Building Regularized Regression Models

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Choosing regression to solve problems

Overfitting and the bias-variance trade-off

Regularization to mitigate overfitting

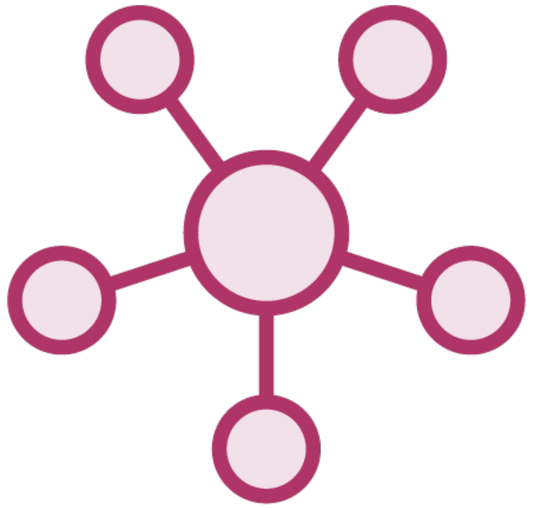
Building and training Ridge, Lasso and ElasticNet regression model

Choosing Regression Algorithms

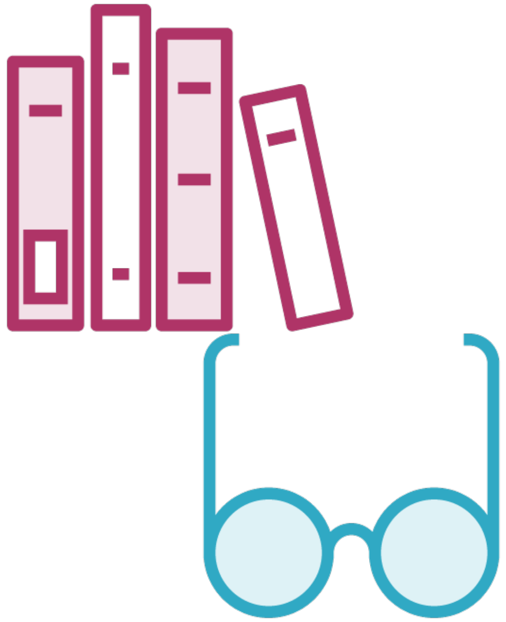
Types of Machine Learning Problems

Classification

Regression



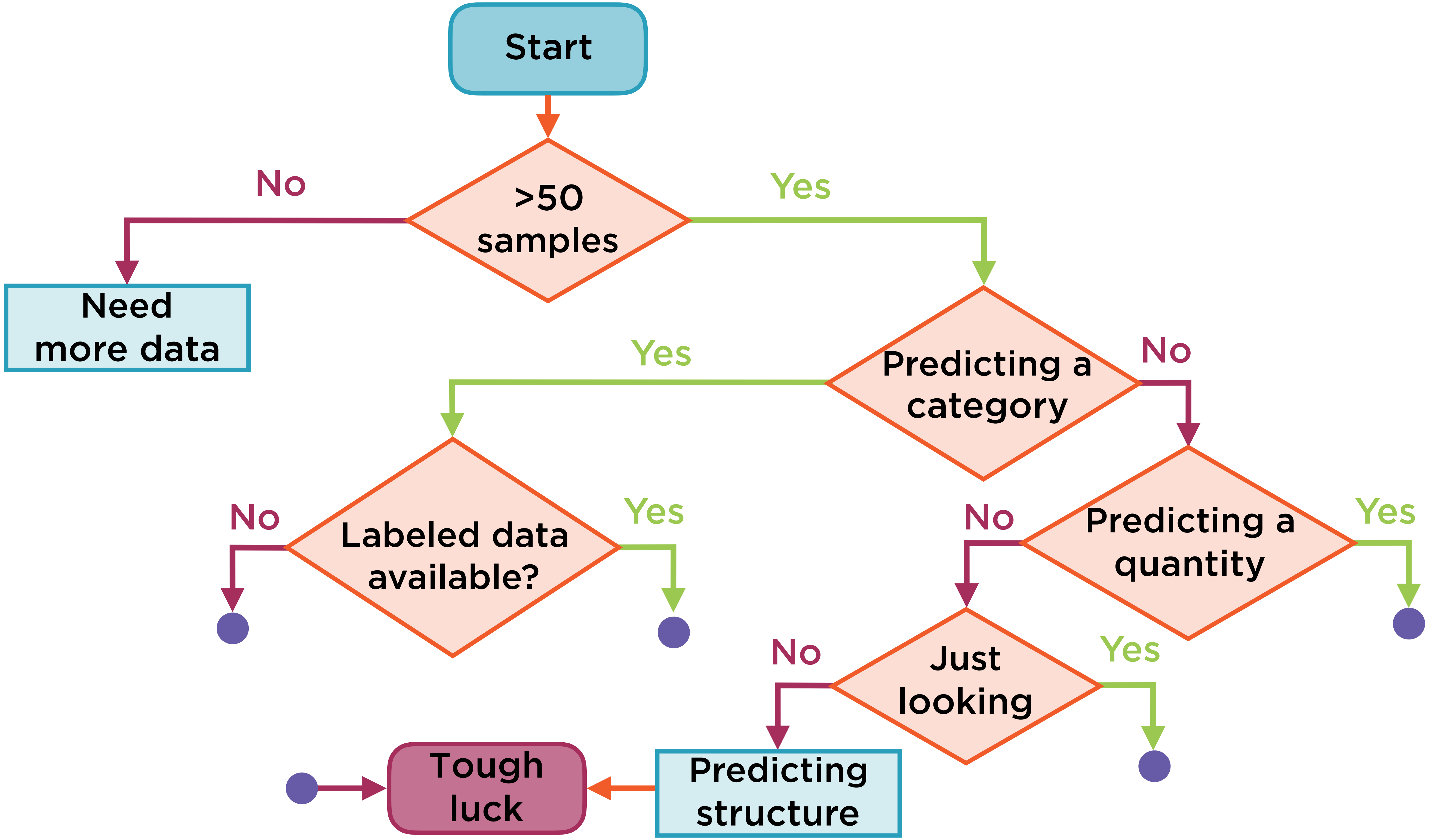
Clustering



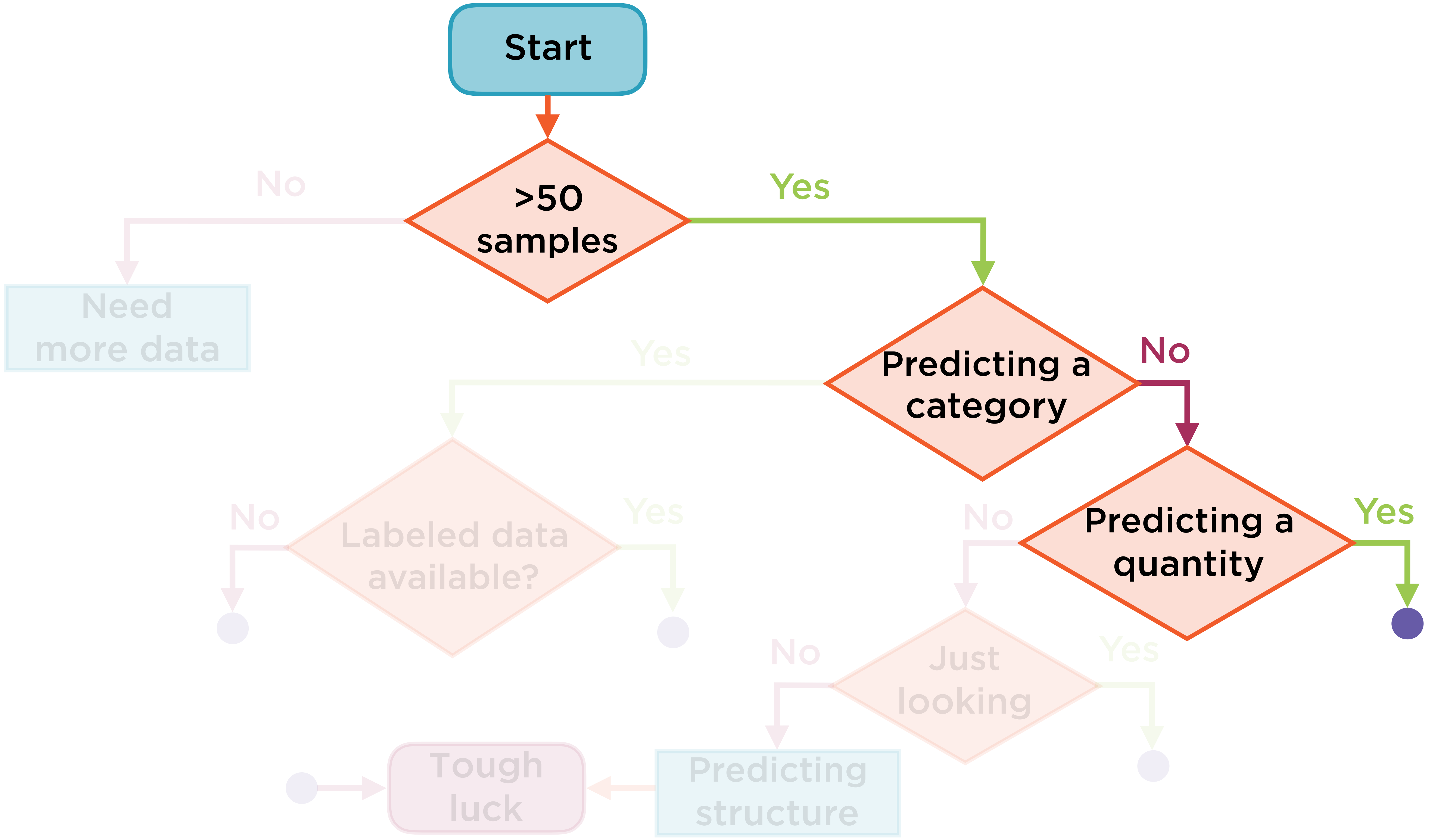
Dimensionality reduction

Focus first on defining the right problem to solve, then on choosing the right estimator to solve it

Choosing the Right Estimator



Choosing the Right Estimator



Types of Machine Learning Problems

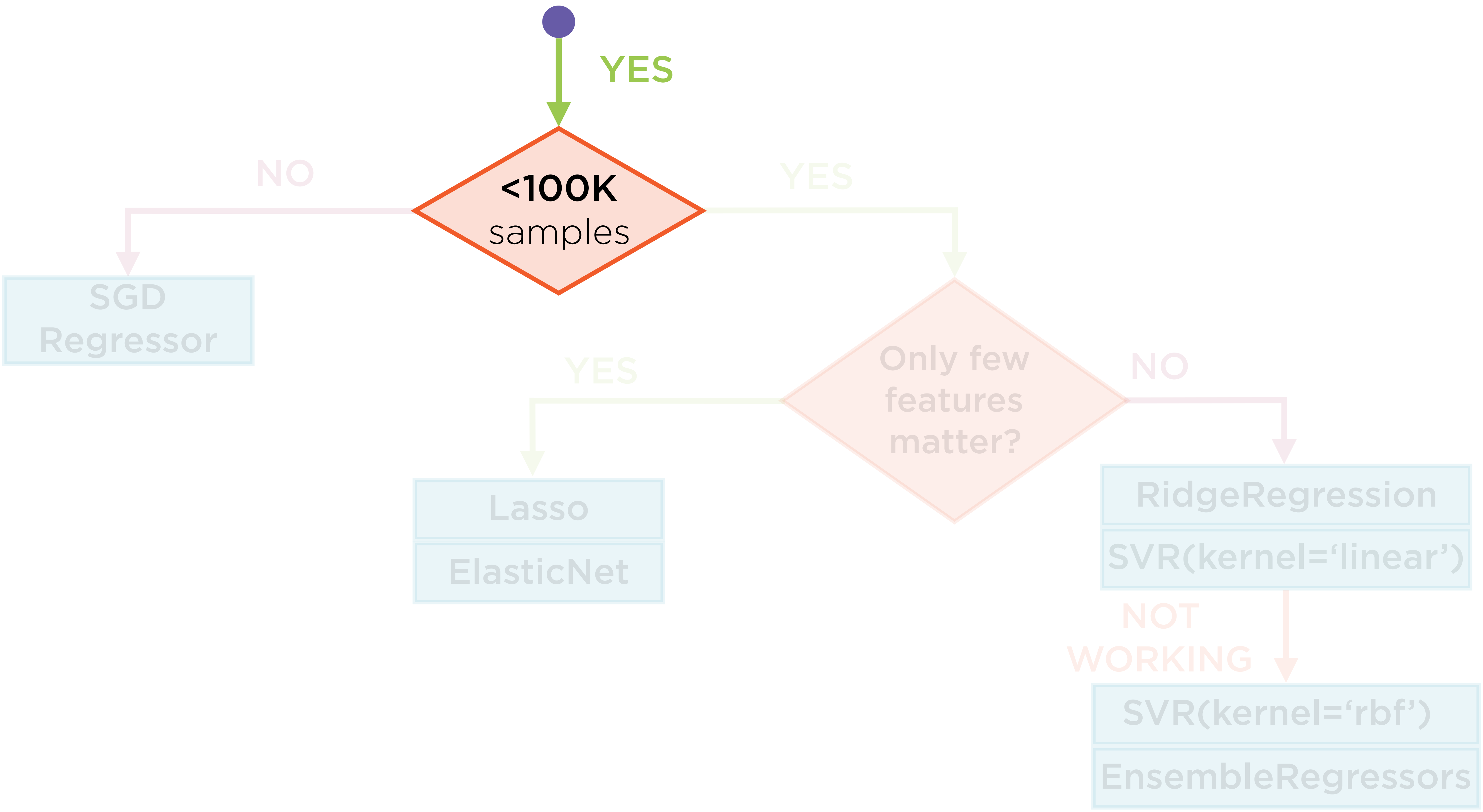
Classification

Regression

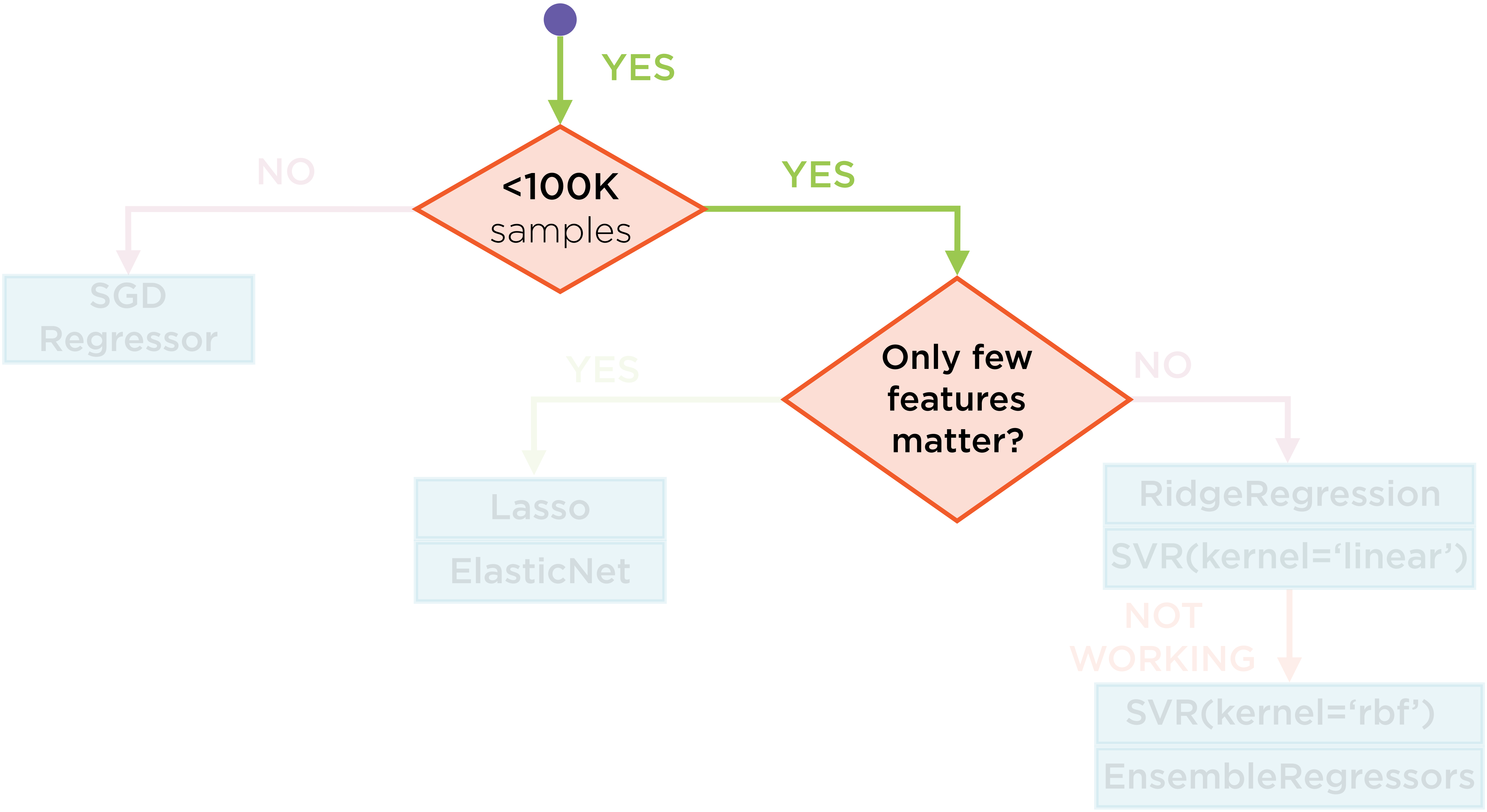
Clustering

Dimensionality reduction

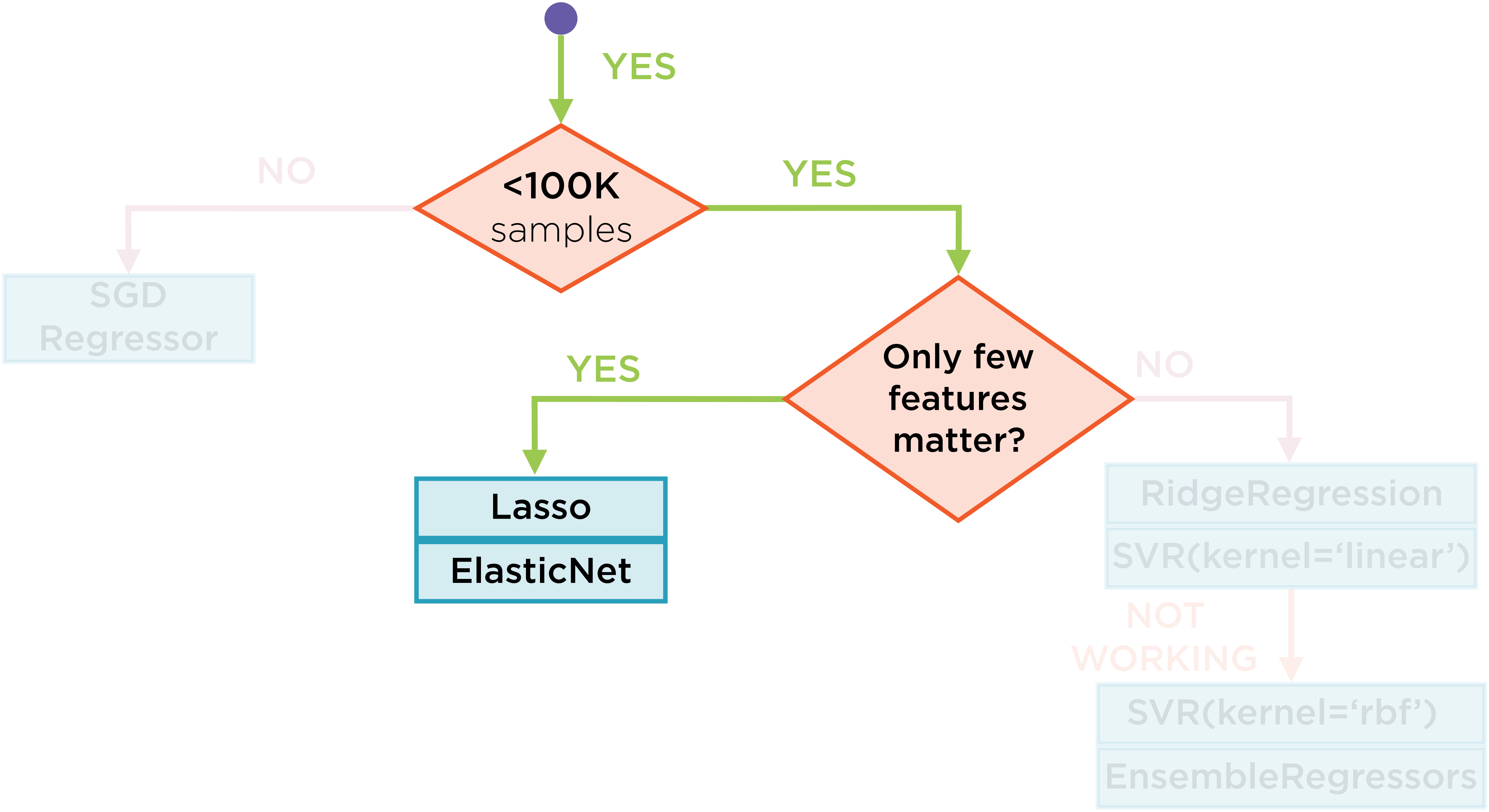
Regression


YES
<100K
samples
NO
SGD
Regressor
YES
Only few
features
matter?
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT
WORKING
SVR(kernel='rbf')
EnsembleRegressors

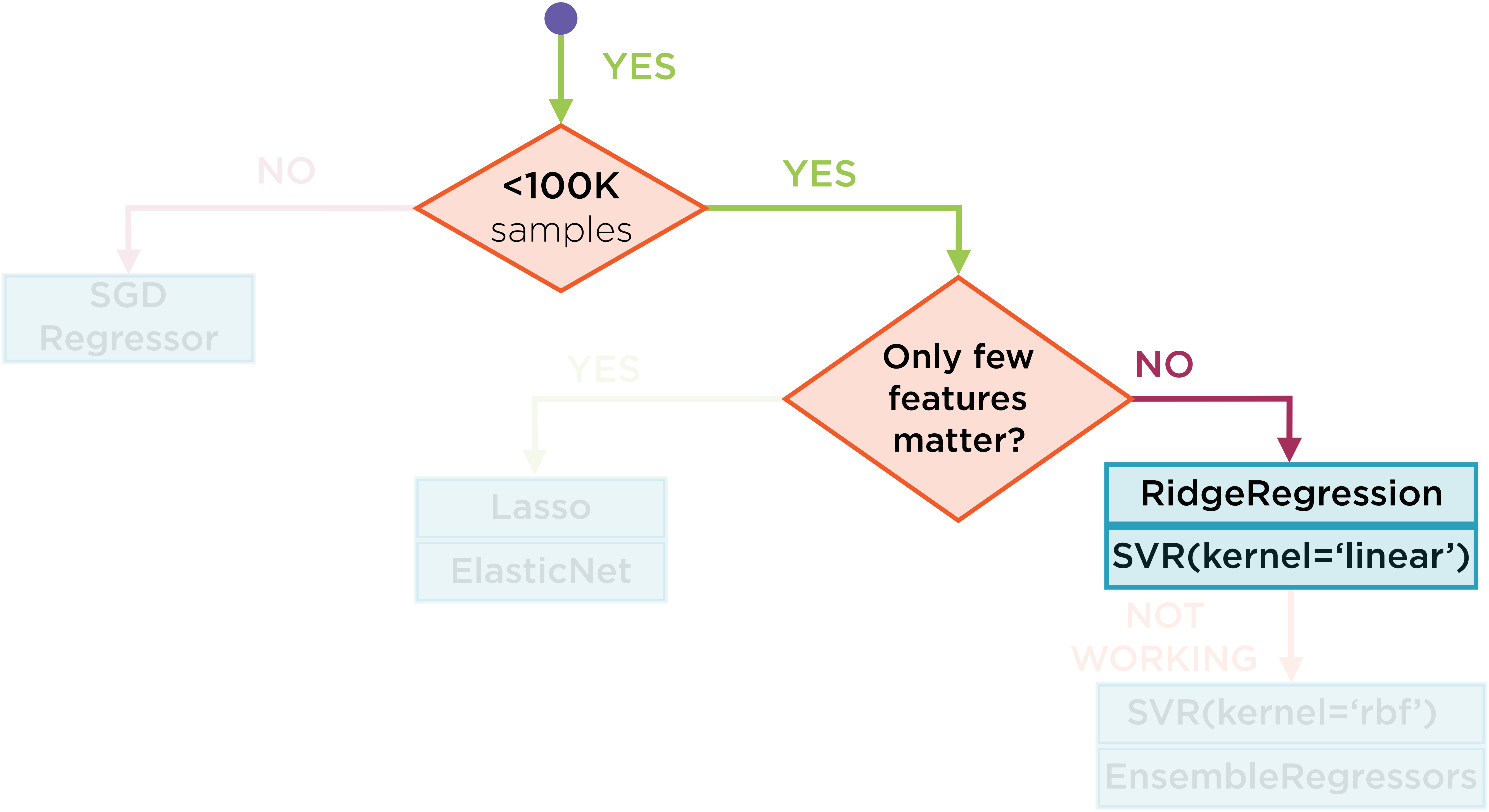
Regression


YES
<100K
samples
NO
YES
SGD
Regressor
Only few
features
matter?
YES
NO
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
NOT
WORKING
SVR(kernel='rbf')
EnsembleRegressors

Regression

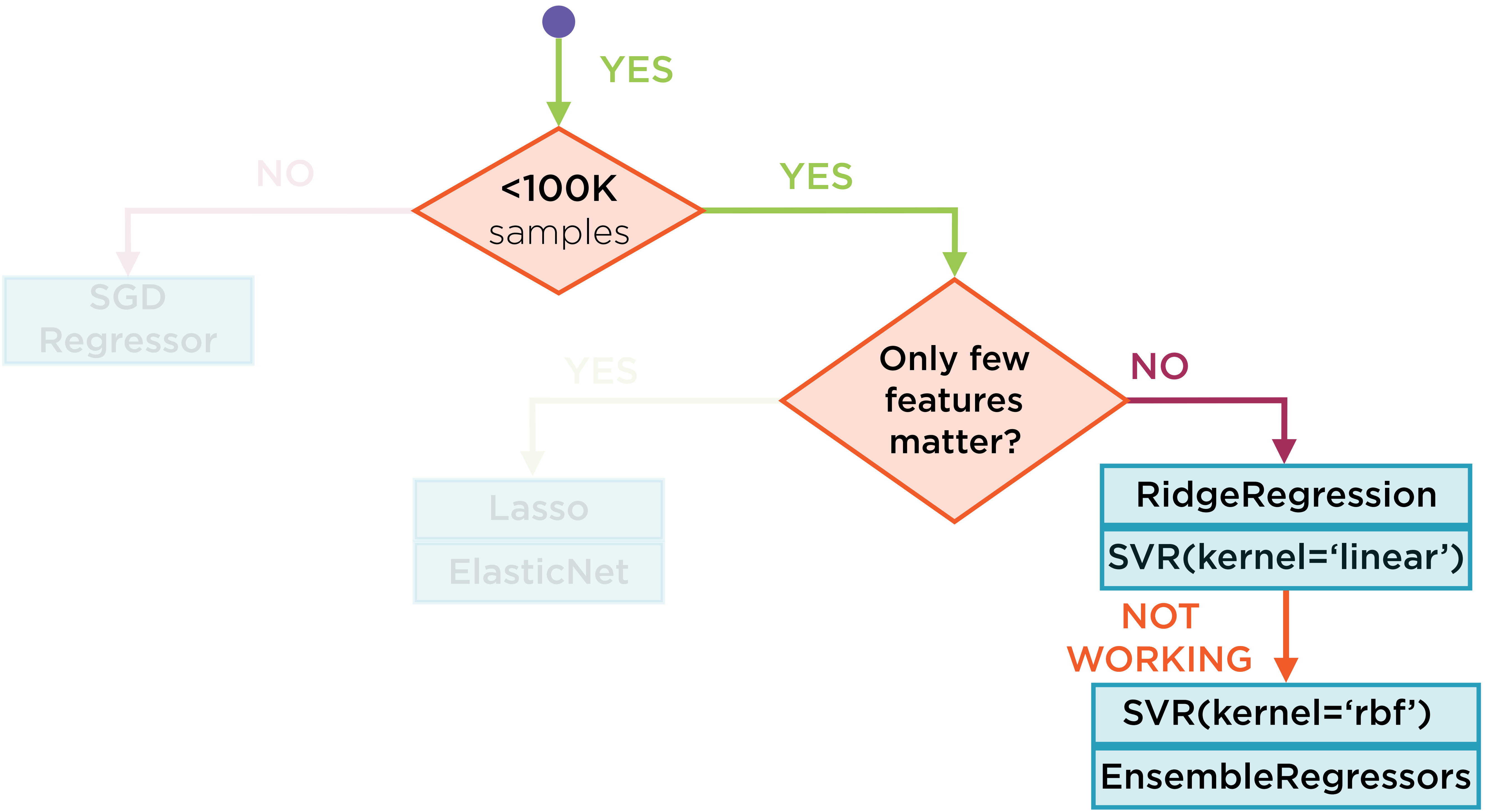

YES
<100K
samples
NO
YES
SGD
Regressor
YES
Only few
features
matter?
NO
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
NOT
WORKING
SVR(kernel='rbf')
EnsembleRegressors

Regression

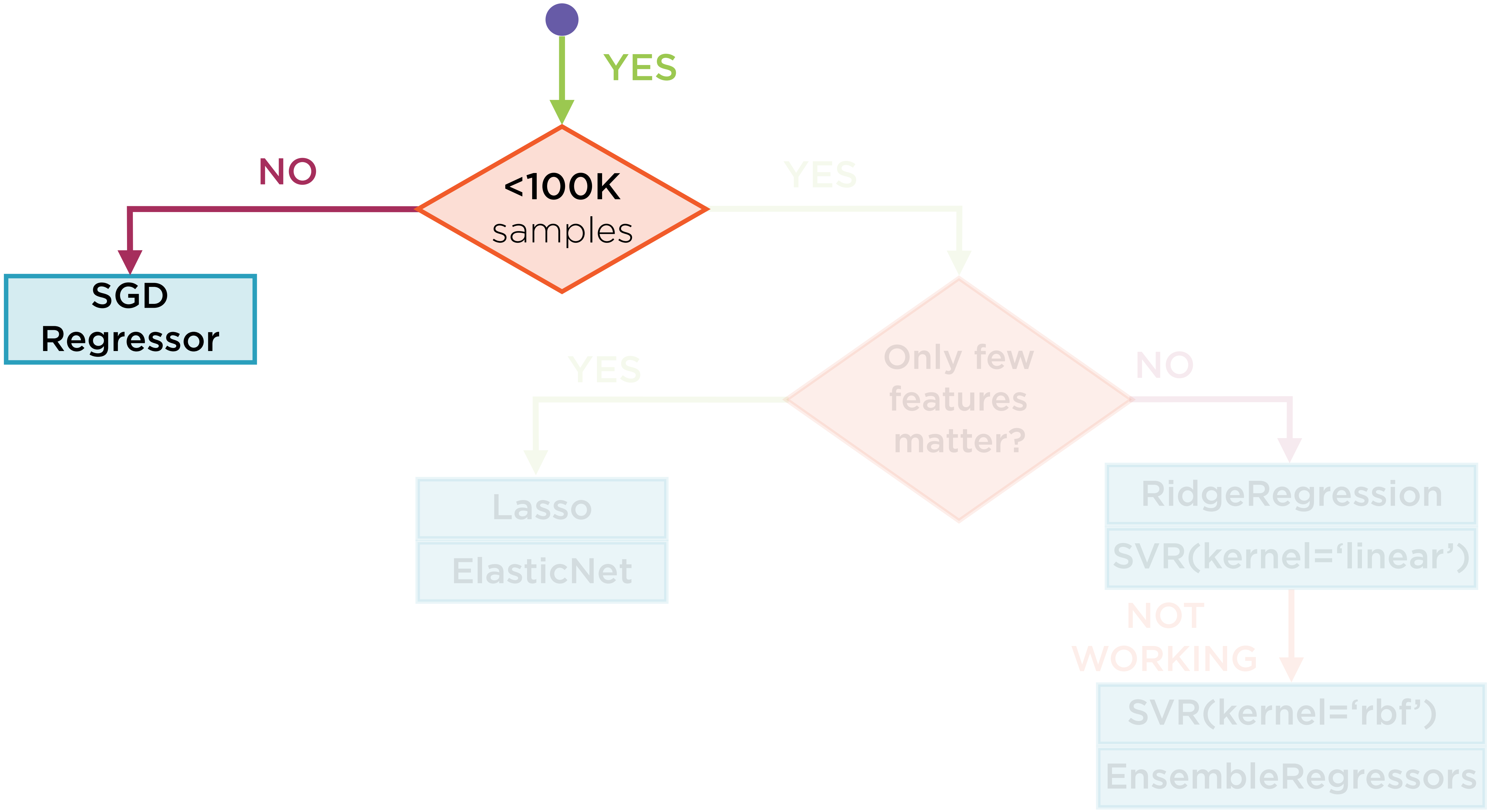


YES
<100K
samples
NO
SGD
Regressor
YES
Only few
features
matter?
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT
WORKING
SVR(kernel='rbf')
EnsembleRegressors

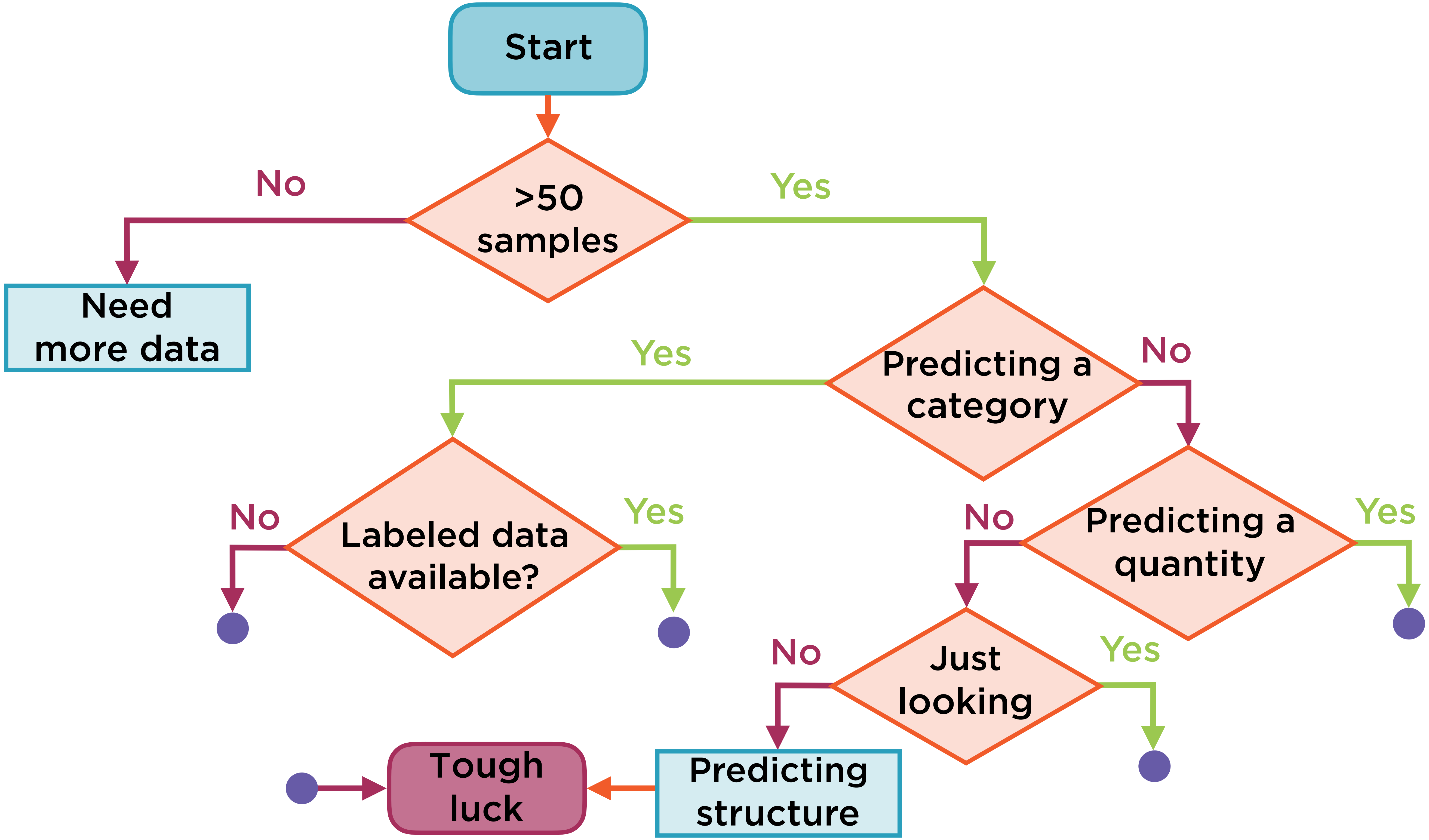
Regression


YES
<100K samples
NO
YES
SGD Regressor
Only few features matter?
YES
NO
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors

Regression

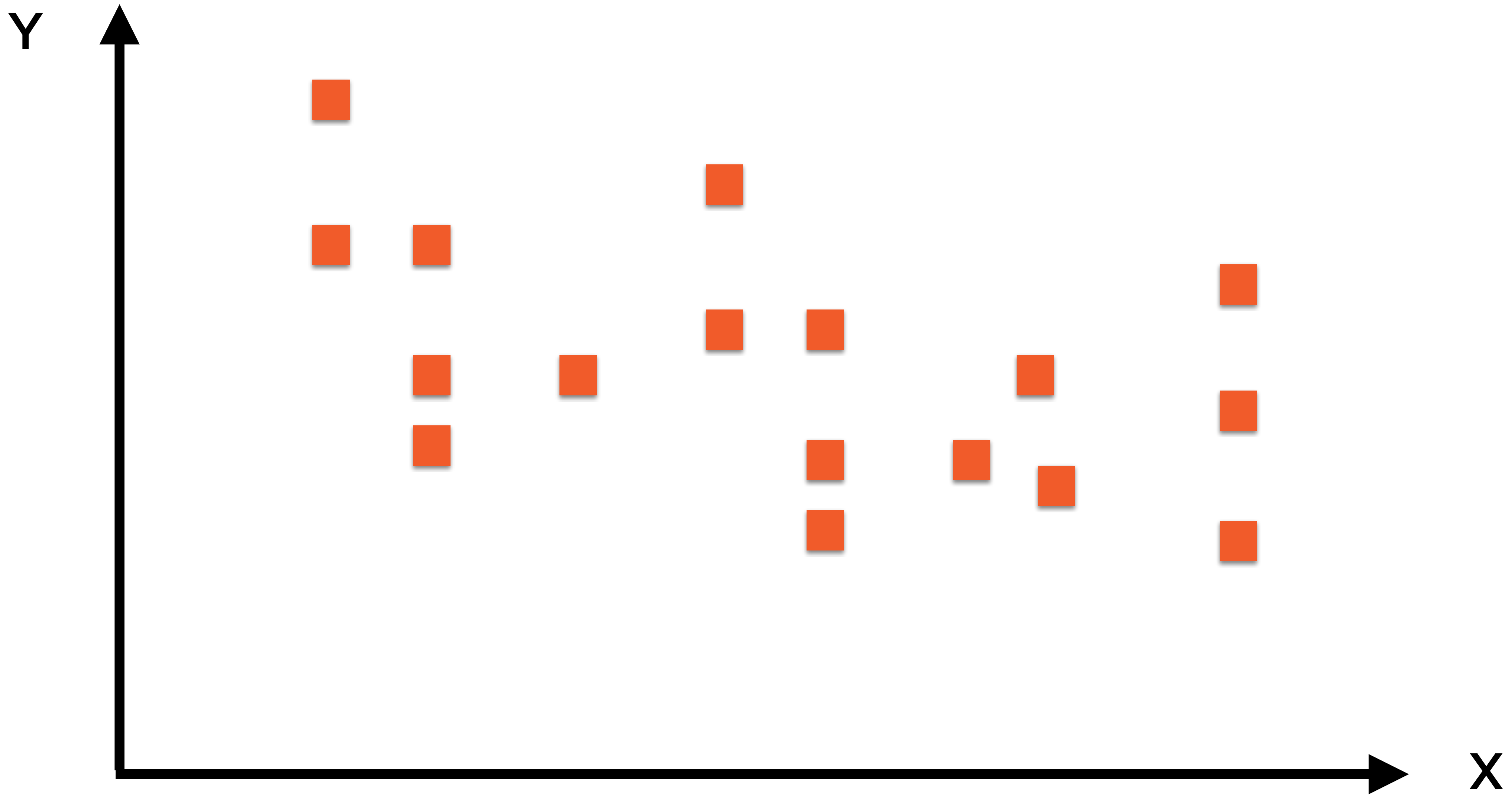

YES
NO
<100K
samples
SGD
Regressor
YES
Only few
features
matter?
YES
NO
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
NOT
WORKING
SVR(kernel='rbf')
EnsembleRegressors

Choosing the Right Estimator



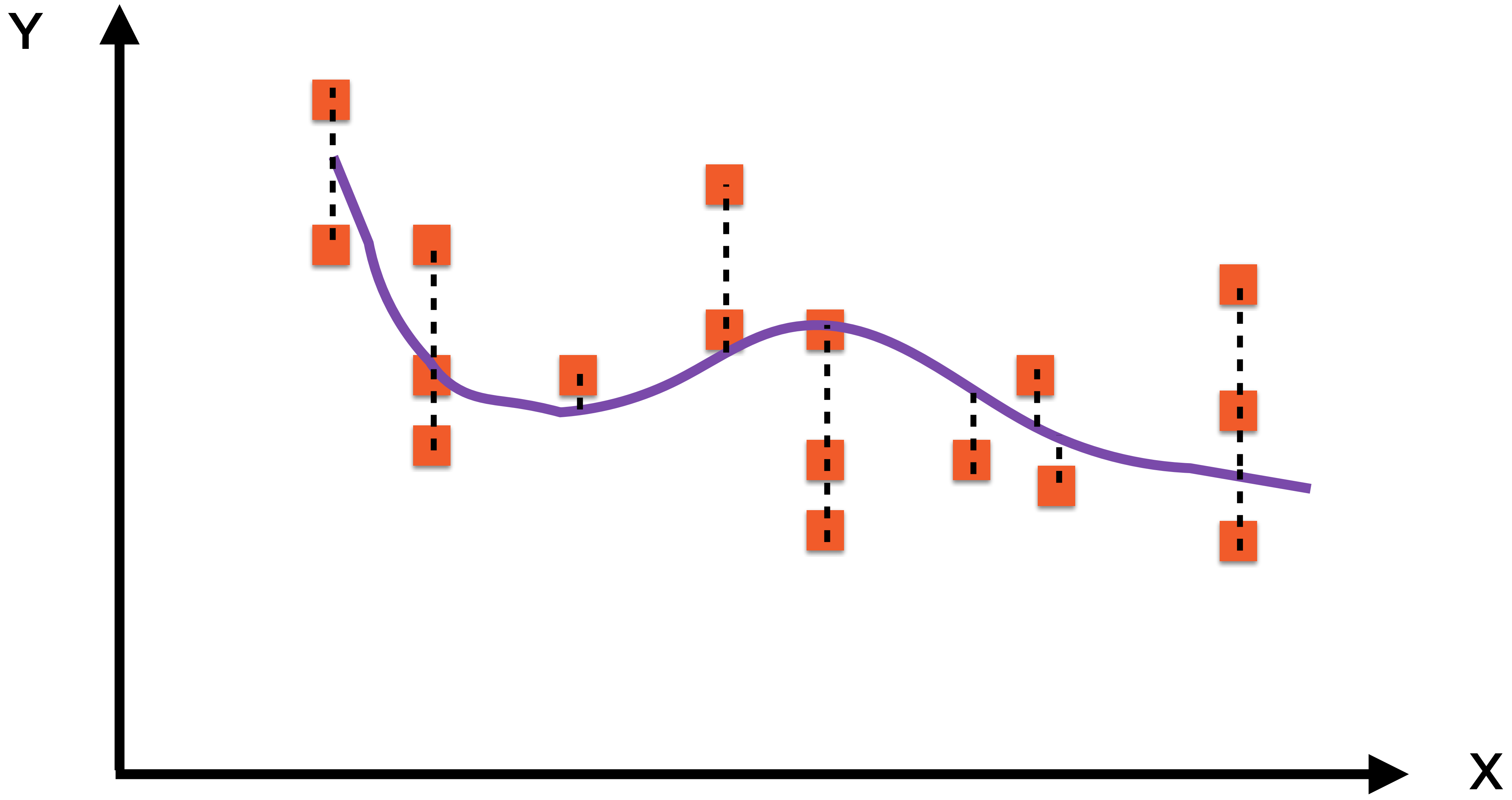
Overfitting and Regularization

Connecting the Dots



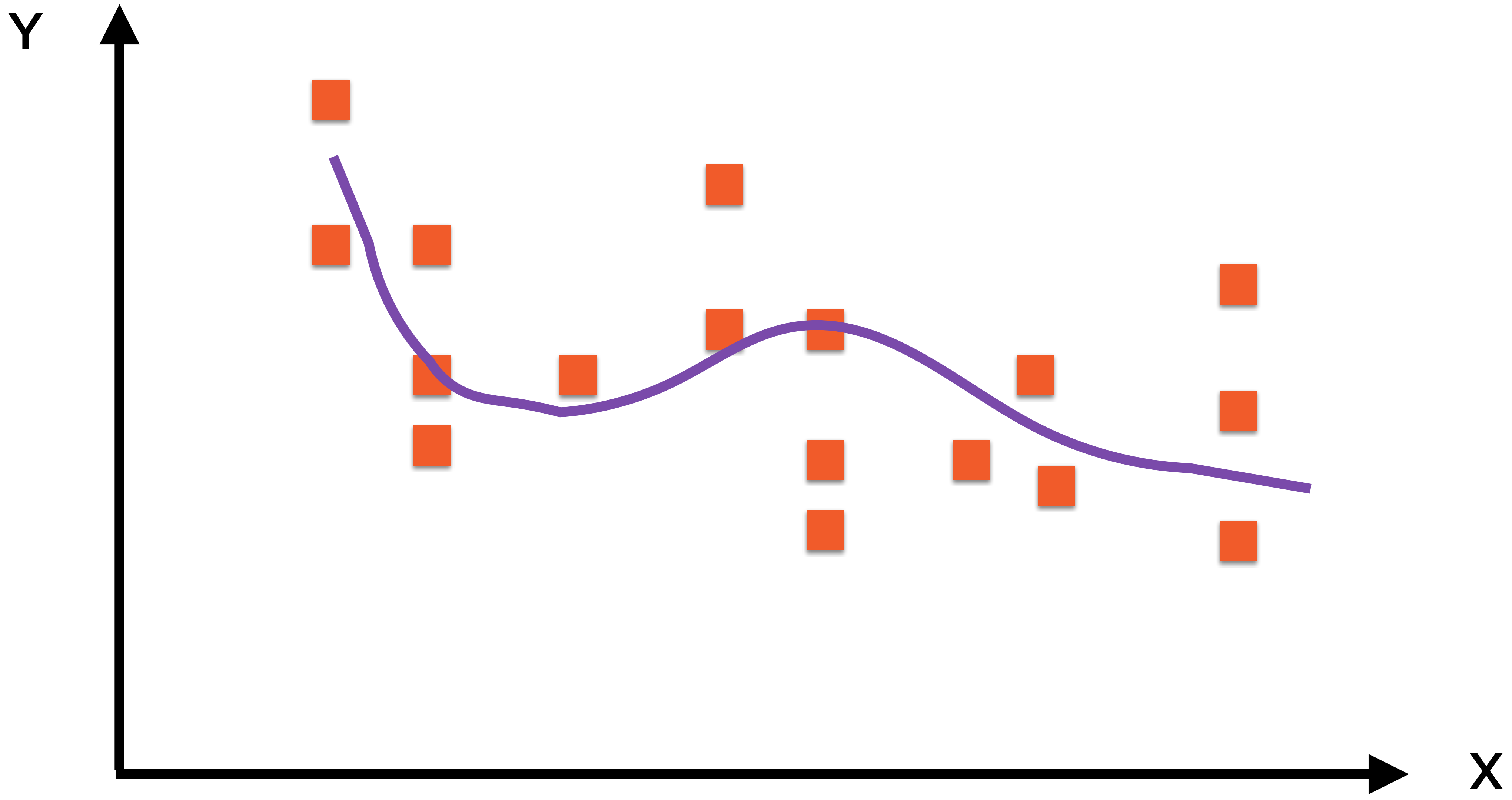
Challenge: Fit the "best" curve through these points

Good Fit?



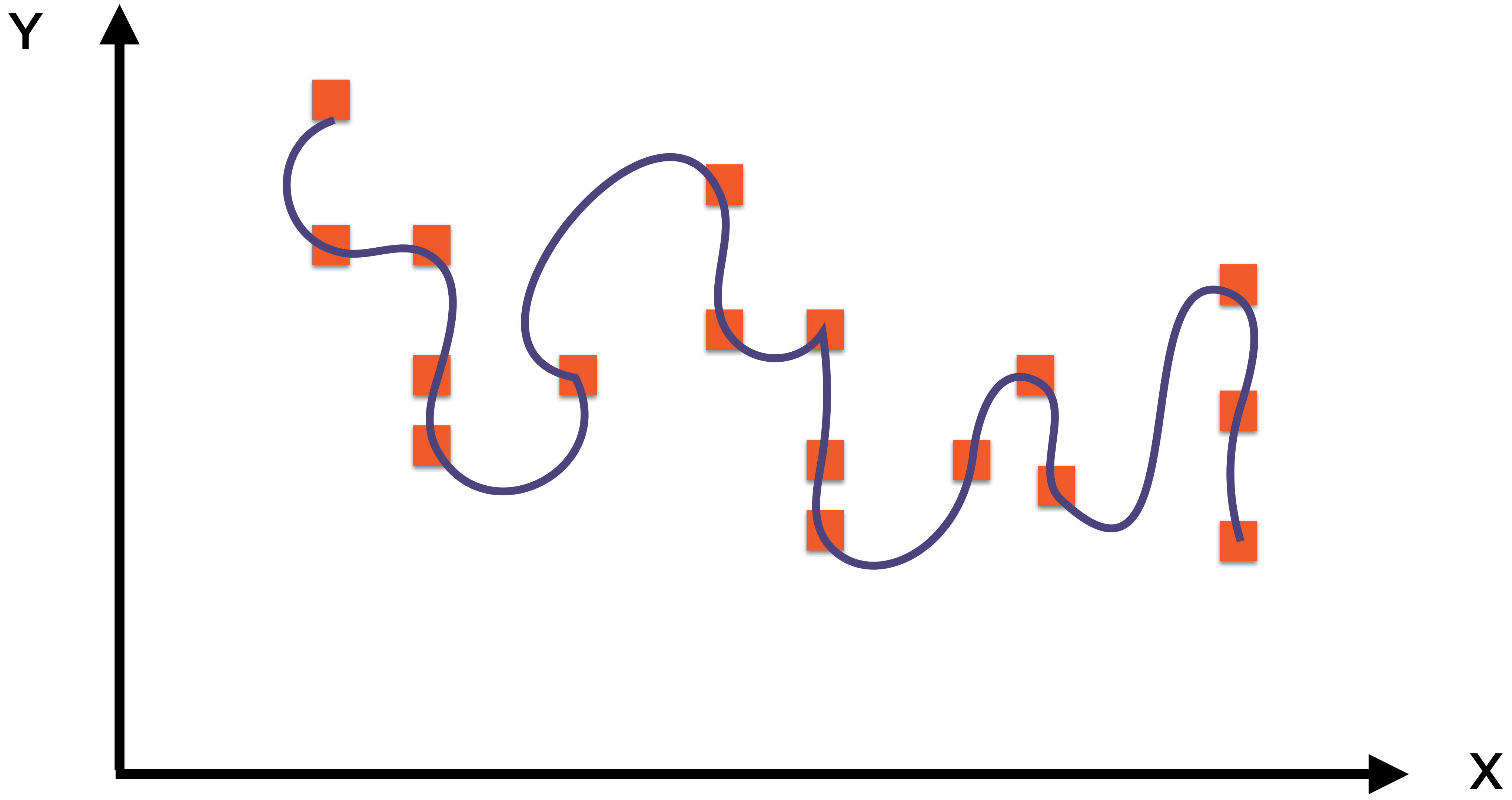
A curve has a "good fit" if the distances of points from the curve are small

Connecting the Dots



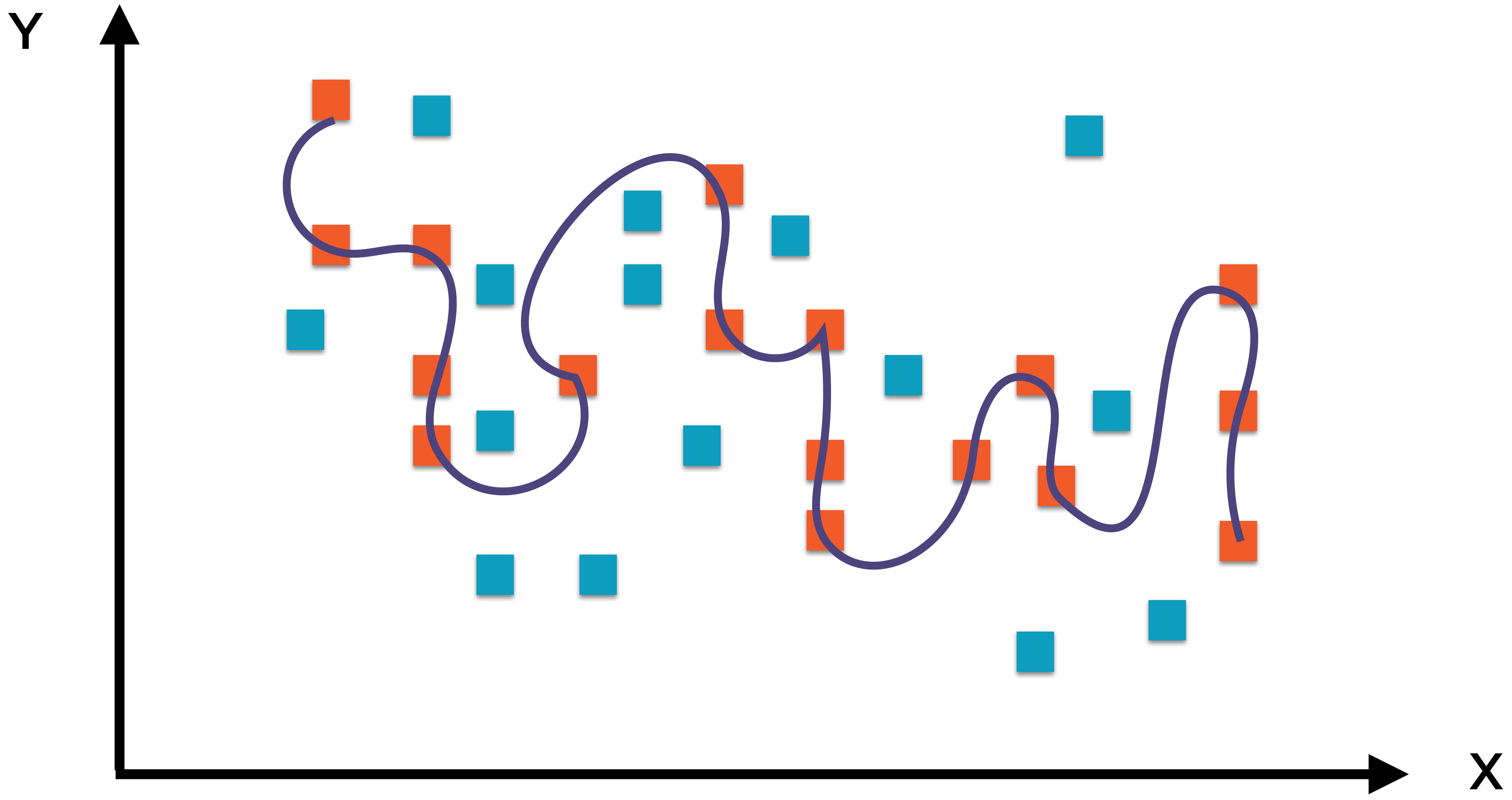
We could draw a pretty complex curve

Connecting the Dots



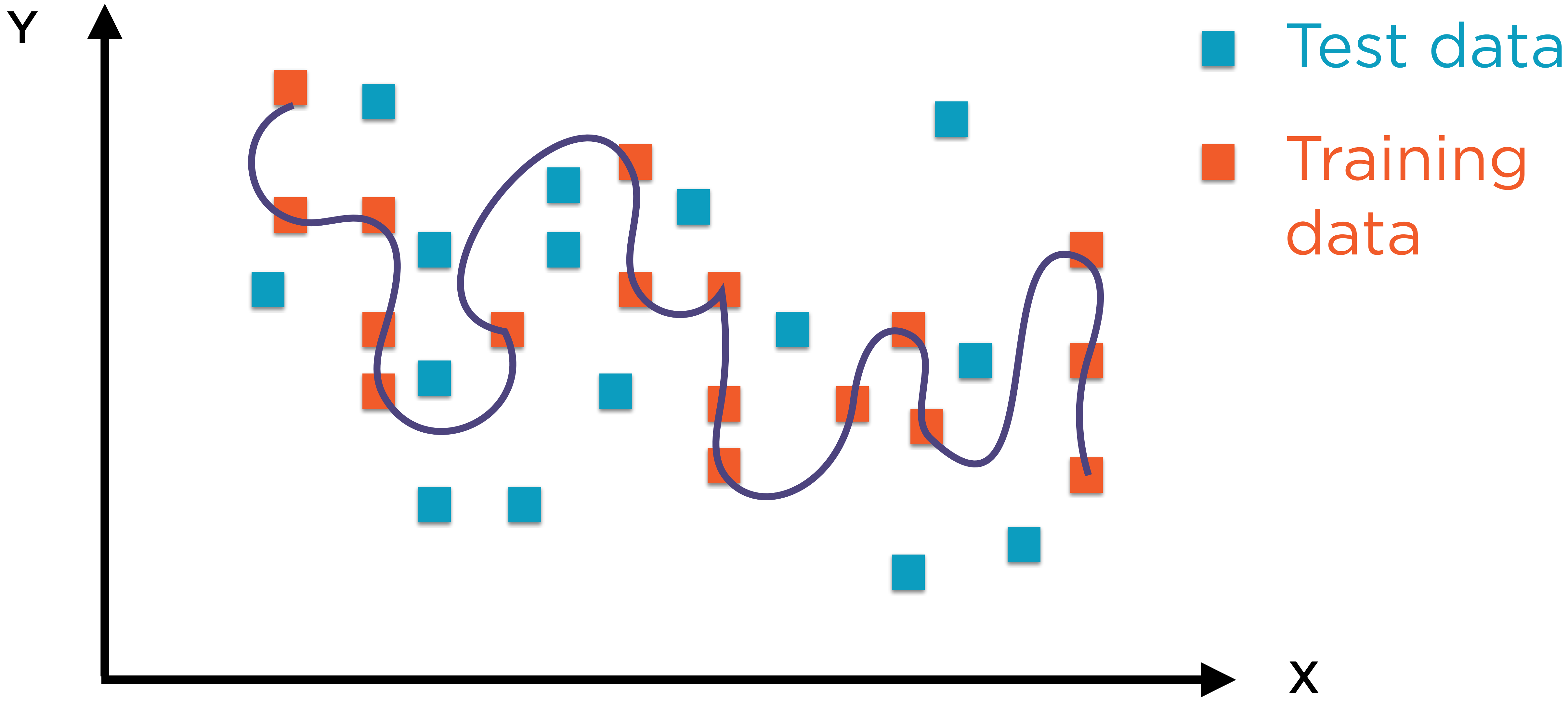
We can even make it pass through every single point

Connecting the Dots



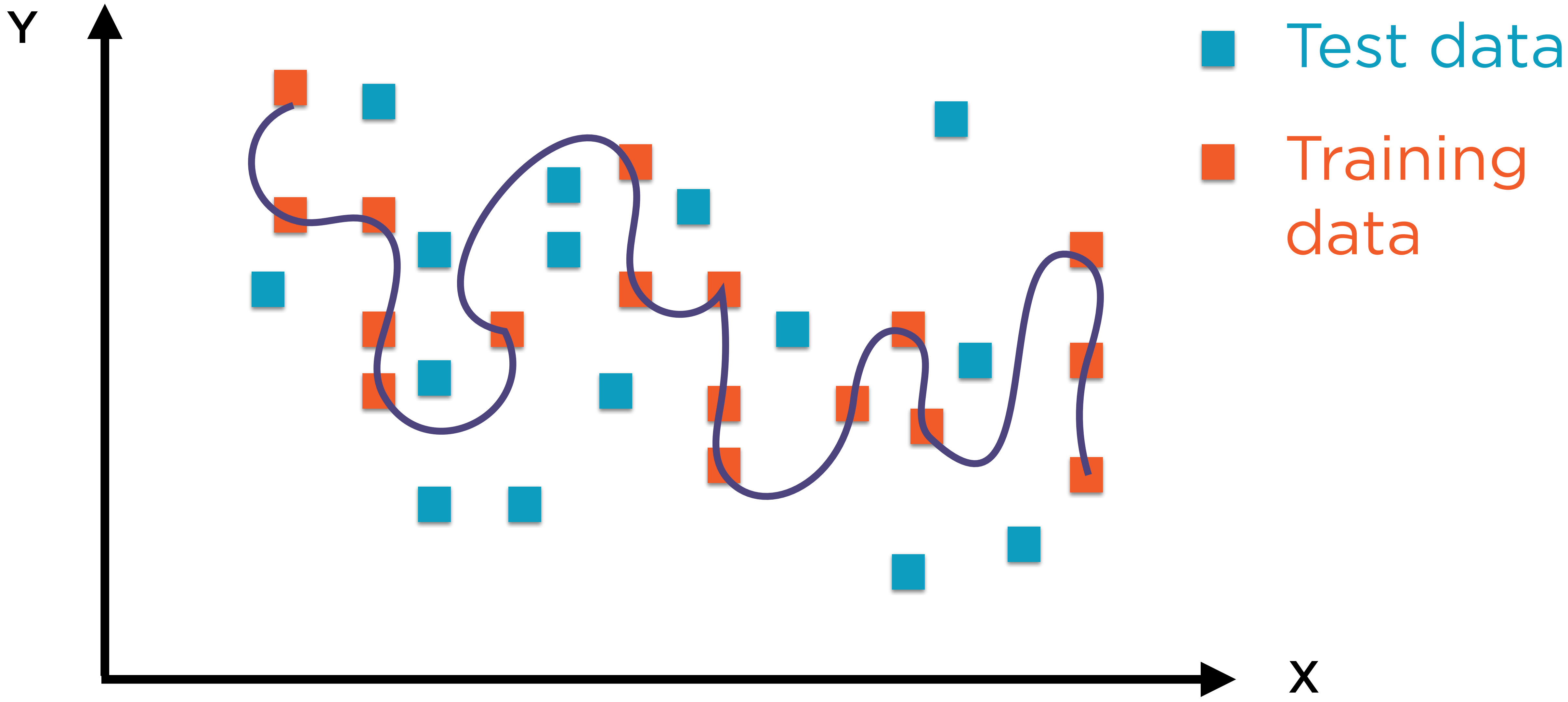
But given a new set of points, this curve might perform quite poorly

Connecting the Dots



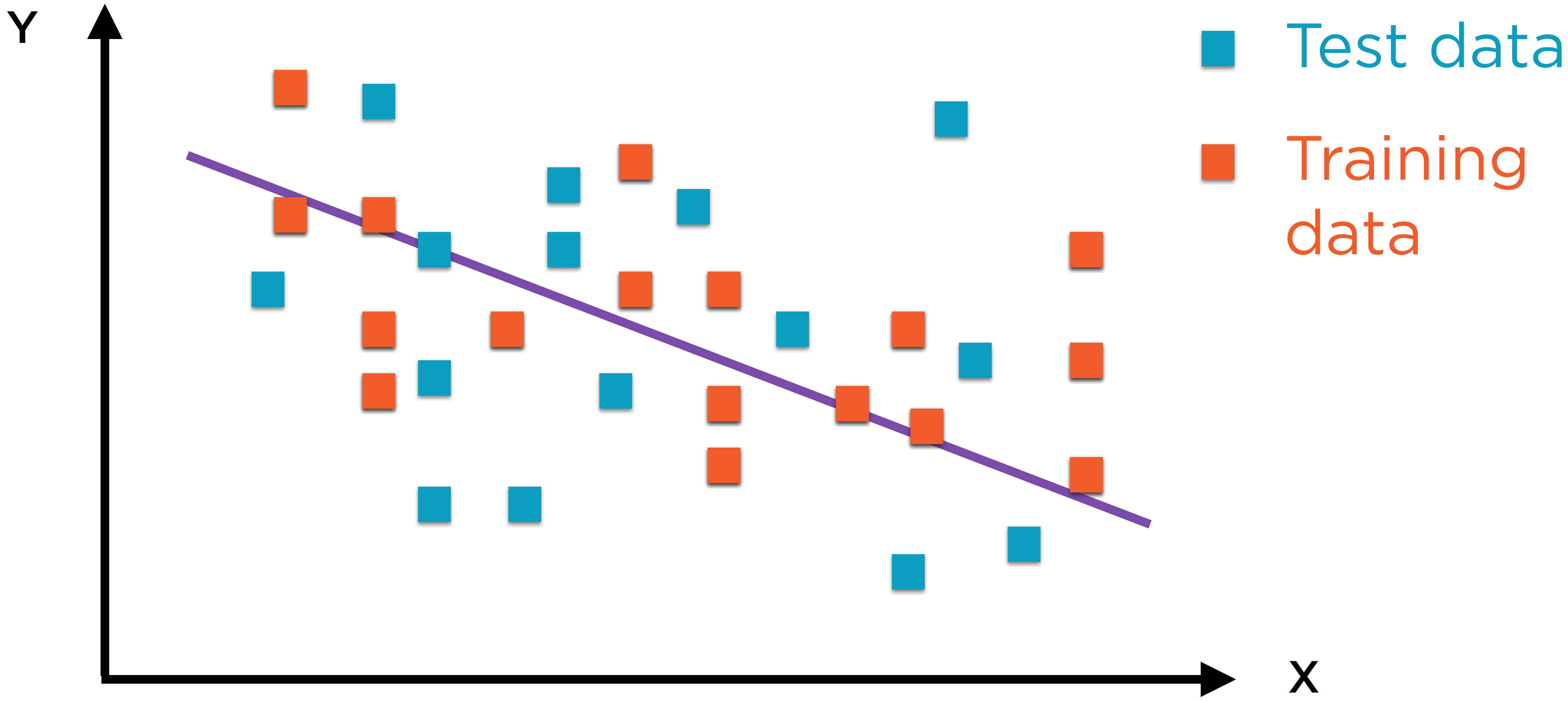
The original points were "training data", the new points are "test data"

Overfitting



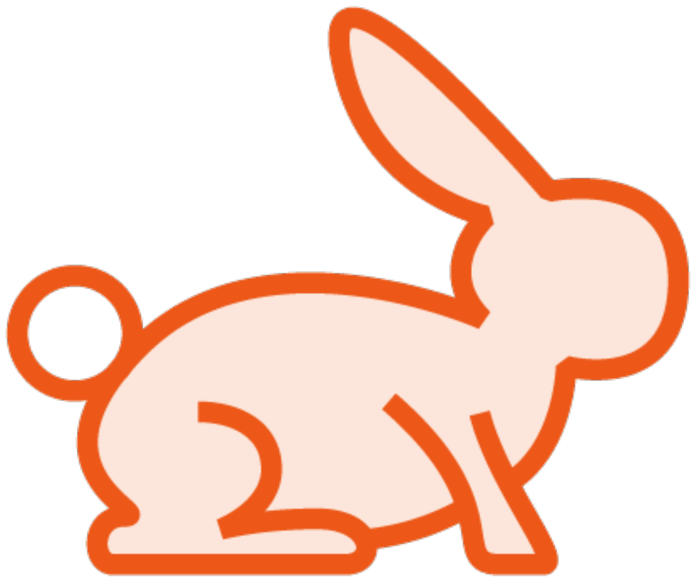
Great performance in training, poor performance in real usage

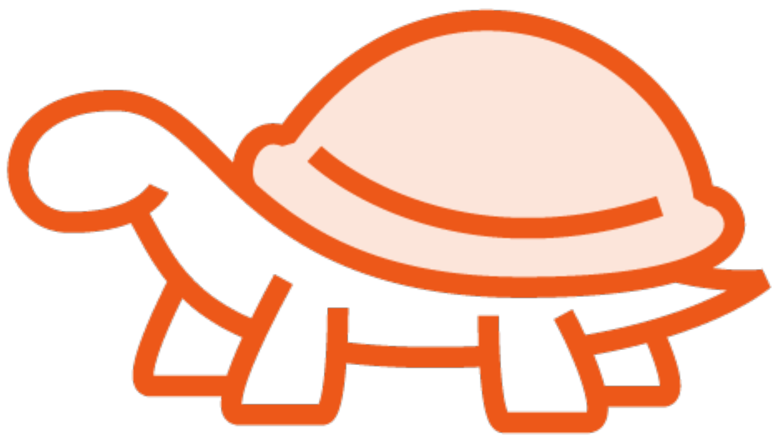
Connecting the Dots



A simple straight line performs worse in training, but better with test data

Overfitting





Low Training Error

Model does very well in training...

High Test Error

...but poorly with real data

Preventing Overfitting

Regularization - Penalize complex models

Cross-validation - Distinct training and validation phases

Dropout (NNs only) - Intentionally turn off some neurons during training

Regularization

Penalize complex models

Add penalty to objective function

Penalty as function of regression coefficients

Forces optimizer to keep it simple

Regularization

Regularization reduces variance error

But increases bias

Lasso, Ridge and Elastic Net

Regularized Regression Models

Lasso Regression

Penalizes large regression coefficients

Ridge Regression

Also penalizes large regression coefficients

Elastic Net Regression

Simply combines lasso and ridge

Ordinary MSE Regression

Minimize

$$\sqrt{(y^{actual} - y^{predicted})^2}$$

To find

A, B

The value of A and B define the "best fit" line

$$y = A + Bx$$

Lasso Regression

Minimize

$$\sqrt{(y_{actual} - y_{predicted})^2} + \alpha\,(|A| + |B|)$$

To find

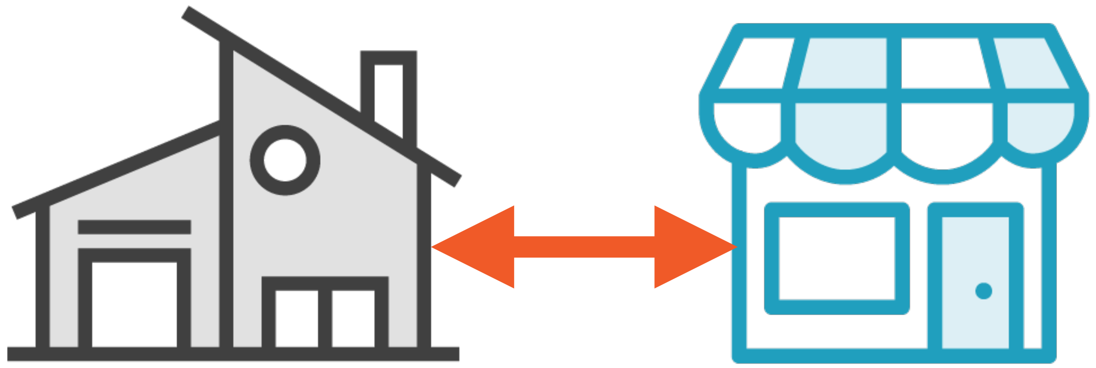
A, B

α is a hyperparameter

The value of A and B still define the "best fit" line

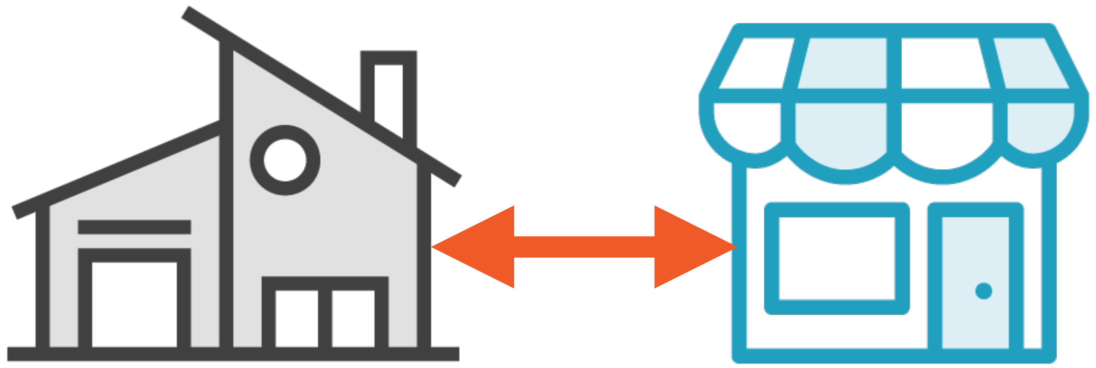
$$y = A + Bx$$

L-1 Norm



$$\text{L1-Norm}(A, B_1, B_2 \ldots B_n) = |A|^1 + |B_1|^1 + |B_2|^1 \ldots + |B_n|^1$$

L-1 Norm



$$L1\text{-}Norm(A, B_1, B_2 \ldots B_n) = |A|^1 + |B_1|^1 + |B_2|^1 \ldots + |B_n|^1$$

Lasso Regression

Minimize

$$\sqrt{(y_{actual} - y_{predicted})^2} + \alpha\,(|A| + |B|)$$

To find

A, B

α is a hyperparameter

The value of A and B still define the “best fit” line

$$y = A + Bx$$

Lasso Regression

Minimize $\sqrt{(y_{actual} - y_{predicted})^2}$ $+ \alpha (|A| + |B|)$

L-1 Norm of regression coefficients

To find

A, B

α is a hyperparameter

The value of A and B still define the "best fit" line

$y = A + Bx$

Ridge Regression

Minimize

$$\sqrt{(y_{actual} - y_{predicted})^2} + \alpha\,(|A|^2 + |B|^2)$$

L-2 Norm of regression coefficients

To find

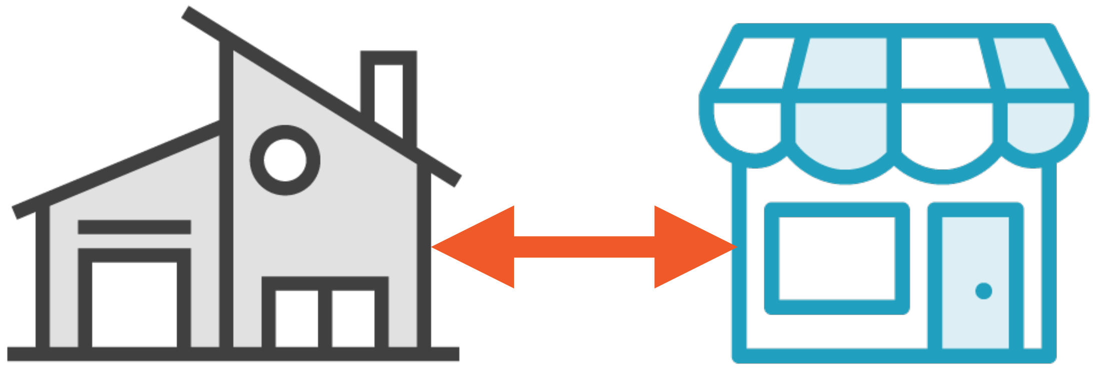
A, B

α is a hyperparameter

The value of A and B still define the “best fit” line

$y = A + Bx$

L-2 Norm



$$L2\text{-}Norm(A, B_1, B_2 \ldots B_n) = |A|^2 + |B_1|^2 + |B_2|^2 \ldots + |B_n|^2$$

Lasso Regression

Add penalty for large coefficients

Penalty term is L-1 norm of coefficients

Penalty weighted by hyperparameter α

Lasso Regression

$\alpha = 0$ ~ Regular (MSE regression)

$\alpha \rightarrow \infty$ ~ Force small coefficients to zero

Model selection by tuning α

Eliminates unimportant features

Lasso Regression

"Lasso" ~ *Least Absolute Shrinkage and Selection Operator*

Math is complex

No closed form, needs numeric solution

Ridge Regression

Minimize

$$\sqrt{(y_{actual} - y_{predicted})^2} + \alpha\,(|A|^2 + |B|^2)$$

L-2 Norm of regression coefficients

To find

A, B

α is a hyperparameter

The value of A and B still define the “best fit” line

$y = A + Bx$

Ridge Regression

Add penalty for large coefficients

Penalty term is L-2 norm of coefficients

Penalty weighted by hyperparameter α

Ridge Regression

Unlike lasso, ridge regression has closed-form solution

Unlike lasso, ridge regression will not force coefficients to 0

- Does not perform model selection

Regularized Regression Models

Lasso Regression

Penalizes large regression coefficients

Ridge Regression

Also penalizes large regression coefficients

Elastic Net Regression

Simply combines lasso and ridge

Demo

Defining helper functions to build, train and evaluate multiple regression models

Demo

Comparing single feature, kitchen sink and parsimonious regression

Demo

Implementing Lasso regression

Demo

Implementing Ridge regression

Demo

Implementing Elastic Net regression

Summary

Choosing regression to solve problems

Overfitting and the bias-variance trade-off

Regularization to mitigate overfitting

Building and training Ridge, Lasso and ElasticNet regression models