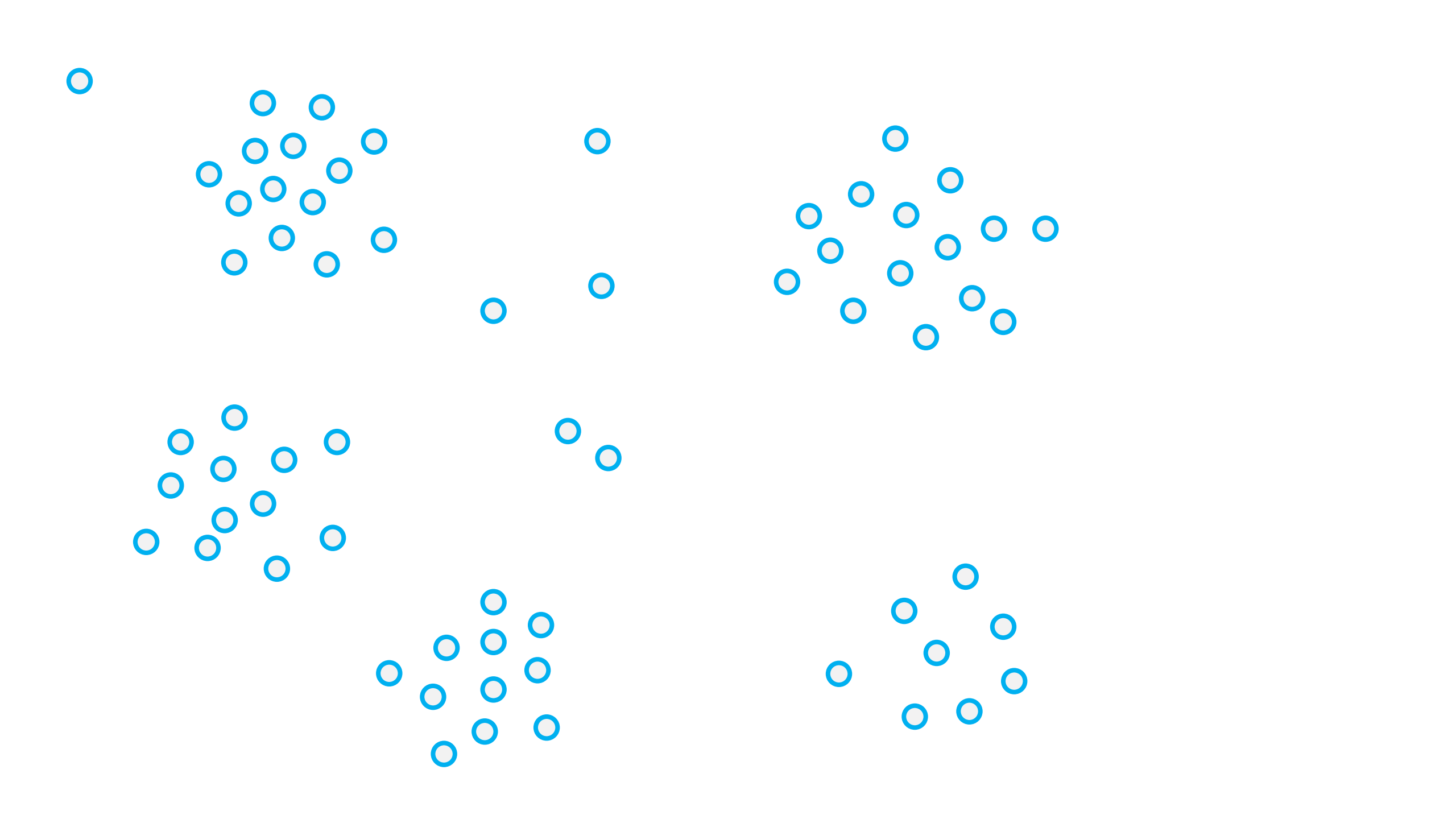
Unpacking the Black Box: Spatial Data Science Methods Explained

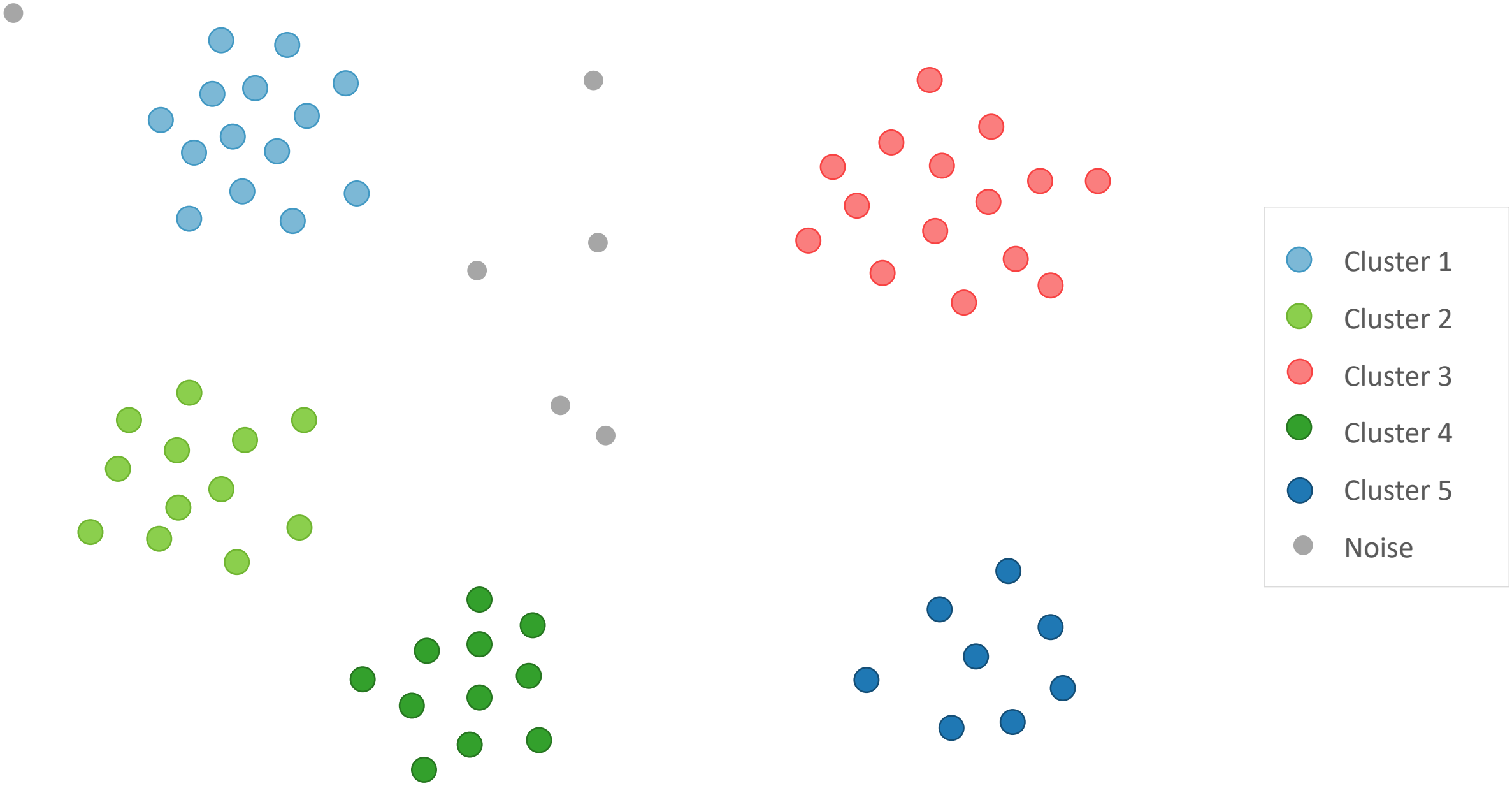
Lauren Bennett
Alberto Nieto

2020 ESRI DEVELOPER SUMMIT | Palm Springs, CA

Density-based Clustering

finds clusters based on feature locations



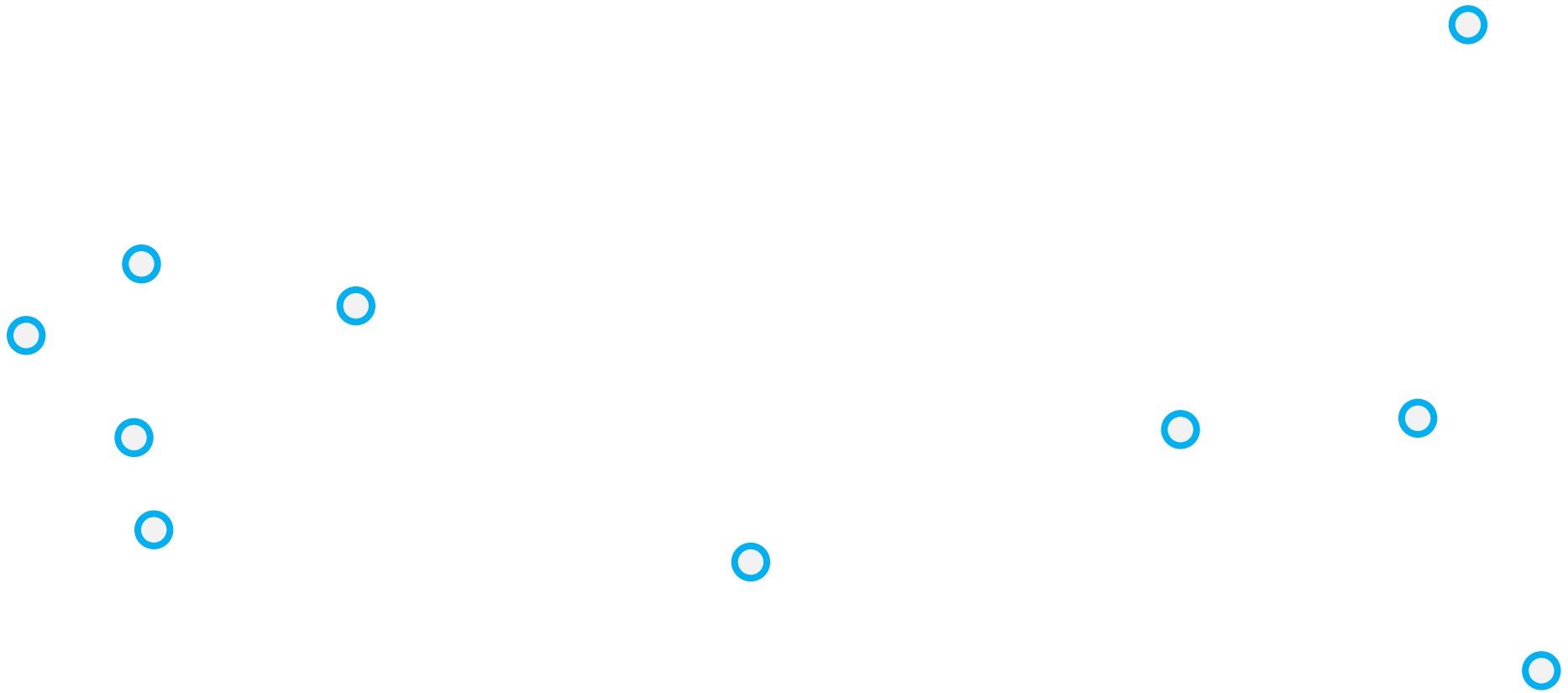

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Noise

DBSCAN – defined distance

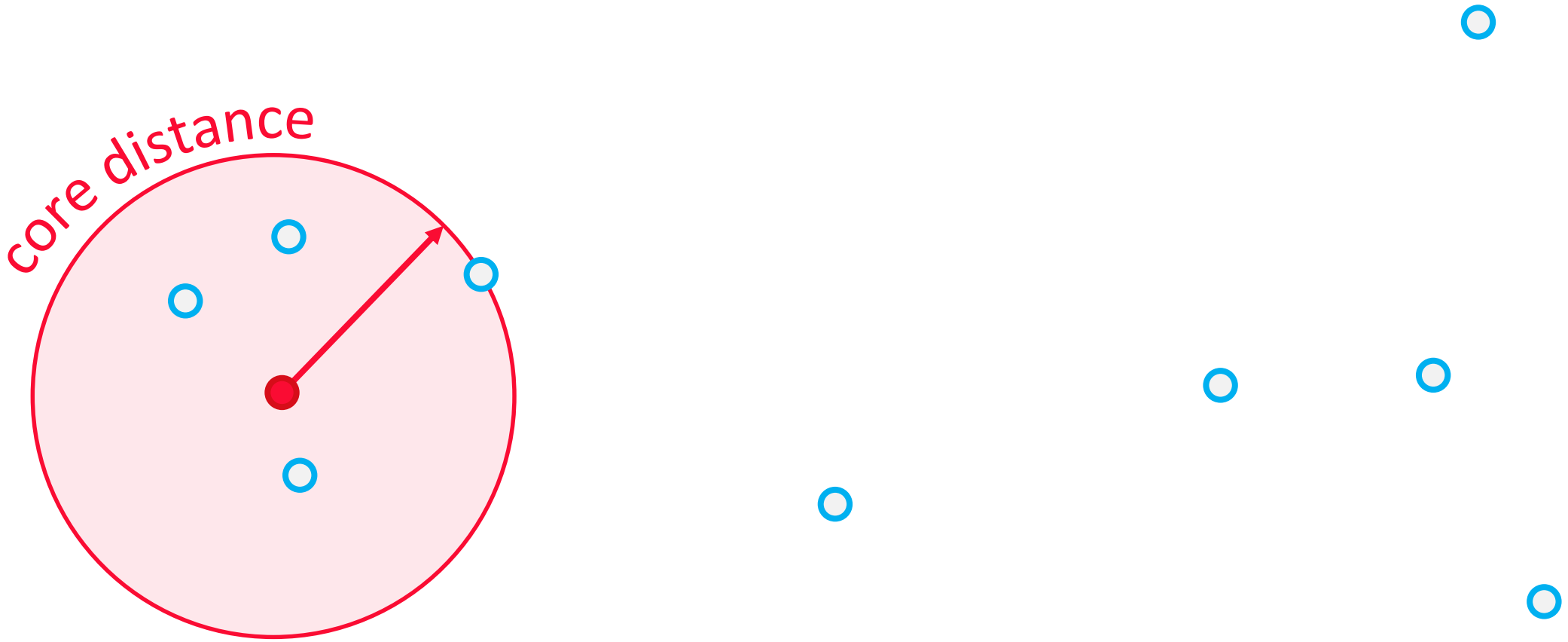
HDBSCAN – self adjusting

OPTICS – multi-scale

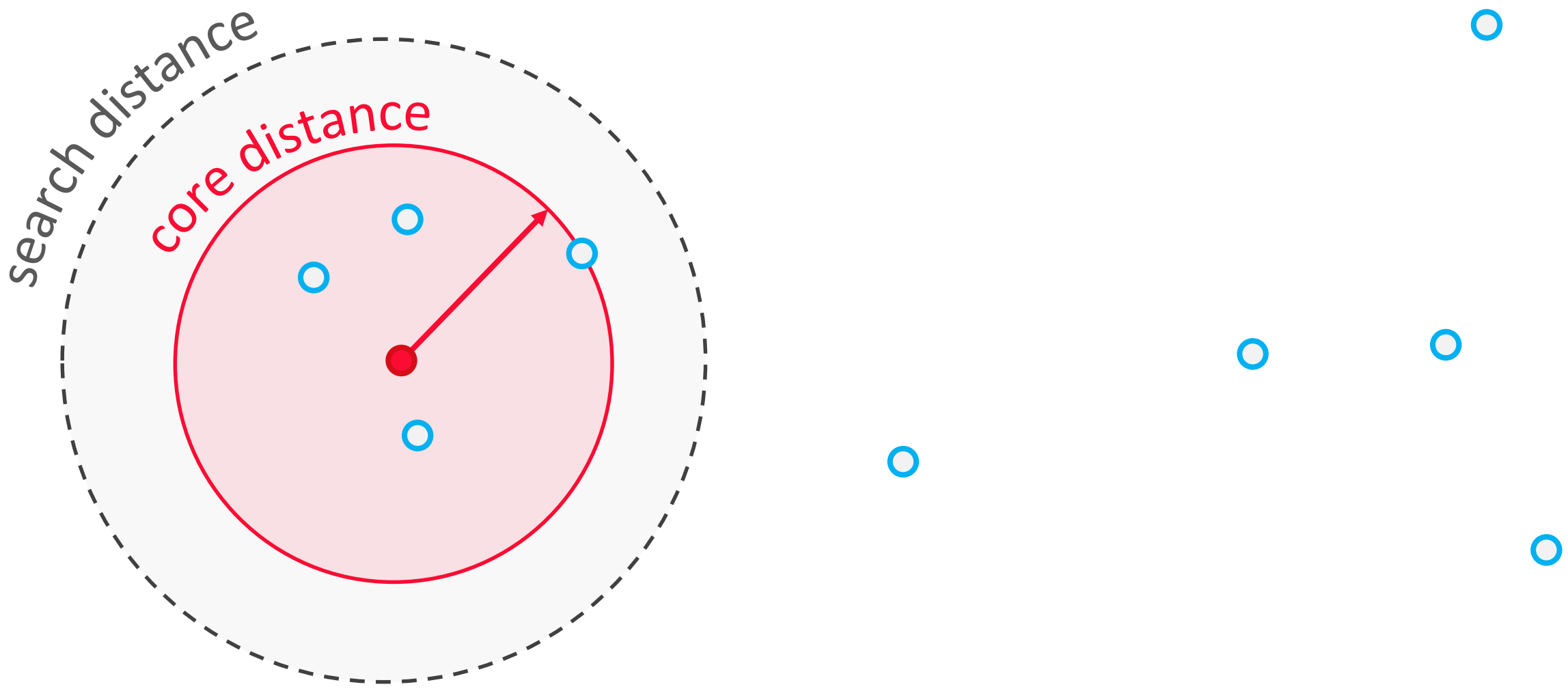
DBSCAN – defined distance



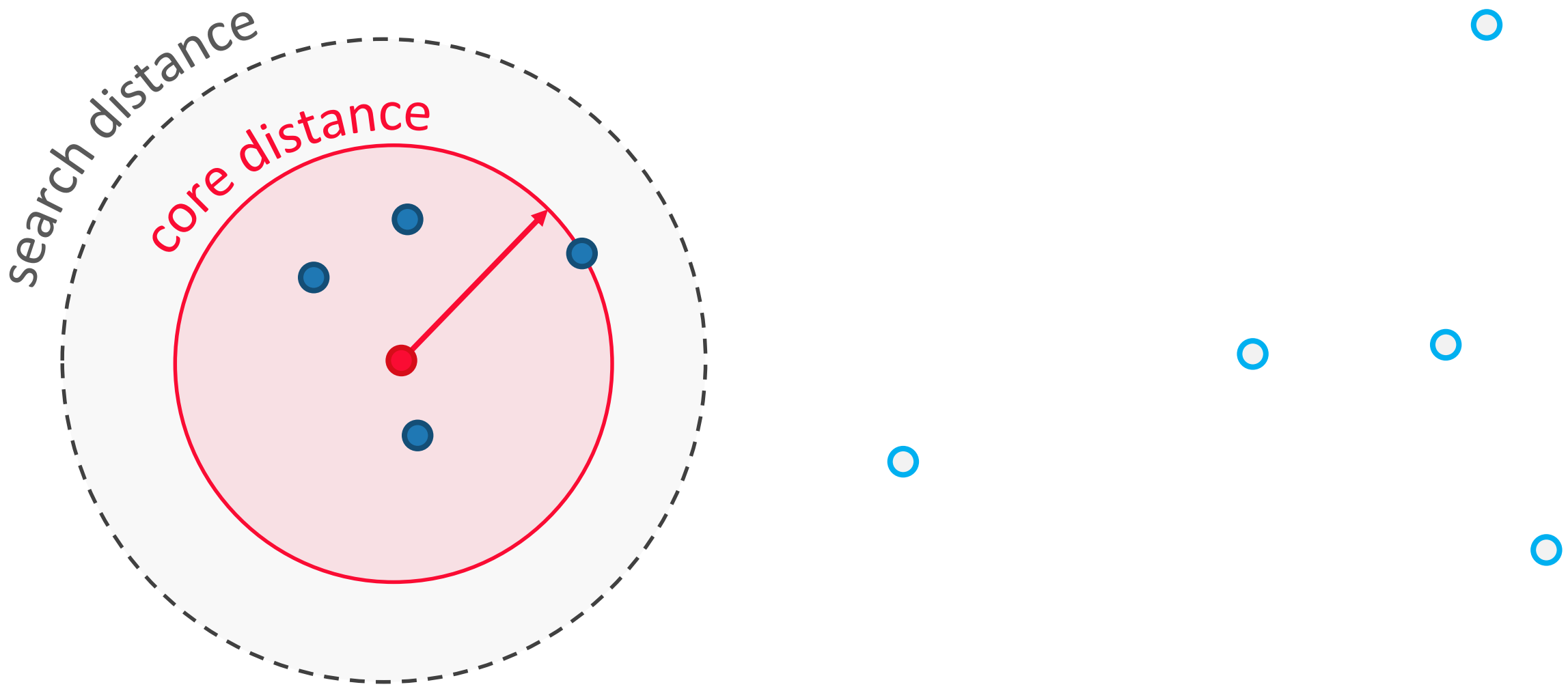
DBSCAN – defined distance



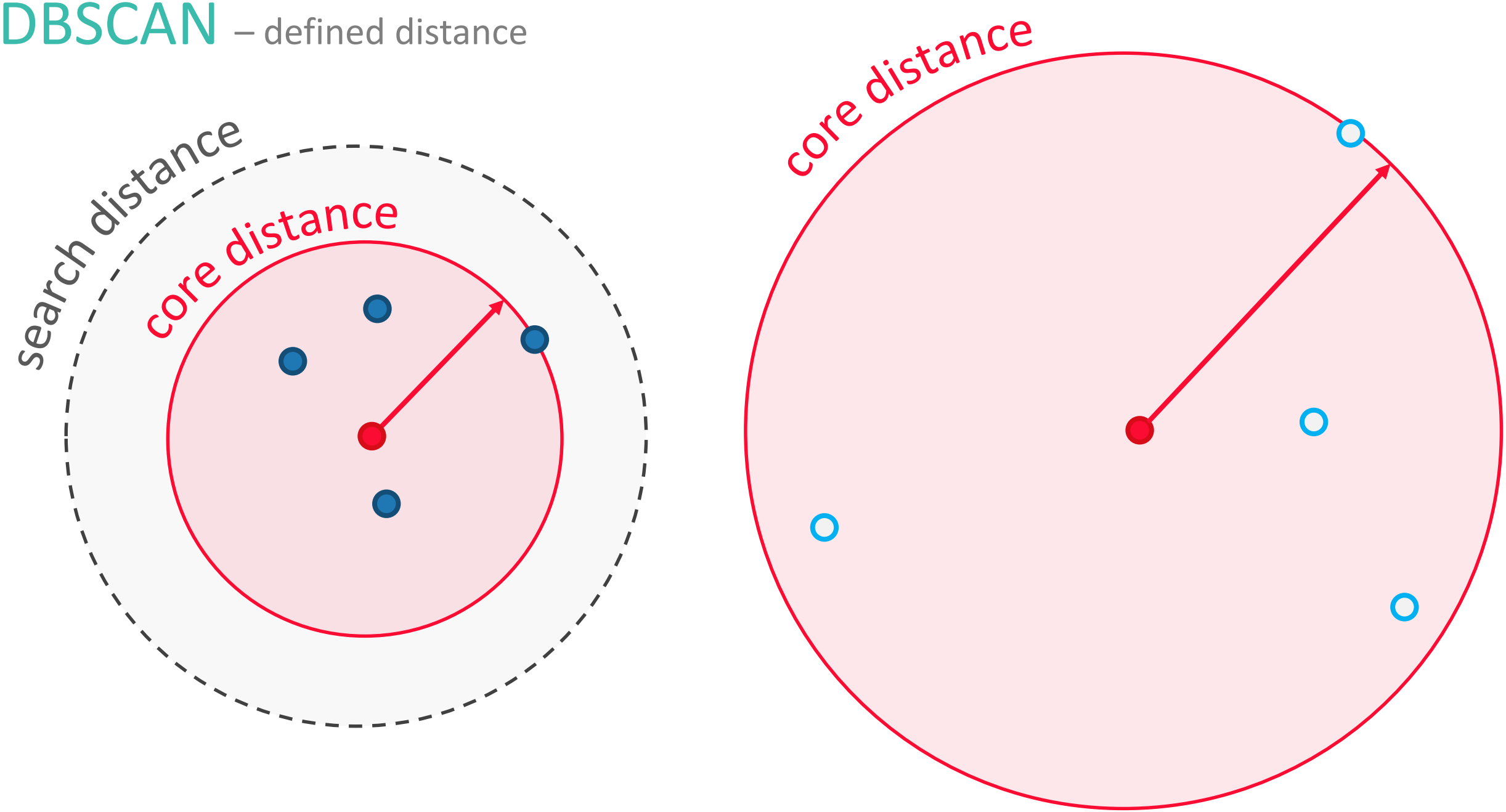
DBSCAN – defined distance



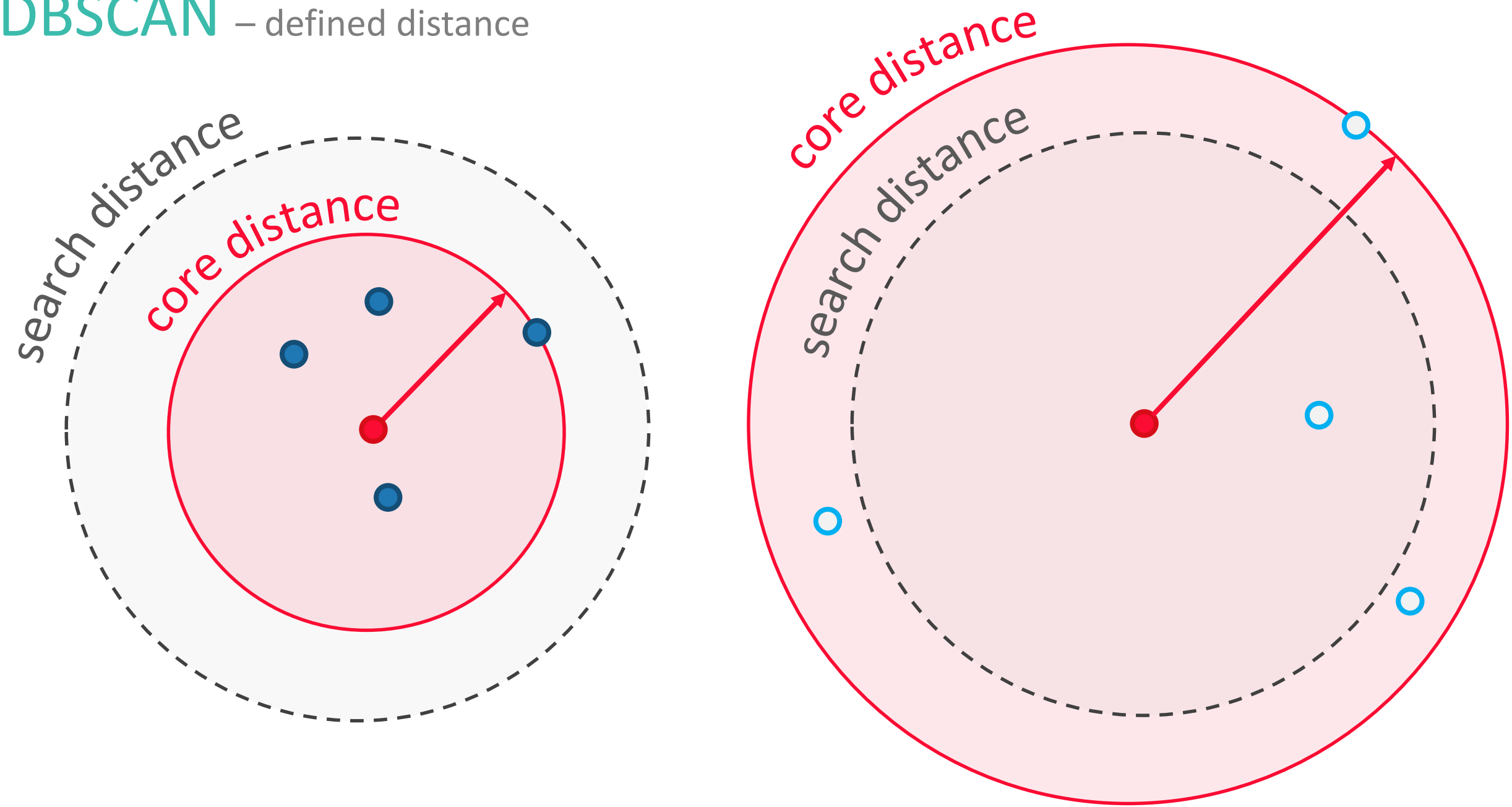
DBSCAN – defined distance



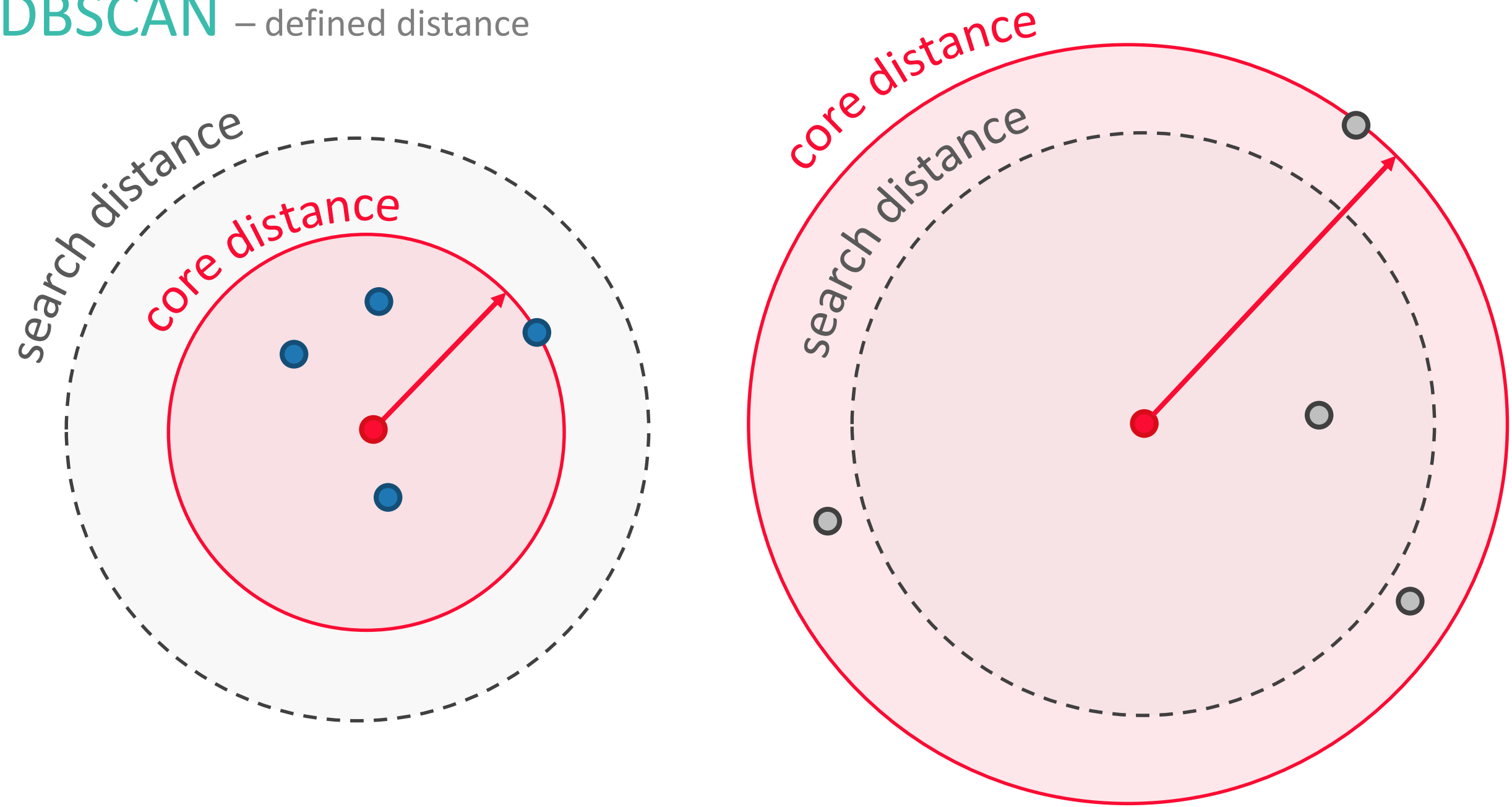
DBSCAN – defined distance



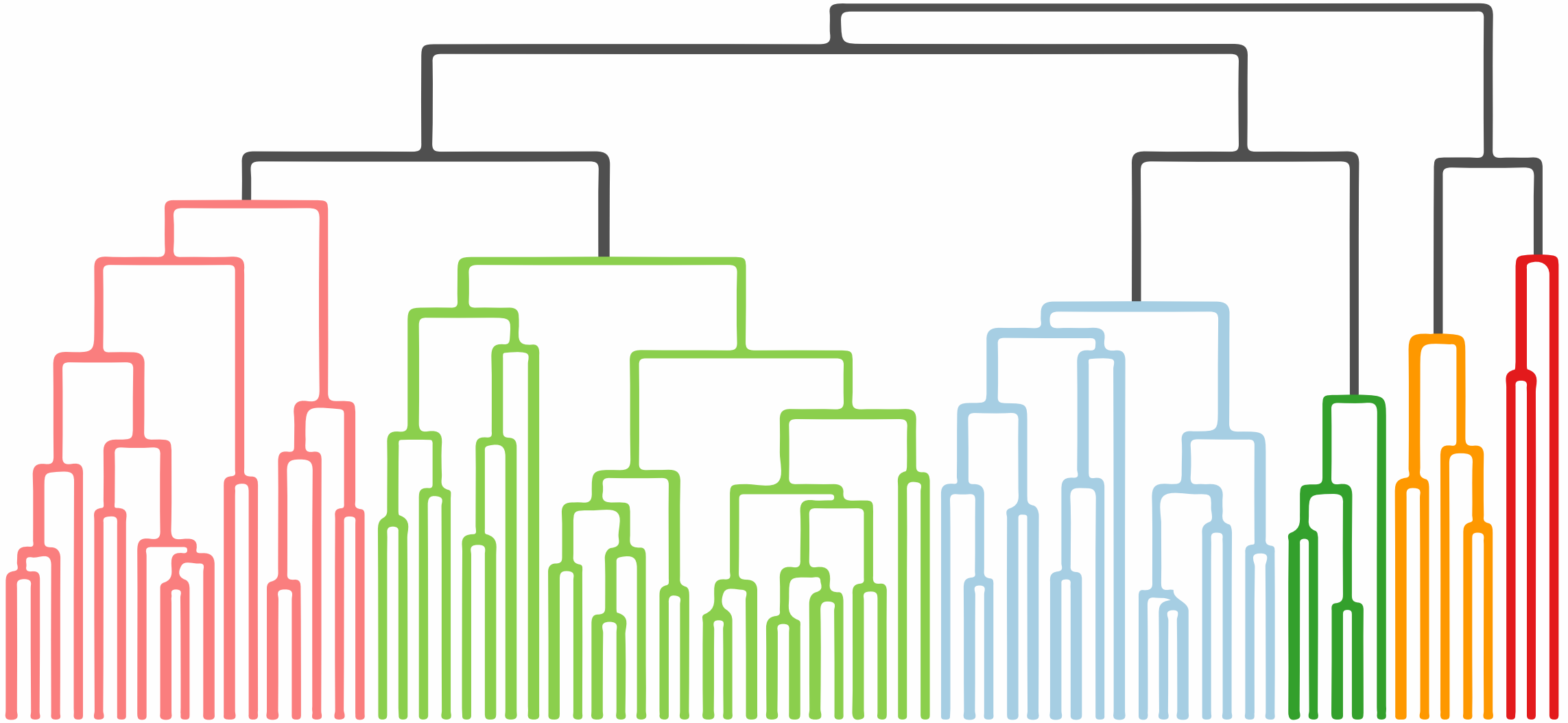
DBSCAN – defined distance



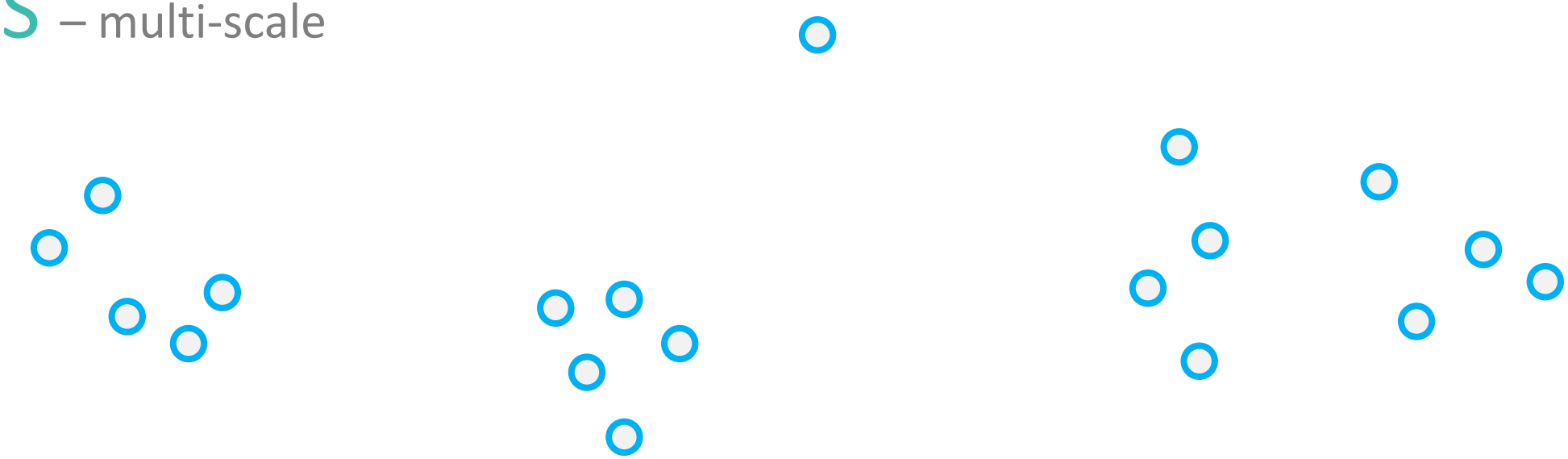
DBSCAN – defined distance



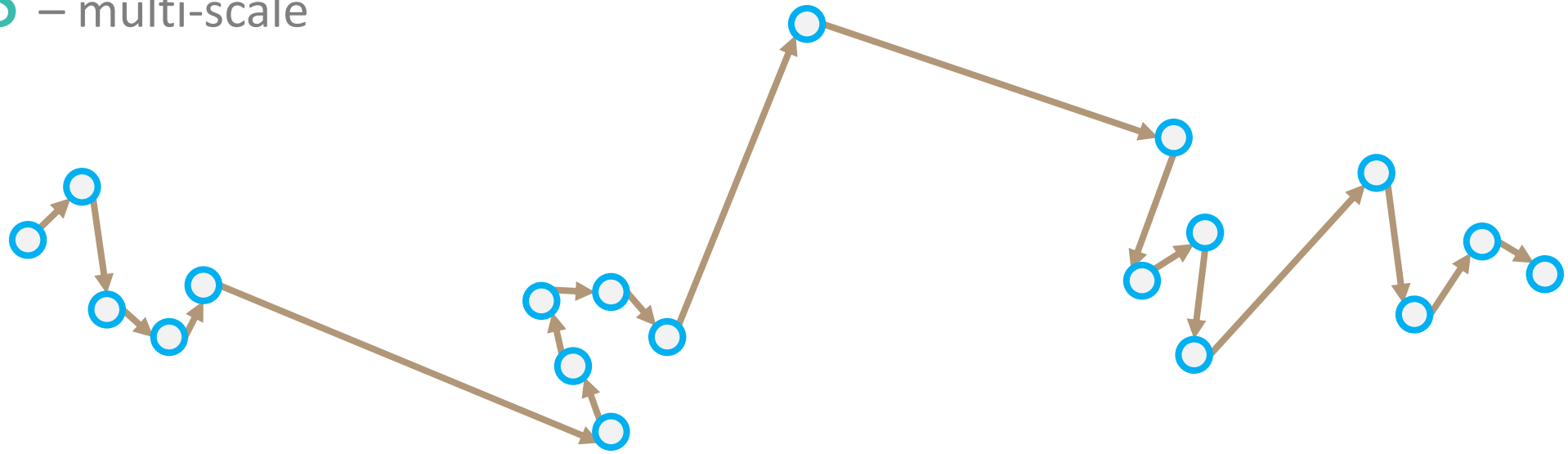
HDBSCAN – self adjusting



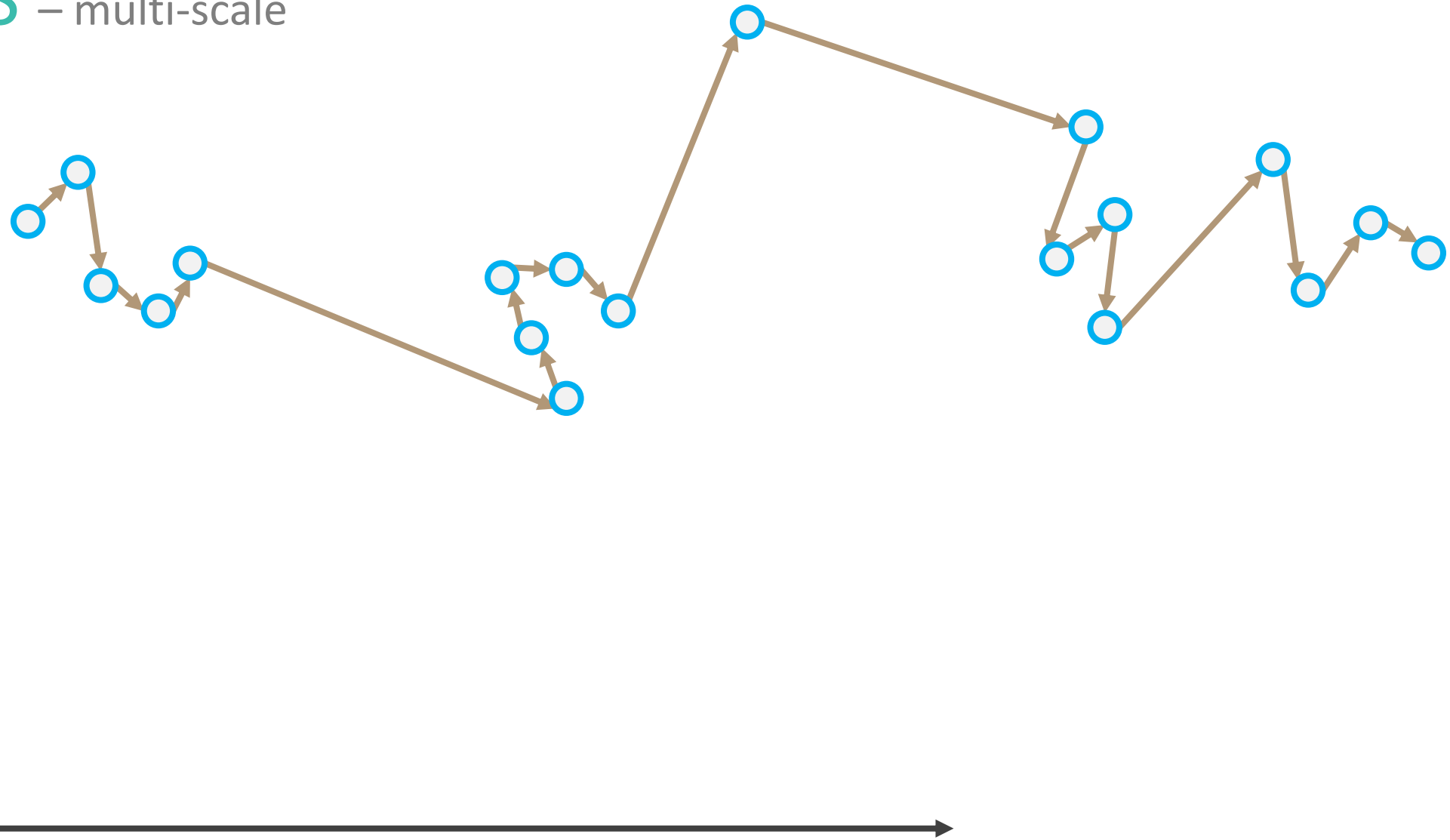
OPTICS – multi-scale



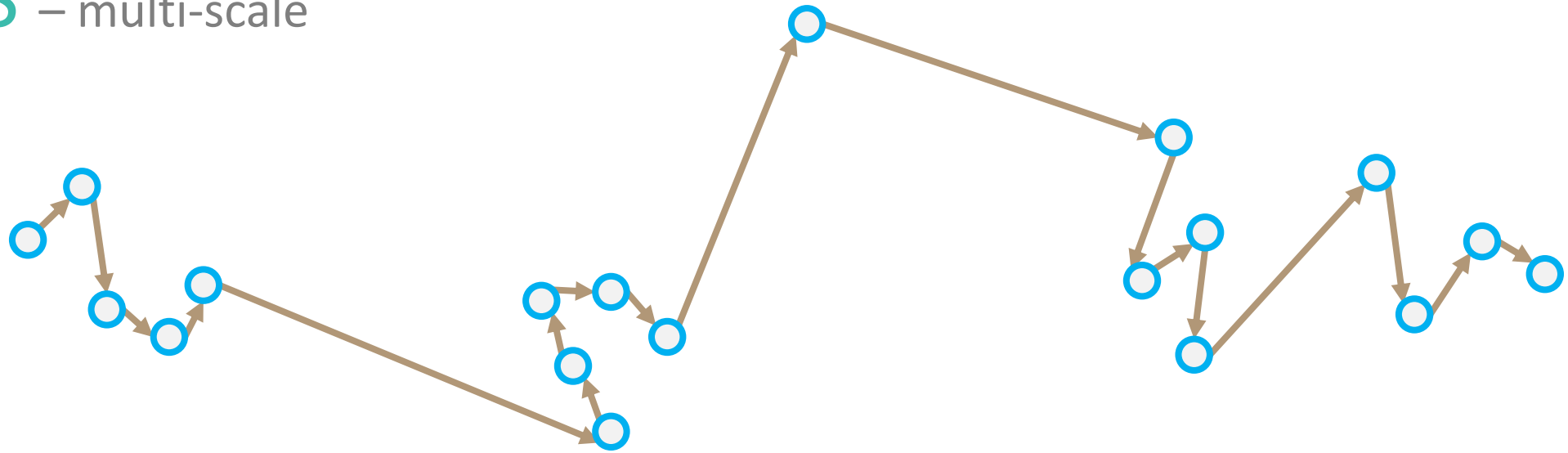
OPTICS – multi-scale

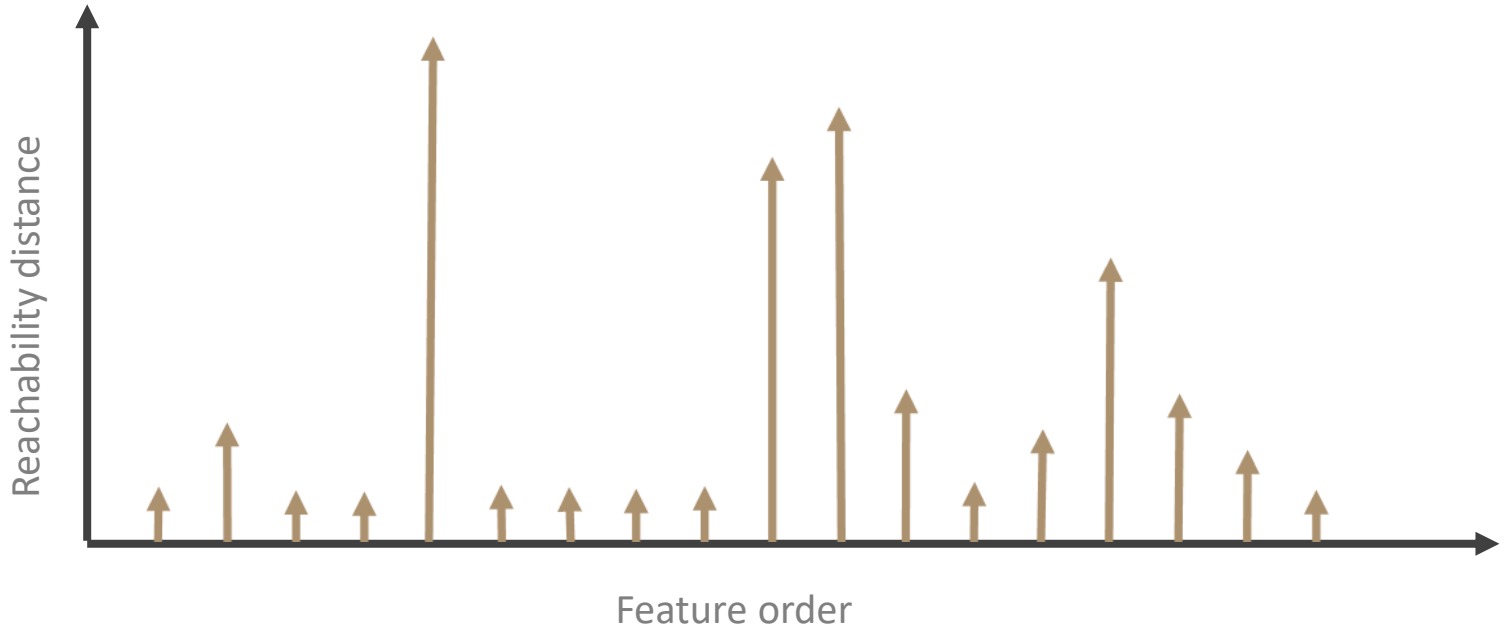


OPTICS – multi-scale

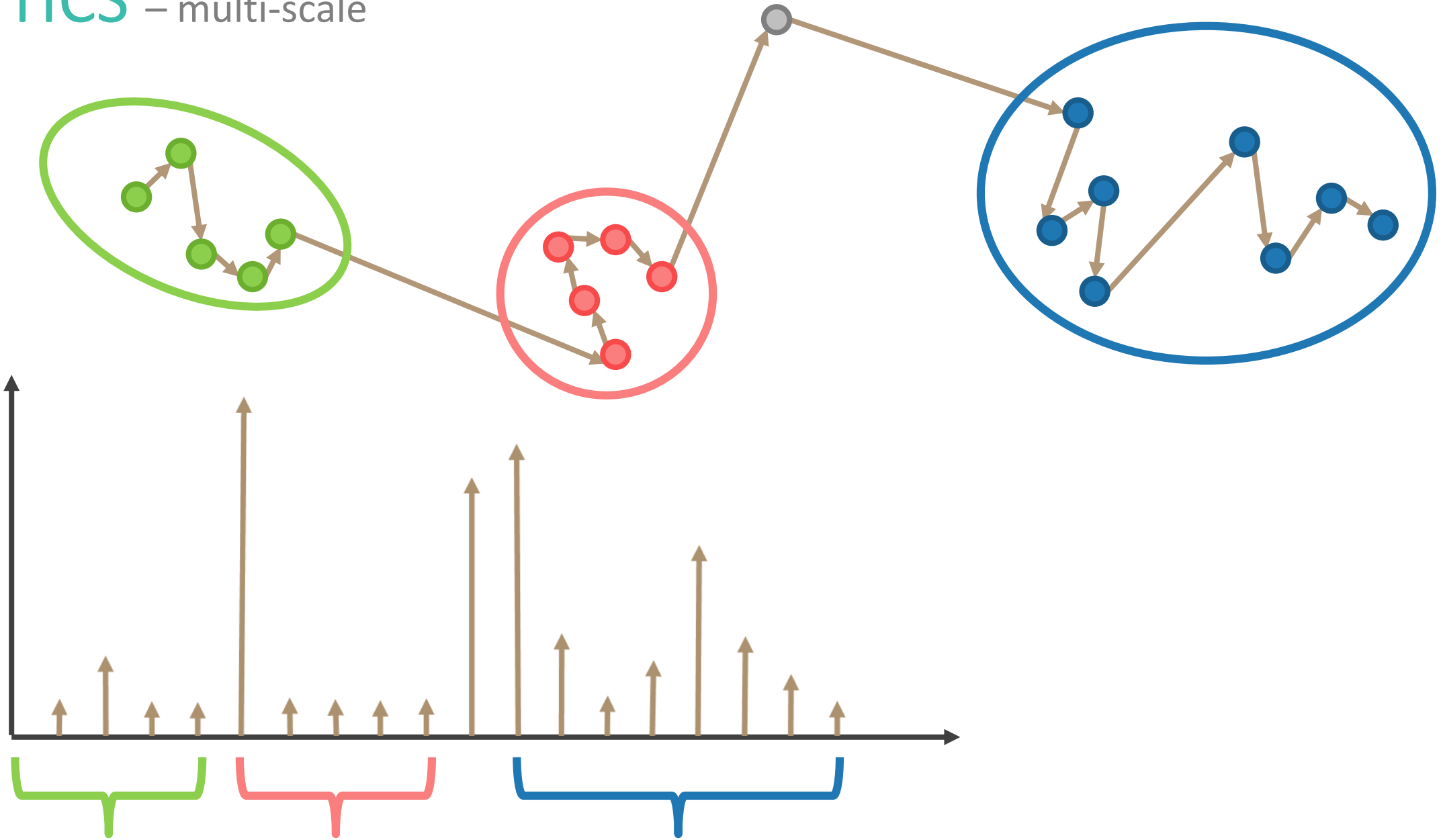


OPTICS – multi-scale

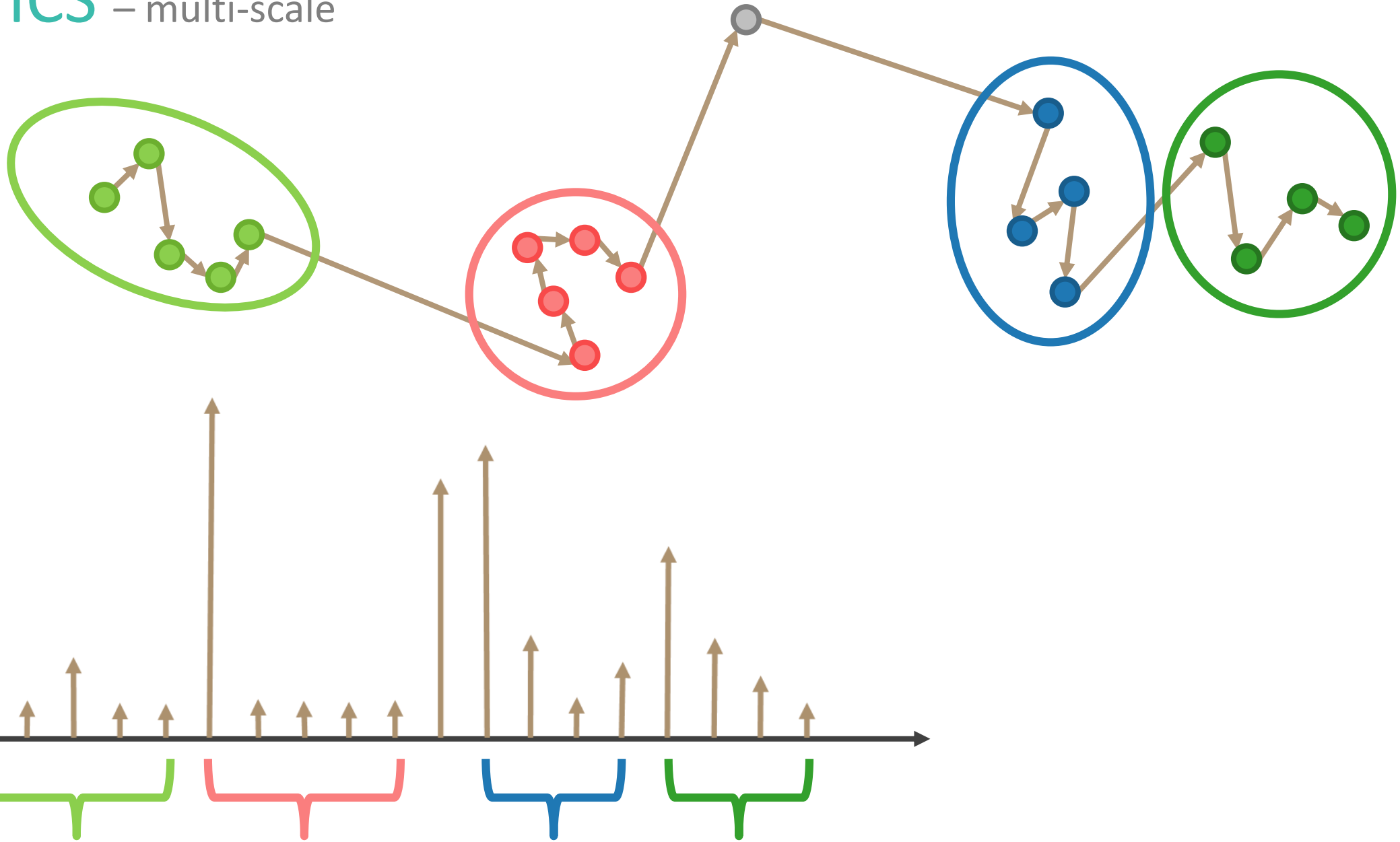




OPTICS – multi-scale



OPTICS – multi-scale



DBSCAN

- Uses fixed search distance
- Clusters of similar densities
- Fast

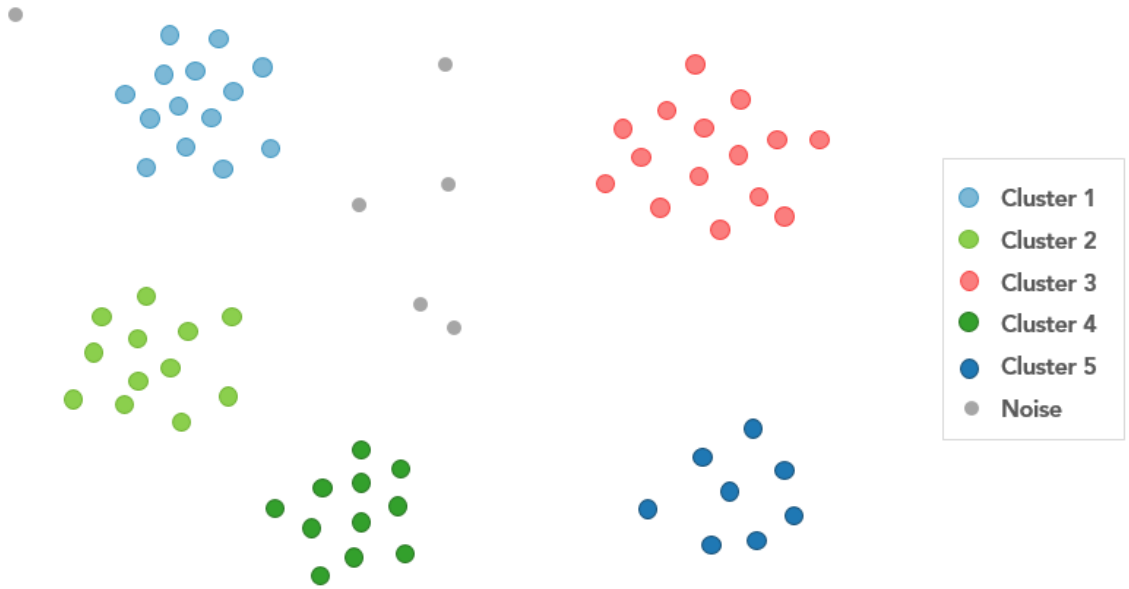
HDBSCAN

- Uses range of search distances to find clusters of varying densities
- Data driven, requires least user input

OPTICS

- Uses neighbor distances to create reachability plot
- Most flexibility for fine tuning
- Can be computationally intensive

Demo



Forest-based Classification & Regression

9945

Predicting using machine learning

Training

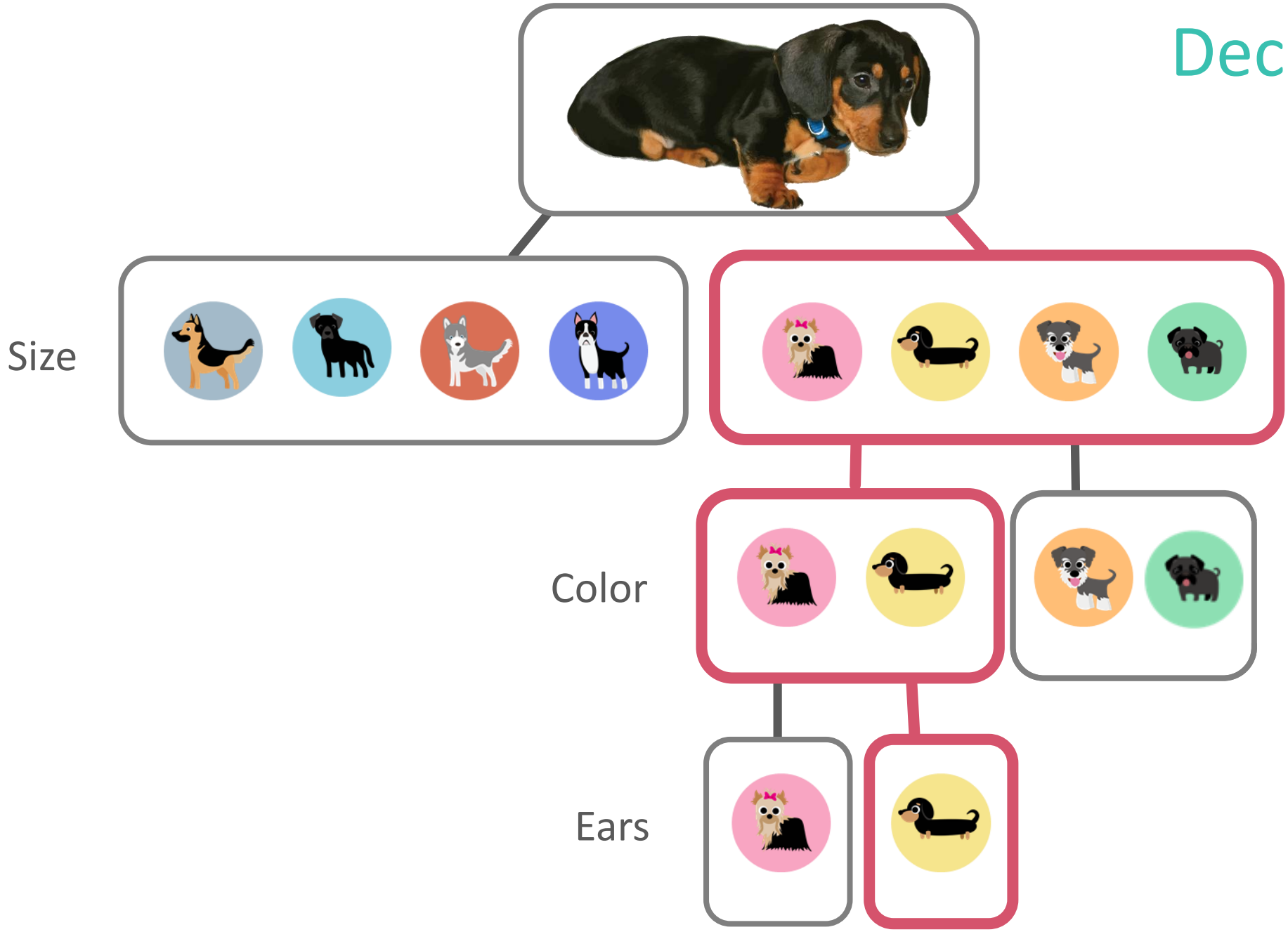
variable to predict

Breed

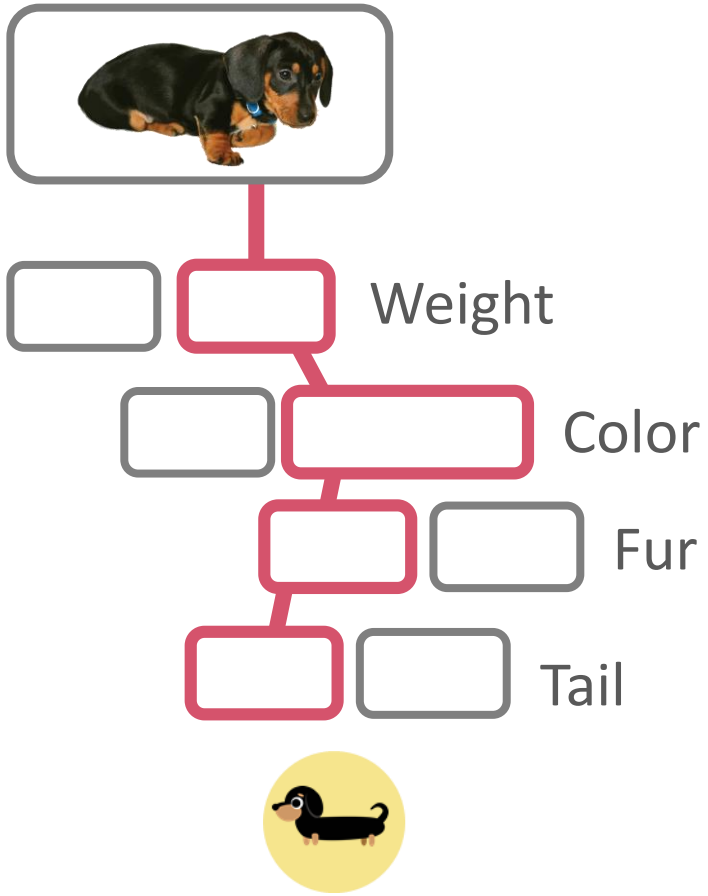
Size
Color
Fur
Ears
Tail
Age
Weight

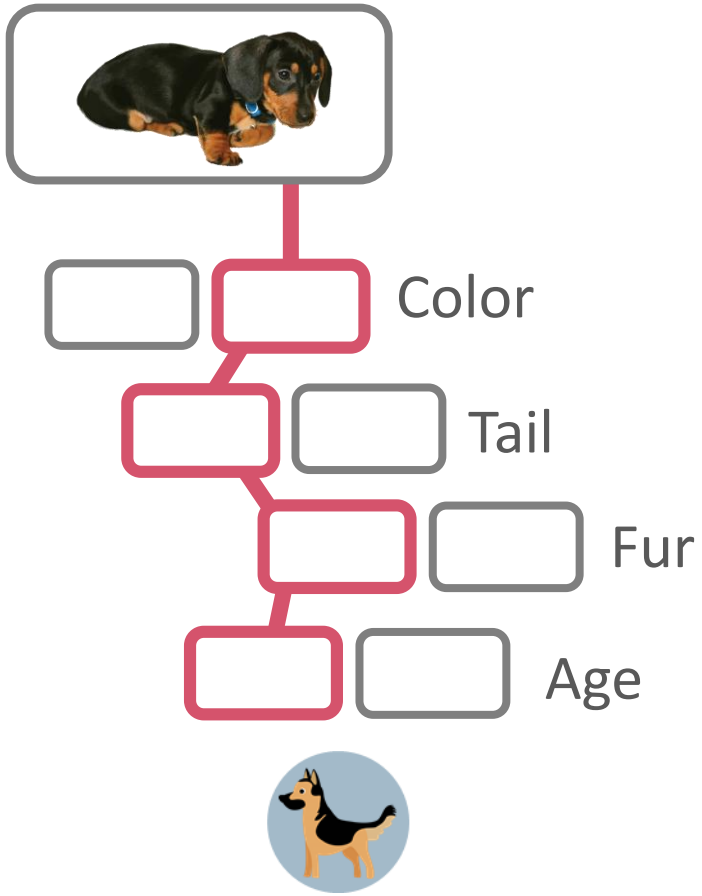
explanatory variables

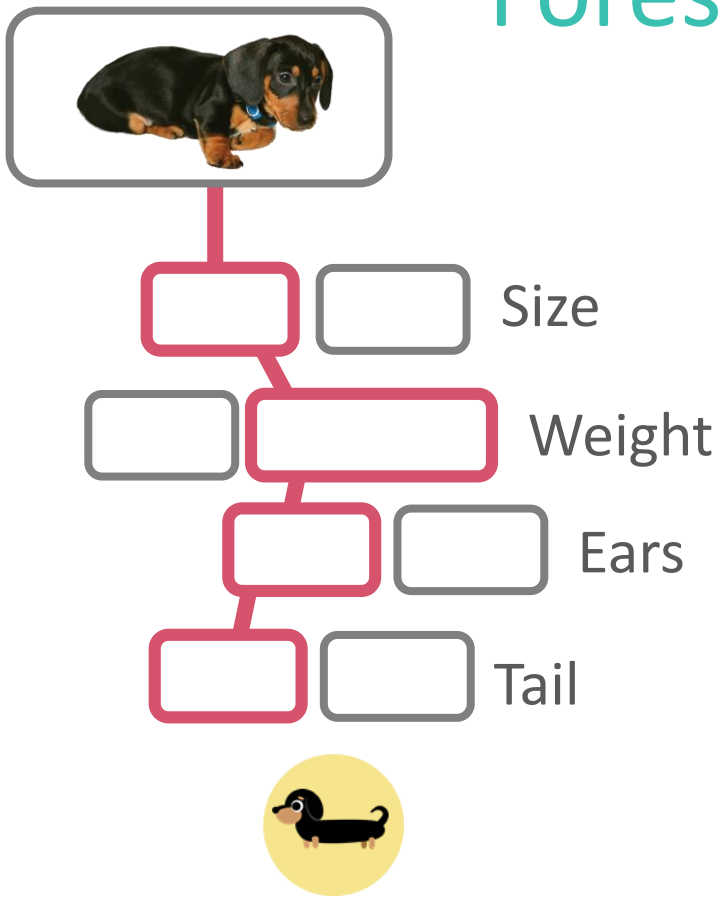
Decision Tree



Forest

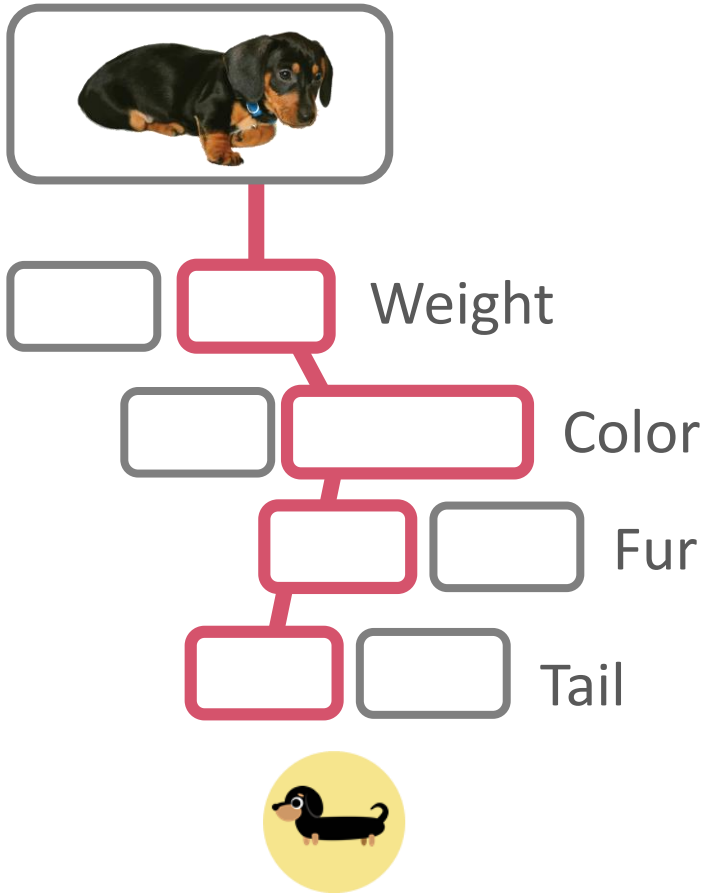


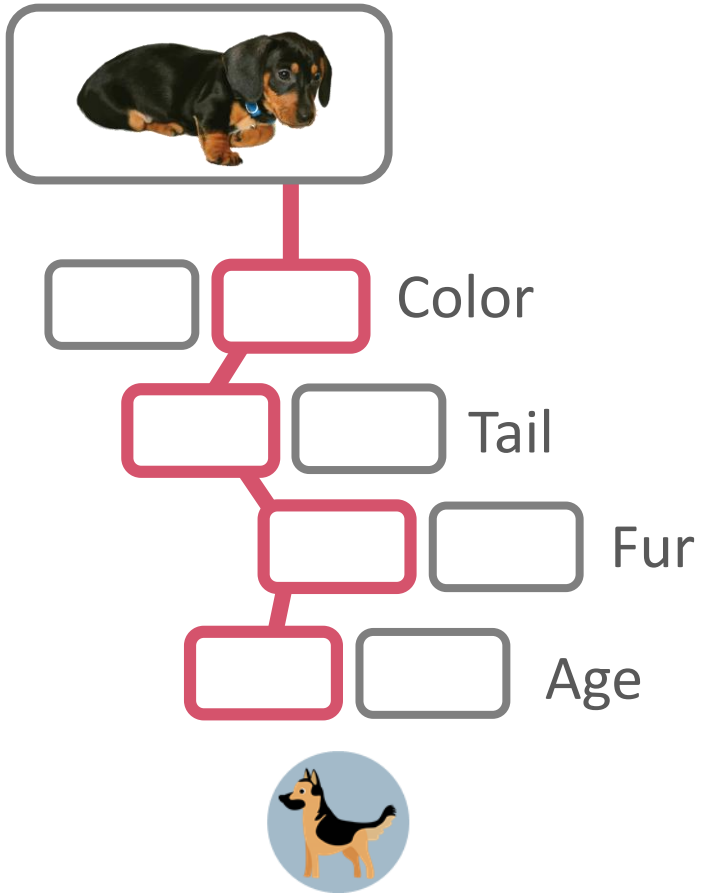


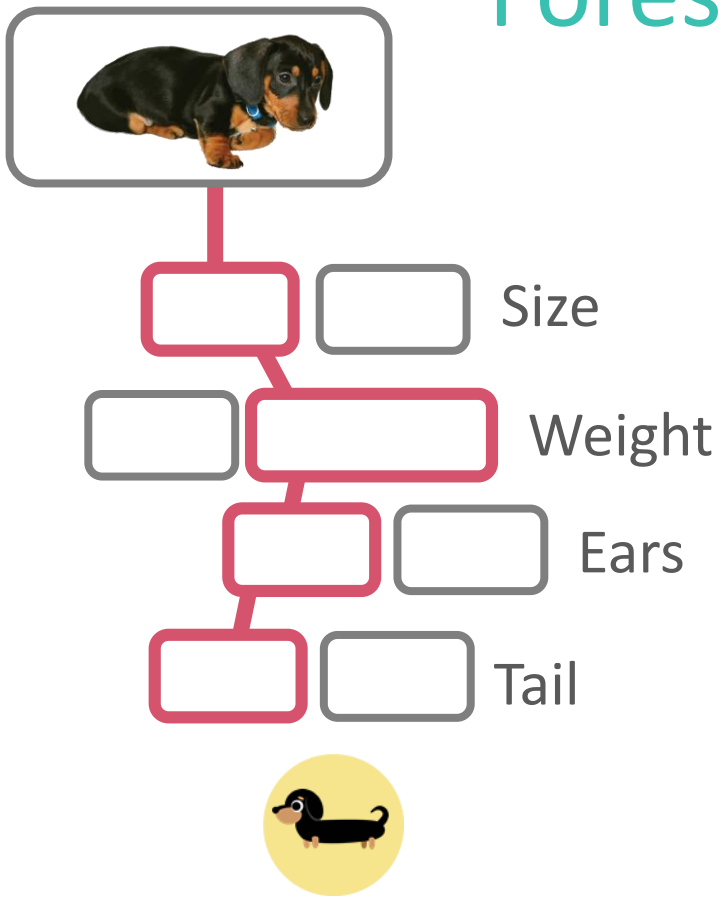


Random subset of data and variables used in each tree

Forest

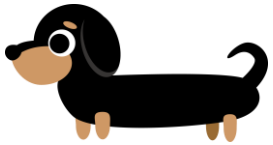






Majority vote wins

=

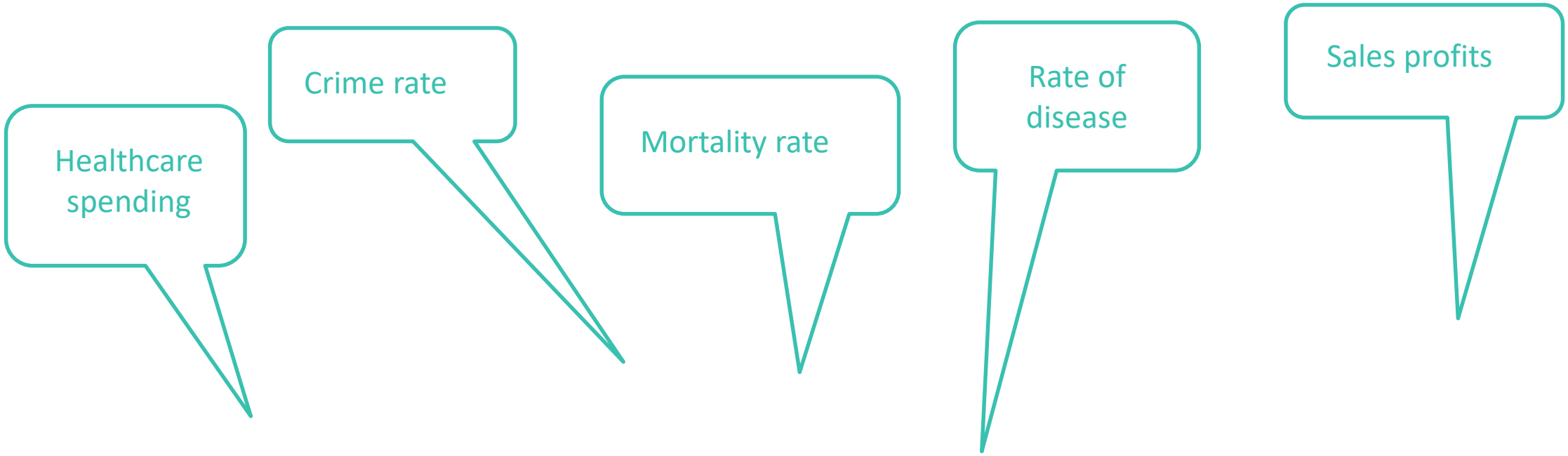


Classification

Predict categorical variable

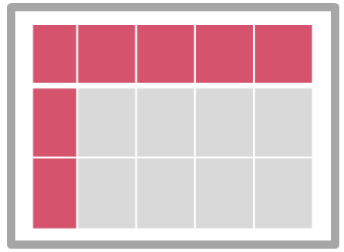
Regression

Predict continuous variable

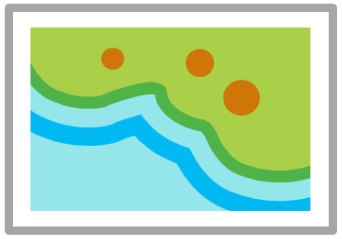


Explanatory Variables

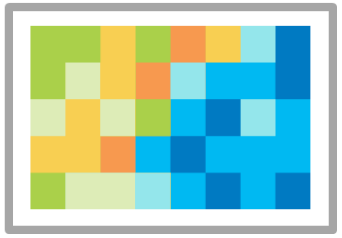
Attributes



Distance features

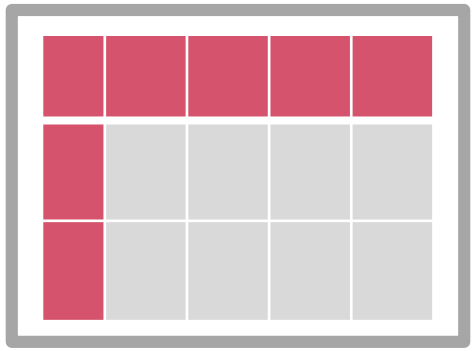


Rasters



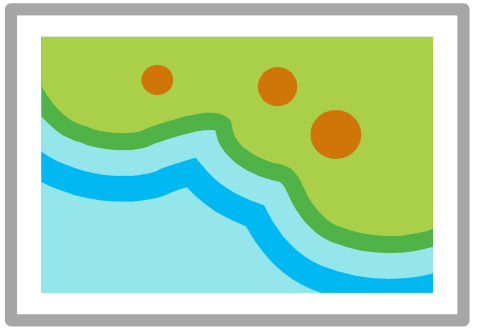
Explanatory Training Variables

Other attributes in the layer containing the Variable to Predict



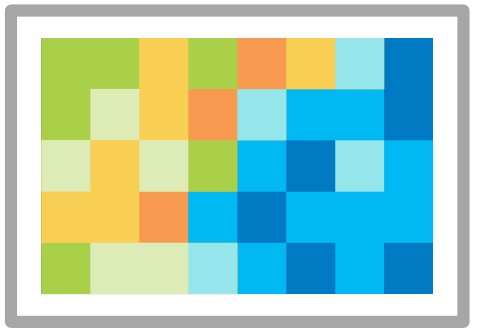
Explanatory Training Distance Features

Features from which distances will be calculated



Explanatory Training Rasters

Rasters from which values will be extracted

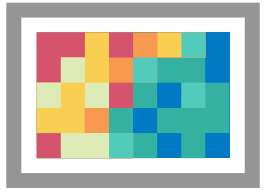


Prediction Type

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

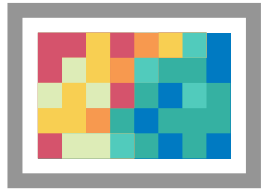
Create a prediction feature class

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

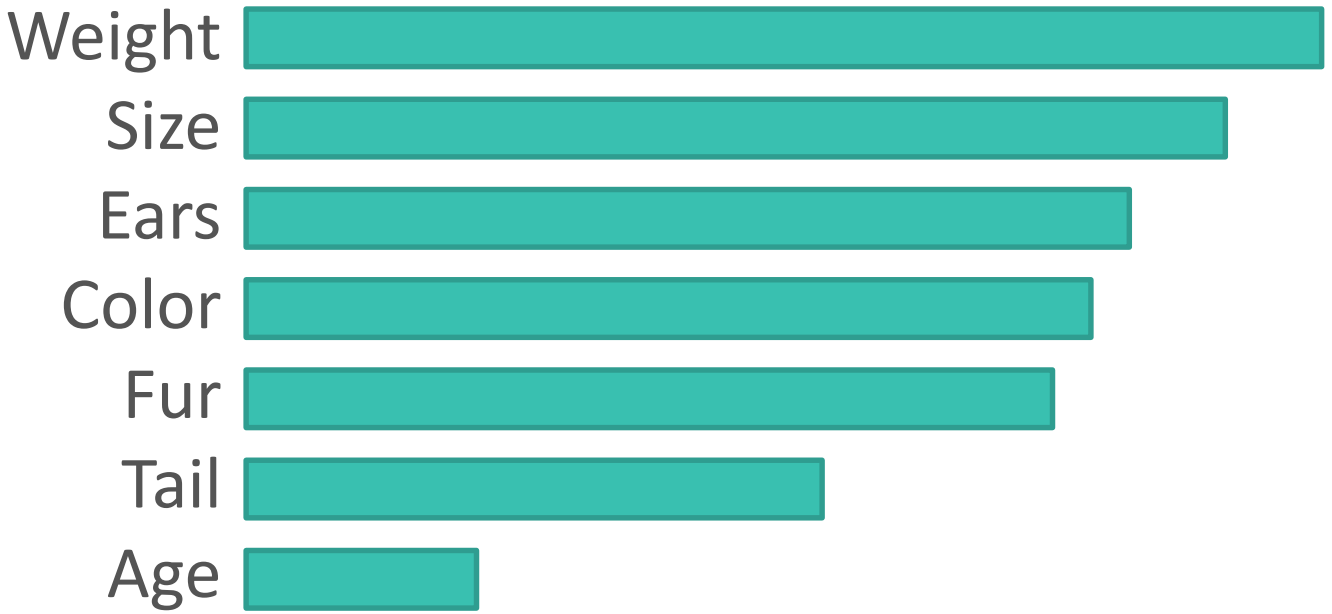
Predict values in a different study area

Predict values in a different time period

Evaluate model performance

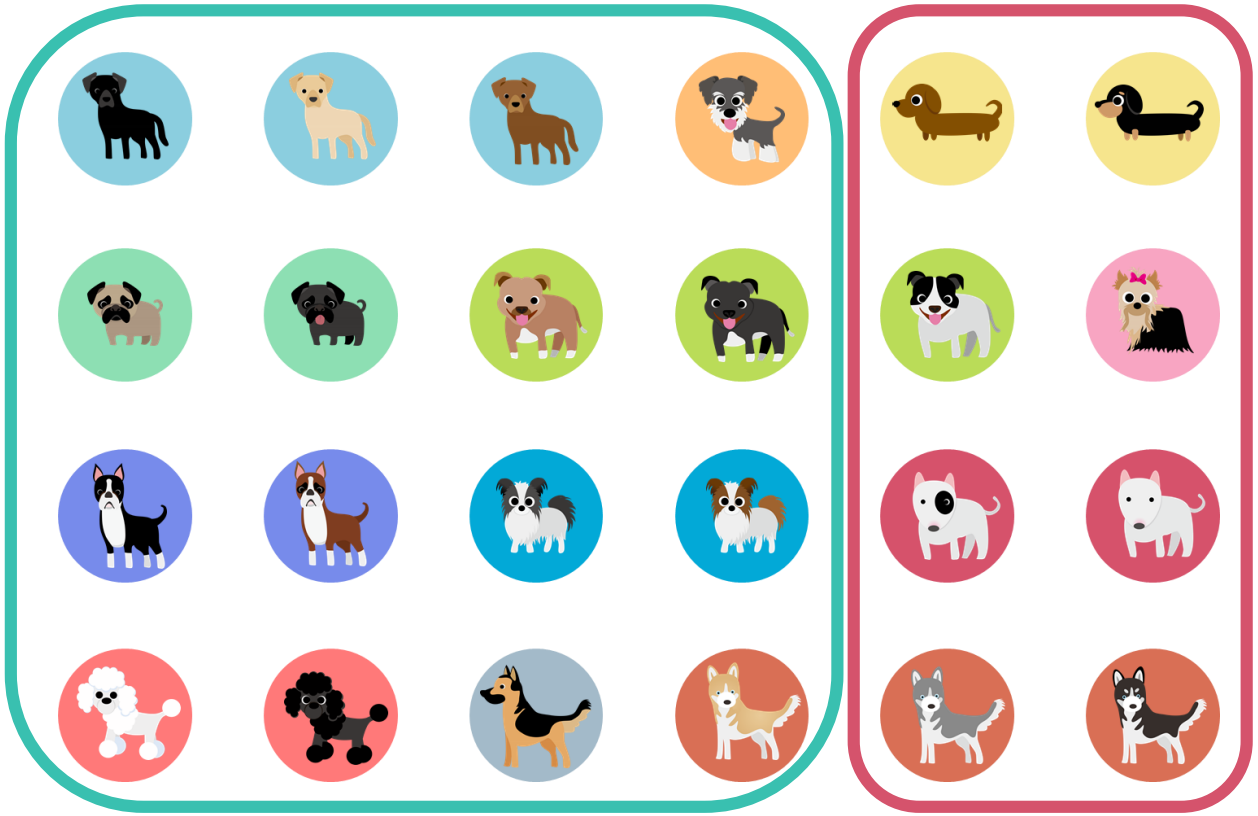
Variable importance

How well does each variable do in splitting the trees?



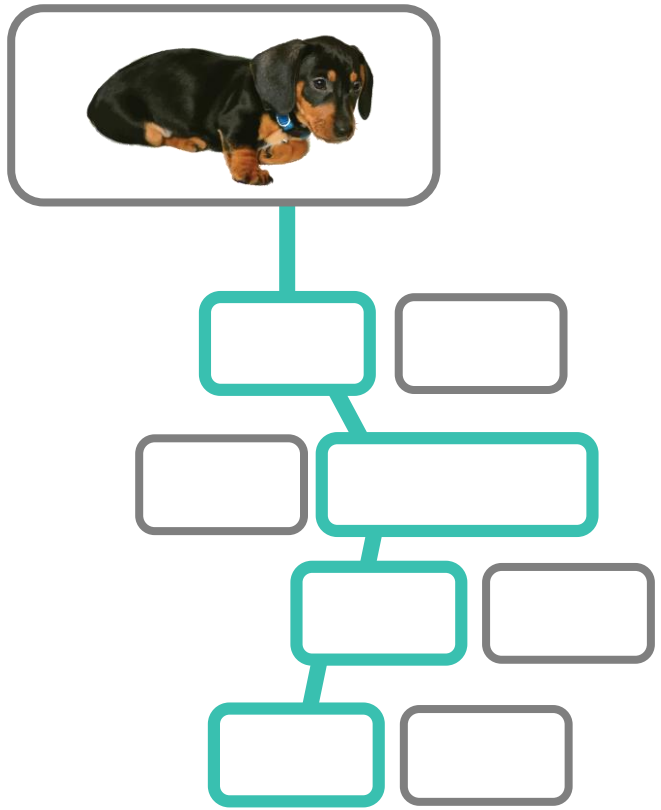
Out Of Bag errors

How well can each tree predict the excluded features?



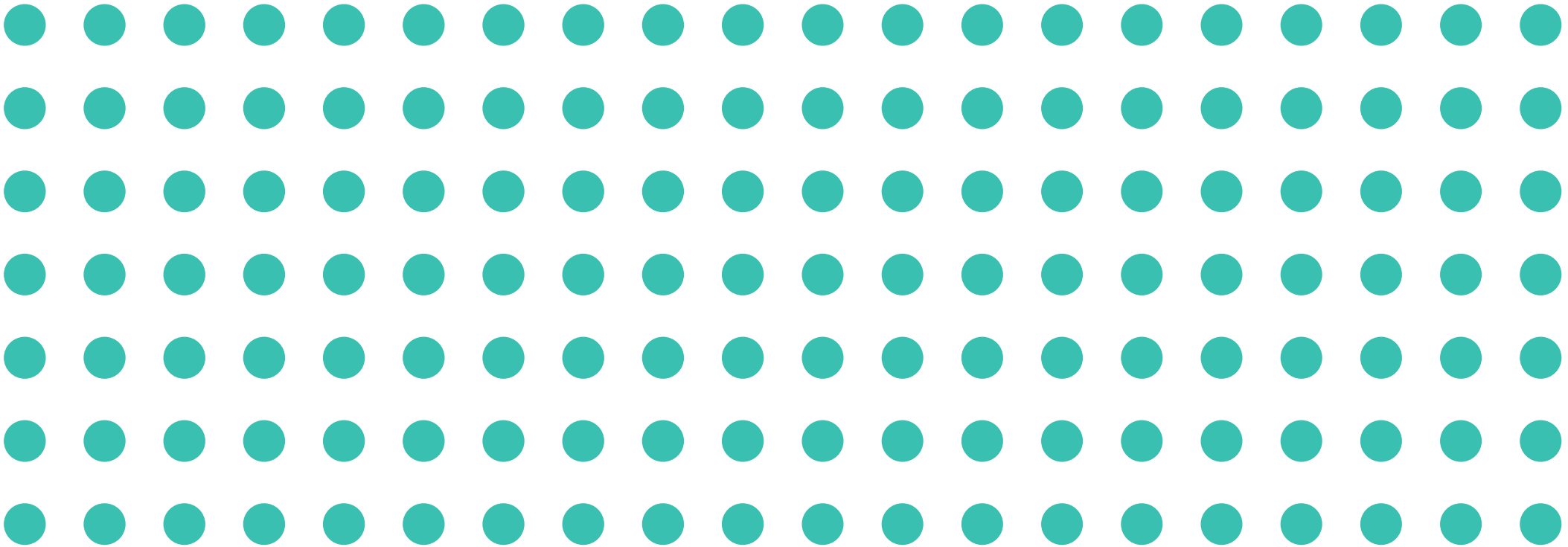
2/3 included (randomly)

1/3 excluded

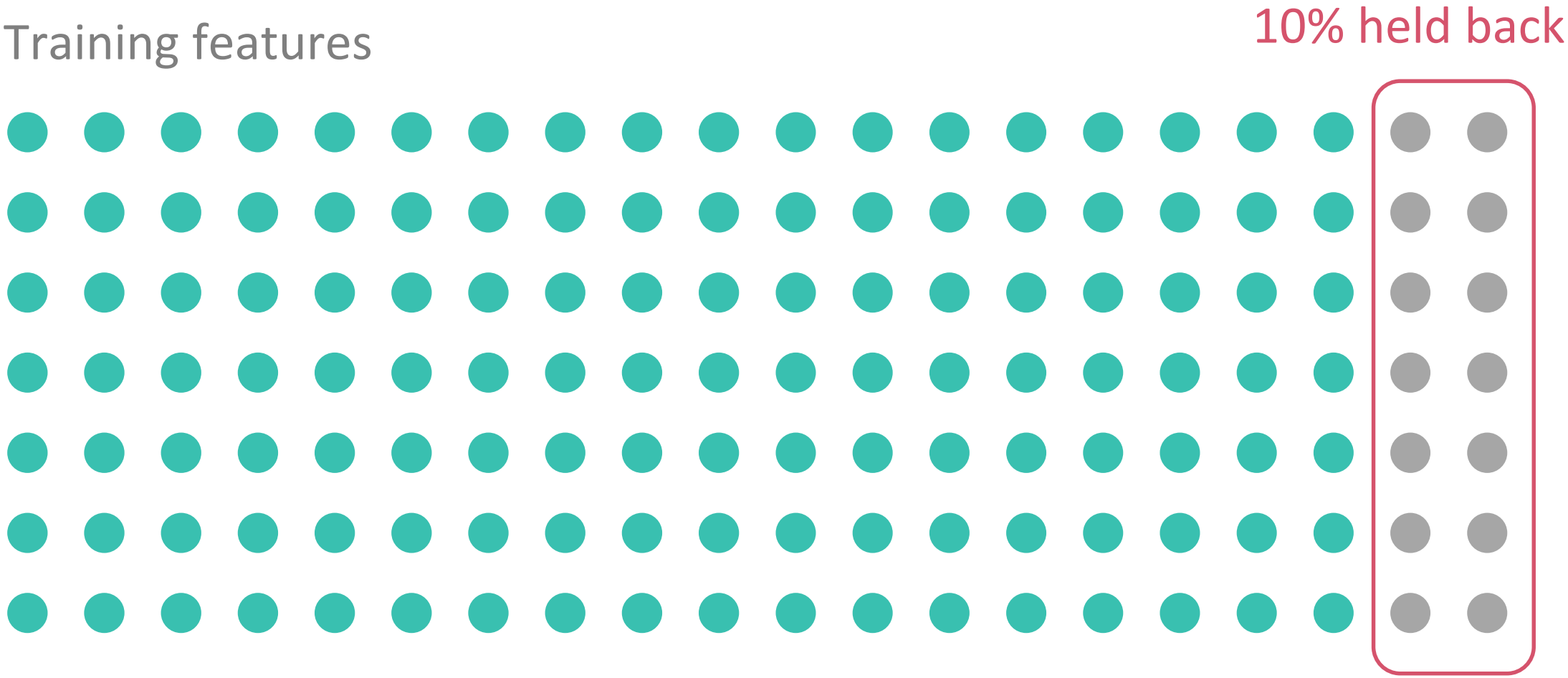


Model Validation

Training features

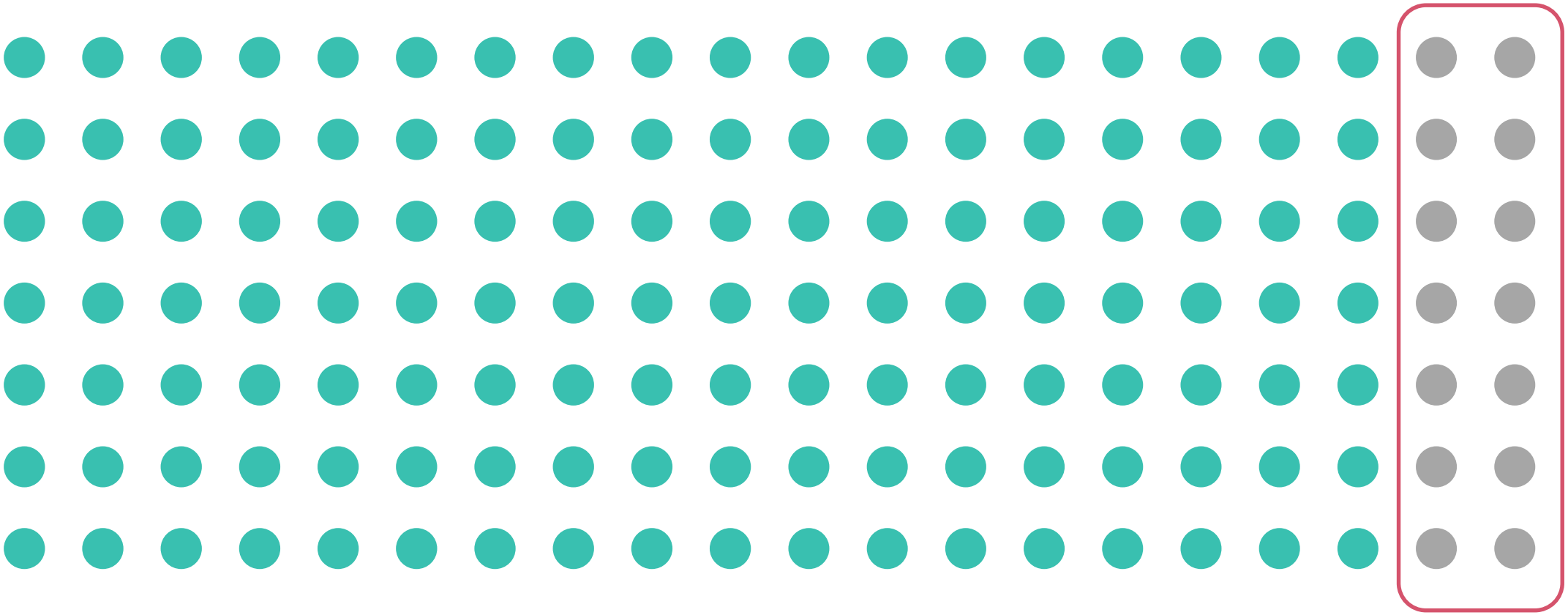


Model Validation

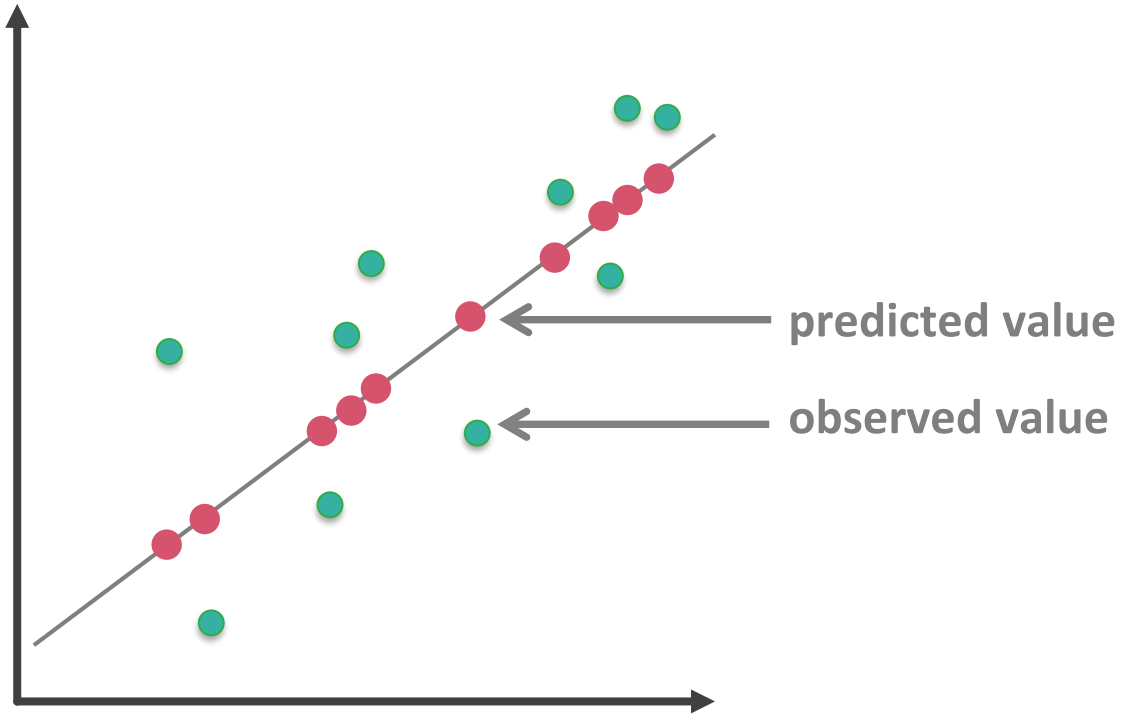


Model Validation

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



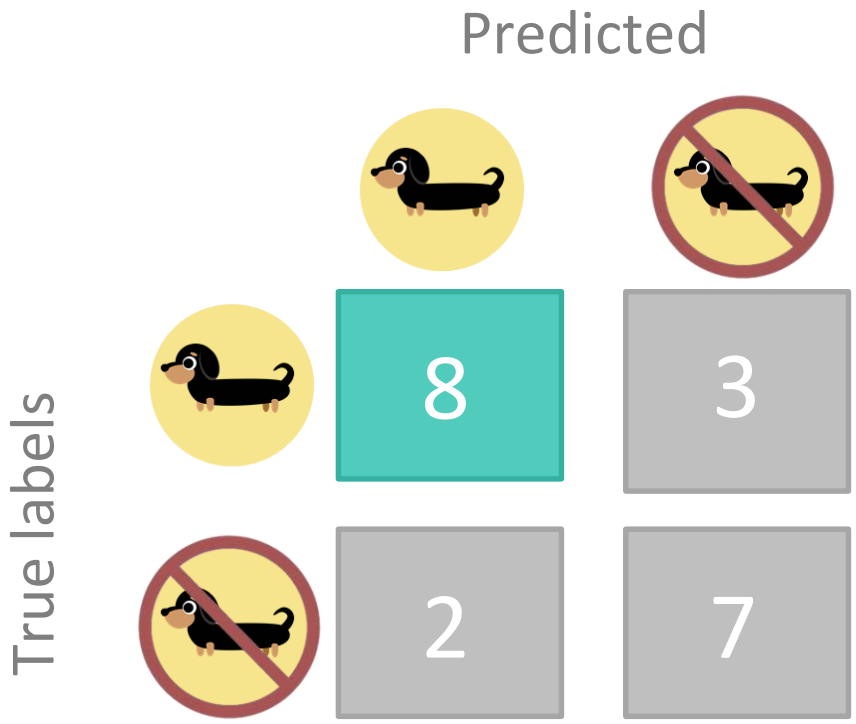
R-squared



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