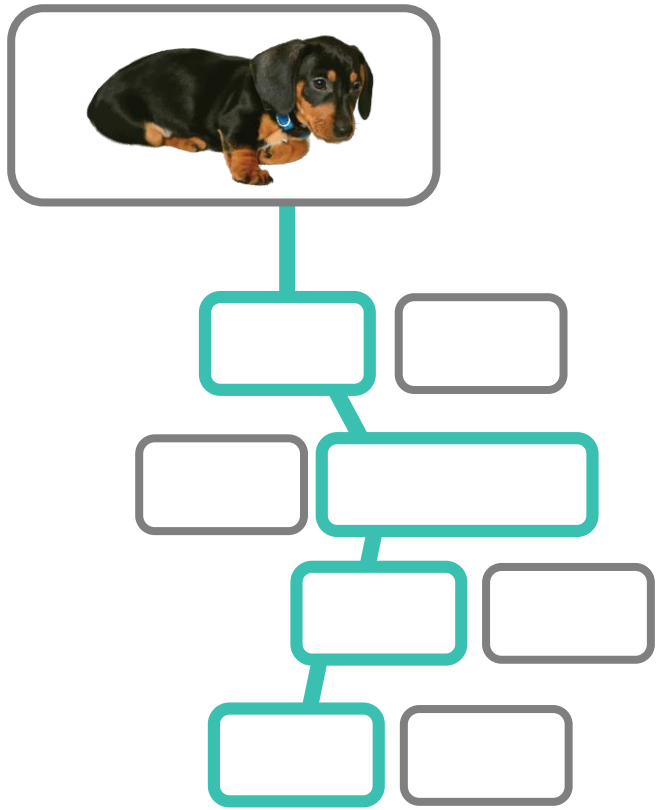
Out Of Bag errors

How well can each tree predict the excluded features?



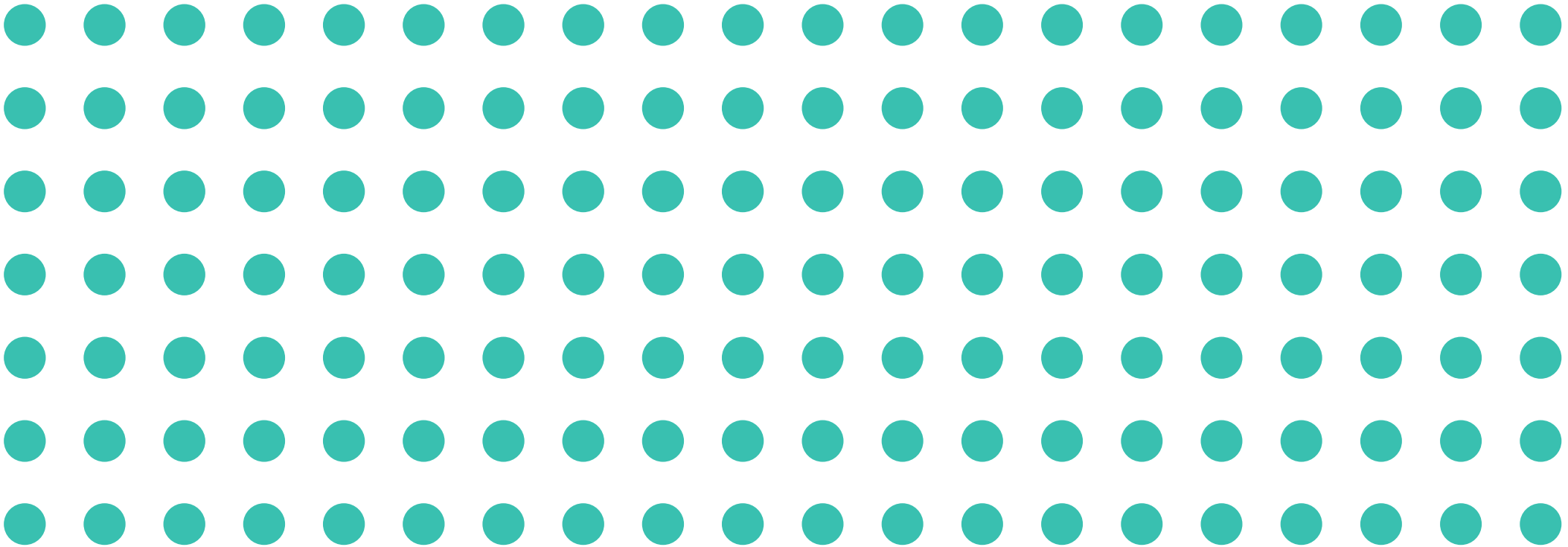
2/3 included (randomly)

1/3 excluded



Model Validation

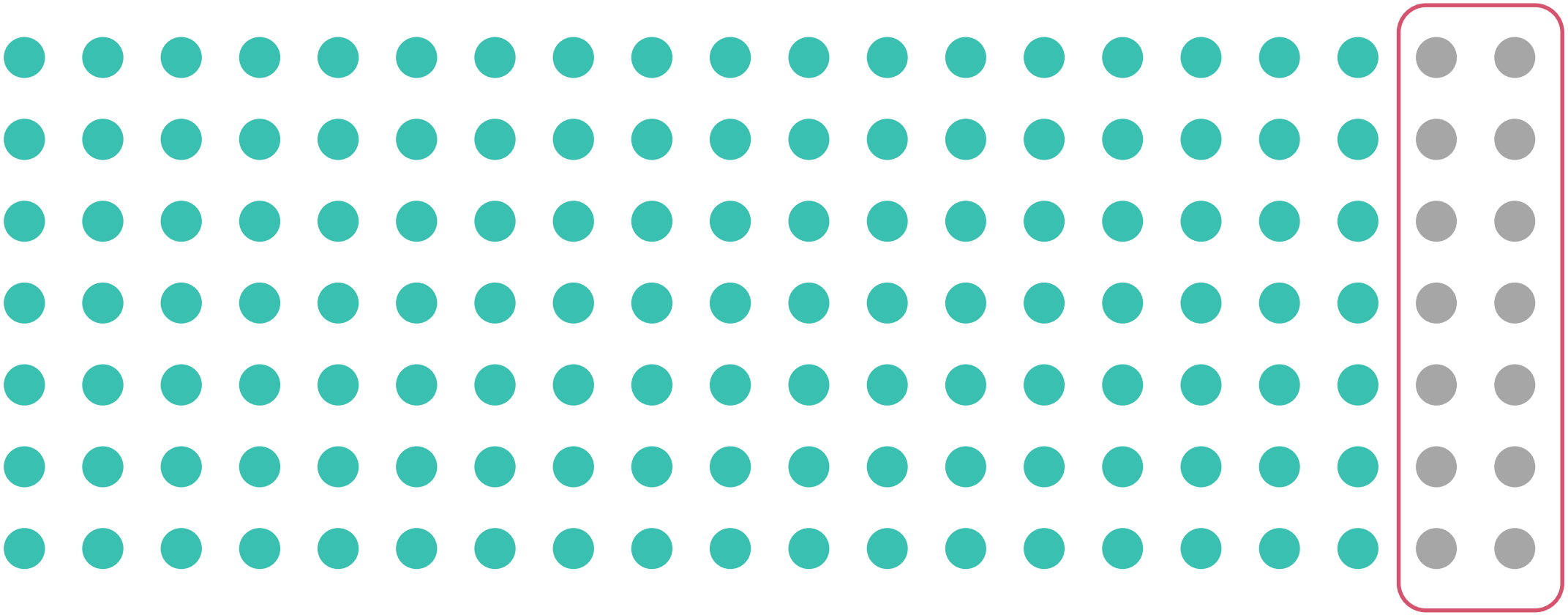
Training features



Model Validation

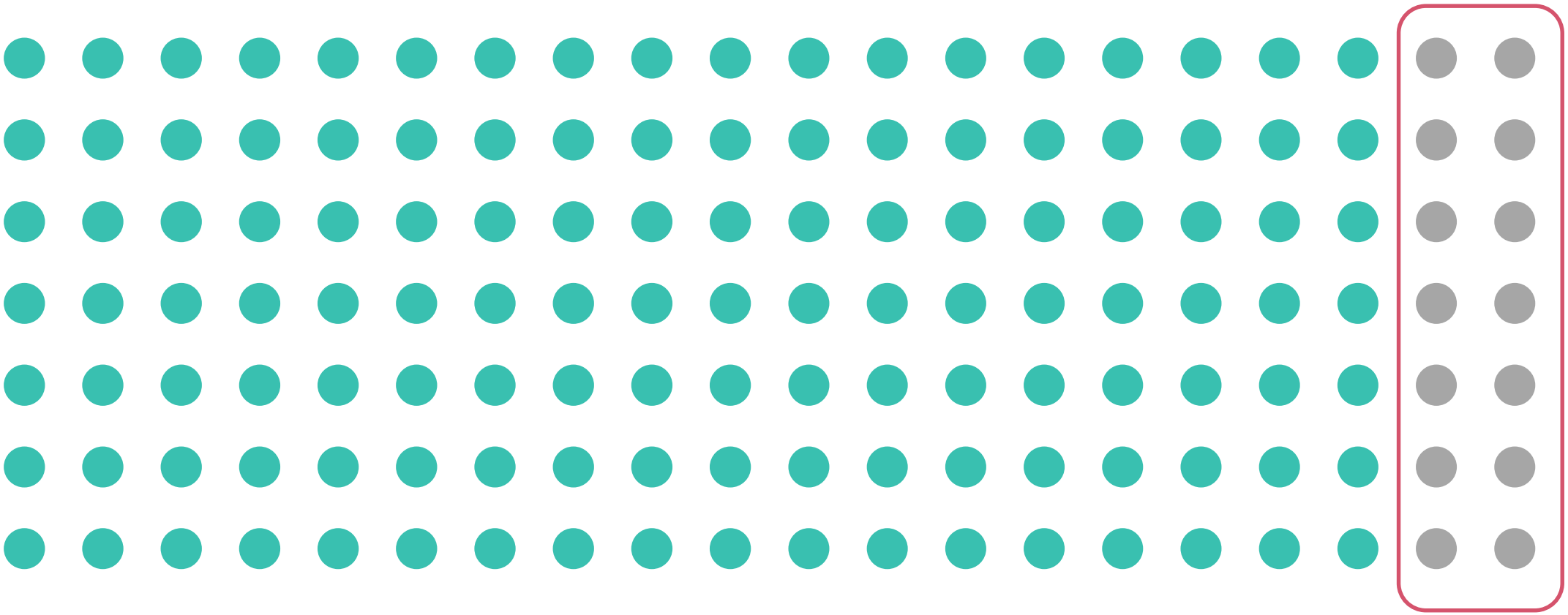
Training features

10% held back

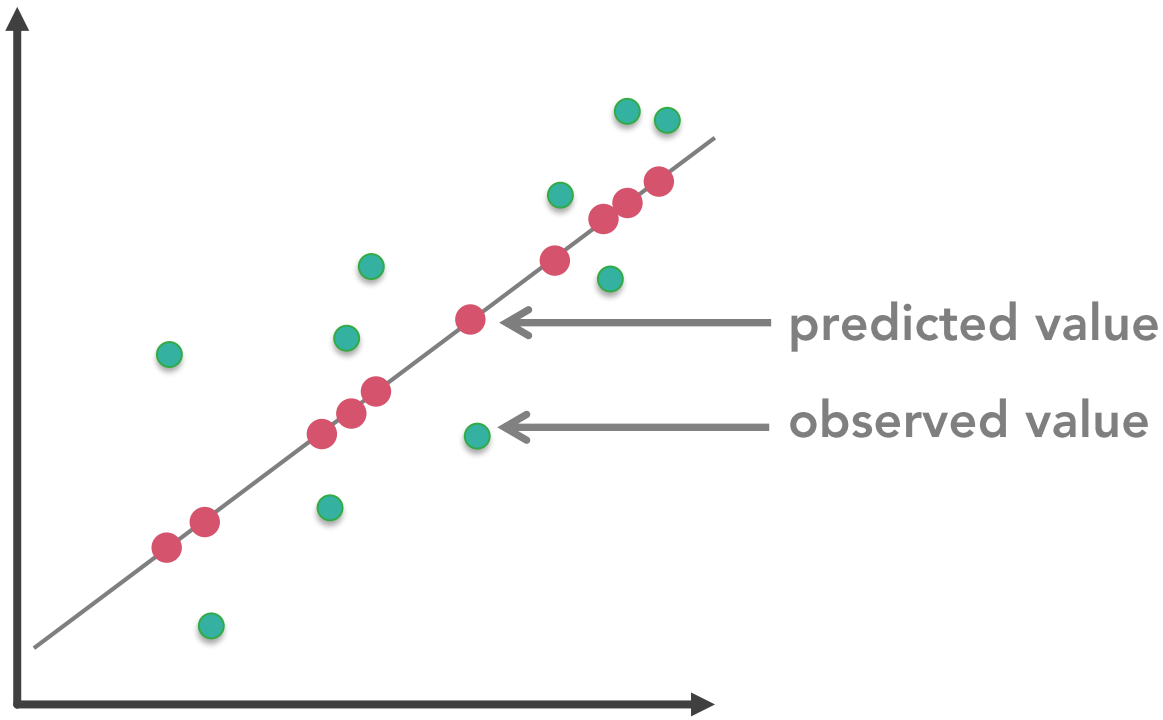


Model Validation

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 

8/(8+2)

Confusion matrix

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

Accuracy for

15/20

Demo

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

- George E. P. Box

Want to learn more???

esriurl.com/spatialstats

TUESDAY

~~8:30a From Means and Medians to Machine Learning: Spatial Statistics Basics and Innovations 15B~~

~~10a Data Visualization for Spatial Analysis 10~~

~~2:30p Spatial Data Mining I: Essentials of Cluster Analysis 15B~~

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

WEDNESDAY

~~10a Spatial Data Mining II: A Deep Dive Into Space-Time Analysis Room 29C~~

~~2:30p Spatial Data Mining I: Essentials of Cluster Analysis Room 15A~~

~~4p Spatial Data Mining II: A Deep Dive Into Space-Time Analysis Room 31B~~

THURSDAY

~~10a Data Visualization for Spatial Analysis 07A/B~~

~~1p Beyond Where: Modeling Spatial Relationships and Making Predictions 17B~~

4p Beyond Where: Modeling Spatial Relationships and Making Predictions 17A

lbennett@esri.com
jdacosta@esri.com
fvale@esri.com

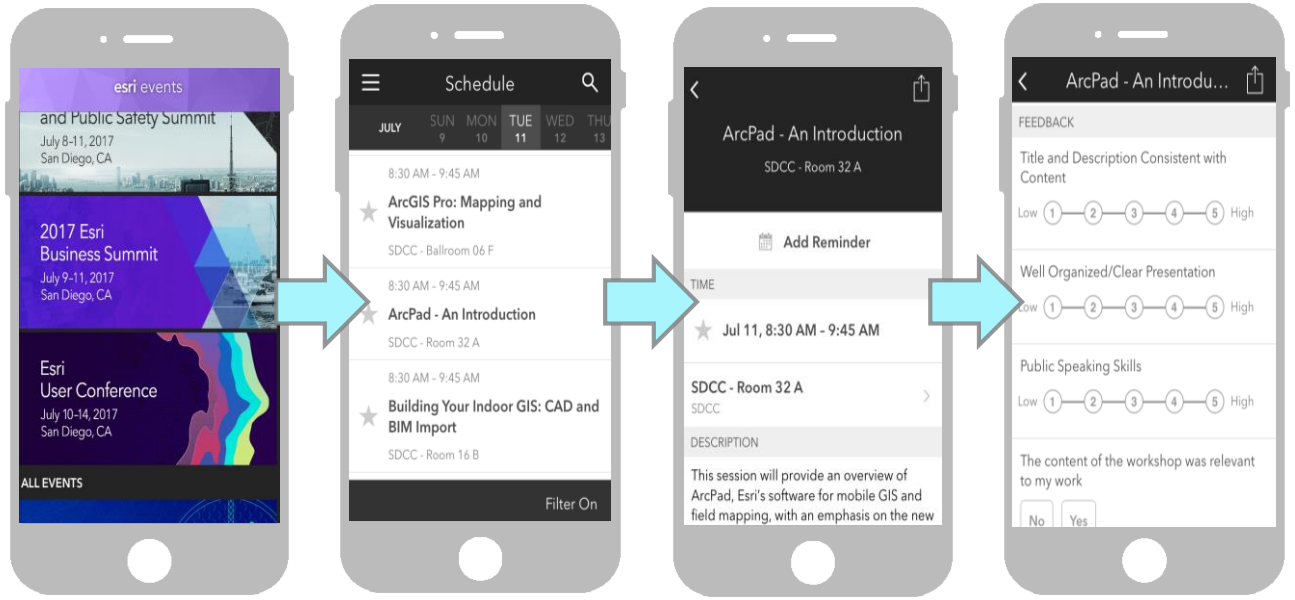
Please fill out a course survey!!

Download the Esri Events app and find your event

Select the session you attended

Scroll down to find the survey

Complete Answers and Select "Submit"



Want to learn more???

esriurl.com/spatialstats

TUESDAY

~~8:30a From Means and Medians to Machine Learning: Spatial Statistics Basics and Innovations 15B~~

~~10a Data Visualization for Spatial Analysis 10~~

~~2:30p Spatial Data Mining I: Essentials of Cluster Analysis 15B~~

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

WEDNESDAY

~~10a Spatial Data Mining II: A Deep Dive Into Space-Time Analysis Room 29C~~

~~2:30p Spatial Data Mining I: Essentials of Cluster Analysis Room 15A~~

~~4p Spatial Data Mining II: A Deep Dive Into Space-Time Analysis Room 31B~~

THURSDAY

~~10a Data Visualization for Spatial Analysis 07A/B~~

~~1p Beyond Where: Modeling Spatial Relationships and Making Predictions 17B~~

~~4p Beyond Where: Modeling Spatial Relationships and Making Predictions 17A~~

esri | THE SCIENCE OF WHERE

lbennett@esri.com
jdacosta@esri.com
fvale@esri.com

Please fill out a course survey!!

Download the Esri Events app and find your event

Select the session you attended

Scroll down to find the survey

Complete Answers and Select "Submit"