Understanding and Implementing Regression Models

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Regression as a form of supervised machine learning

Ordinary Least Squares (OLS) regression

Evaluating regression models using R^2

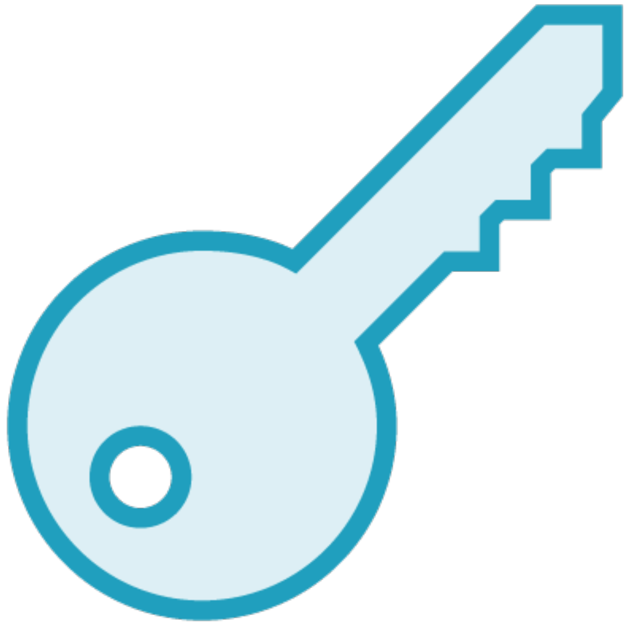
Choosing the right regression algorithm based on features and data

Lasso and Ridge regression

Gradient Descent in regression

Building Regression Models

X Causes Y



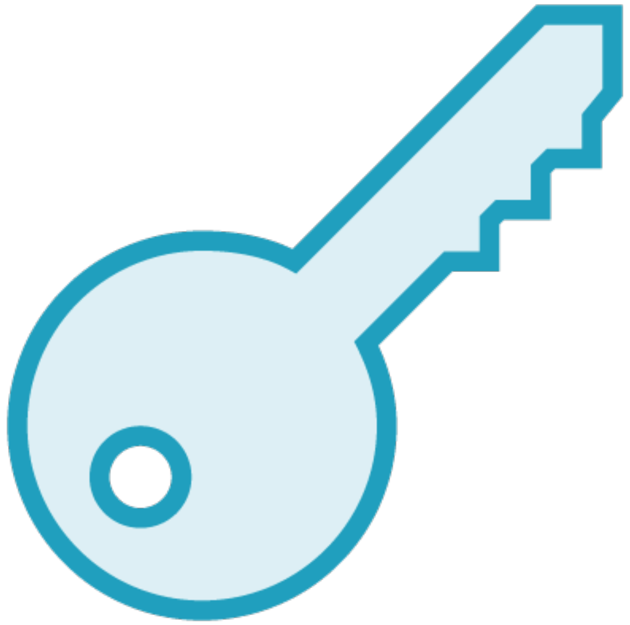
Cause

Independent variable

Effect

Dependent variable

X Causes Y



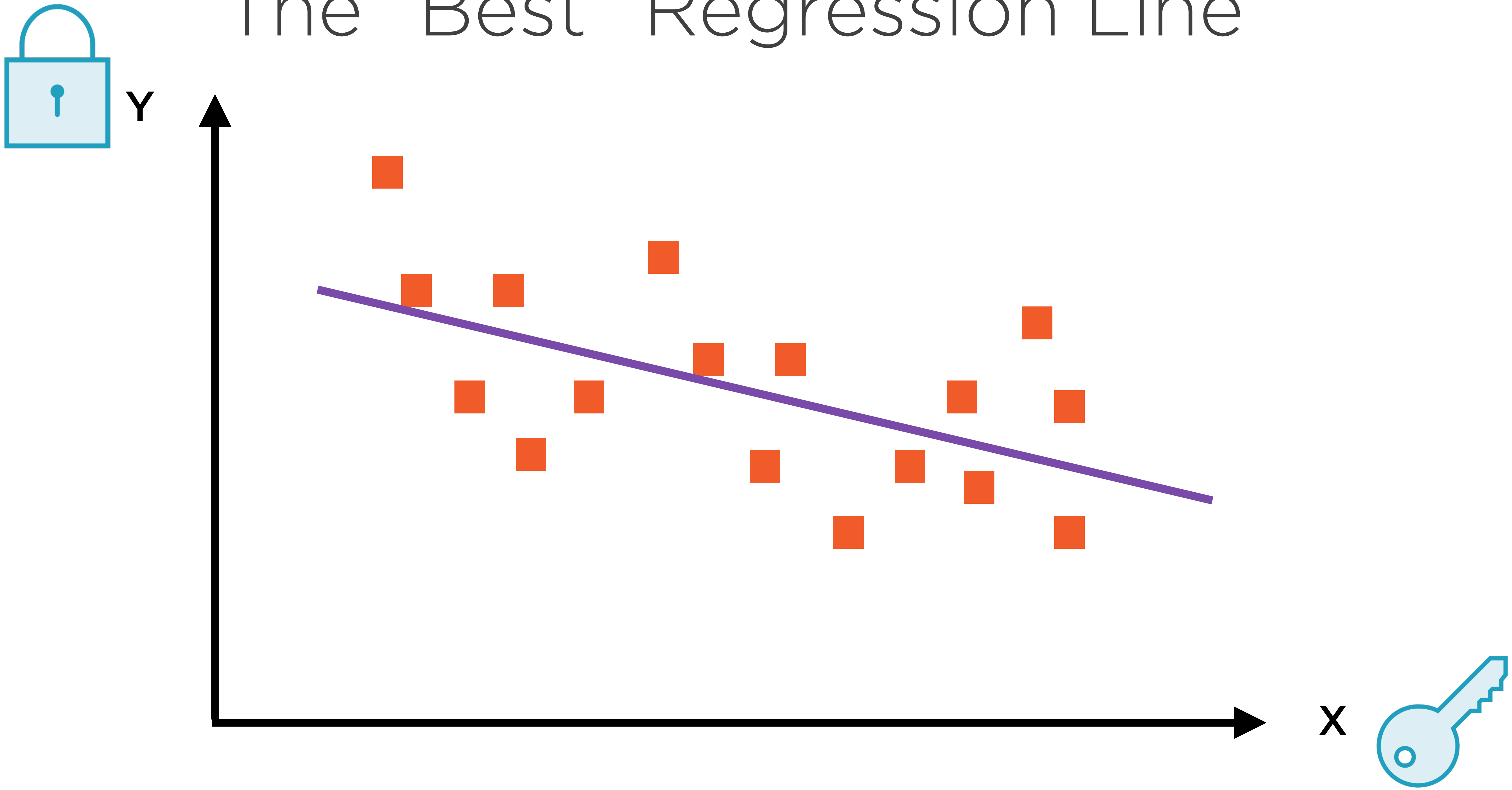
Cause

Explanatory variable

Effect

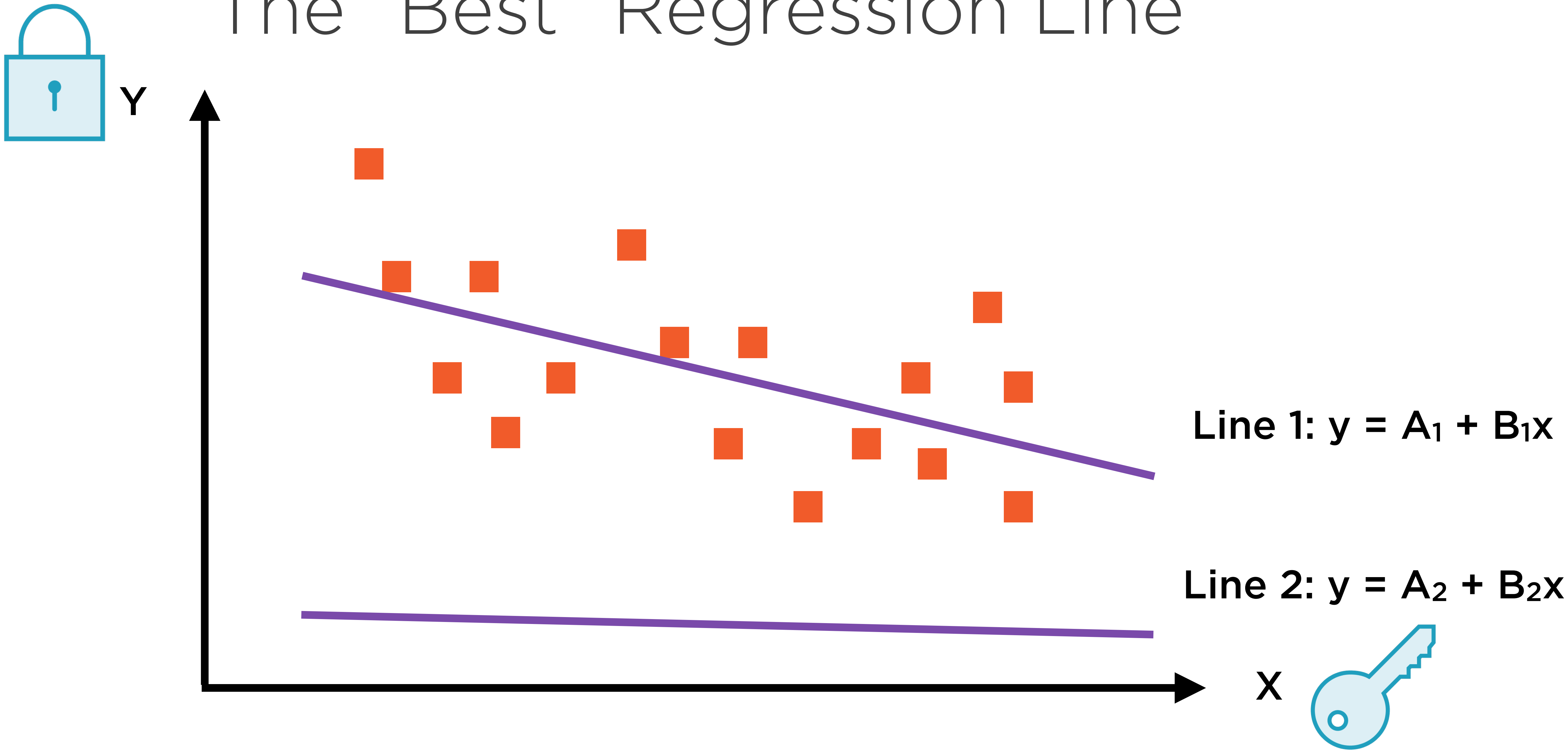
Dependent variable

The “Best” Regression Line



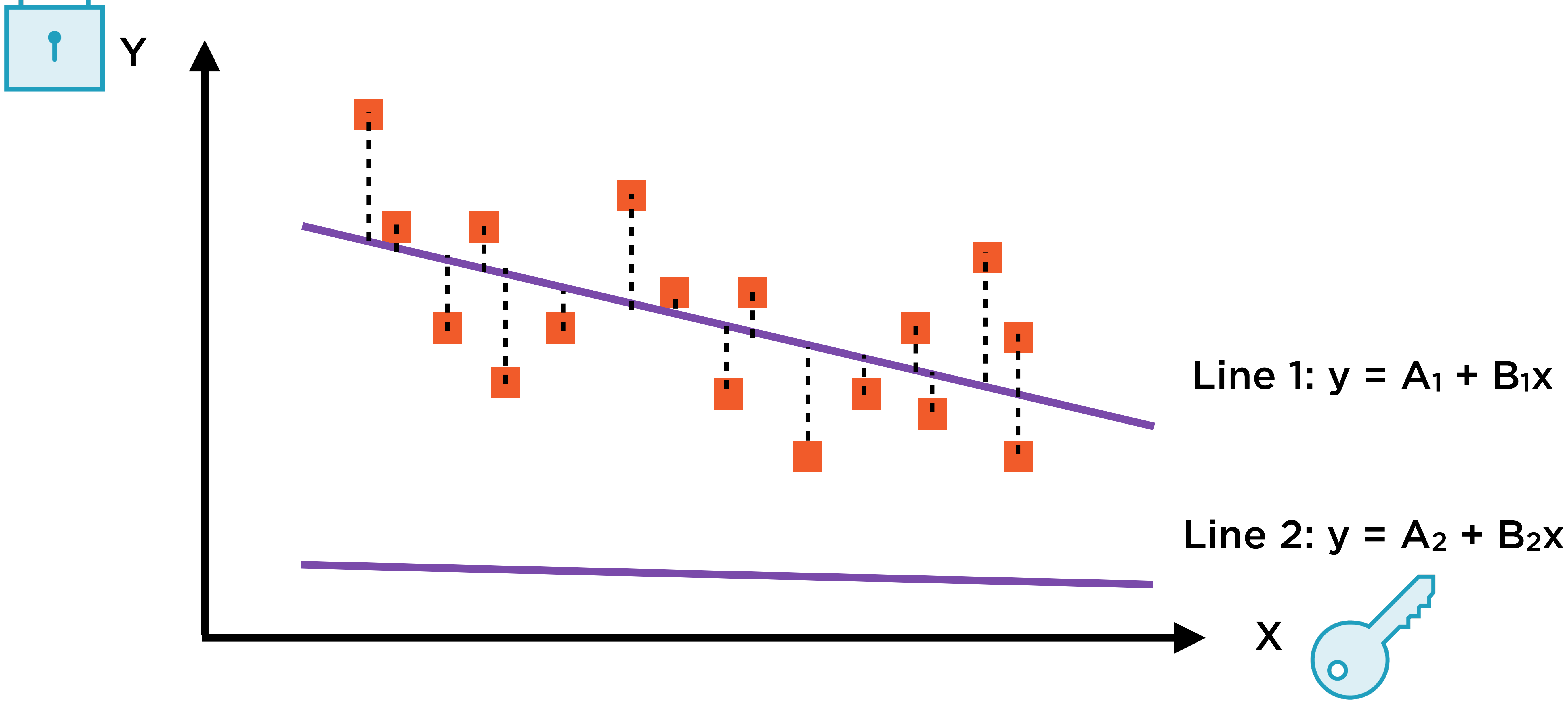
Linear Regression involves finding the “best fit” line

The “Best” Regression Line



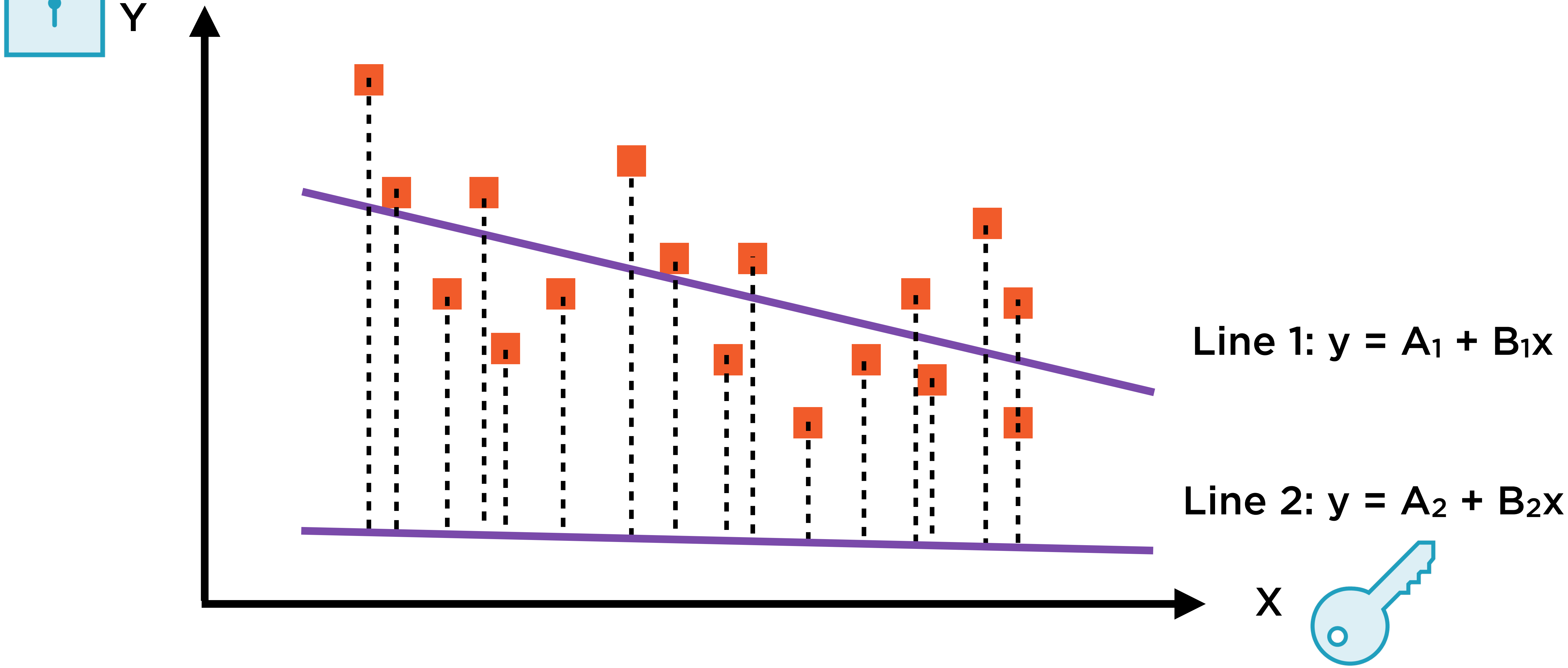
Let’s compare two lines, Line 1 and Line 2

Minimising Least Square Error



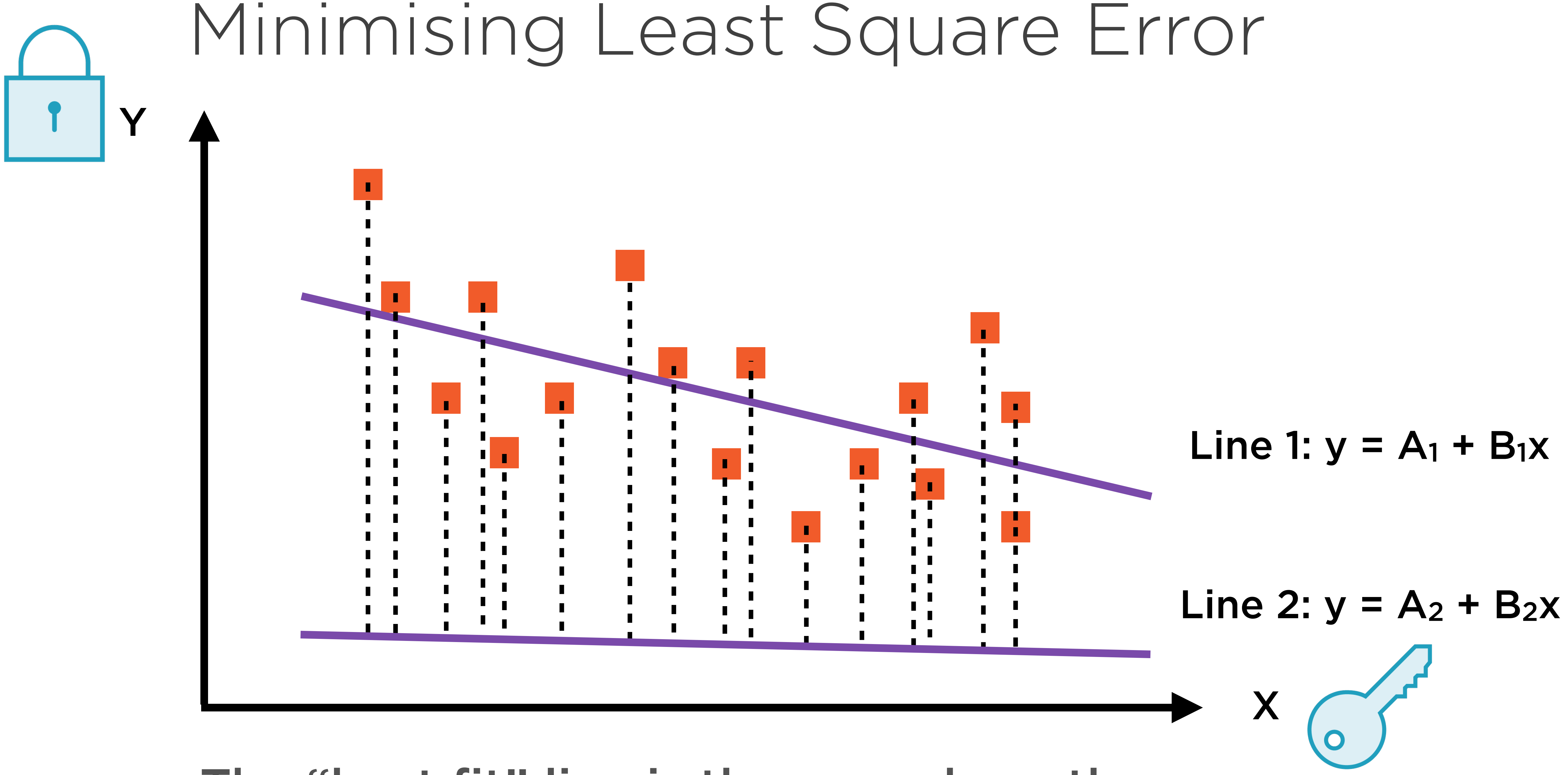
Drop vertical lines from each point to the lines 1 and 2

Minimising Least Square Error



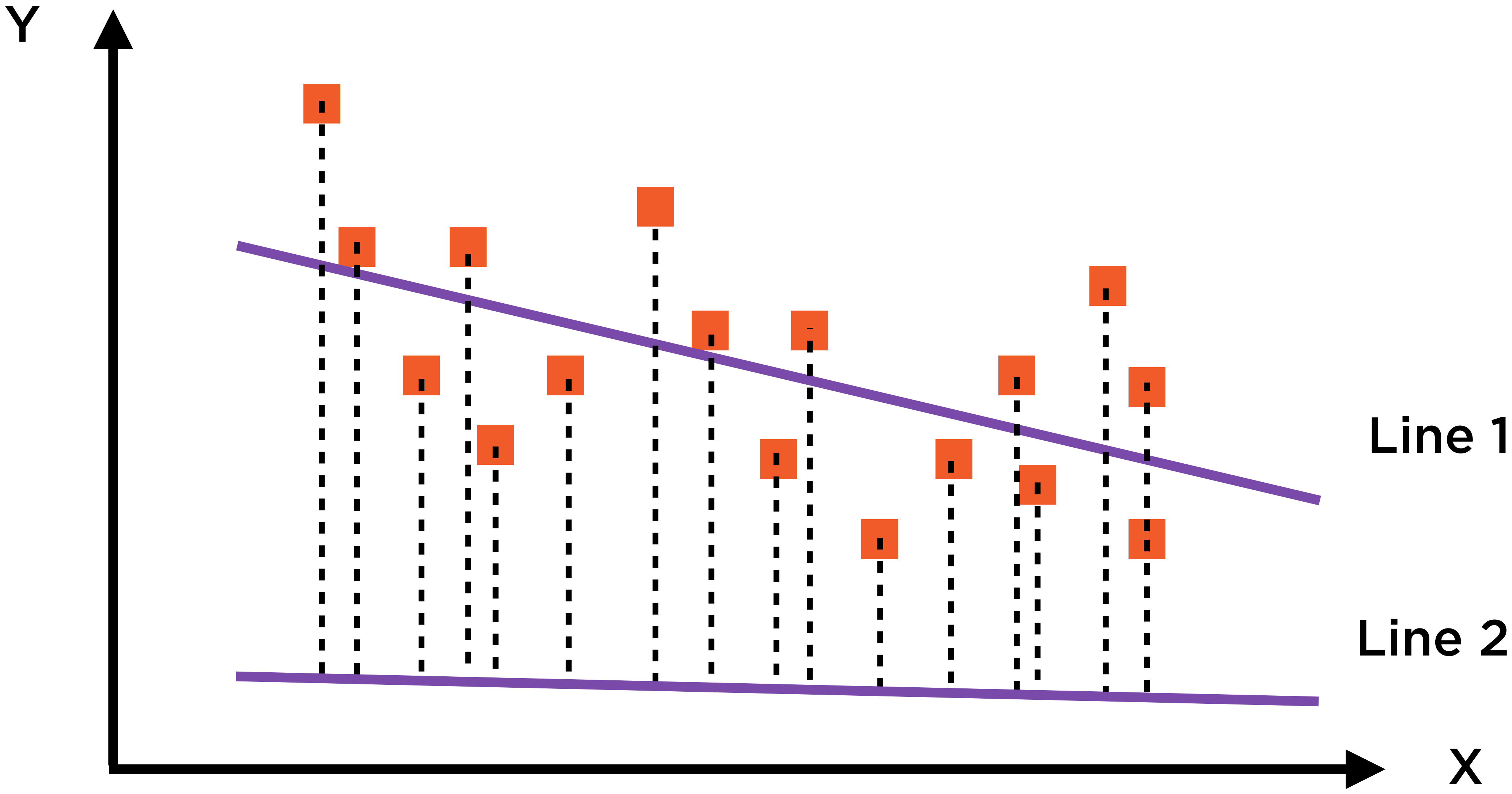
Drop vertical lines from each point to the lines 1 and 2

Minimising Least Square Error



The “best fit” line is the one where the sum of the squares of the lengths of these dotted lines is minimum

Minimising Least Square Error

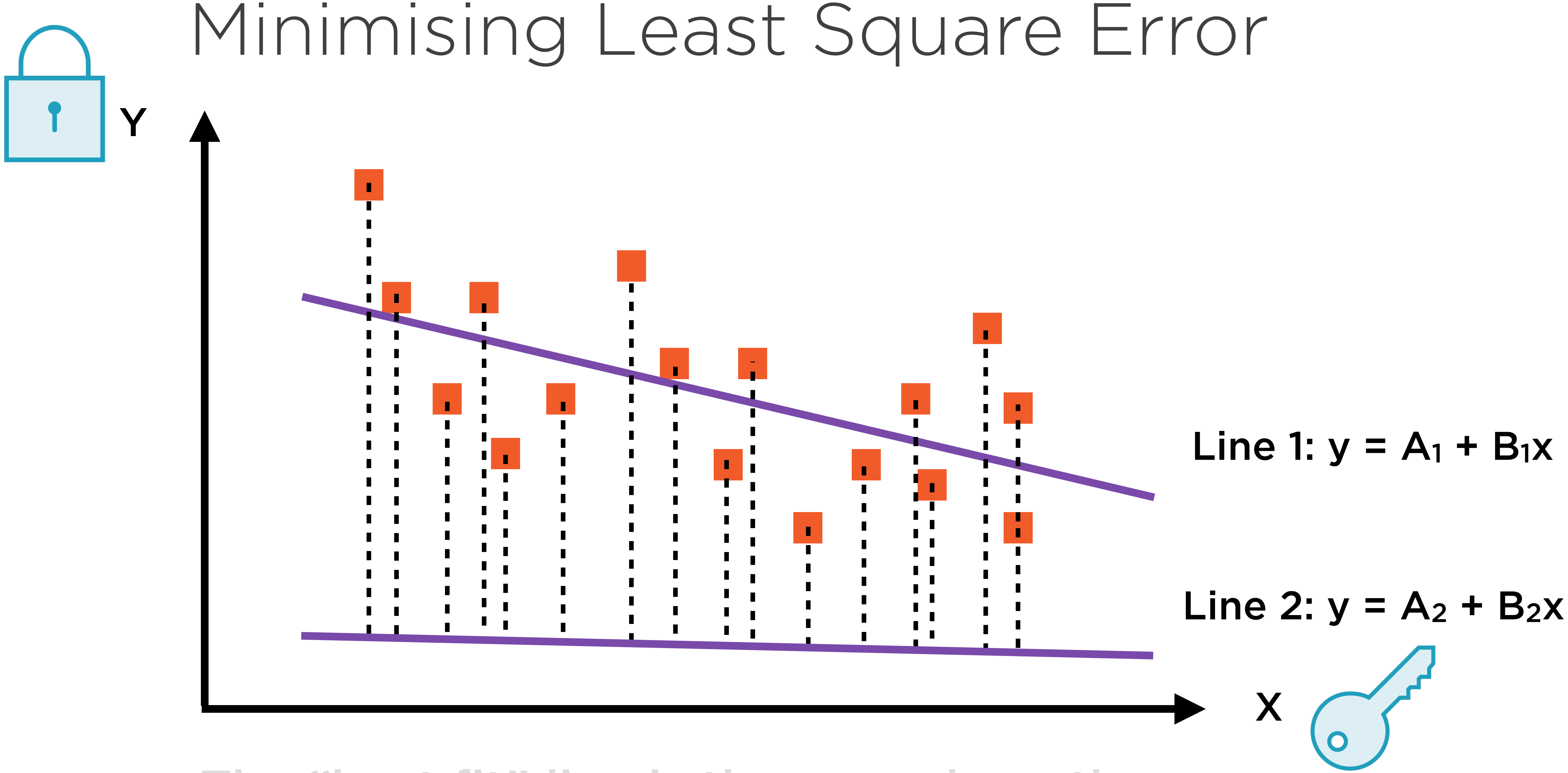


Line 1: $y = A_1 + B_1x$

Line 2: $y = A_2 + B_2x$

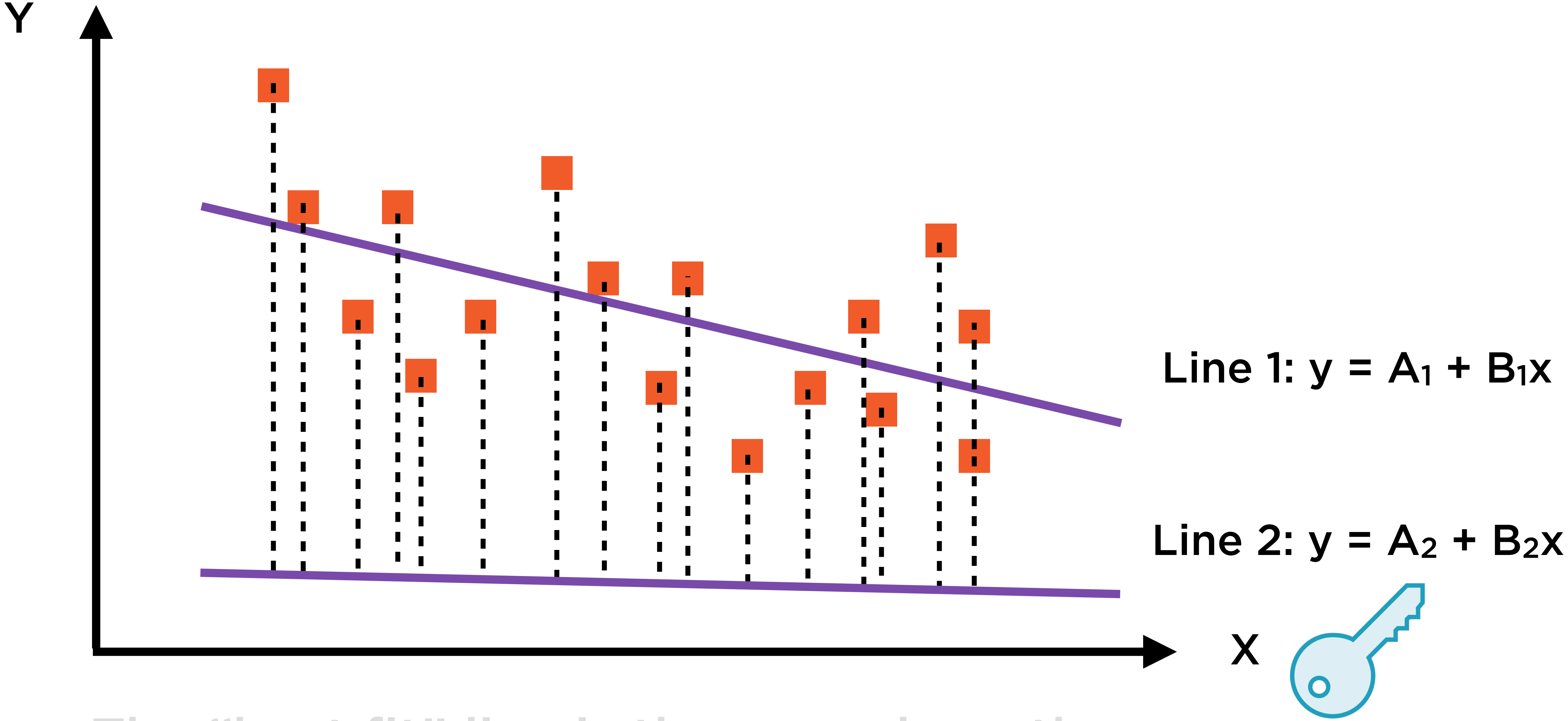
The “best fit” line is the one where the sum of the squares of the **lengths of these dotted lines** is minimum

Minimising Least Square Error



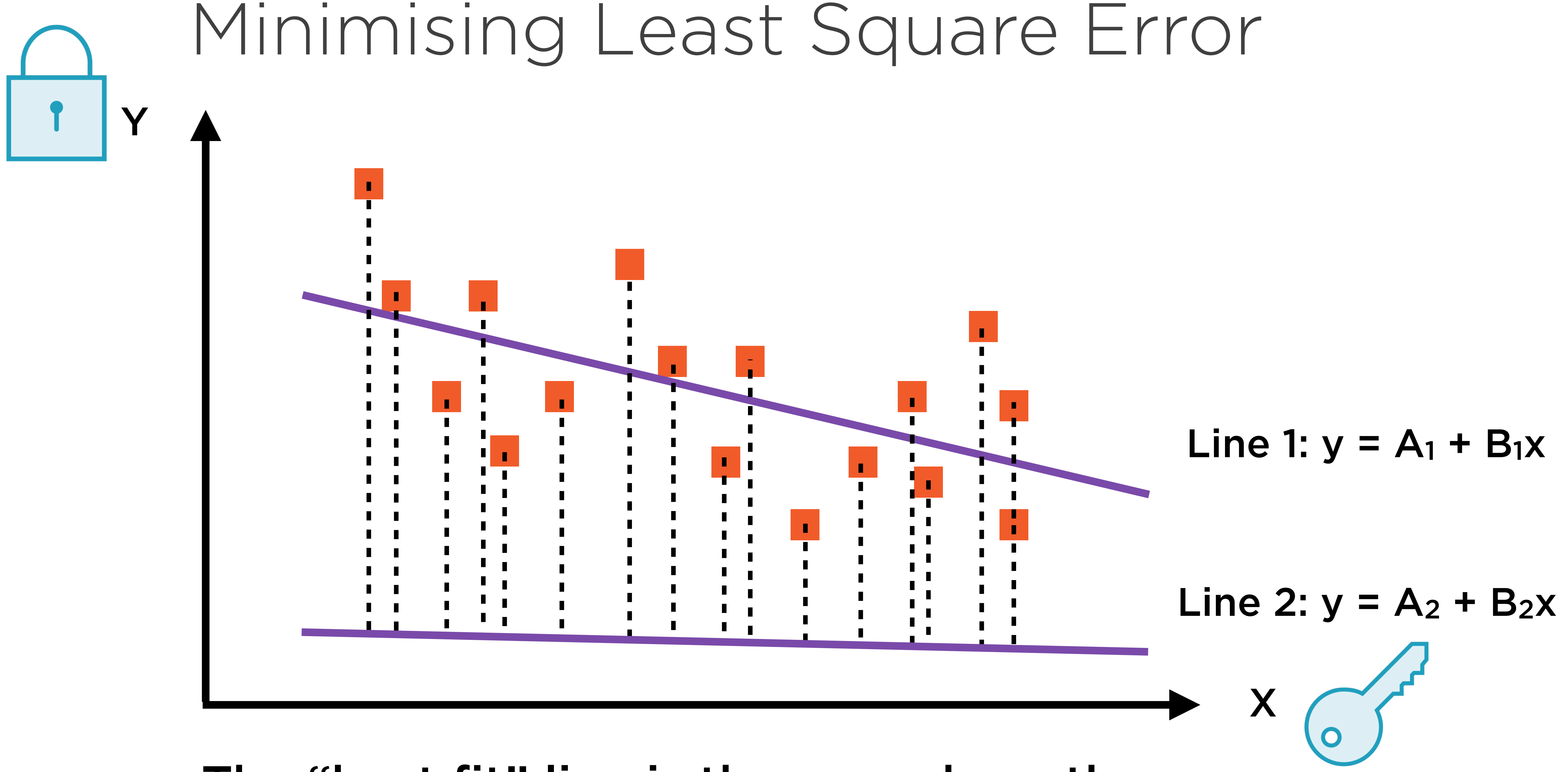
The “best fit” line is the one where the **sum of the squares** of the lengths of these dotted lines is minimum

Minimising Least Square Error



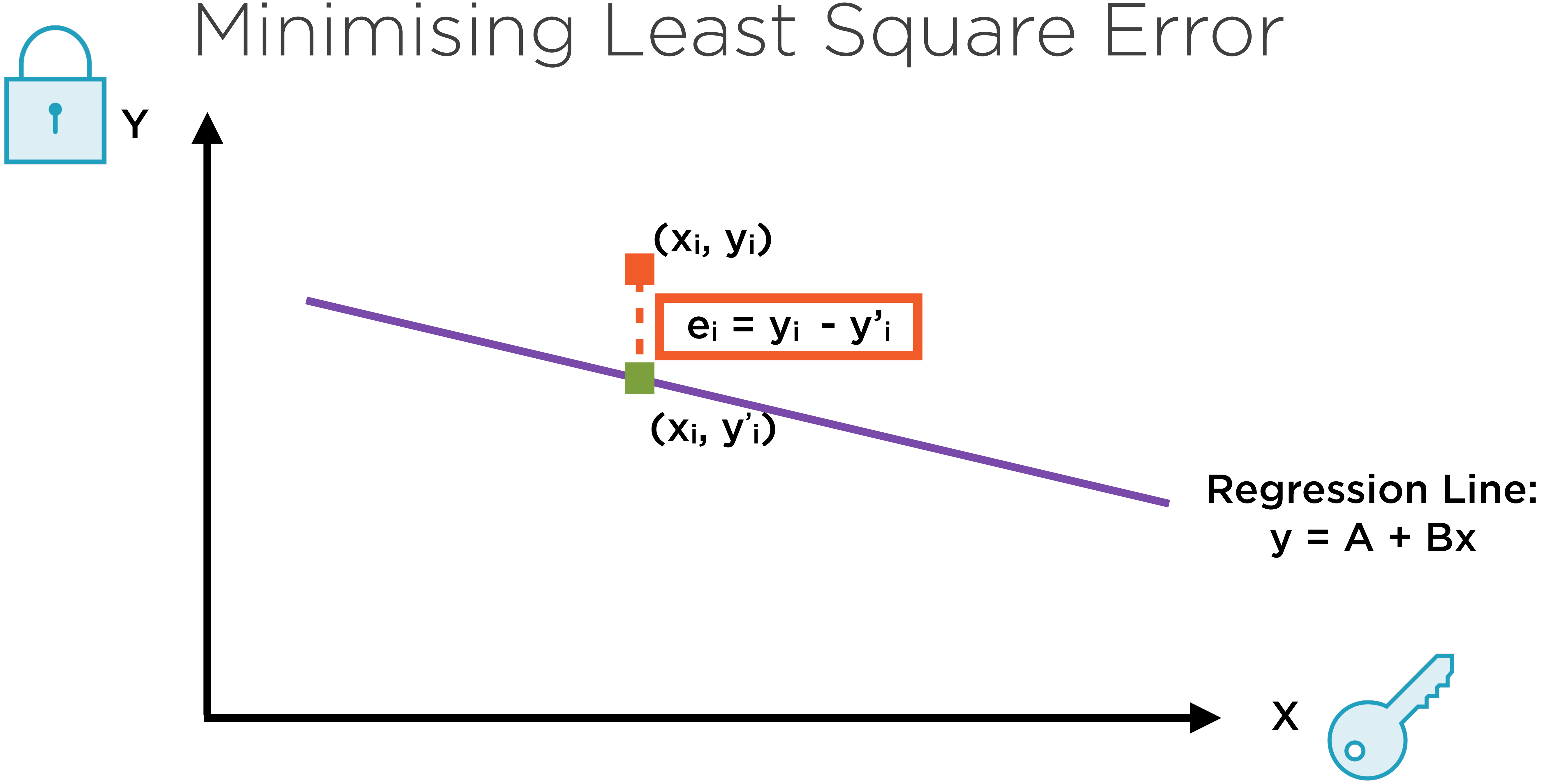
The “best fit” line is the one where the sum of the squares of the lengths of the errors is minimum

Minimising Least Square Error



The “best fit” line is the one where the sum of the squares of the lengths of the errors is minimum

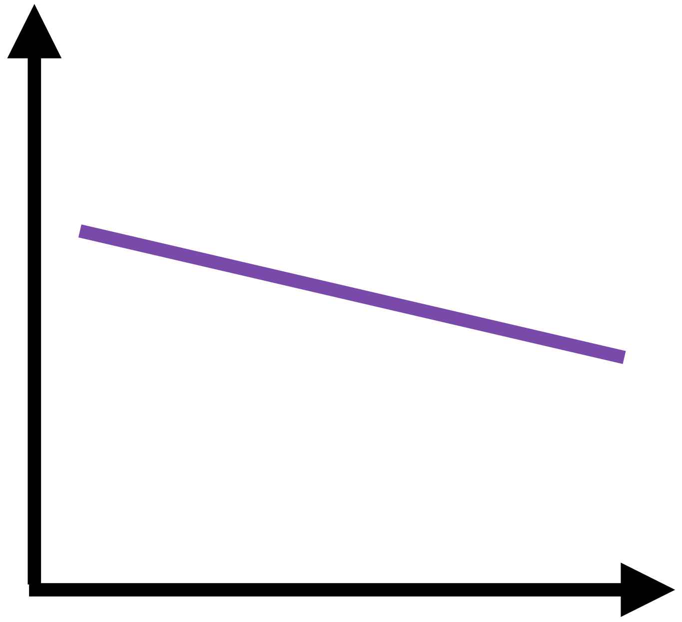
Minimising Least Square Error



Residuals of a regression are the difference between actual and fitted values of the dependent variable

The regression line is that line which minimizes the variance of the residuals (MSE)

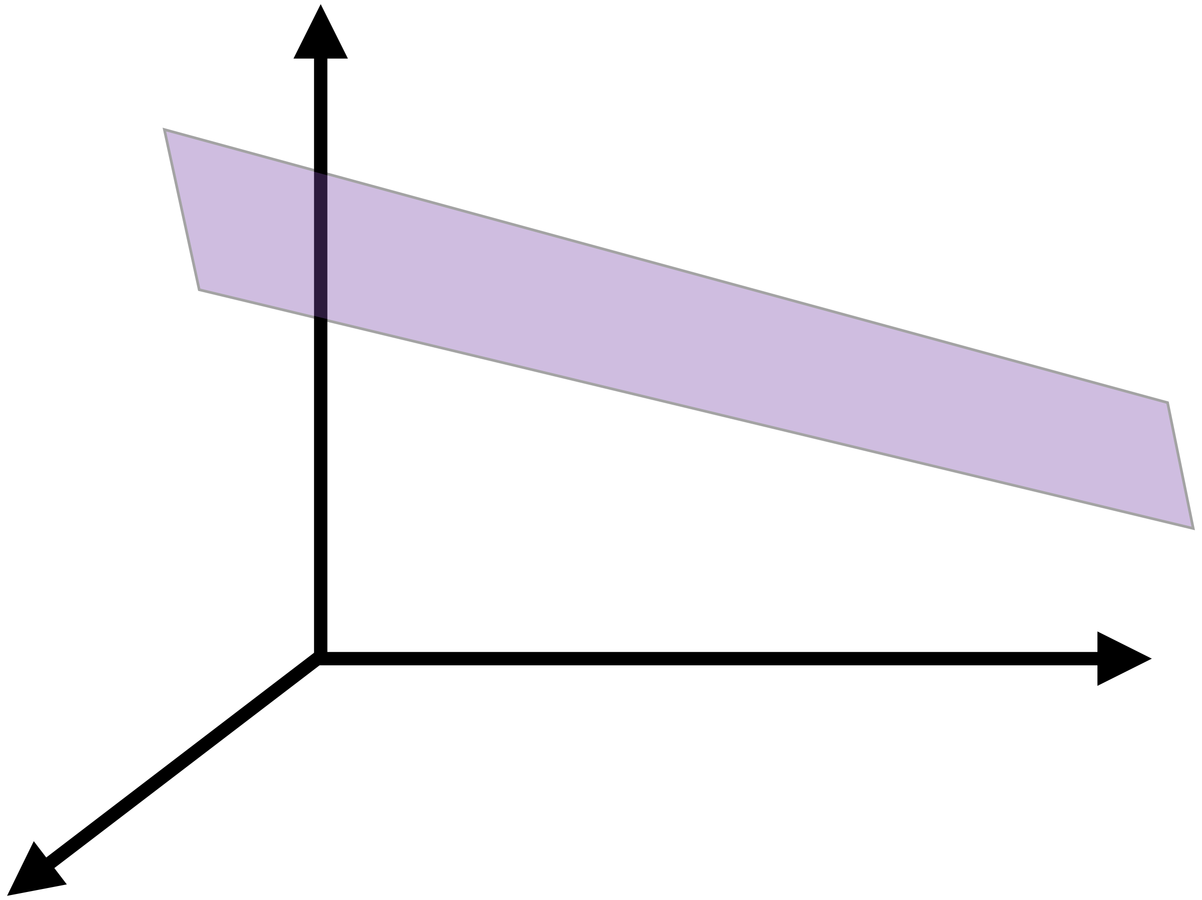
Simple and Multiple Regression



Simple Regression

One independent variable

$y = A + Bx$

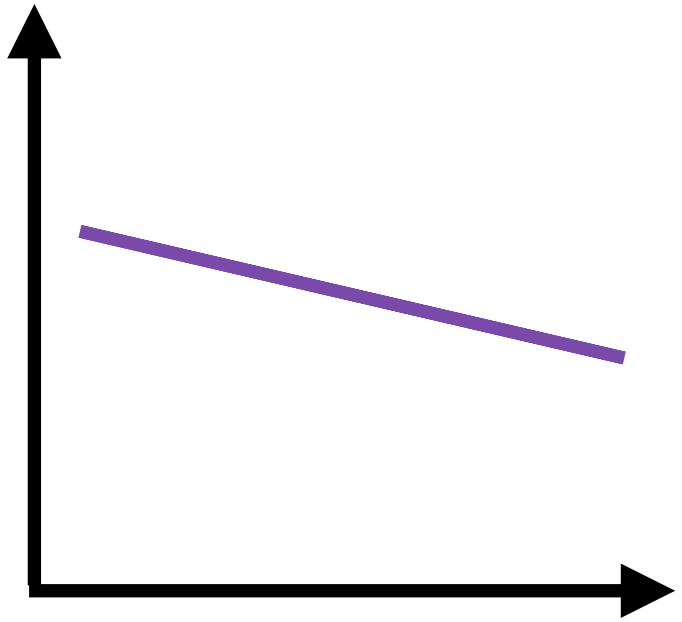


Multiple Regression

Multiple independent variables

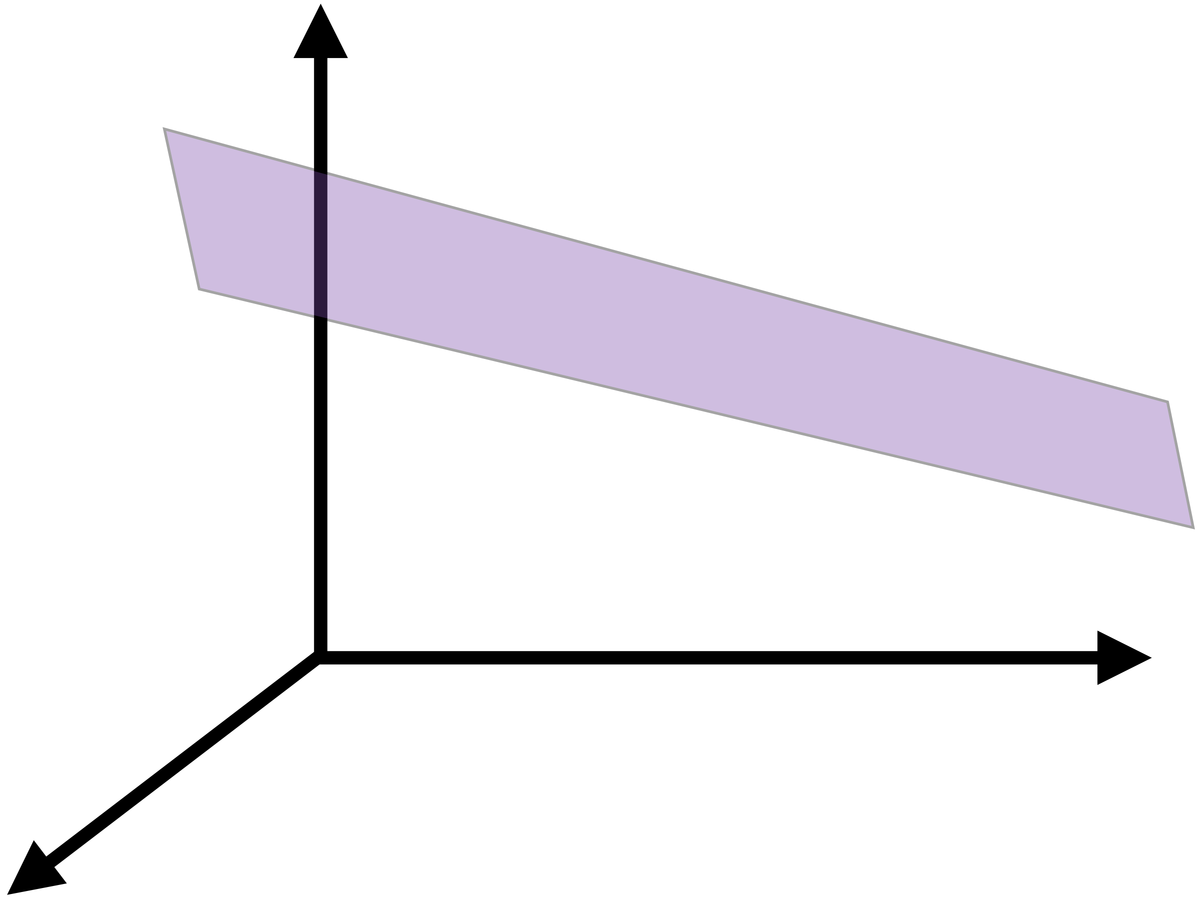
$y = A + B_1x_1 + B_2x_2 + B_3x_3$

MSE Minimization Extends To Multiple Regression



Simple Regression

One independent variable



Multiple Regression

Multiple independent variables

$R^2 = ESS / TSS$

R^2

R^2 = Explained Sum of Squares / Total Sum of Squares

R^2

ESS - Variance of fitted values

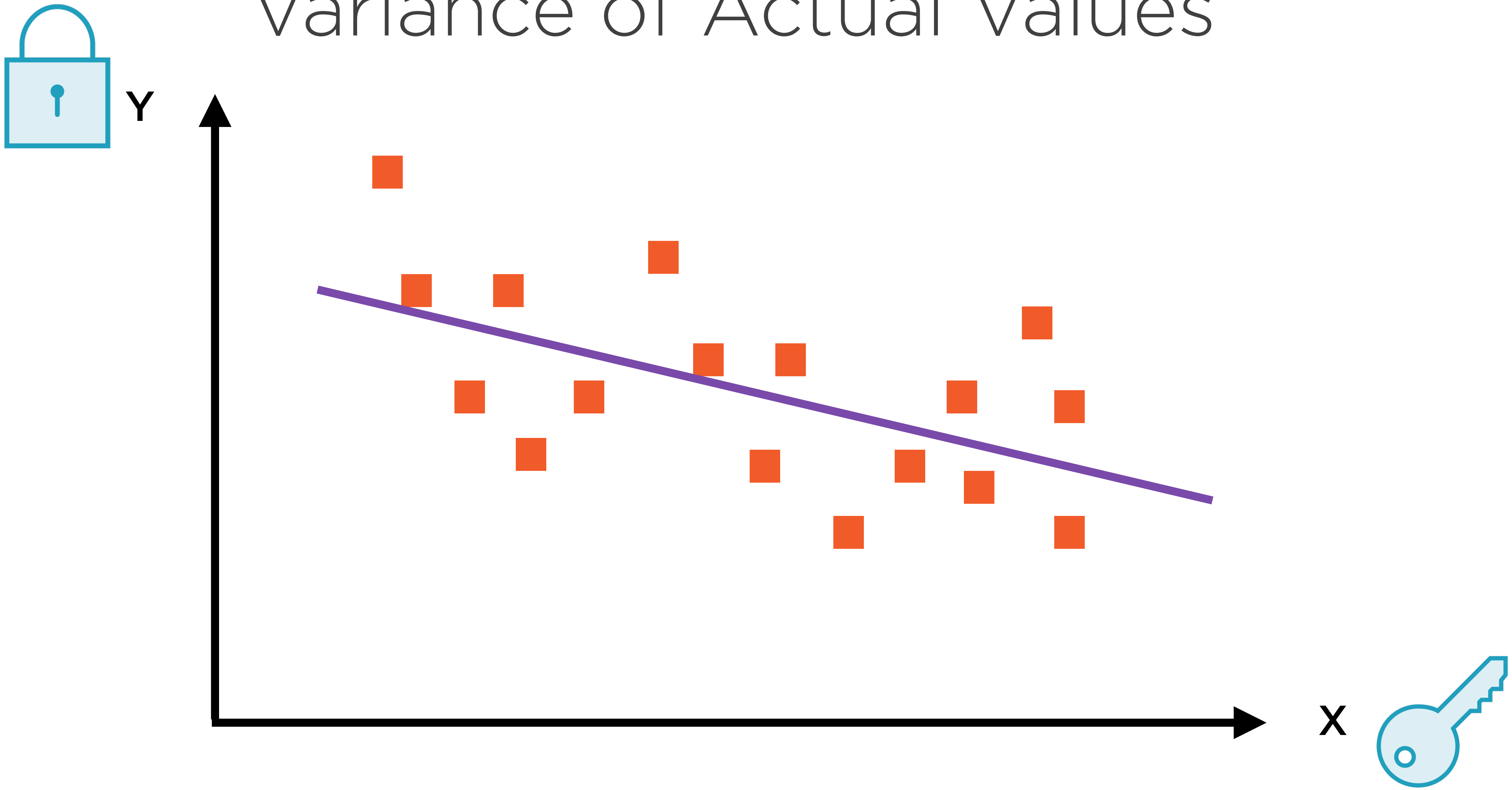
TSS - Variance of actual values

`R² = Explained Sum of Squares / Total Sum of Squares`

R^2

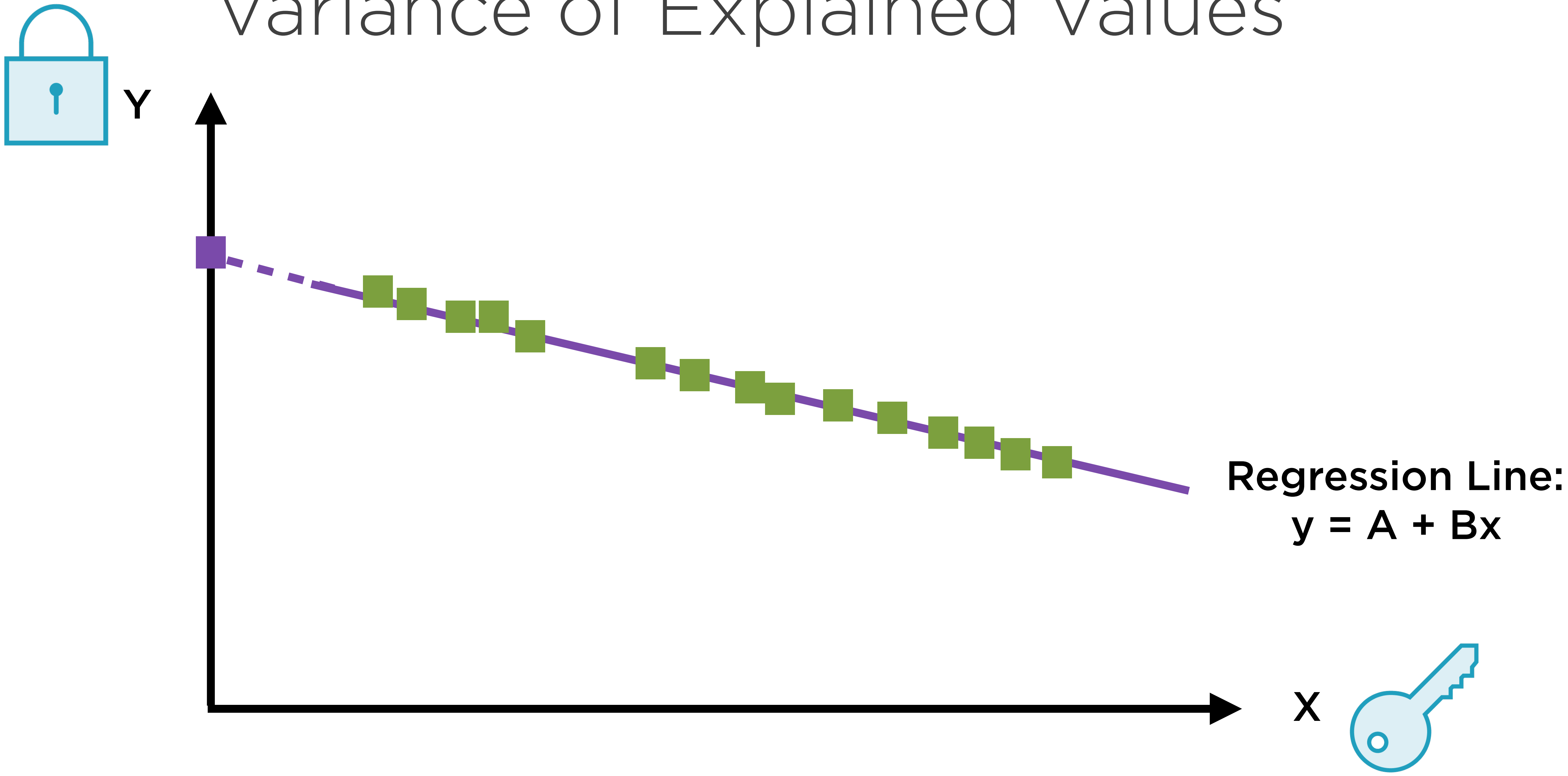
The percentage of total variance explained by the regression. Usually, the higher the R^2, the better the quality of the regression (upper bound is 100%)

Variance of Actual Values



The original data points have some variance (TSS)

Variance of Explained Values



The fitted data points have their own variance (ESS)

$$R^2 = ESS / TSS$$

R^2

How much of the original variance is captured in the fitted values?

Generally, higher this number the better the regression

Adjusted-R^2 = R^2 x (Penalty for adding irrelevant variables)

Adjusted-R^2

Increases if irrelevant* variables are deleted

(*irrelevant variables = any group whose F-ratio < 1)

The regression line found by minimizing variance of residuals (MSE) is the line with the **best R^2**

Demo

Performing linear regression with numeric features

Demo

Preprocessing numeric and categorical data and fitting a regression model

Choosing Regression Algorithms

Choosing Regression Algorithms

Size of Dataset

	Small	Medium	Large
Many			
Moderate			
Few			

Number of Features

Choosing Regression Algorithms

Size of Dataset

	Small	Medium	Large
Many			
Moderate			
Few			

Number of Features

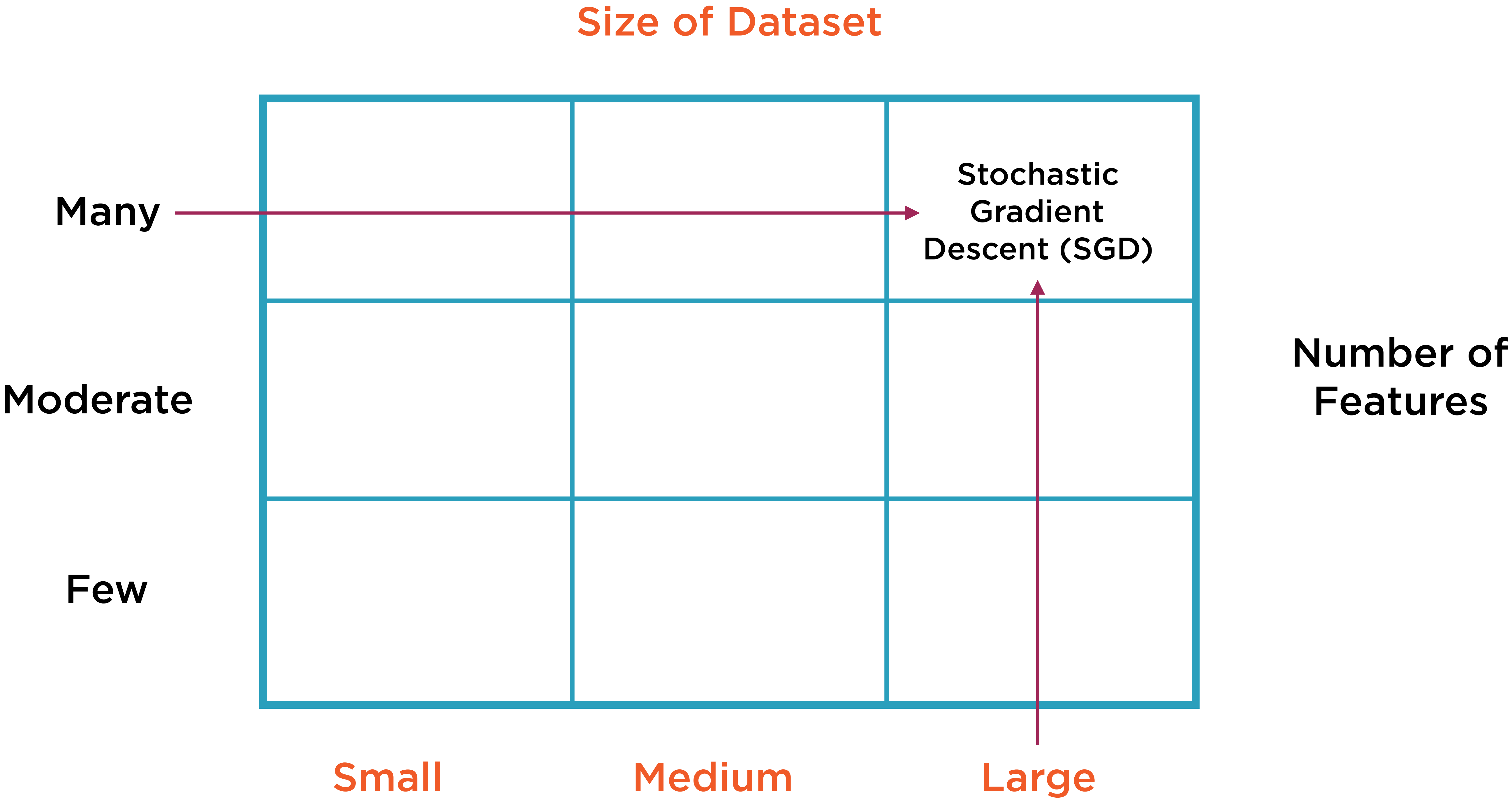
Choosing Regression Algorithms

Size of Dataset

	Small	Medium	Large
Many			
Moderate			
Few			

Number of Features

100K+ Data Points: Use SGD



Size of Dataset

Number of Features

	Small	Medium	Large
Many			Stochastic Gradient Descent (SGD)
Moderate			
Few			

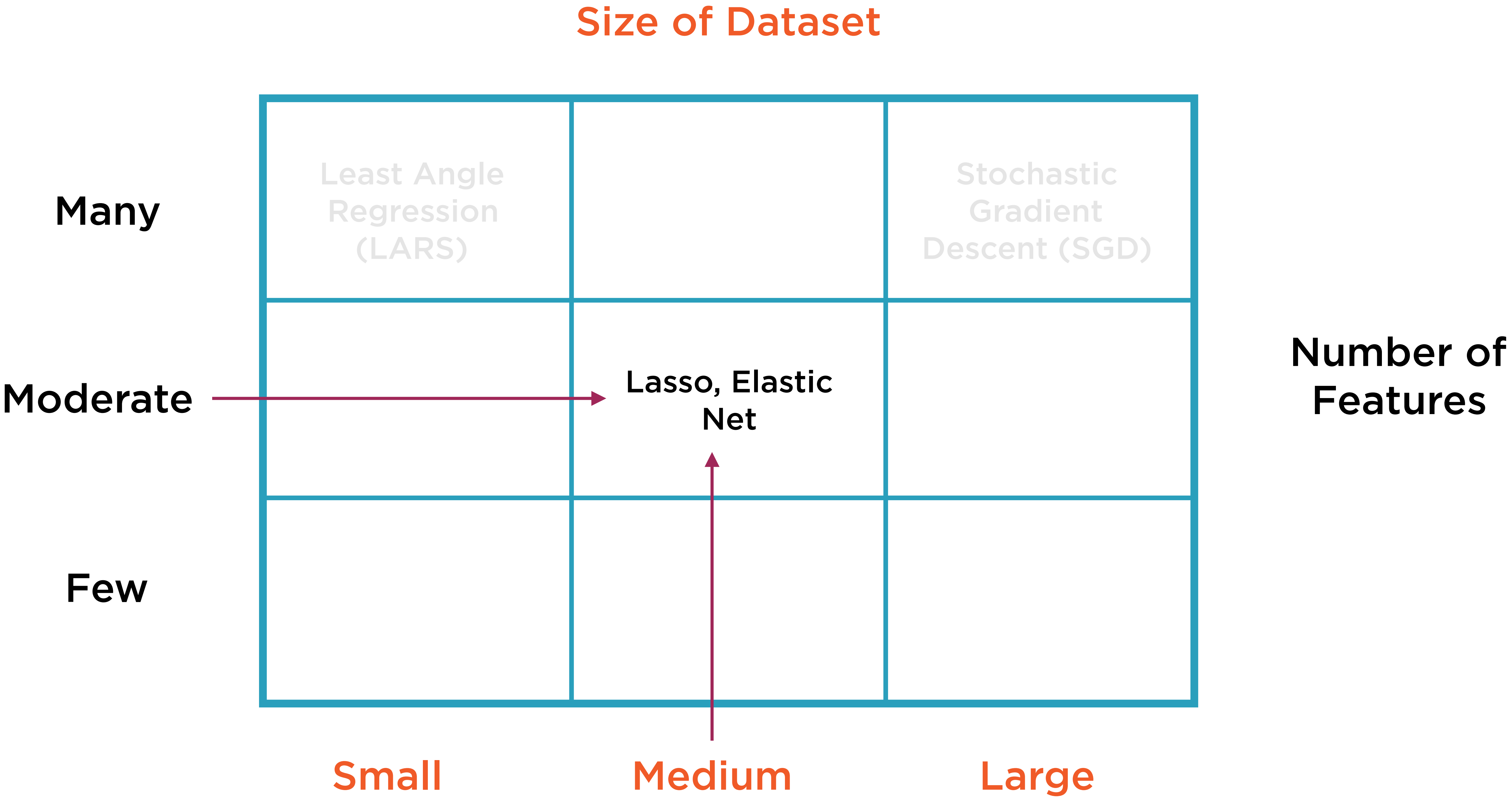
More Features Than Samples: Use LARS

Size of Dataset

	Small	Medium	Large
Many	**Least Angle Regression (LARS)**		Stochastic Gradient Descent (SGD)
Moderate			
Few			

Number of Features

Many Features, Few Useful: Lasso, ElasticNet

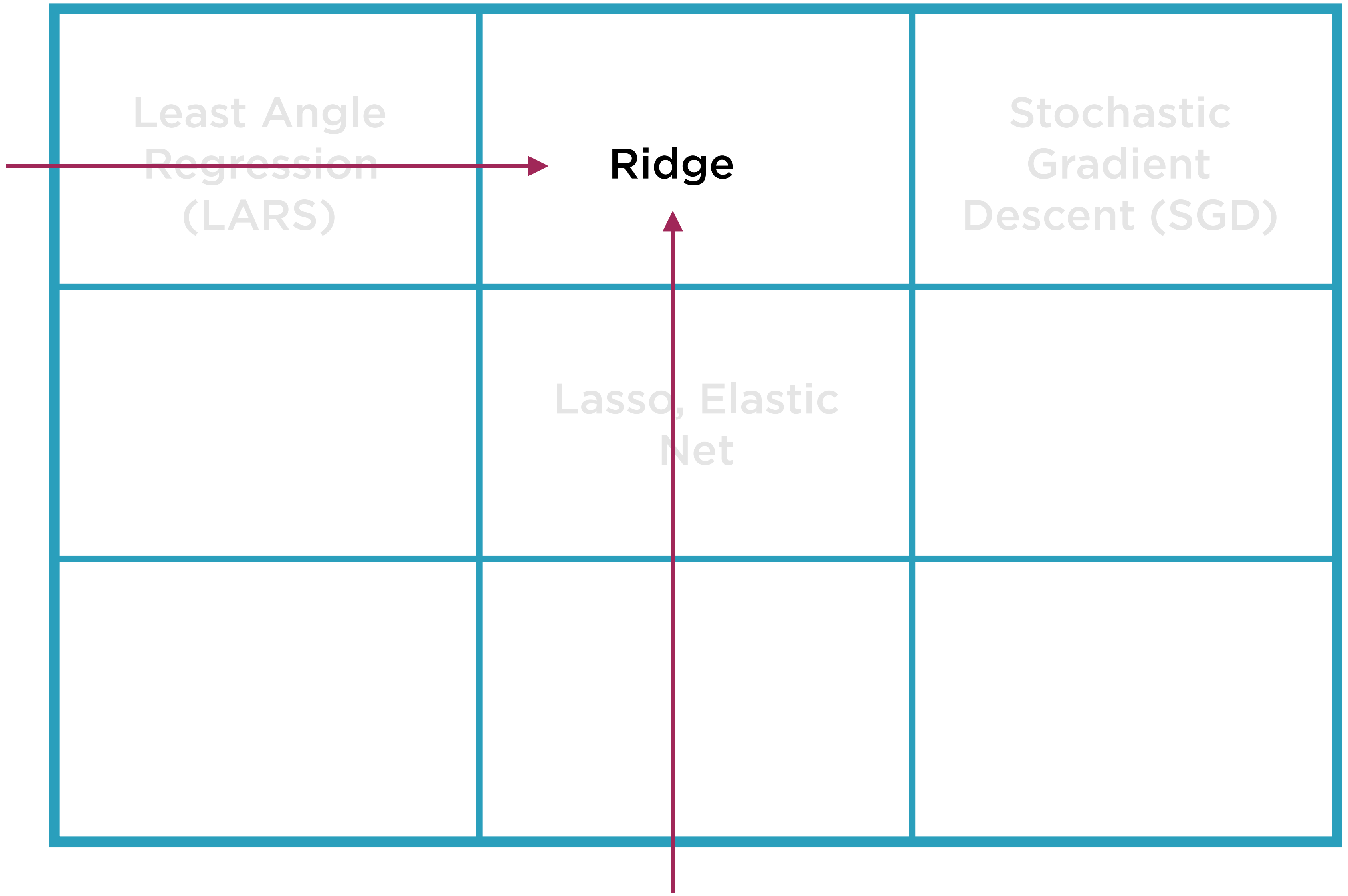


Size of Dataset

Number of Features

	Small	Medium	Large
Many	Least Angle Regression (LARS)		Stochastic Gradient Descent (SGD)
Moderate		Lasso, Elastic Net	
Few			

Many Features, Most Useful: Ridge



Size of Dataset

Number of Features

	Small	Medium	Large
Many	Least Angle Regression (LARS)	Ridge	Stochastic Gradient Descent (SGD)
Moderate		Lasso, Elastic Net	
Few			

Medium-sized Data with Non-linearity: SVR

Size of Dataset

	Small	Medium	Large
Many	Least Angle Regression (LARS)	Ridge	Stochastic Gradient Descent (SGD)
Moderate	**Support Vector Regression (Linear Kernel)**	Lasso, Elastic Net	
Few			

Number of Features

Small Data with Non-linearity: SVR with RBF

Size of Dataset

Number of Features

	Small	Medium	Large
Many	Least Angle Regression (LARS)	Ridge	Stochastic Gradient Descent (SGD)
Moderate	Support Vector Regression (Linear Kernel)	Lasso, Elastic Net	Support Vector Regression (Linear Kernel)
Few	**Support Vector Regression (RBF Kernel)**		

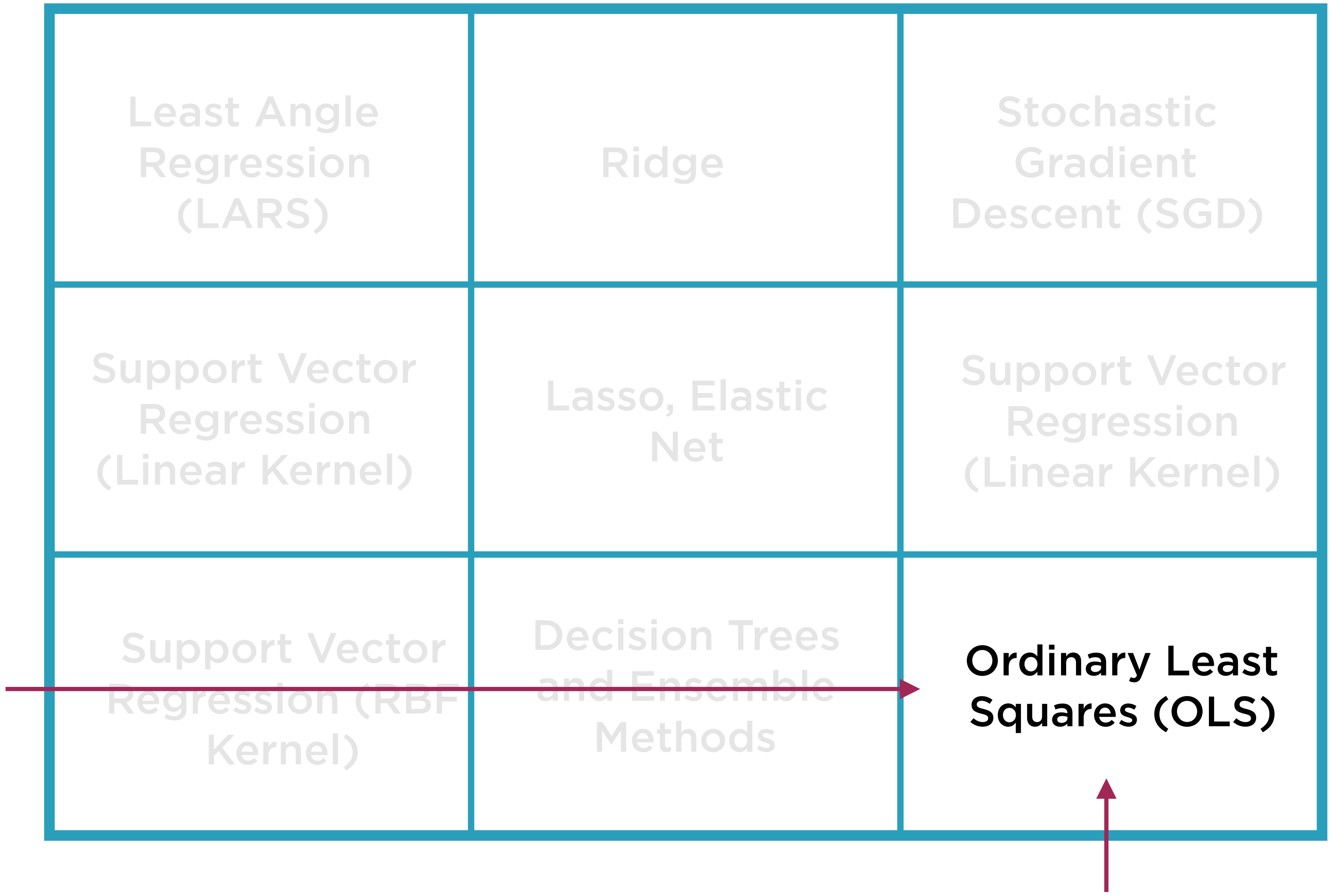
Many Features, Few Useful: Decision Trees

Size of Dataset

	Small	Medium	Large
Many	Least Angle Regression (LARS)	Ridge	Stochastic Gradient Descent (SGD)
Moderate	Support Vector Regression (Linear Kernel)	Lasso, Elastic Net	Support Vector Regression (Linear Kernel)
Few	Support Vector Regression (RBF Kernel)	**Decision Trees and Ensemble Methods**	

Number of Features

Many Samples, Few Features: OLS



Size of Dataset

Number of Features

	Small	Medium	Large
Many	Least Angle Regression (LARS)	Ridge	Stochastic Gradient Descent (SGD)
Moderate	Support Vector Regression (Linear Kernel)	Lasso, Elastic Net	Support Vector Regression (Linear Kernel)
Few	Support Vector Regression (RBF Kernel)	Decision Trees and Ensemble Methods	Ordinary Least Squares (OLS)

Choosing Regression Algorithms

Size of Dataset

	Small	Medium	Large
Many	Least Angle Regression (LARS)	Ridge	Stochastic Gradient Descent (SGD)
Moderate	Support Vector Regression (Linear Kernel)	Lasso, Elastic Net	Support Vector Regression (Linear Kernel)
Few	Support Vector Regression (RBF Kernel)	Decision Trees and Ensemble Methods	Ordinary Least Squares (OLS)

Number of Features

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 still define the “best fit” line

$$y = A + Bx$$

Lasso Regression

Minimize

$$\sqrt{(y^{actual} - y^{predicted})^2} + \boxed{\alpha \; (|A| + |B|)}$$

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

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

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

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)

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

Model selection by tuning α

Eliminates unimportant features

Lasso Regression

“Lasso” ~ *L*east *A*bsolute *S*hrinkage and *S*election *O*perator

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

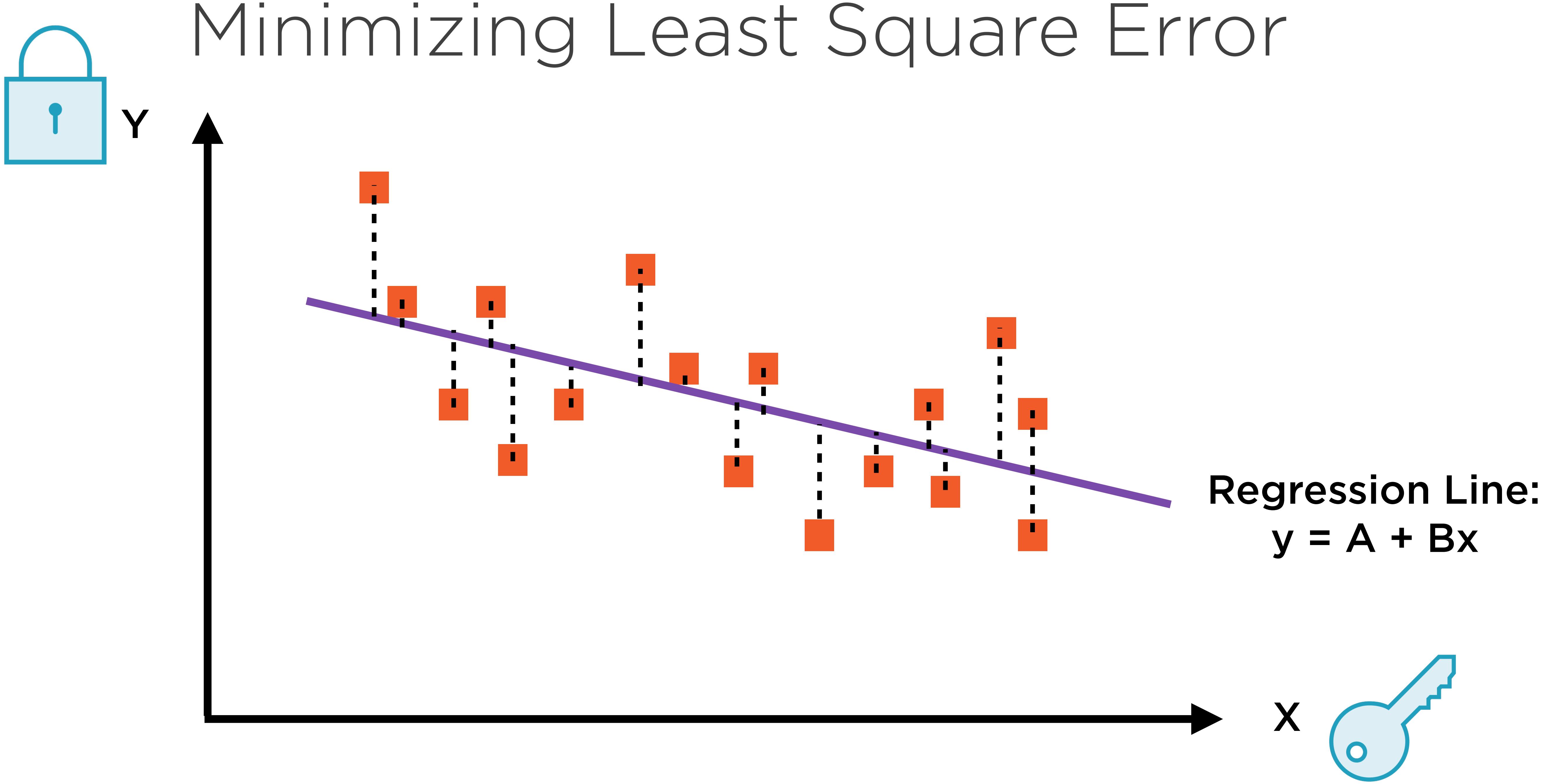
Also penalizes large regression coefficients

Elastic Net Regression

Simply combines lasso and ridge

SGD Regression

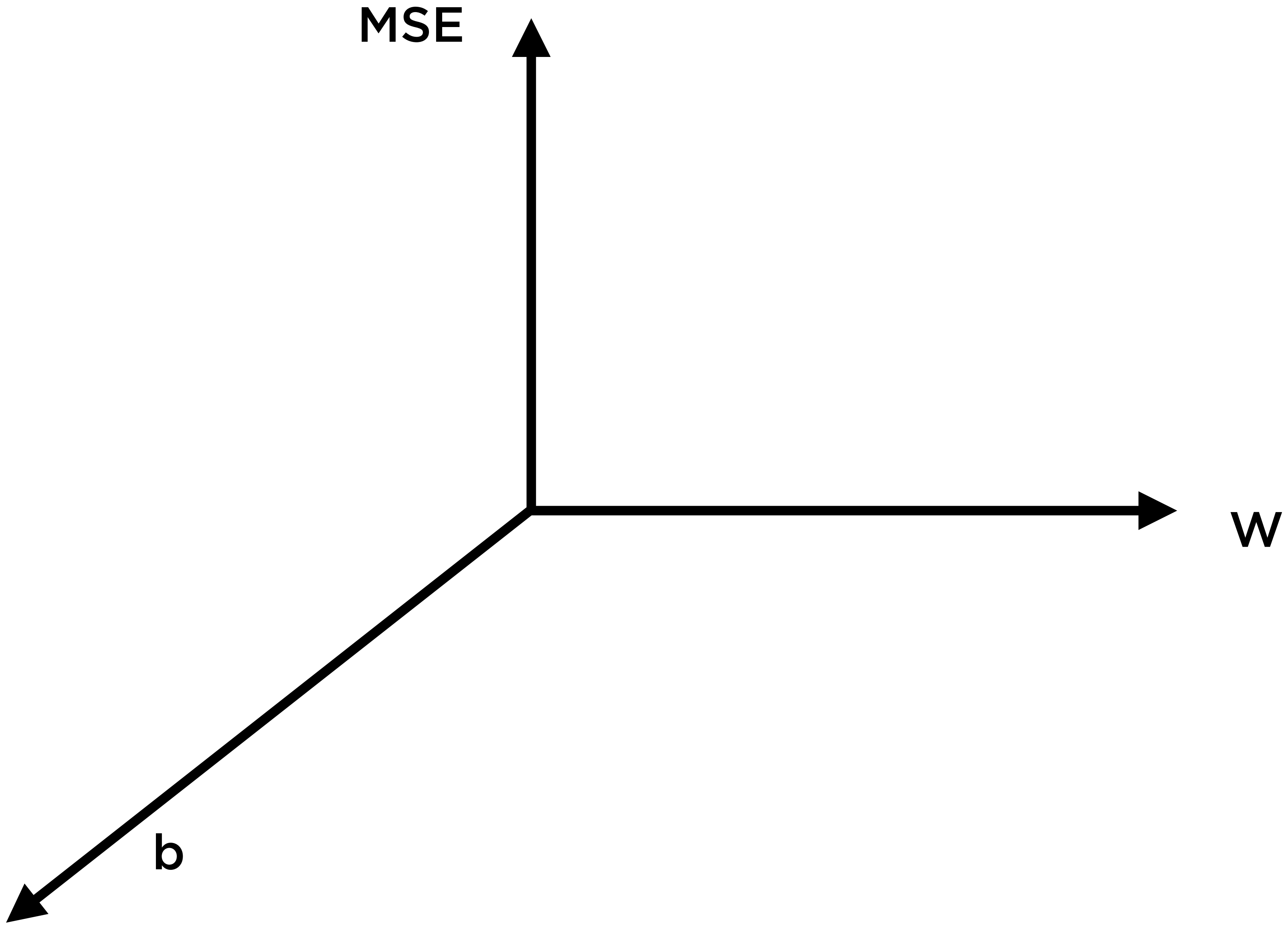
Minimizing Least Square Error



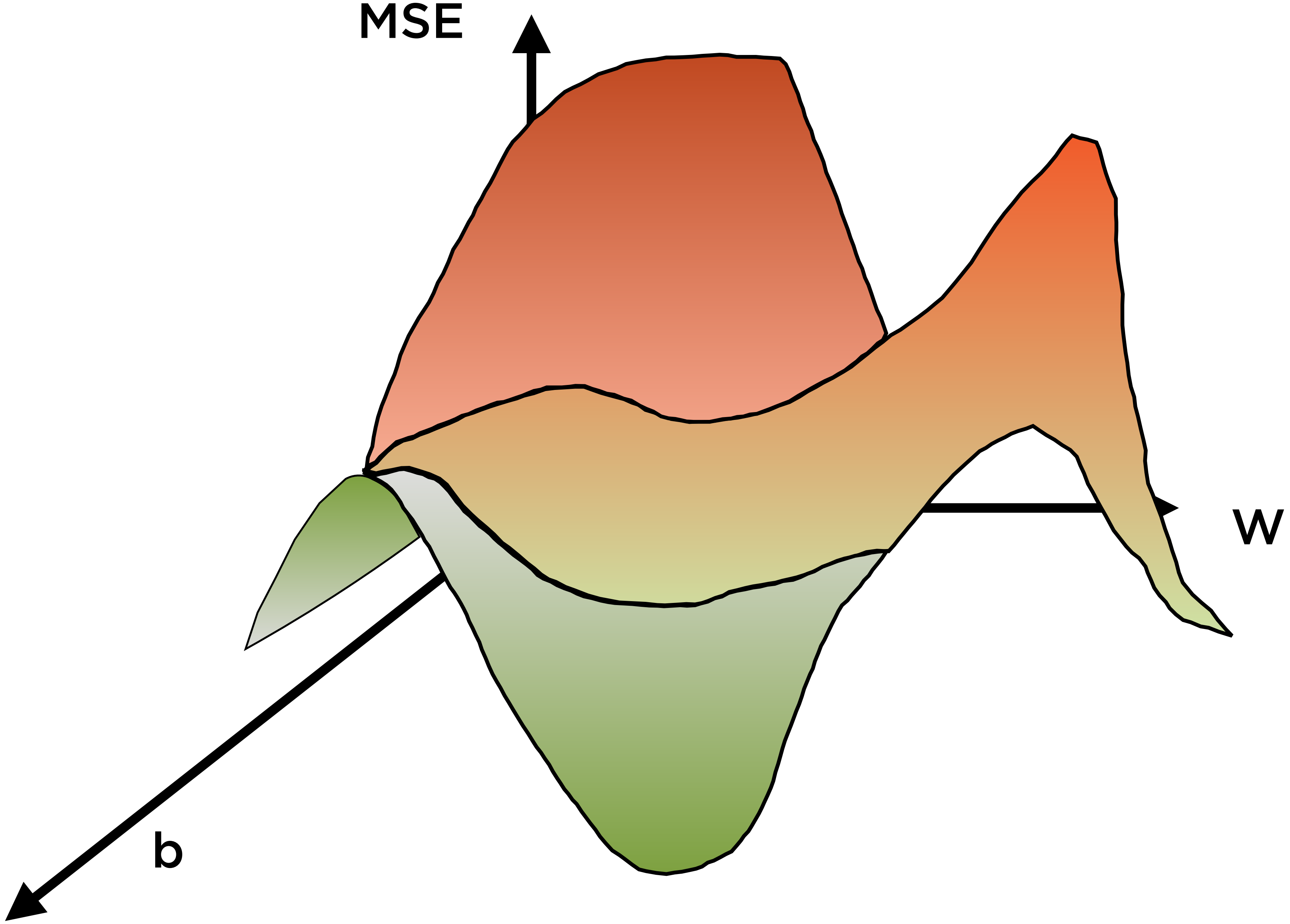
Regression Line:
$y = A + Bx$

The “best fit” line is called the regression line

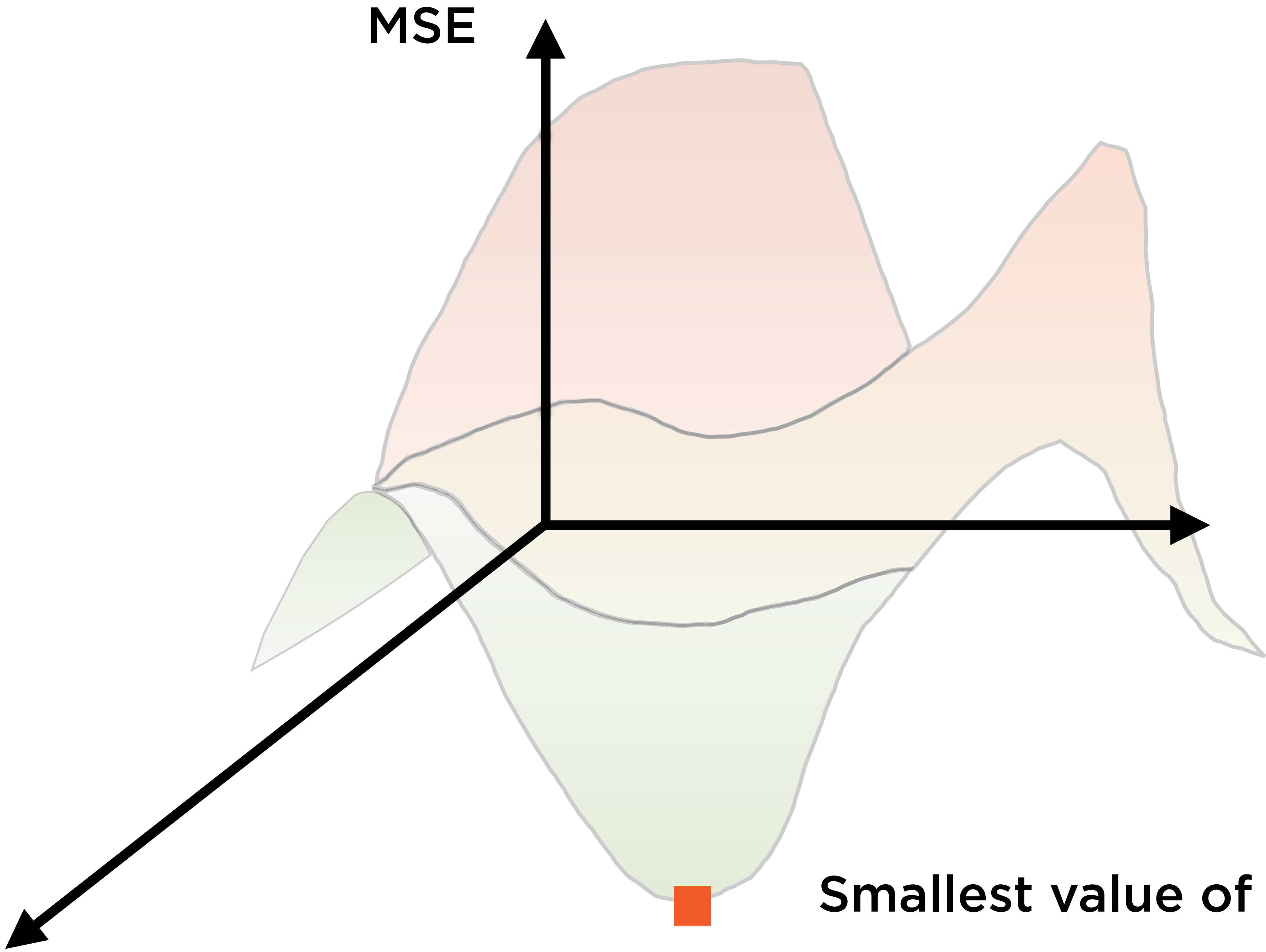
Minimizing MSE



Minimizing MSE

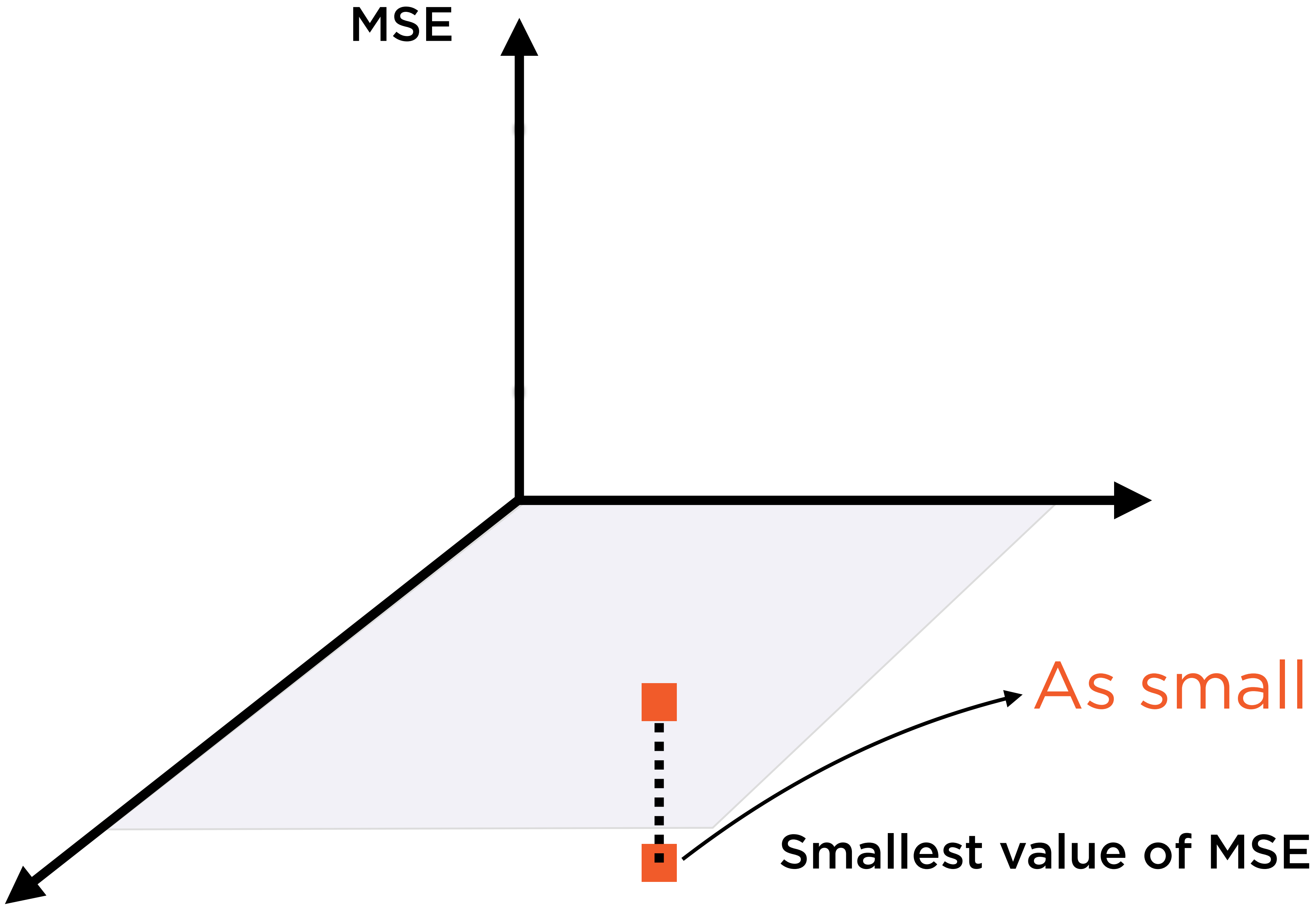


Minimizing MSE



Smallest value of MSE

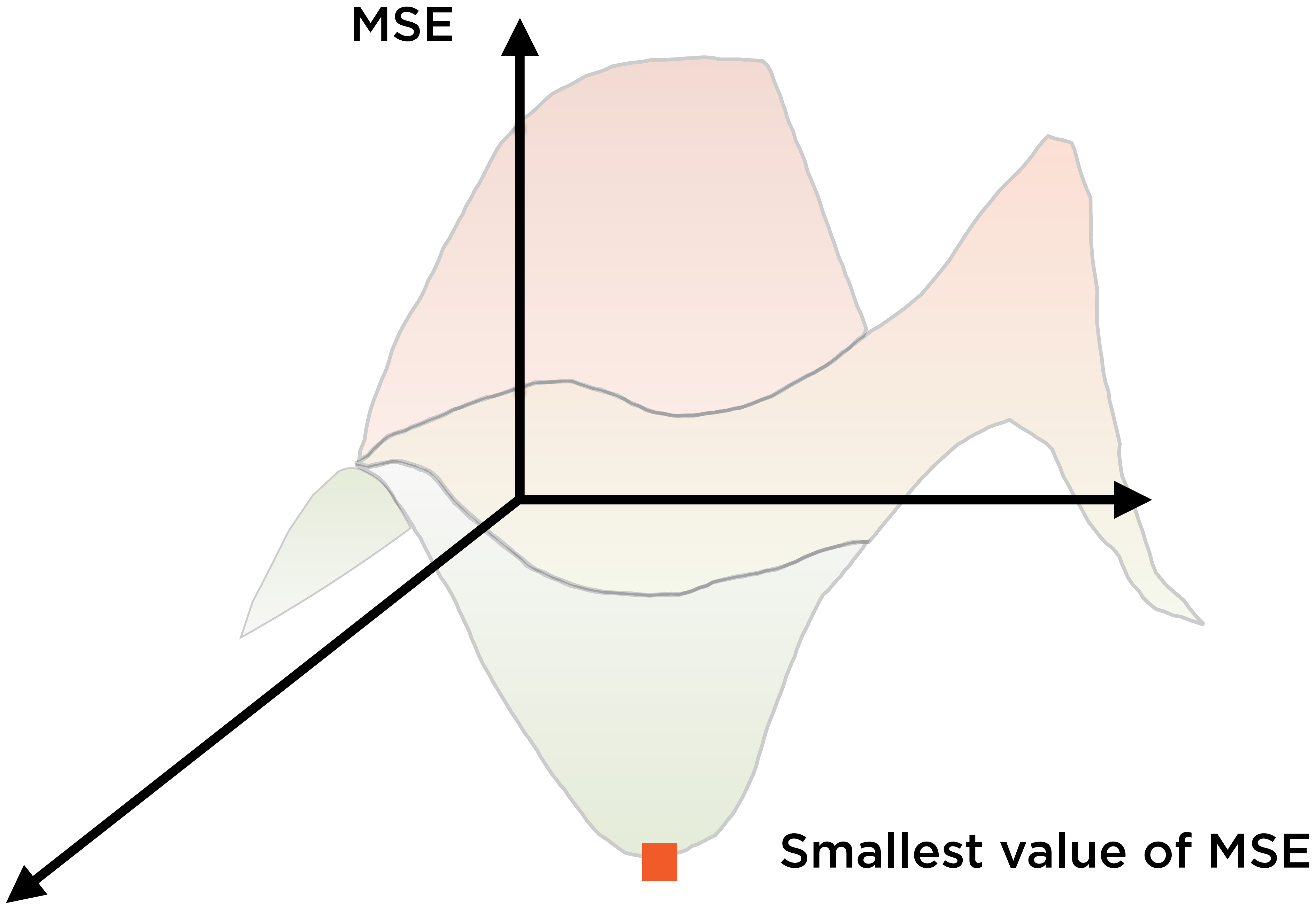
Minimizing MSE



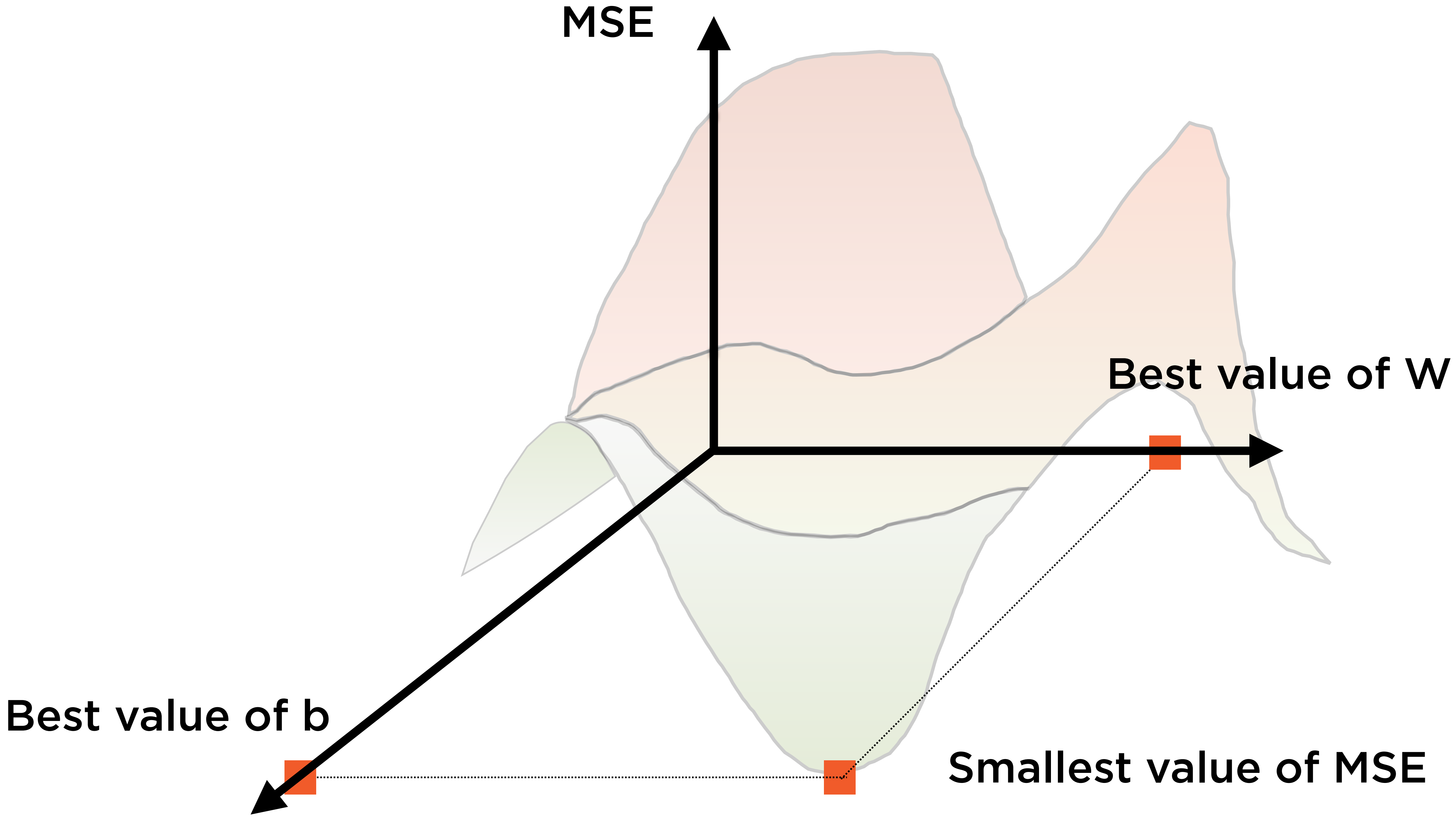
As small as possible!

Smallest value of MSE

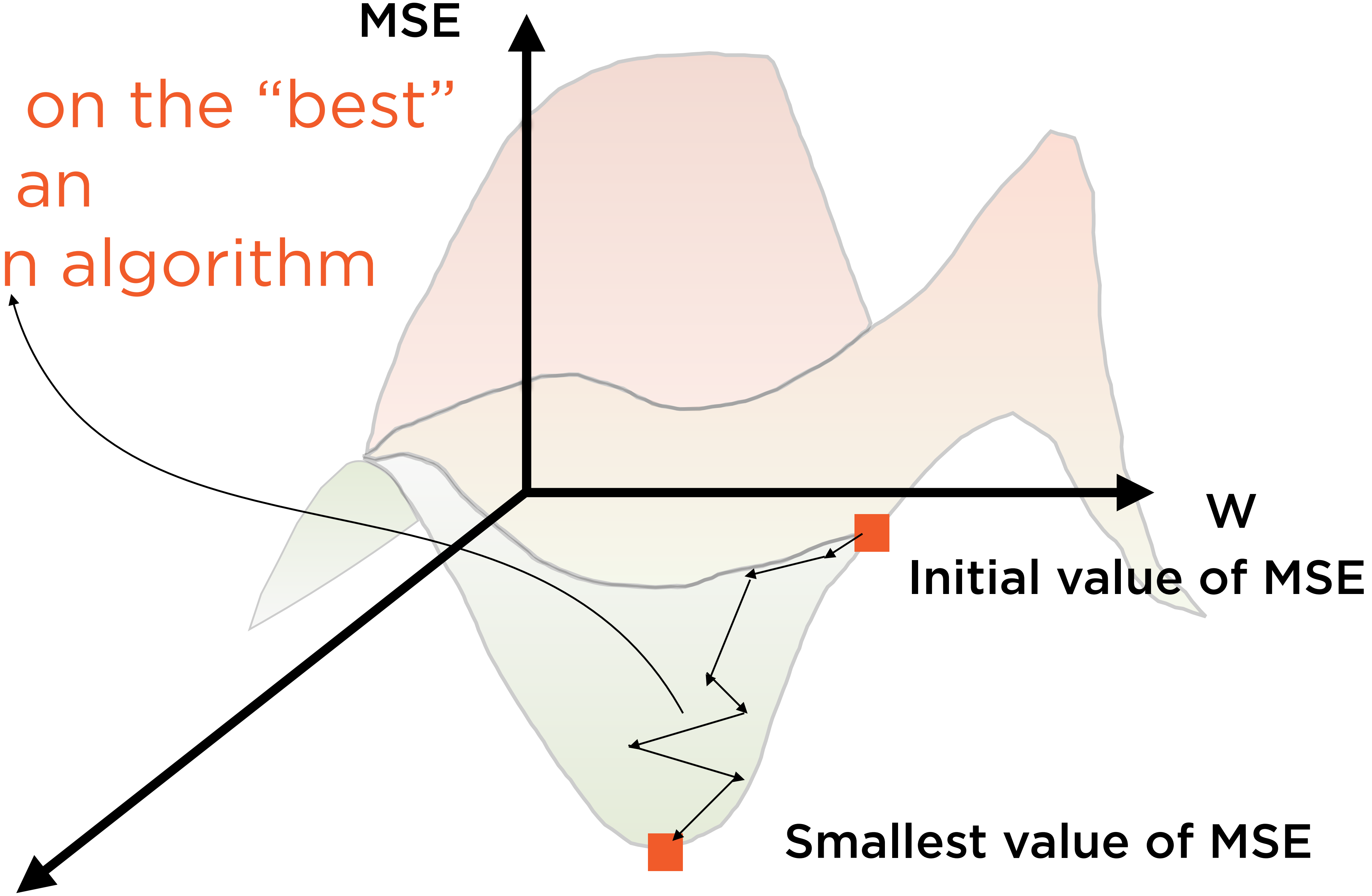
Minimizing MSE



Minimizing MSE

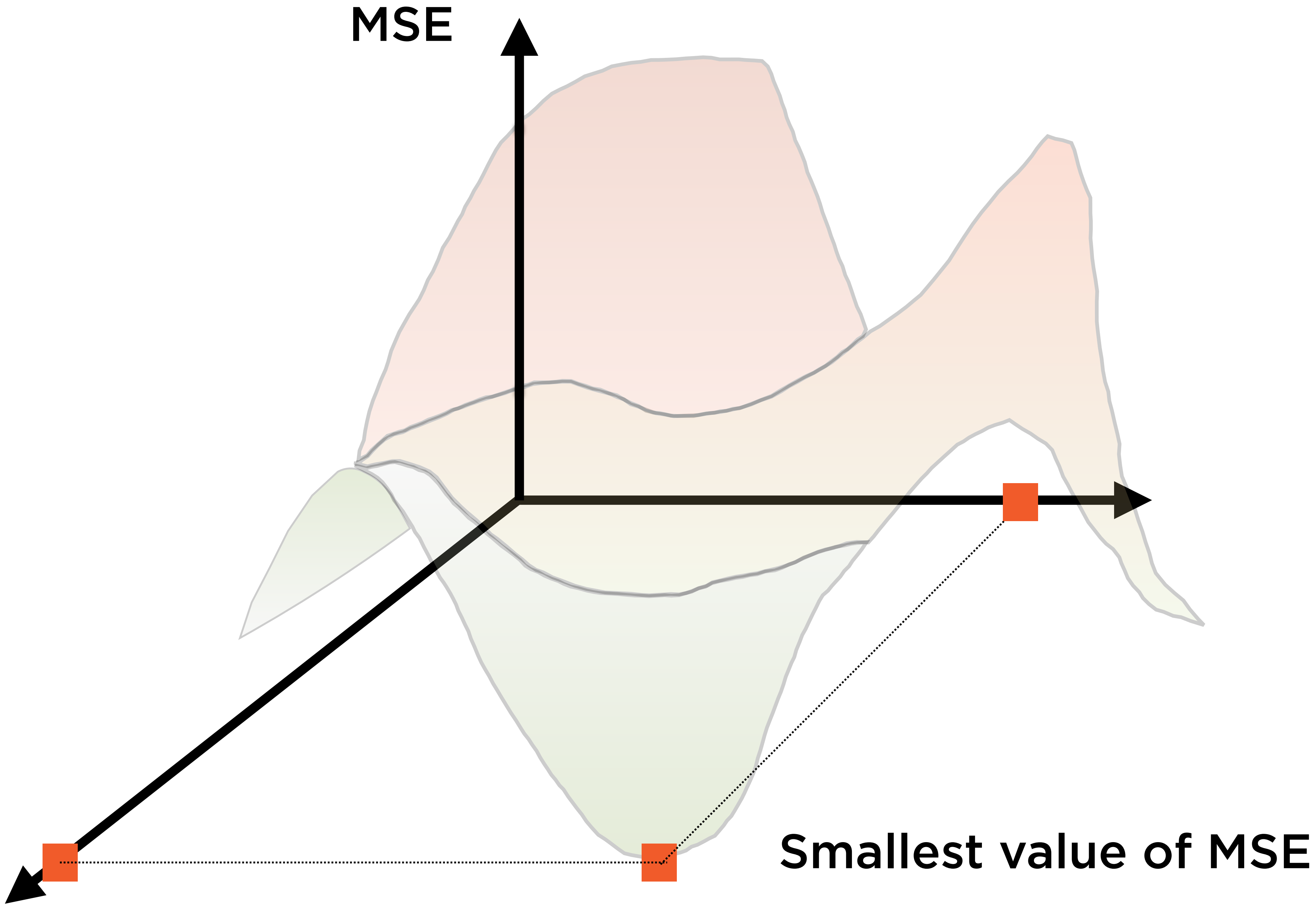


“Gradient Descent”

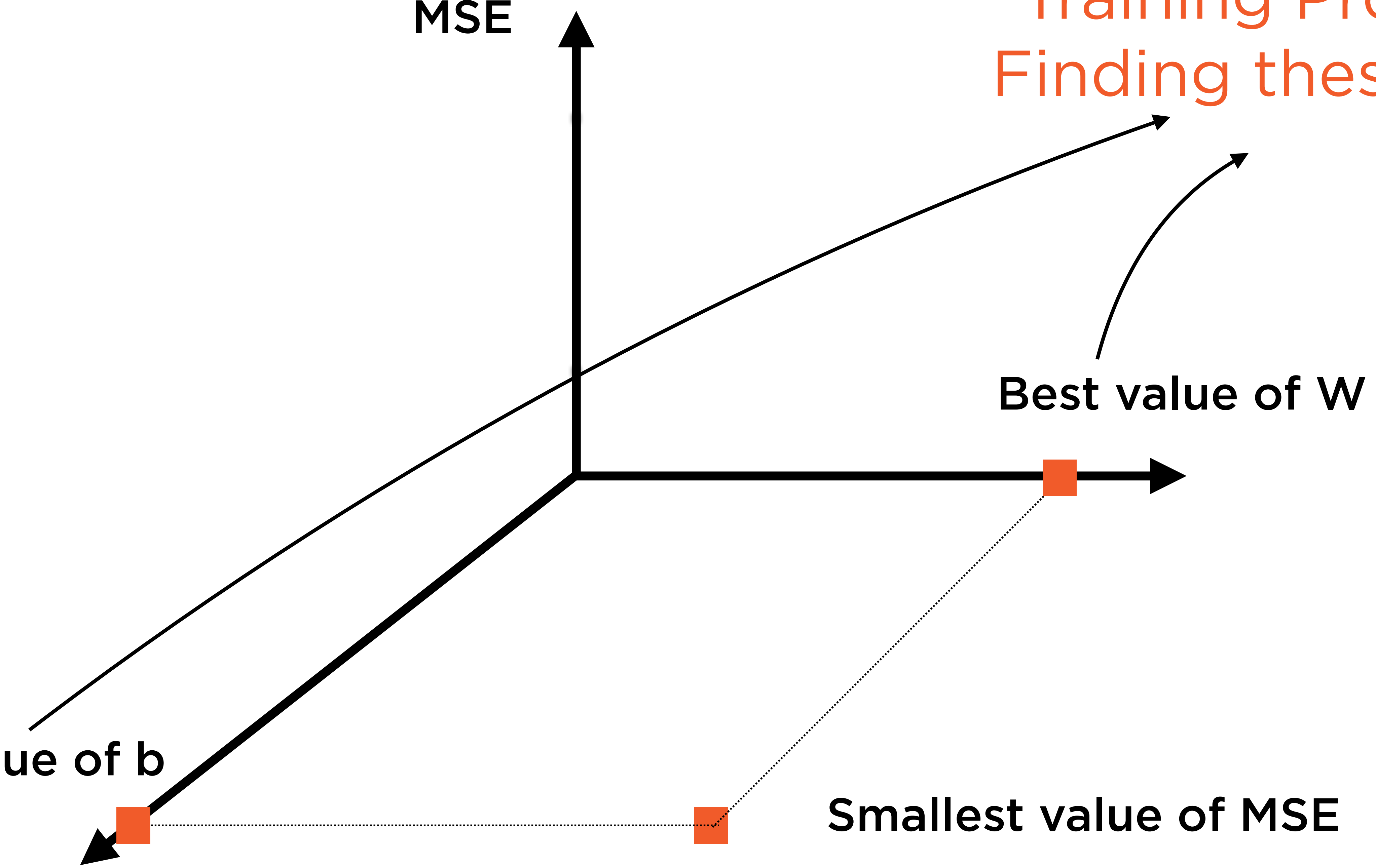


Converging on the “best” value using an optimization algorithm

Minimizing MSE



“Training” the Algorithm



MSE

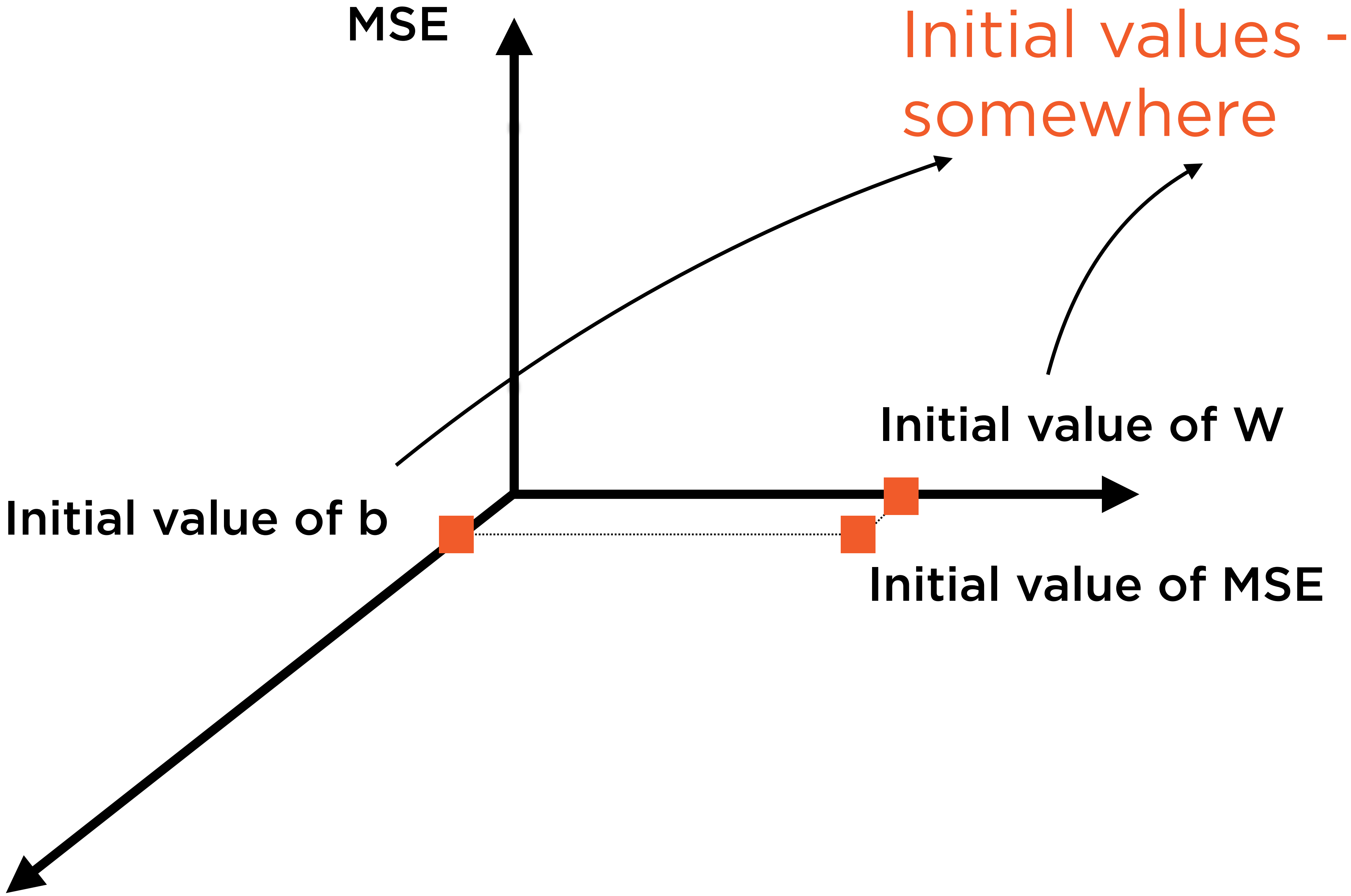
“Training Process” =
Finding these best values

Best value of W

Best value of b

Smallest value of MSE

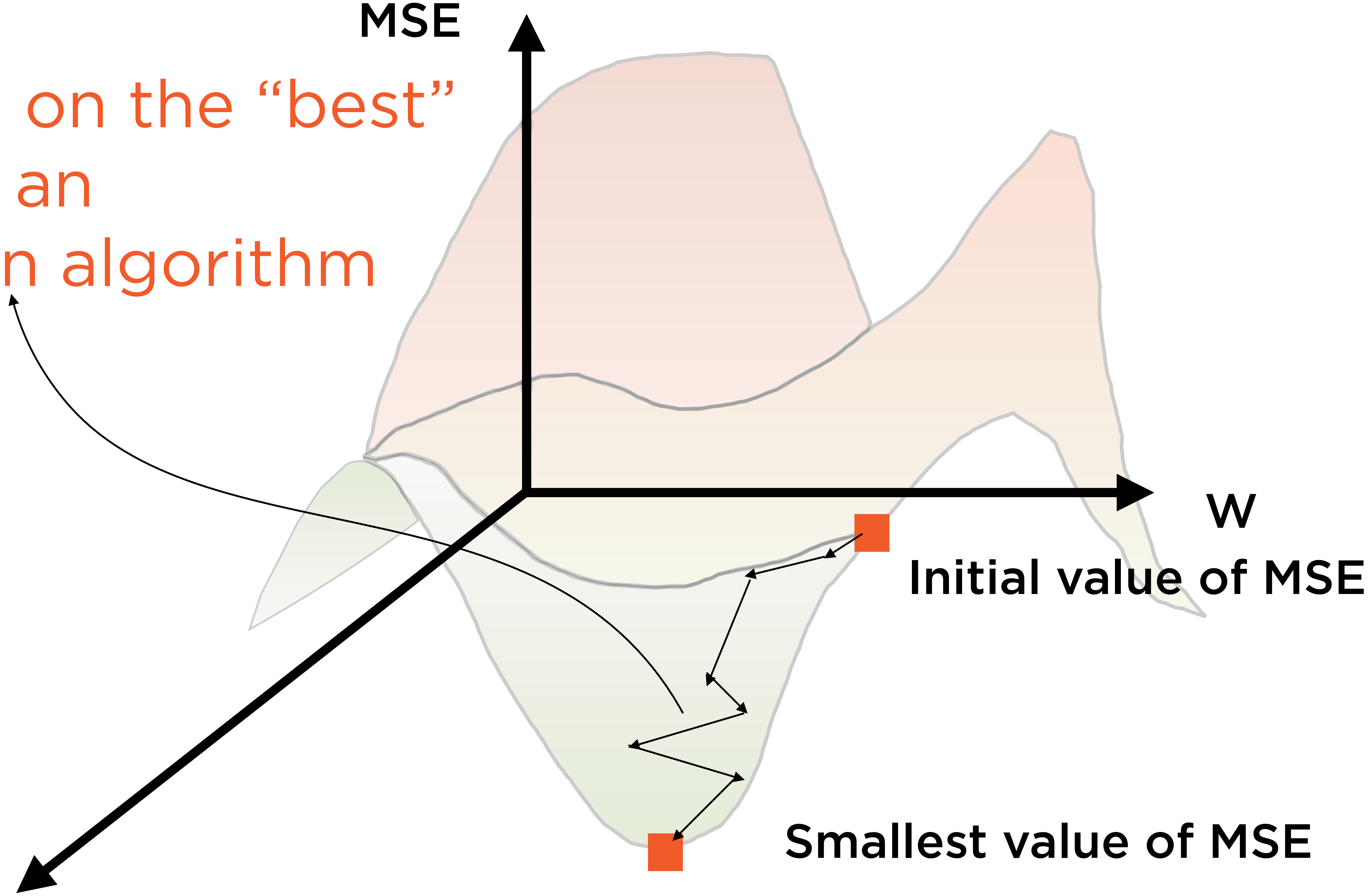
Start Somewhere



Initial values - have to start somewhere

"Gradient Descent"

Converging on the "best" value using an optimization algorithm



Stochastic Gradient Descent **iteratively** converges to the best model

Works very well for training on large datasets

Demo

Performing regression using multiple techniques such as Lasso, Ridge, and Stochastic Gradient Descent

Summary

Regression as a form of supervised machine learning

Ordinary Least Squares (OLS) regression

Evaluating regression models using R^2

Choosing the right regression algorithm based on features and data

Lasso and Ridge regression

Gradient Descent in regression