Understanding Simple Regression Models

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Set up the regression problem and describe its solution

Introduce simple regression models that have a single explanatory variable

Use simple regression models

- to explain variance
- to make forecasts

Understand the assumptions underlying regression

Setting Up The Regression Problem

X Causes Y

Cause

Independent variable

Effect

Dependent variable

X Causes Y

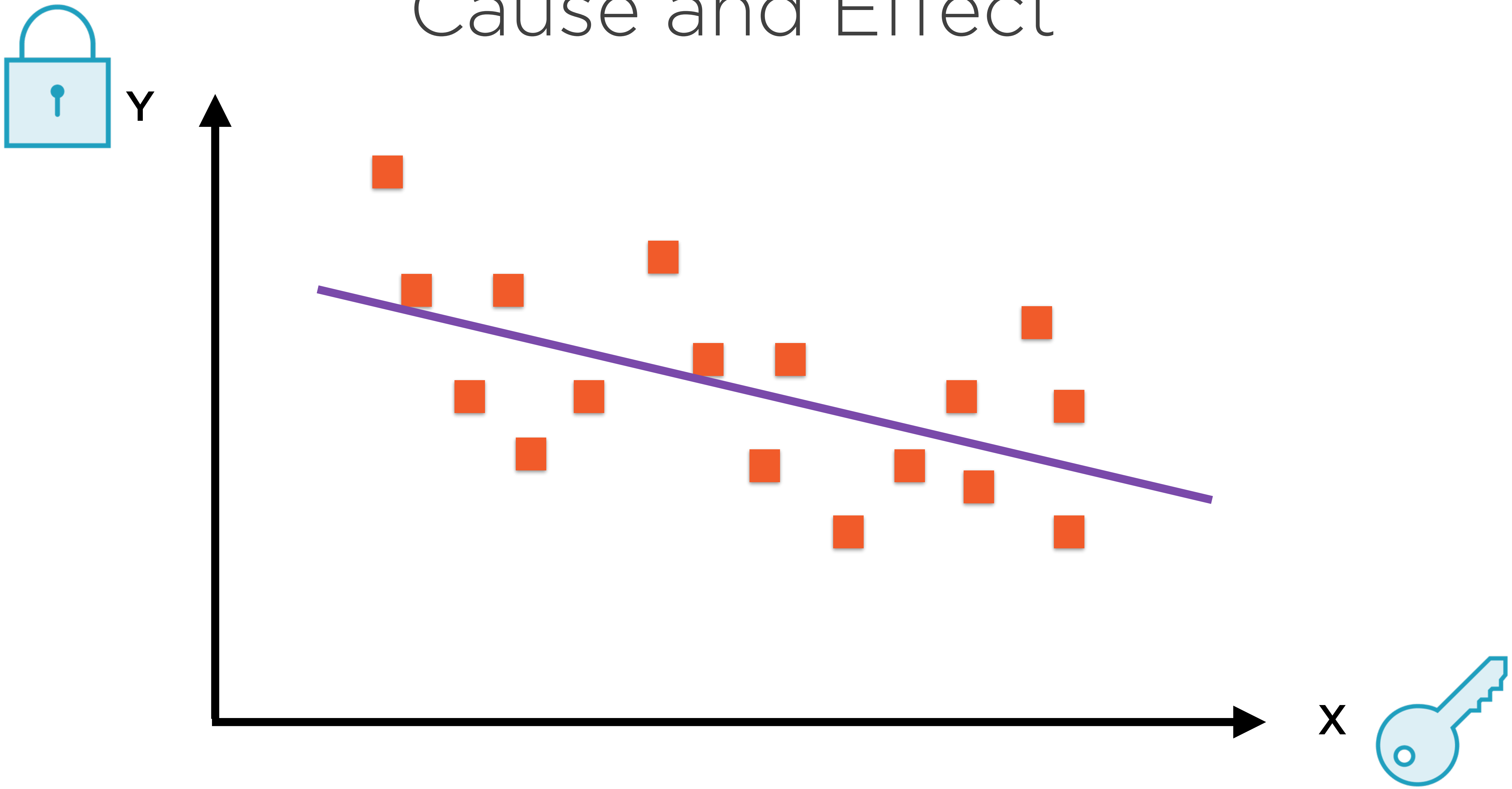
Cause

Explanatory variable

Effect

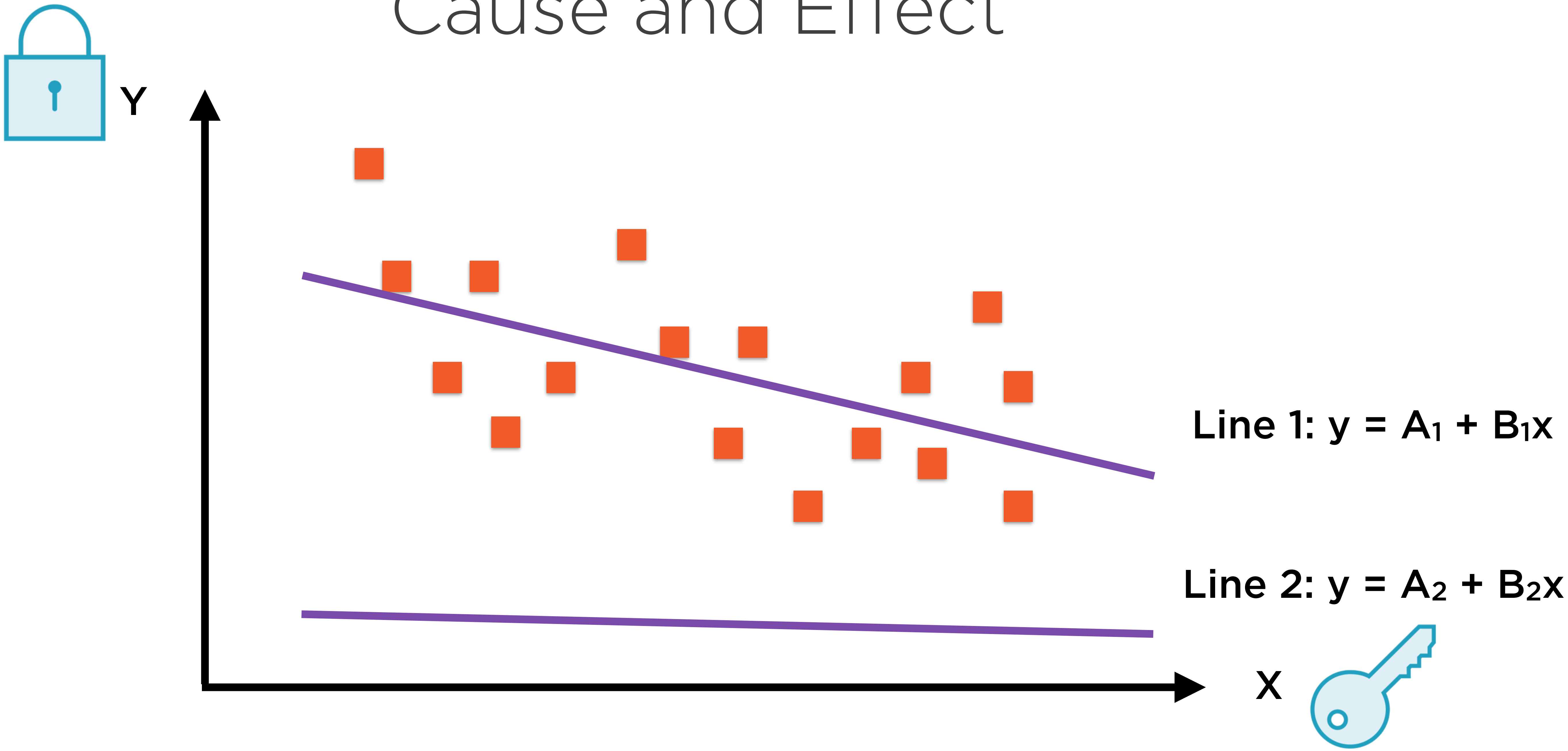
Dependent variable

Cause and Effect



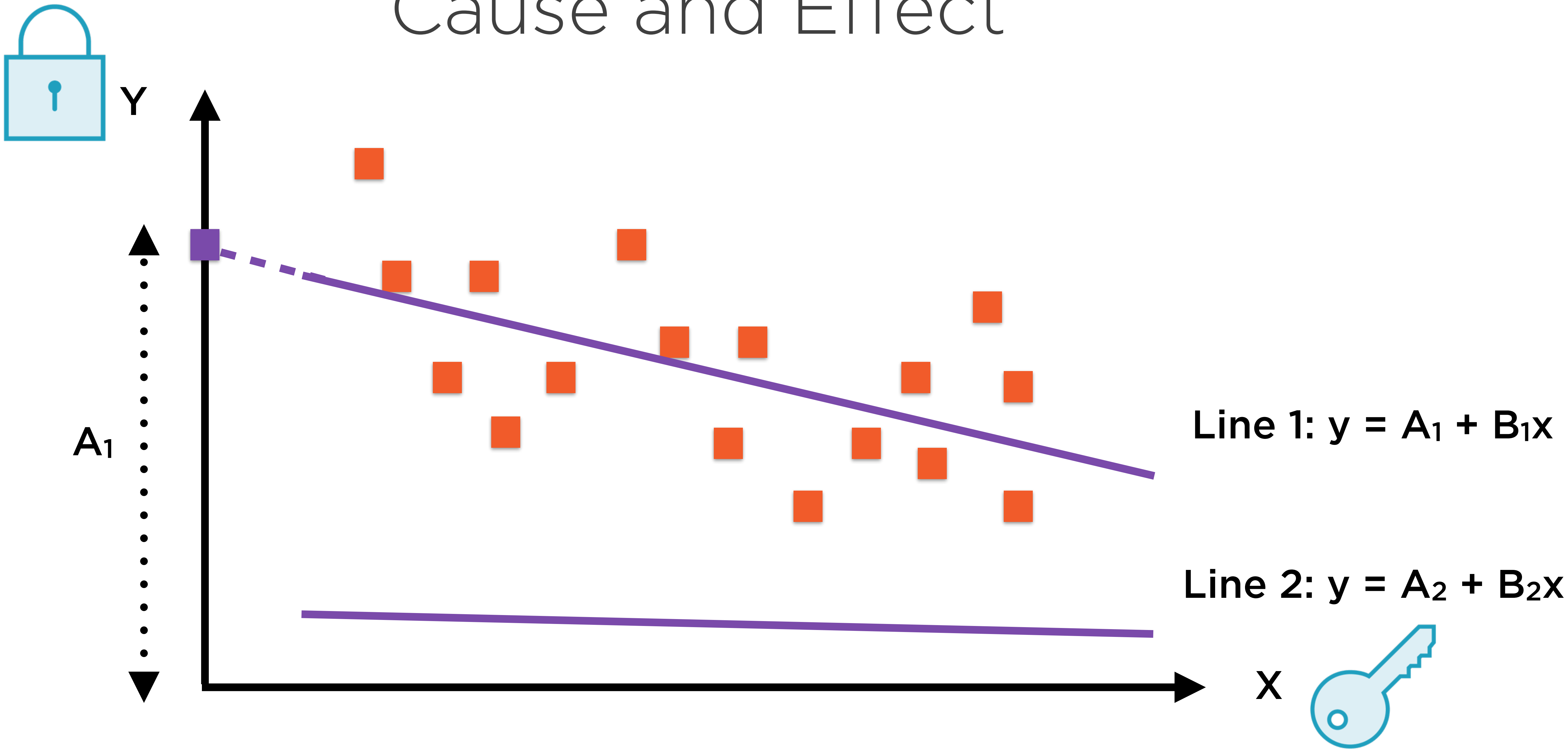
Linear Regression involves finding the "best fit" line

Cause and Effect



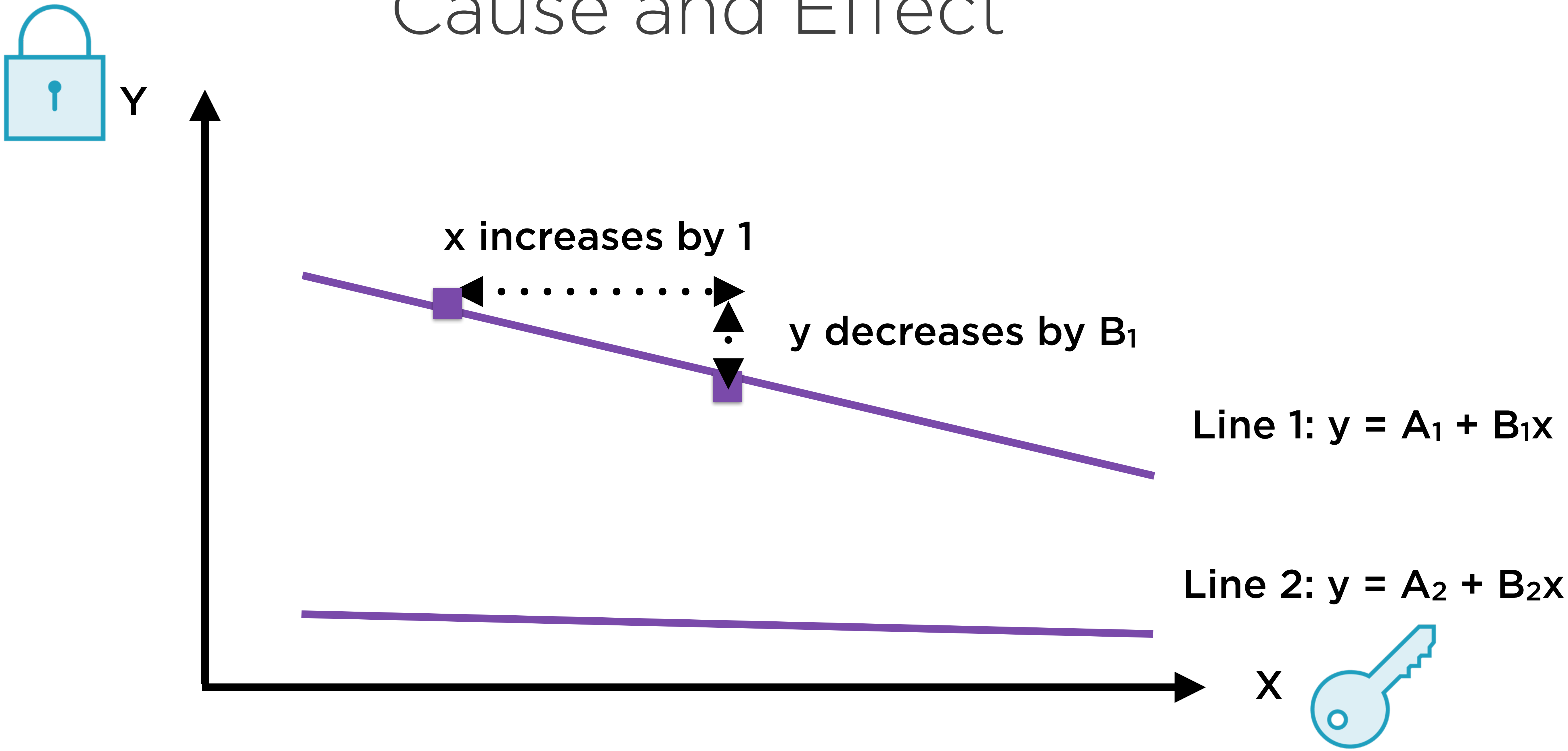
Let’s compare two lines, Line 1 and Line 2

Cause and Effect



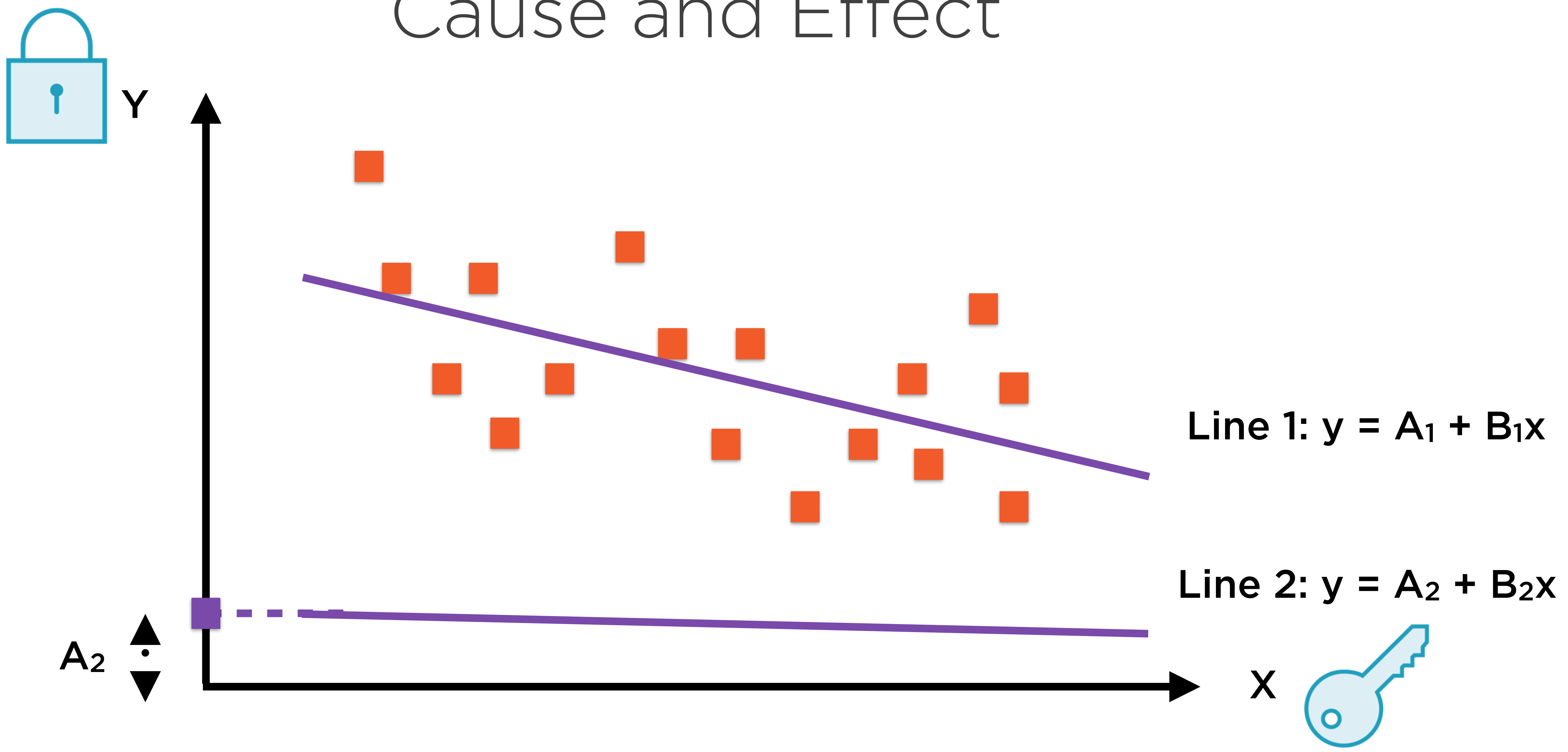
The first line has y-intercept A_1

Cause and Effect



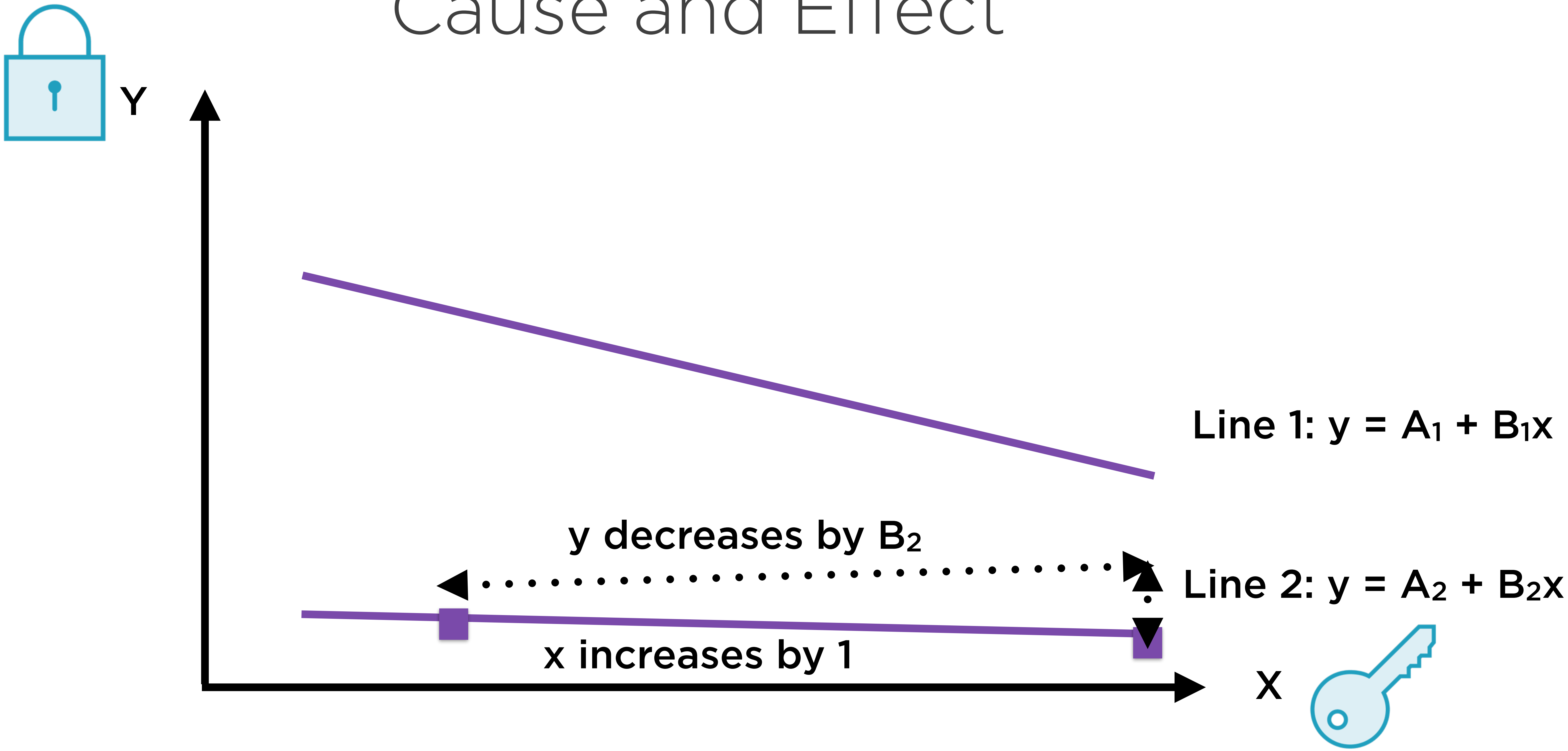
In the first line, if x increases by 1 unit, y decreases by B_1 units

Cause and Effect



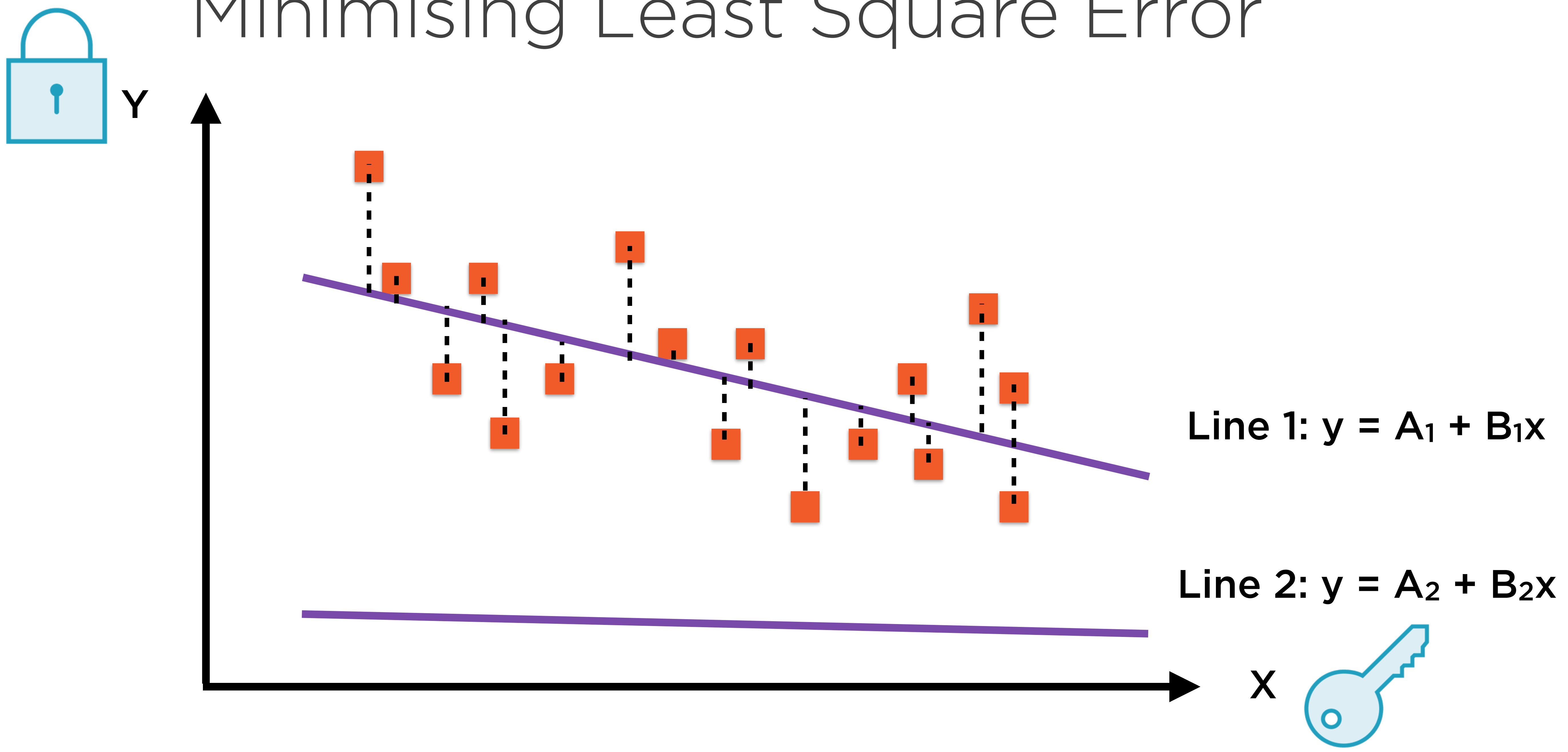
The second line has y-intercept A_2

Cause and Effect



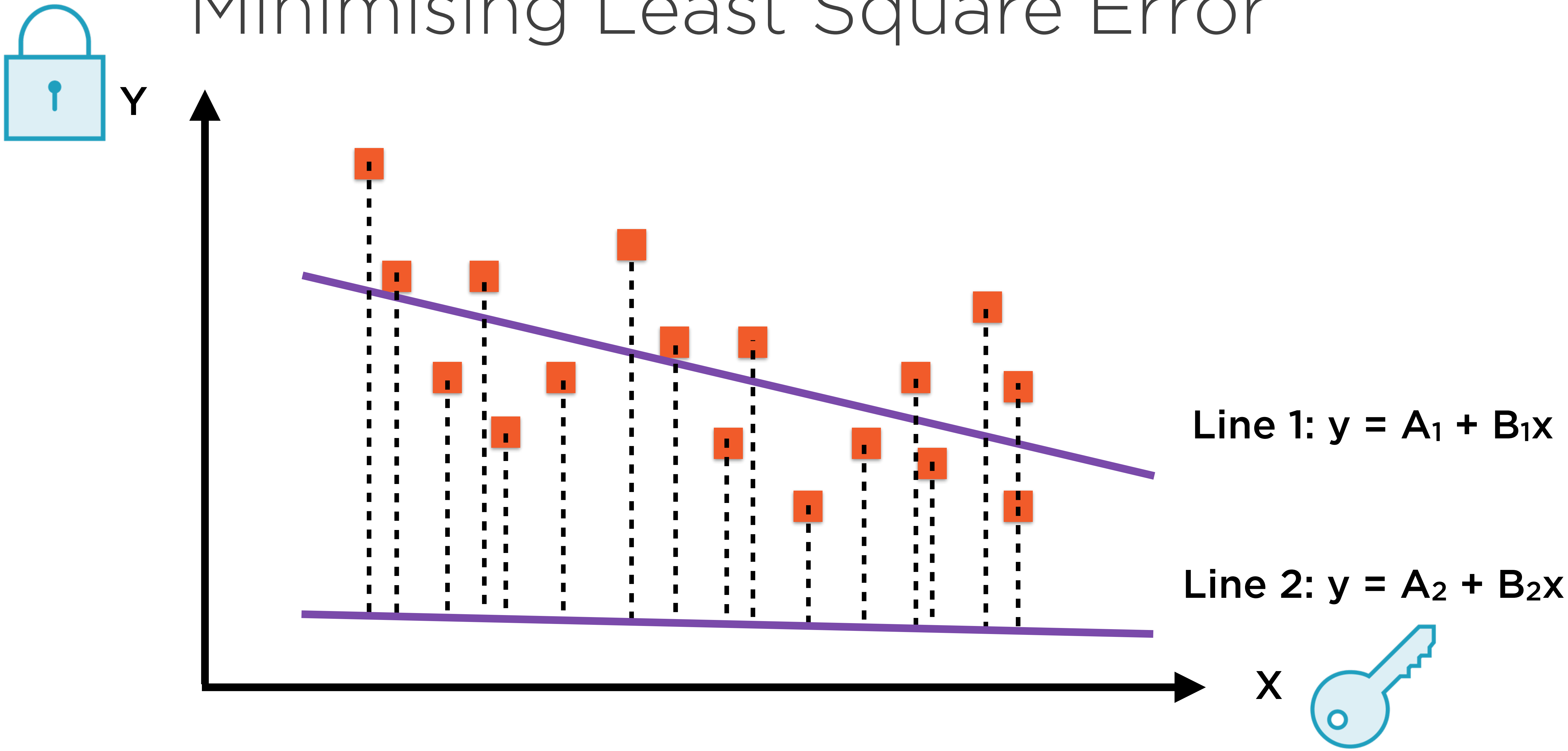
In the second line, if x increases by 1 unit, y decreases by B_2 units

Minimising Least Square Error



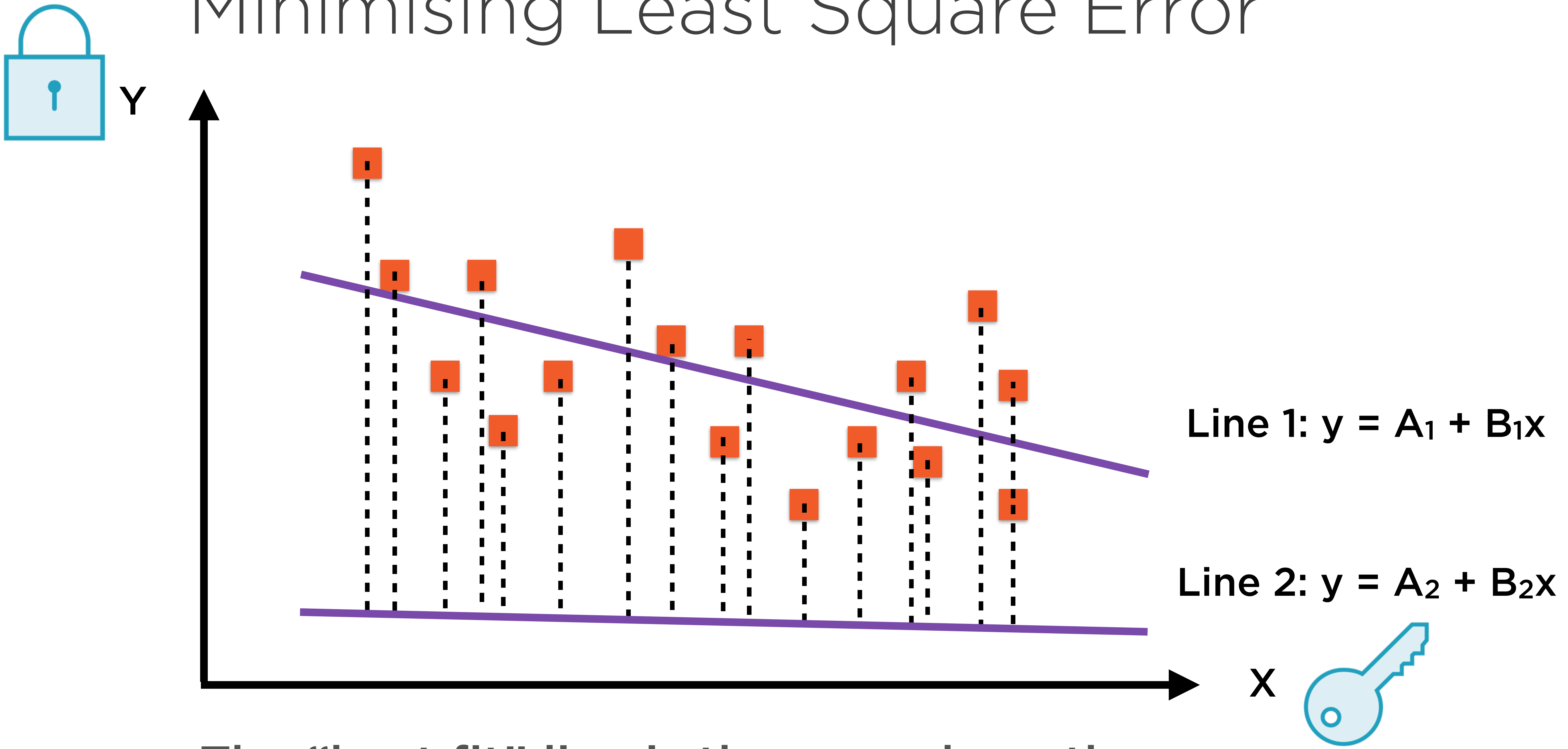
Drop vertical lines from each point to the lines A and B

Minimising Least Square Error



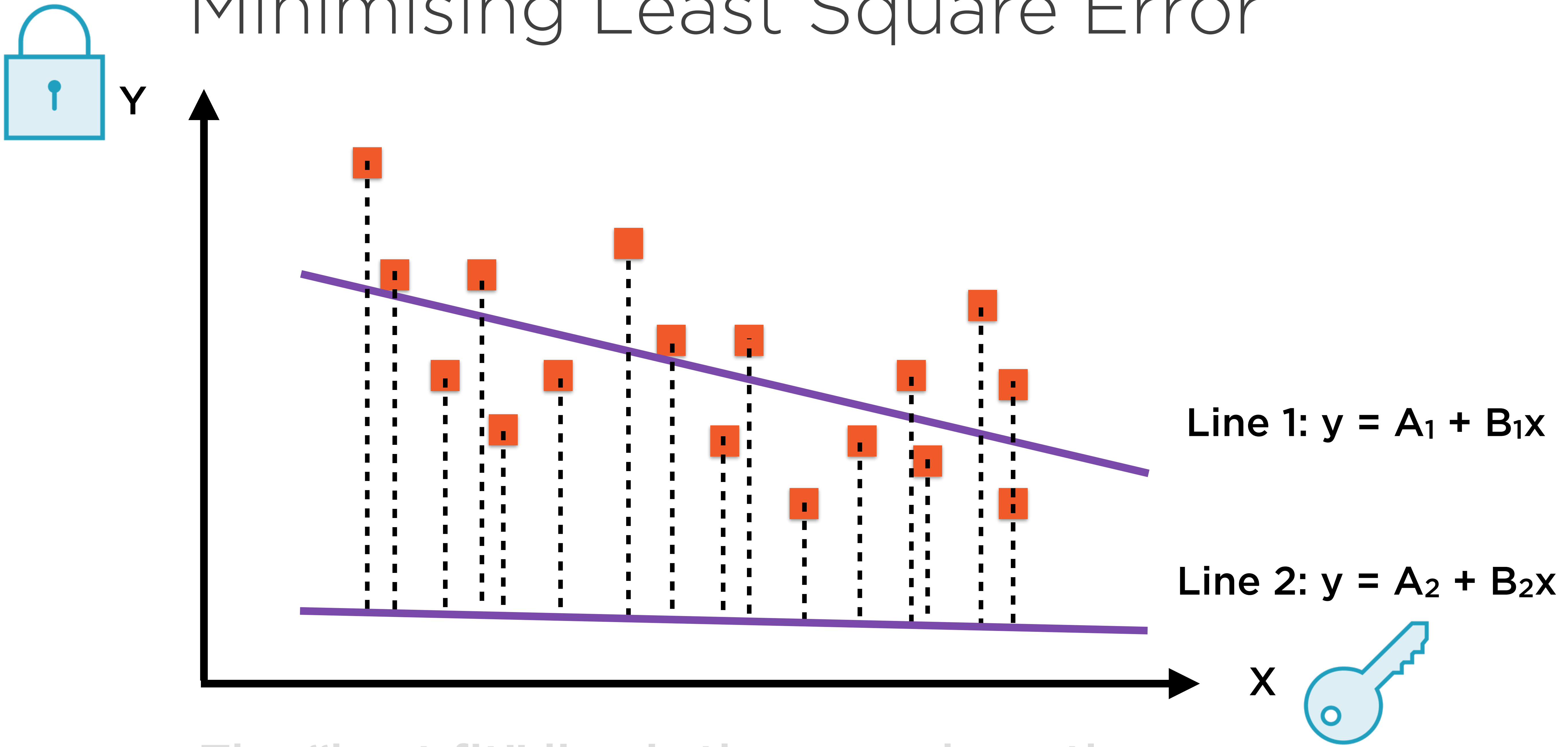
Drop vertical lines from each point to the lines A and B

Minimising Least Square Error



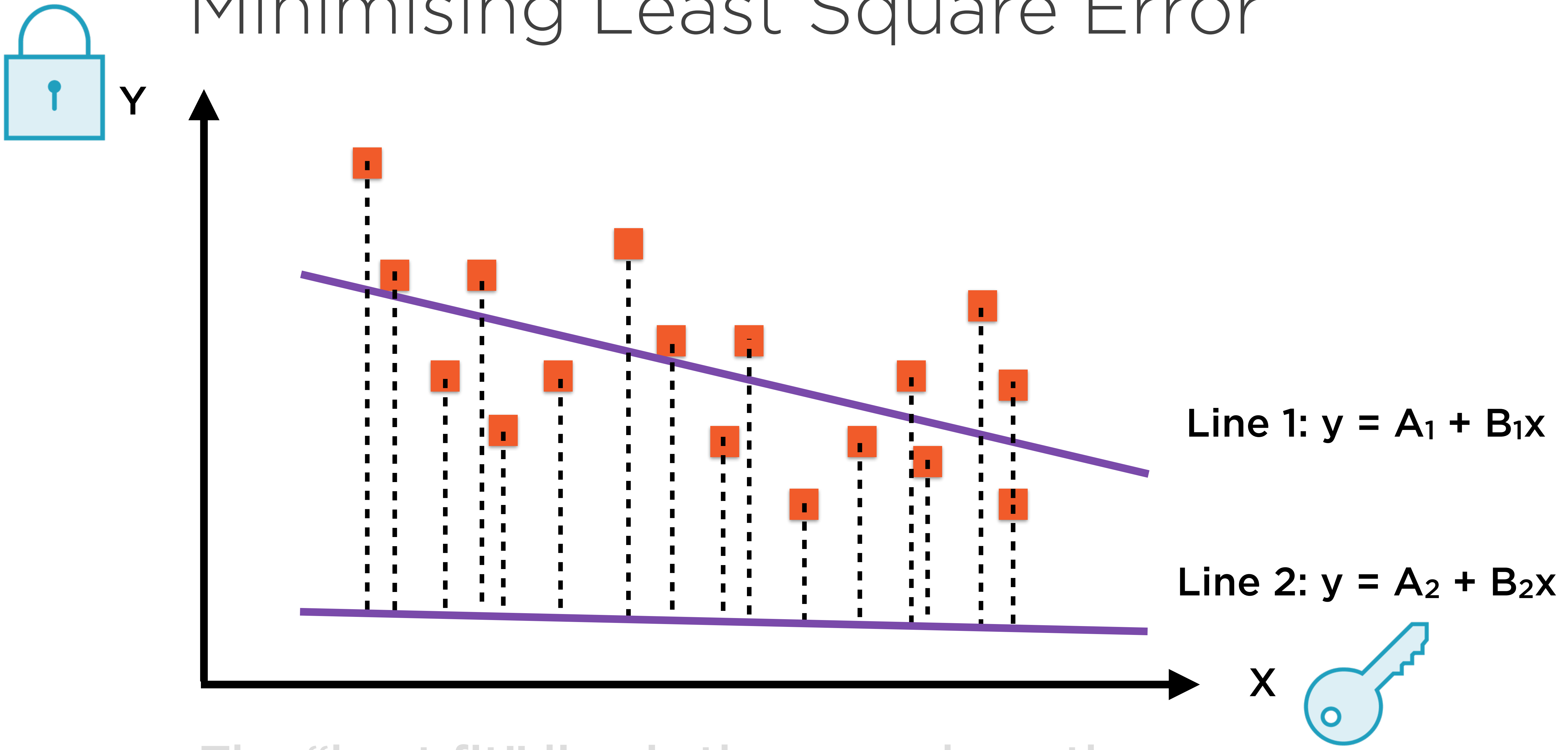
The “best fit” line is the one where the sum of the squares of the lengths of these dotted lines is minimum

Minimising Least Square Error



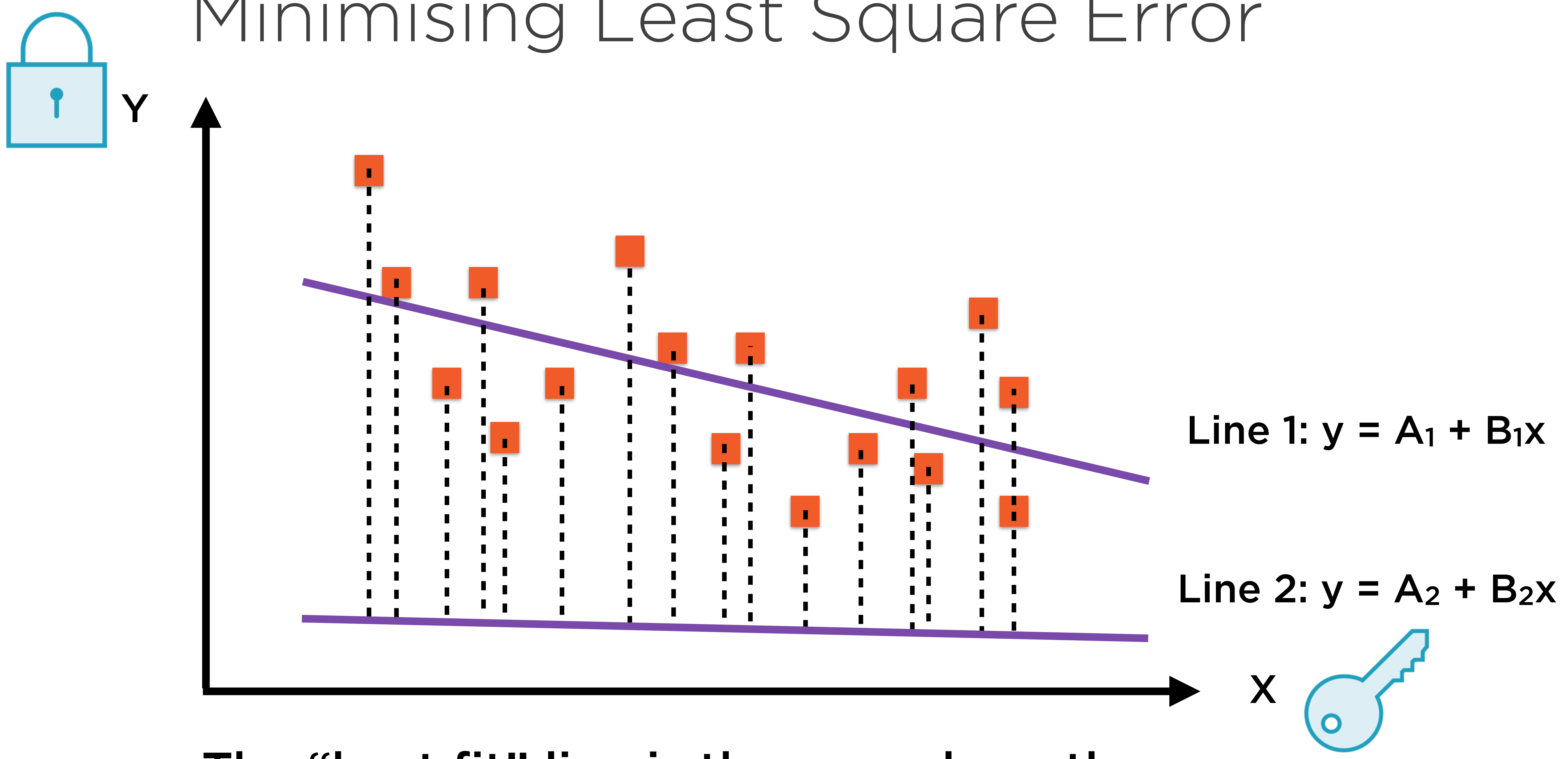
The "best fit" line is the one where the sum of the squares of the lengths of **these dotted lines** is minimum

Minimising Least Square Error



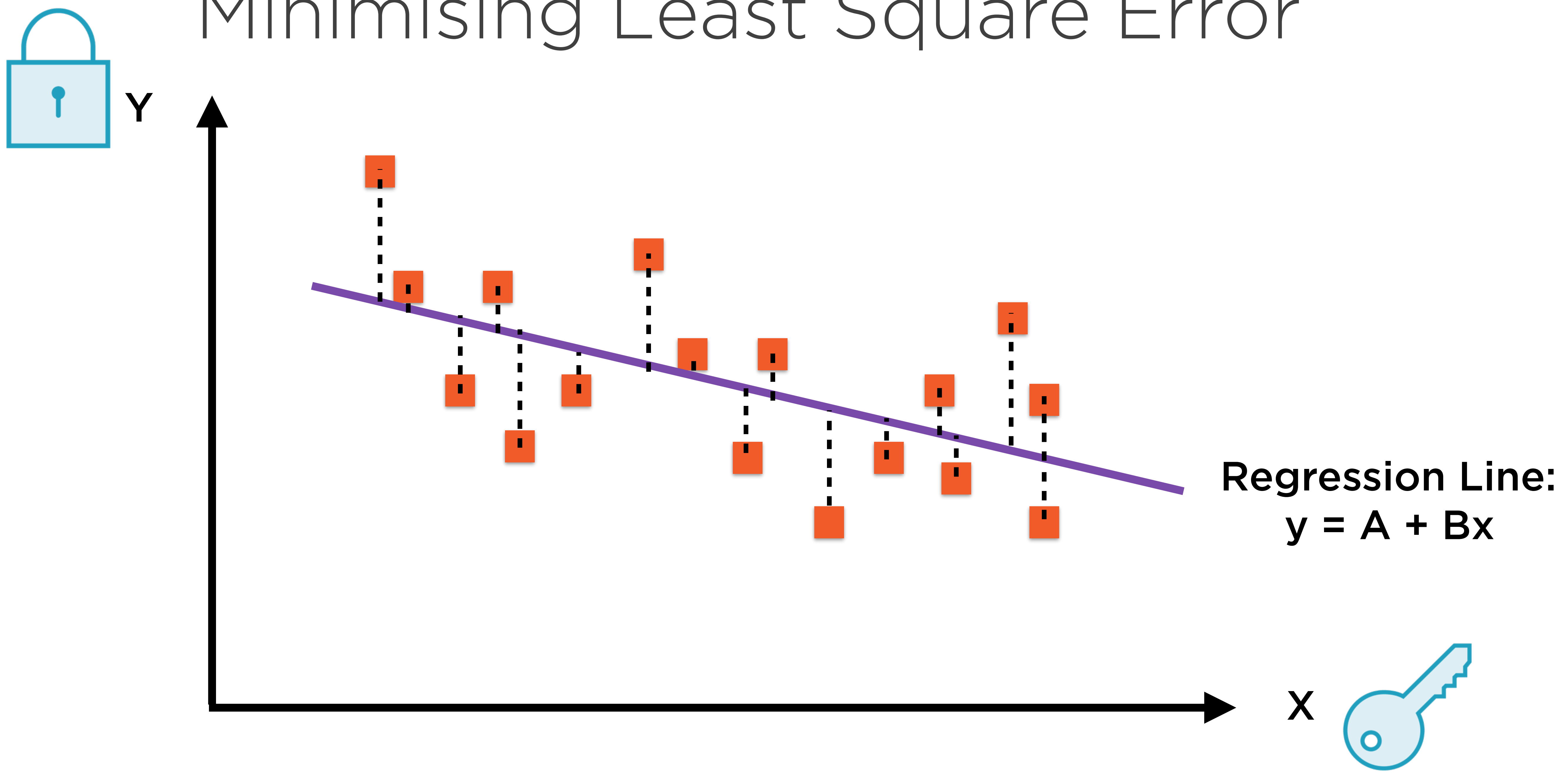
The "best fit" line is the one where the sum of the squares of the lengths of **the errors** is minimum

Minimising Least Square Error



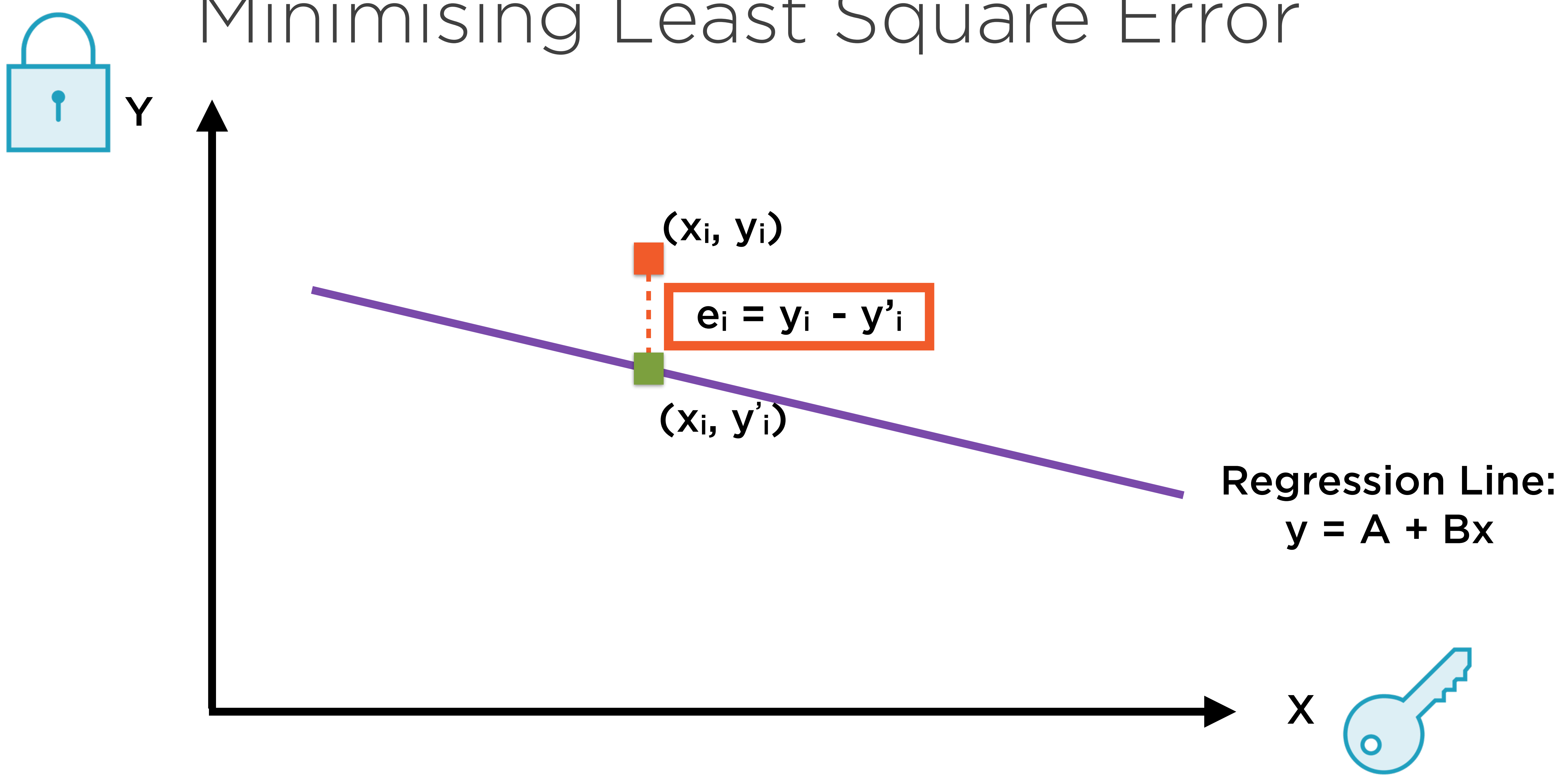
The "best fit" line is the one where the sum of the squares of the lengths of the errors is minimum

Minimising Least Square Error

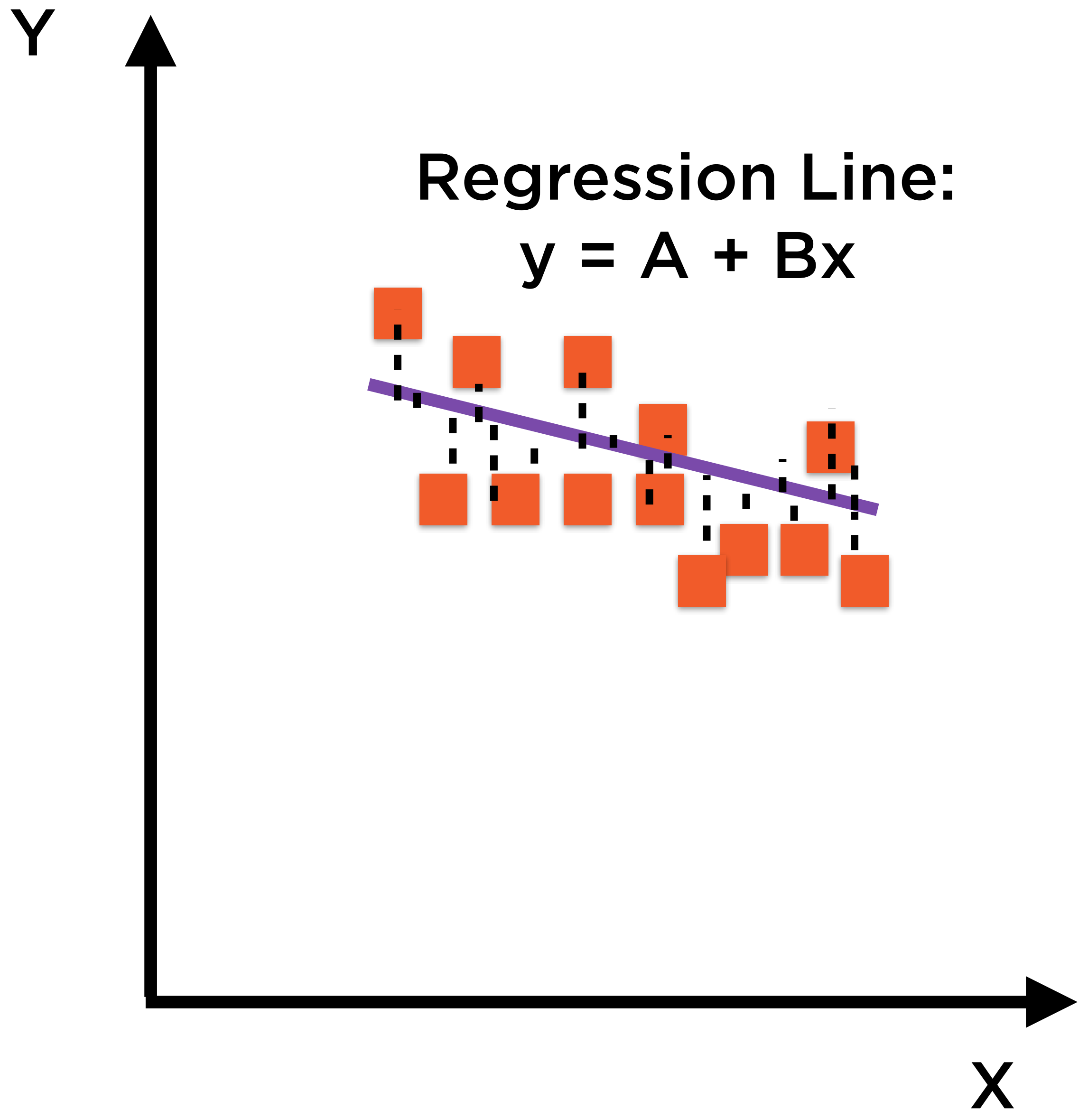


The "best fit" line is called the regression line

Minimising Least Square Error



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



Ideally, residuals should

- have zero mean
- common variance
- be independent of each other
- be independent of x
- be normally distributed

Solving the Regression Problem

Three Estimation Methods

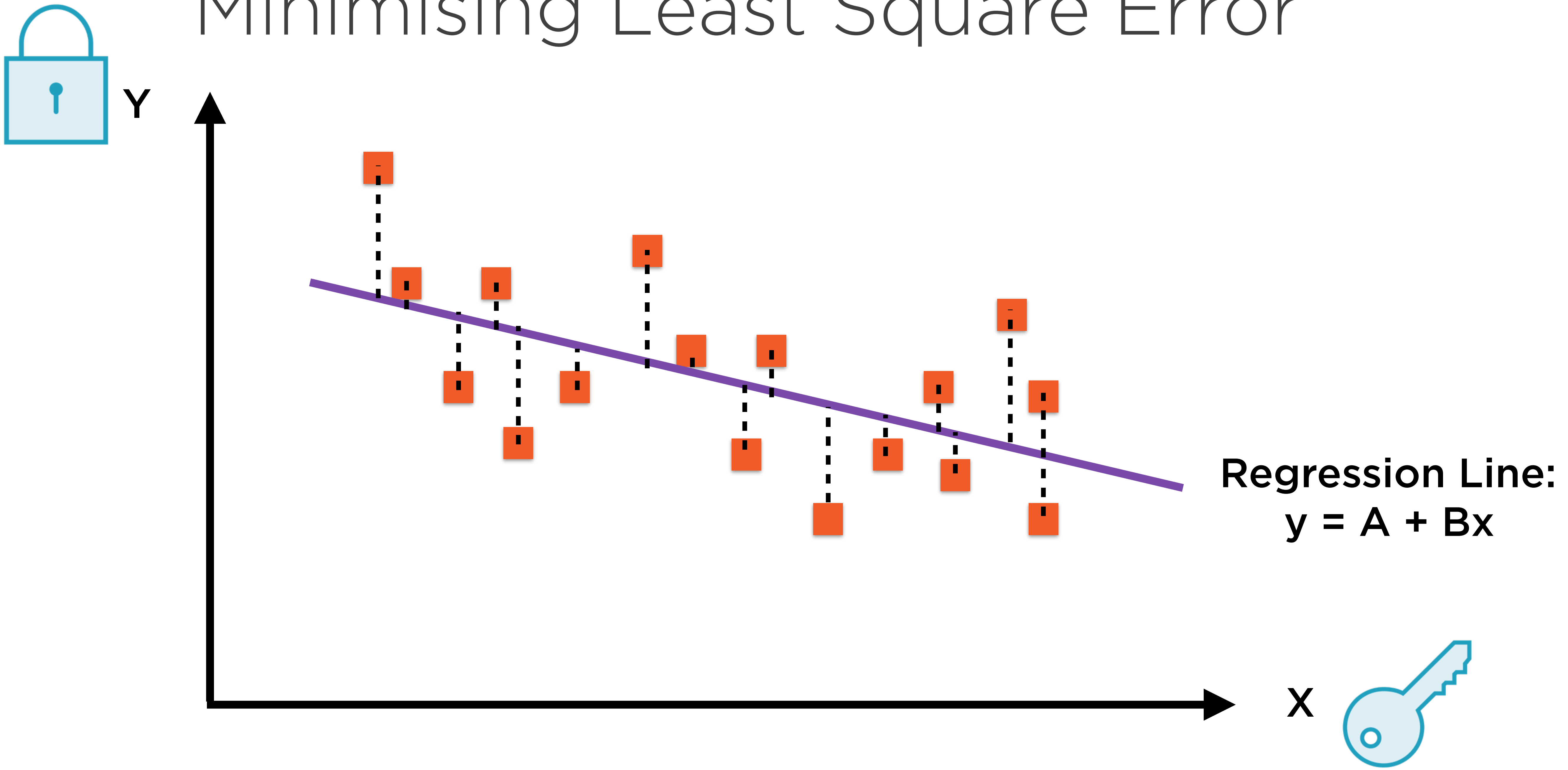
Method of moments

Method of least squares

Maximum likelihood estimation

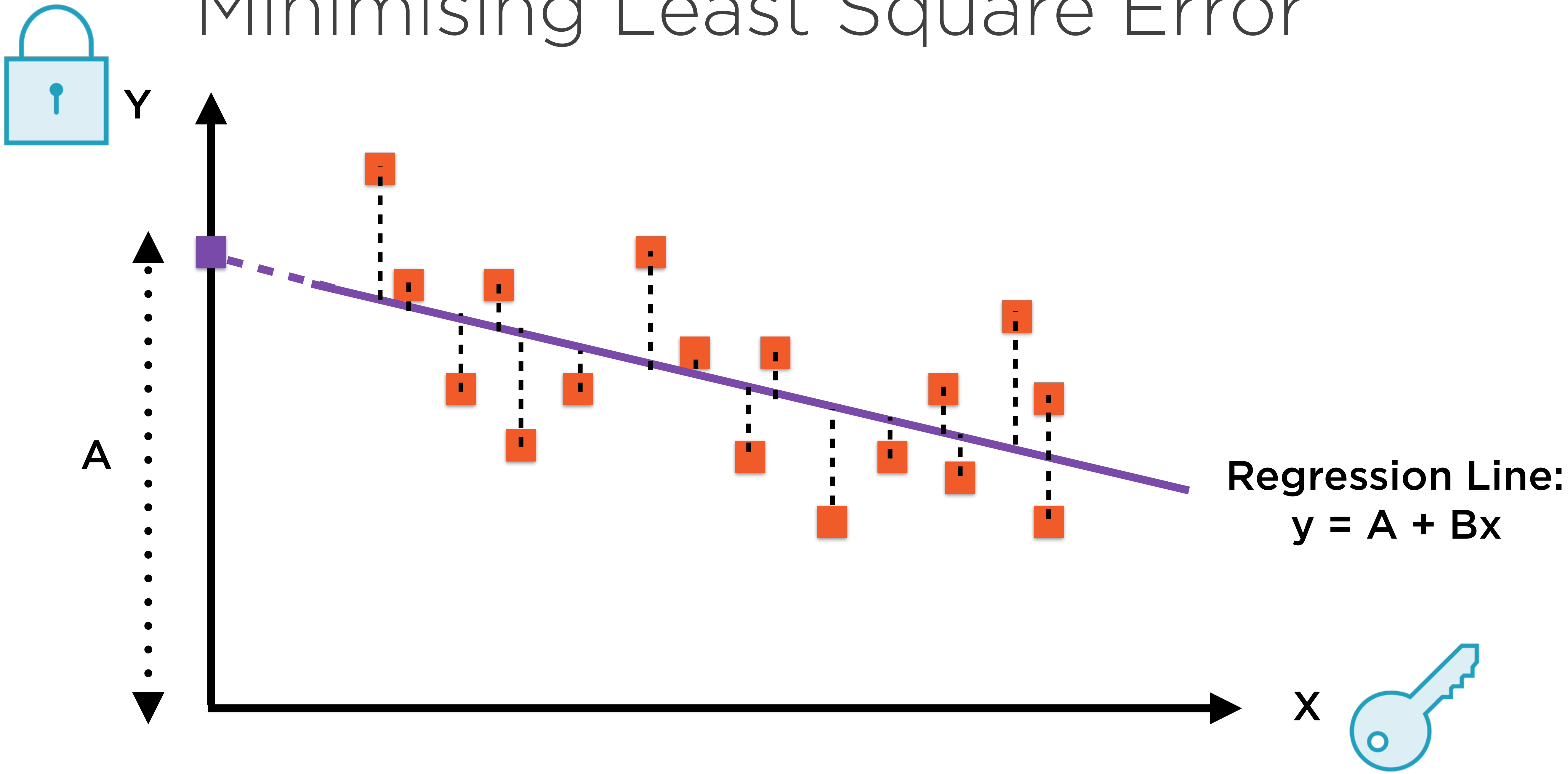
Cookie cutter techniques to determine the values of A and B (regression coefficients)

Minimising Least Square Error



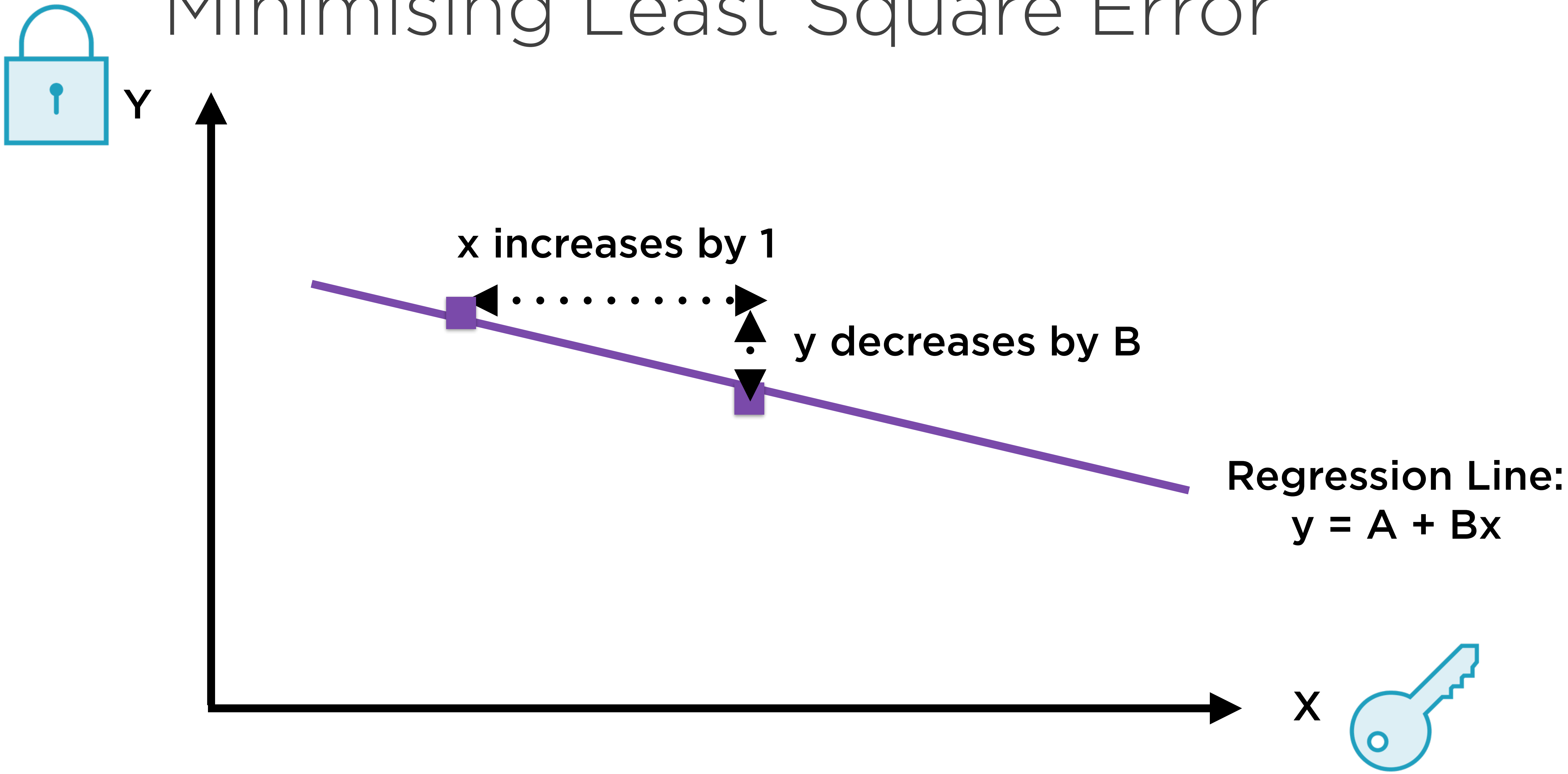
The “best fit” line is called the regression line

Minimising Least Square Error



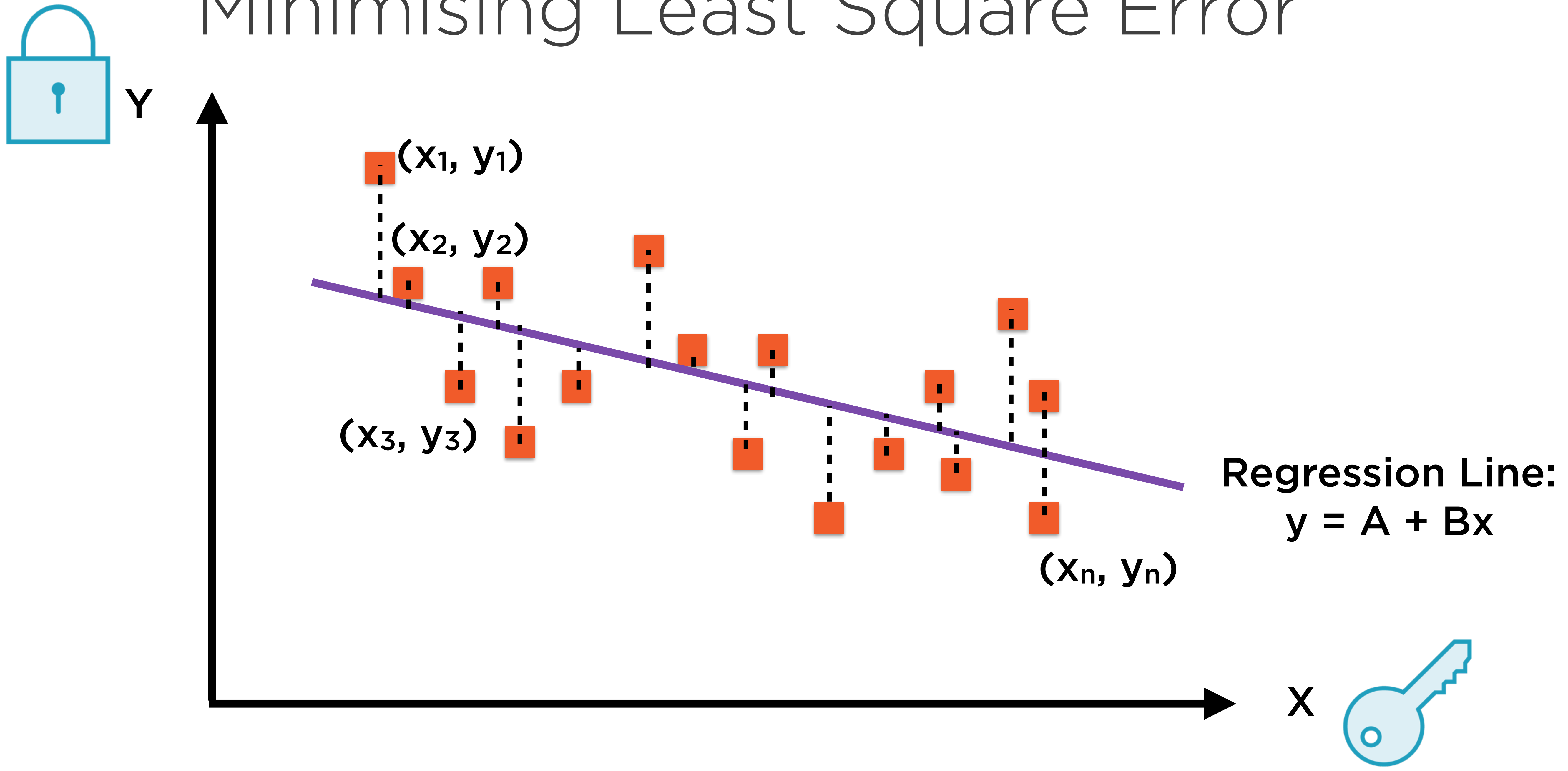
The term A in the equation of the line is the y-intercept

Minimising Least Square Error



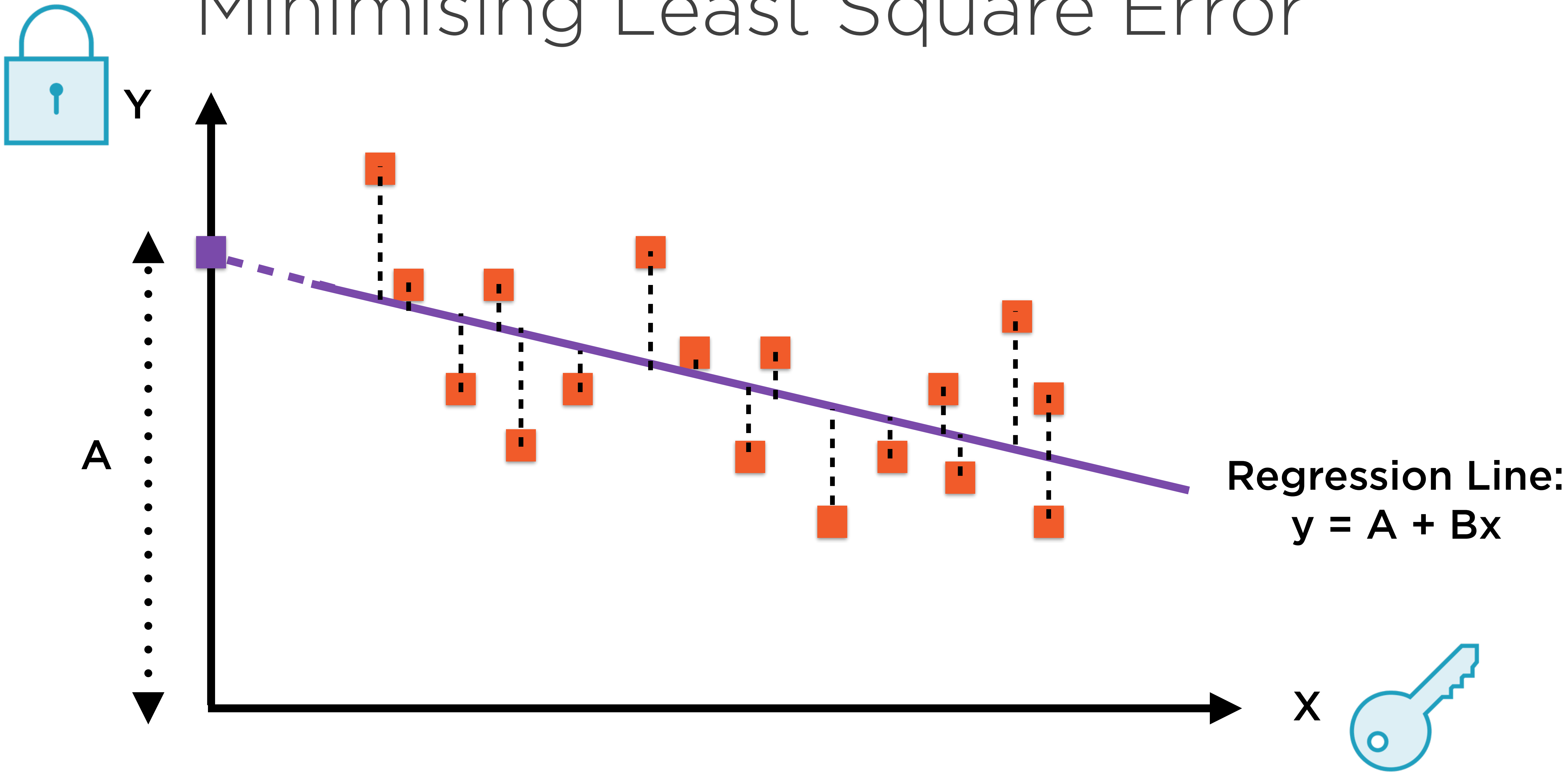
The term B is the slope, and gives the sensitivity of y to a change of 1 unit in x

Minimising Least Square Error



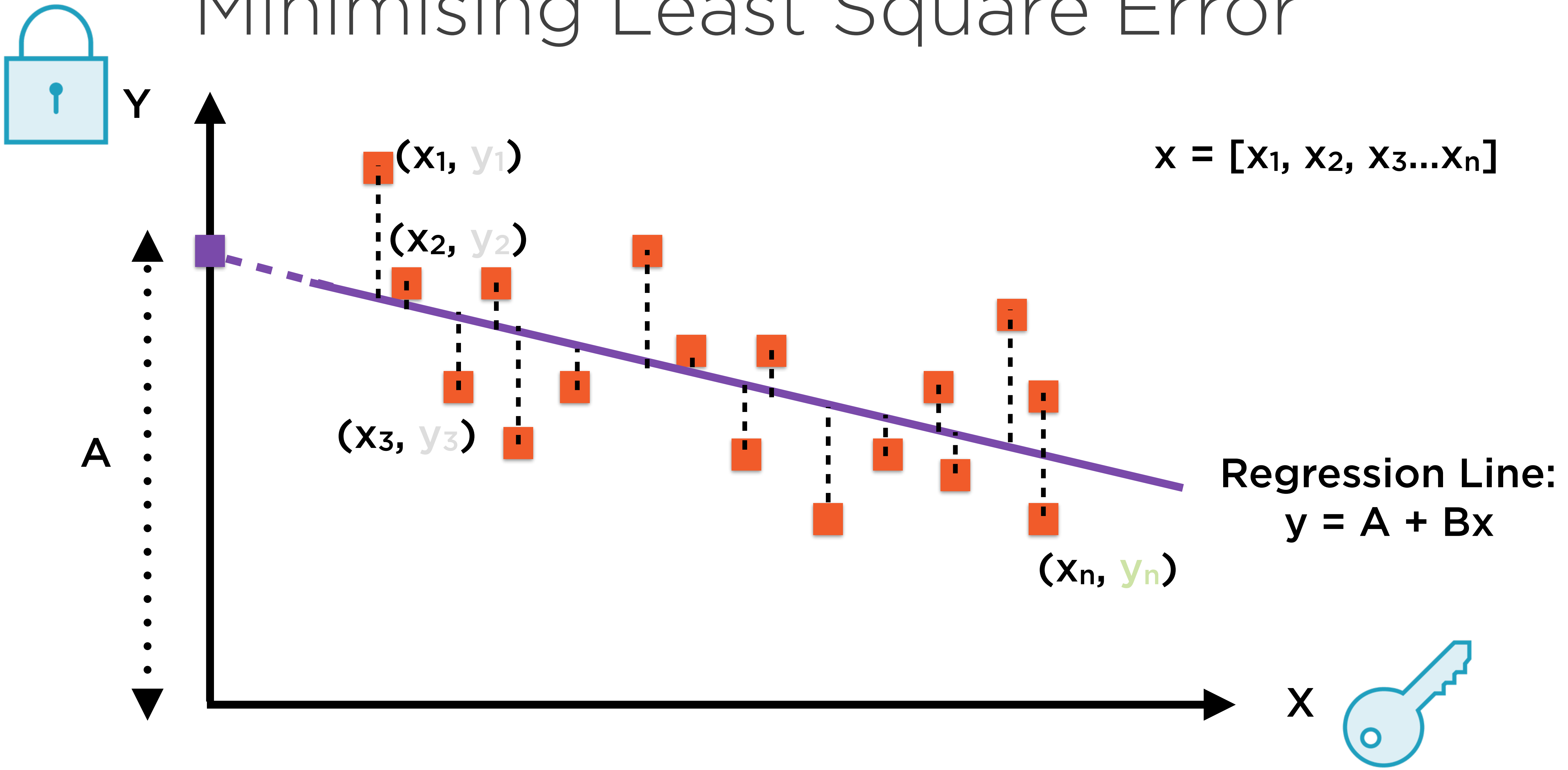
Represent all n points as
(x_i, y_i), where i = 1 to n

Minimising Least Square Error



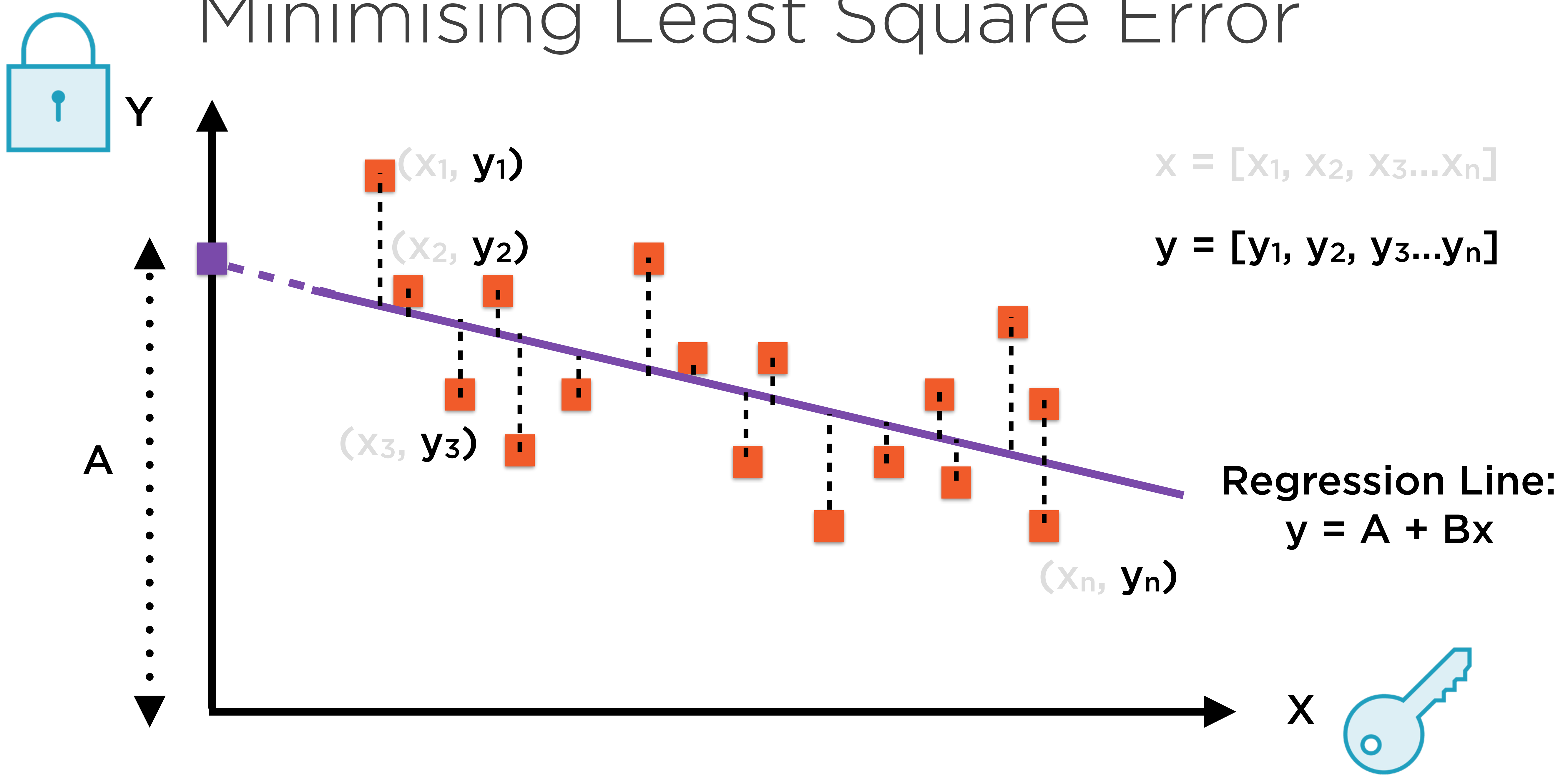
The "best fit" line is called the regression line

Minimising Least Square Error



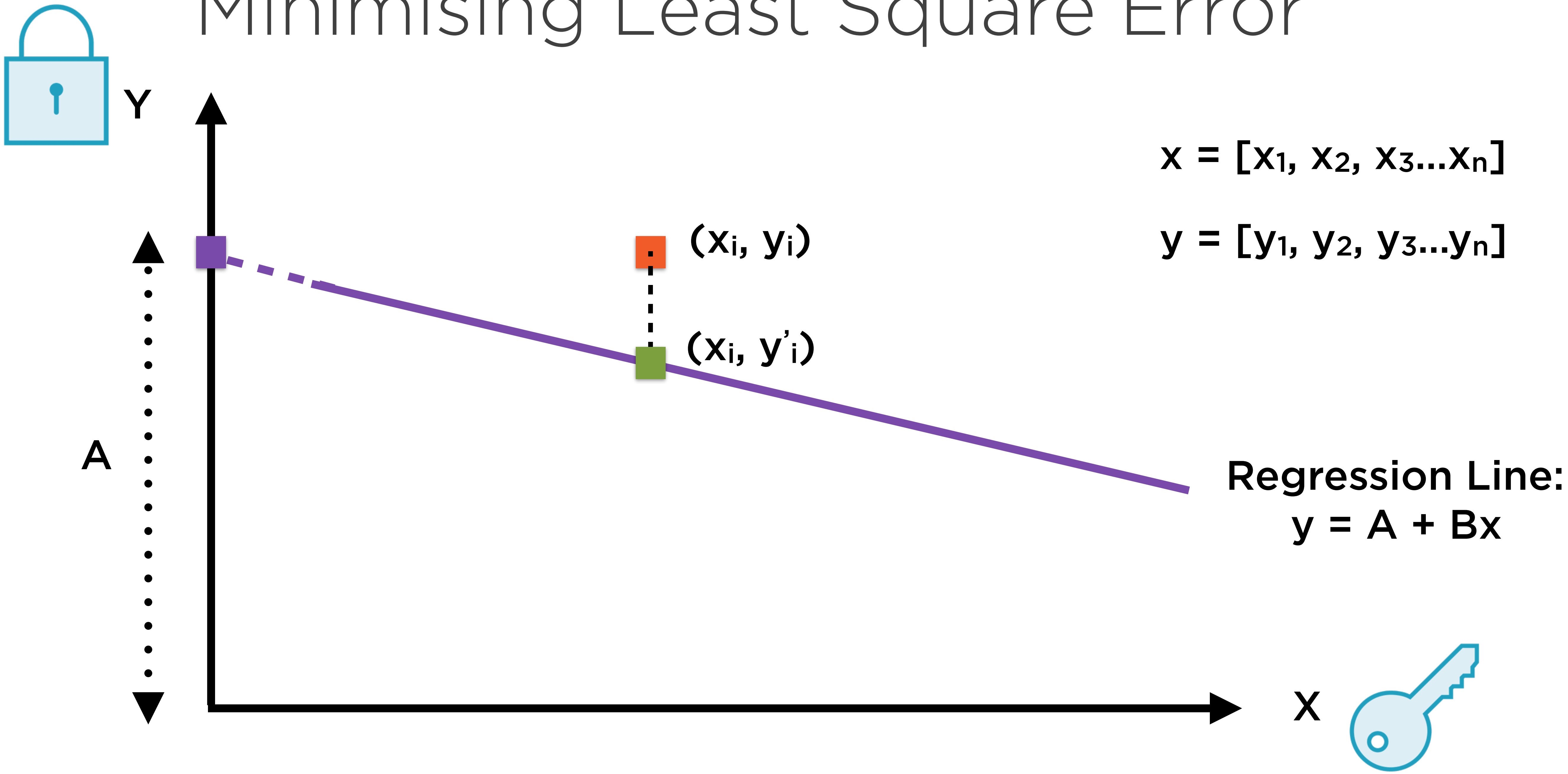
x in the regression line refers to the vector of all x coordinates

Minimising Least Square Error



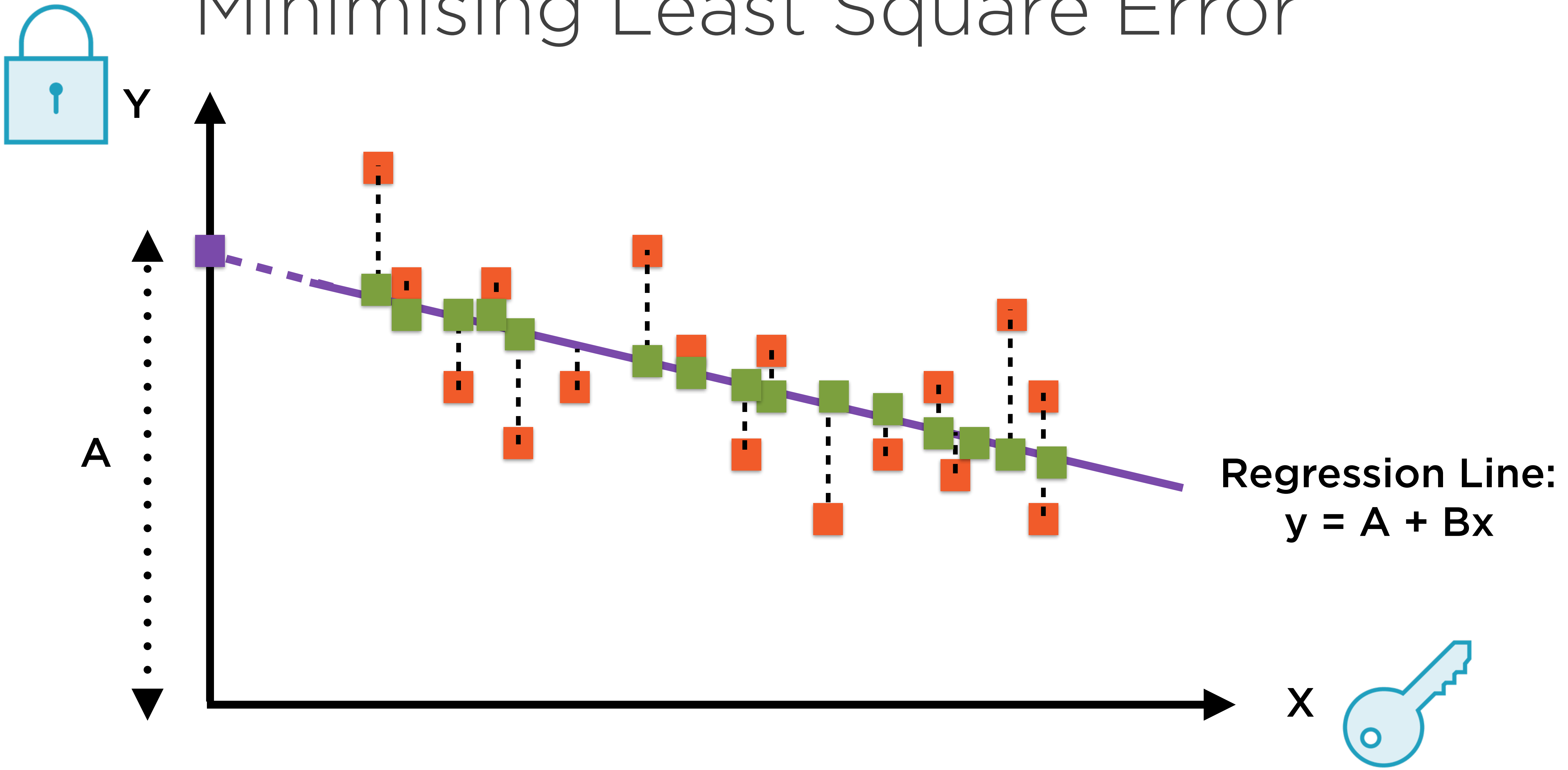
y in the regression line refers to the vector of all y coordinates

Minimising Least Square Error



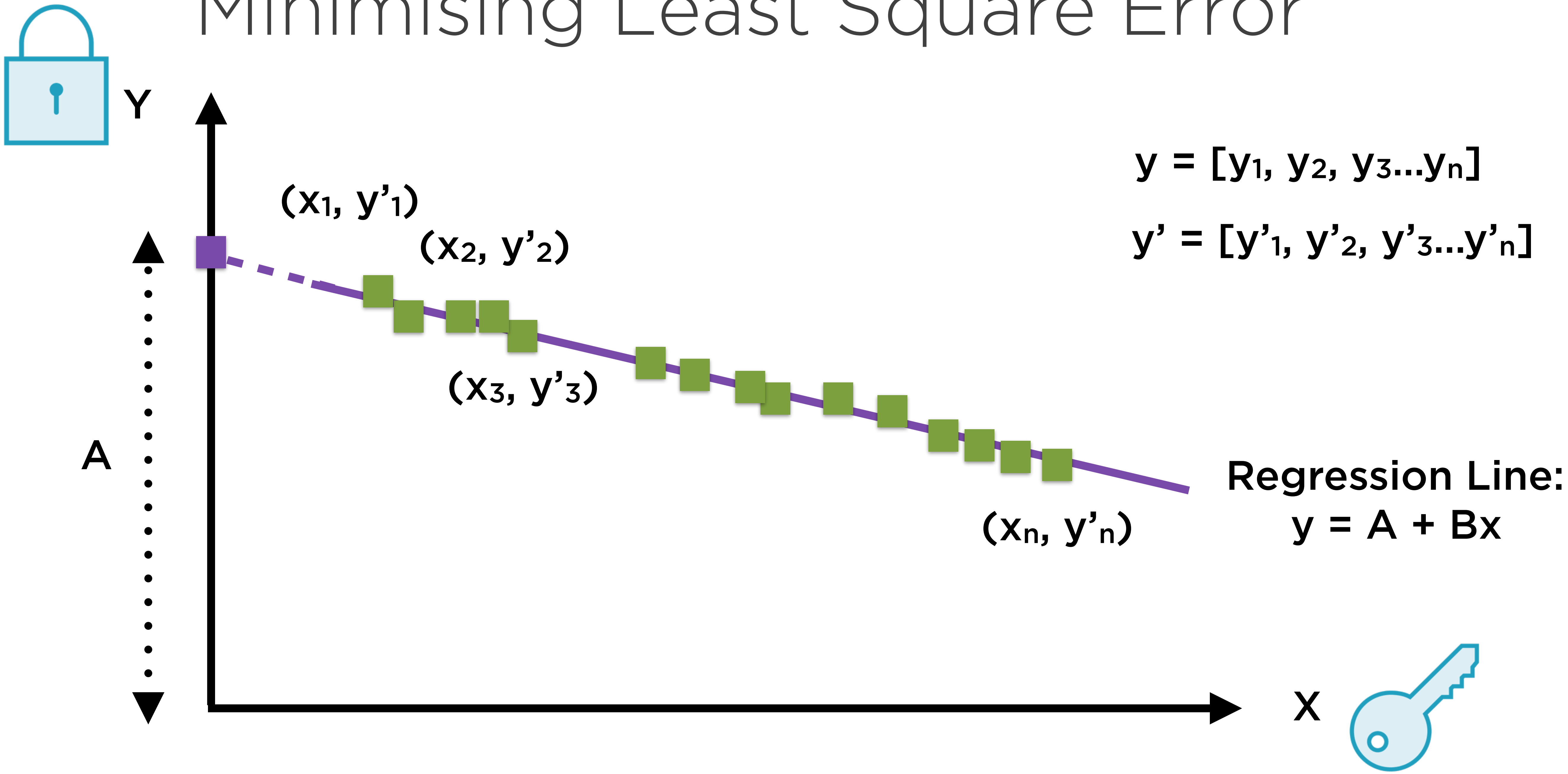
Each point (x_i, y_i) has a corresponding point (x_i, y'_i) on the regression line

Minimising Least Square Error



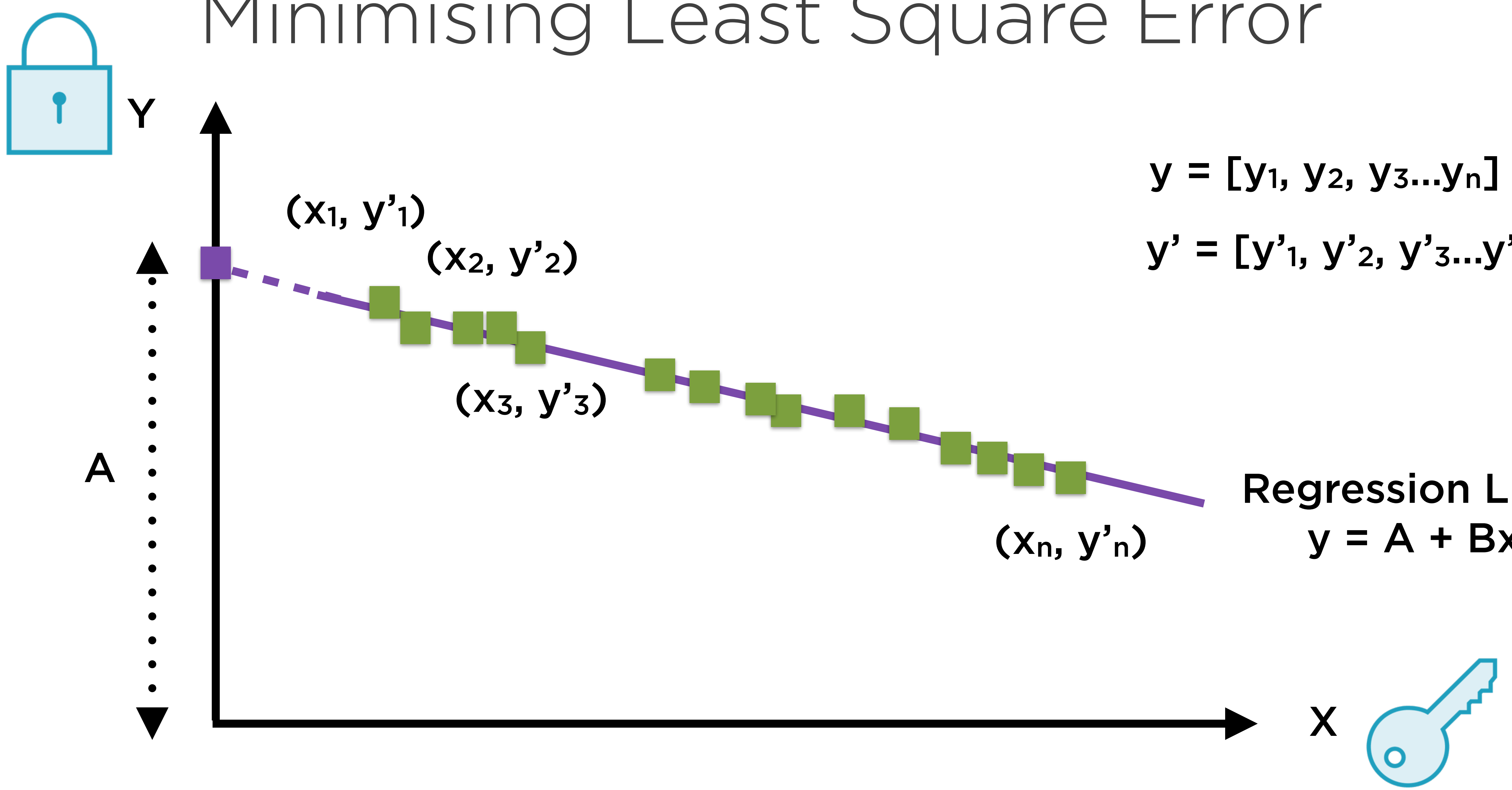
Find all such points (x_i,y'_i) on the the regression line

Minimising Least Square Error



Find all such points (x_i, y'_i) on the regression line

Minimising Least Square Error



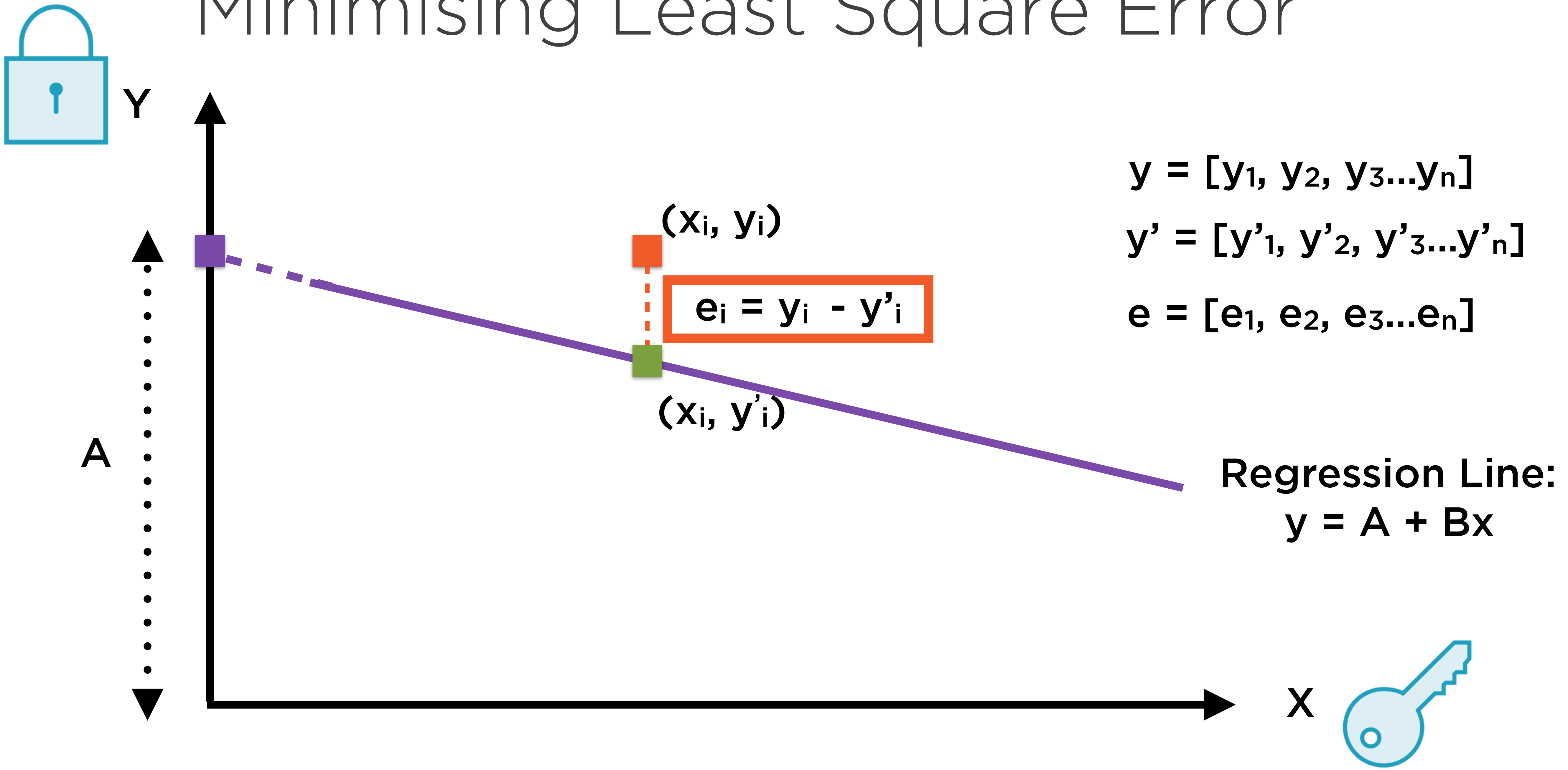
$y = [y_1, y_2, y_3 \ldots y_n]$

$y' = [y'_1, y'_2, y'_3 \ldots y'_n]$

Regression Line:
$y = A + Bx$

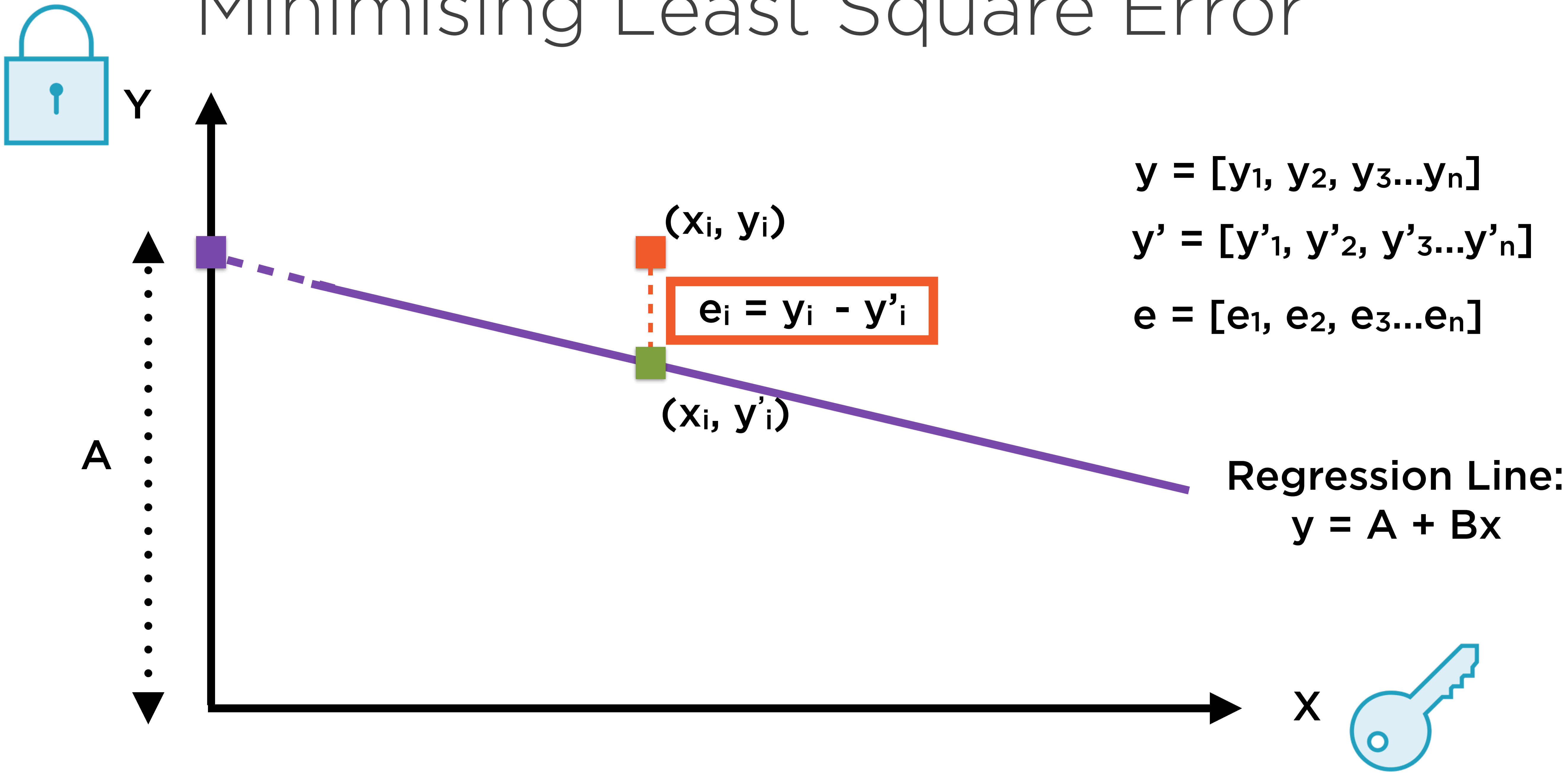
The corresponding values of y'_i are called the fitted values

Minimising Least Square Error



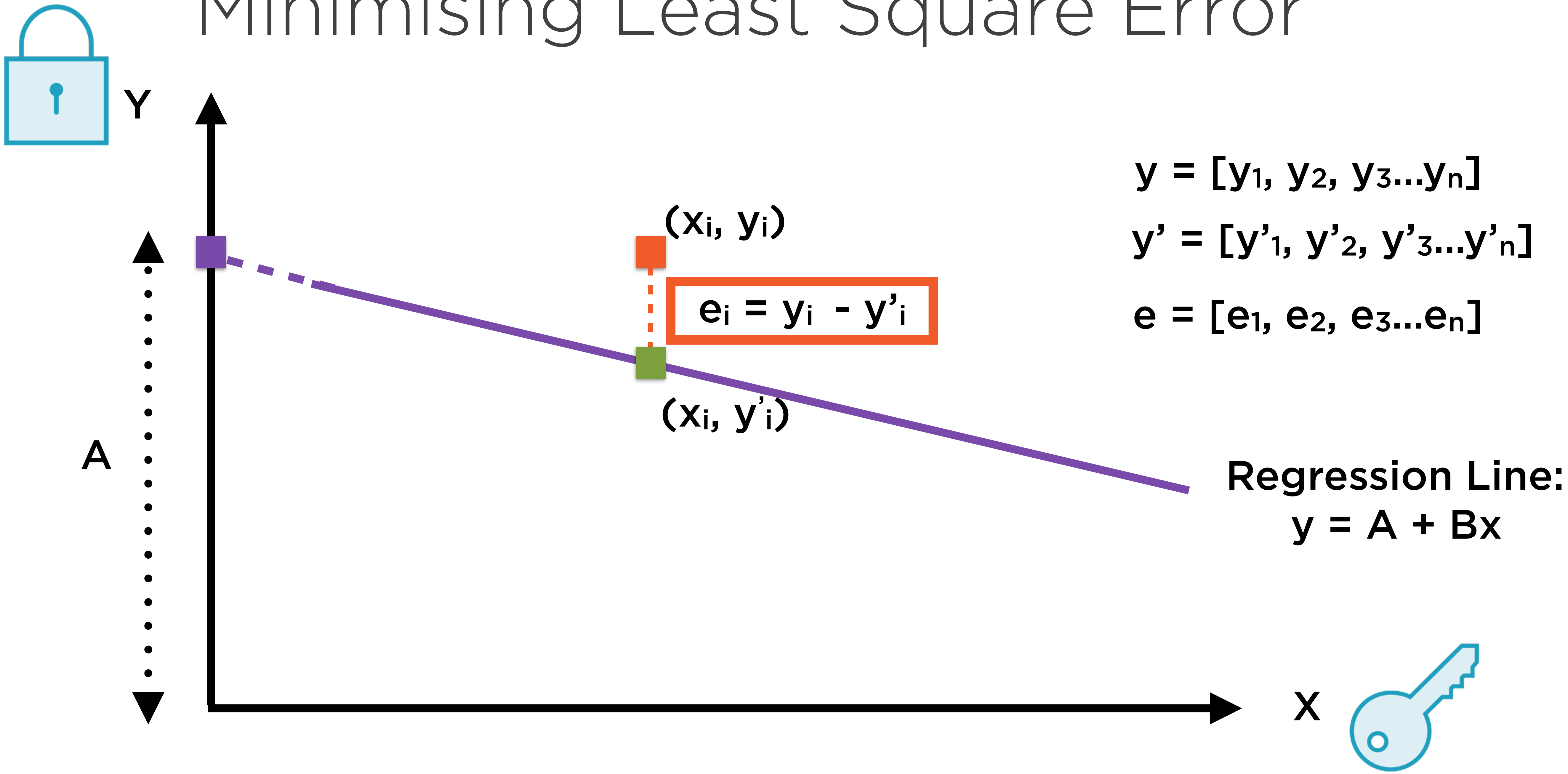
For each point, the difference between y_i and y'_i is called e_i, the residual or the error

Minimising Least Square Error



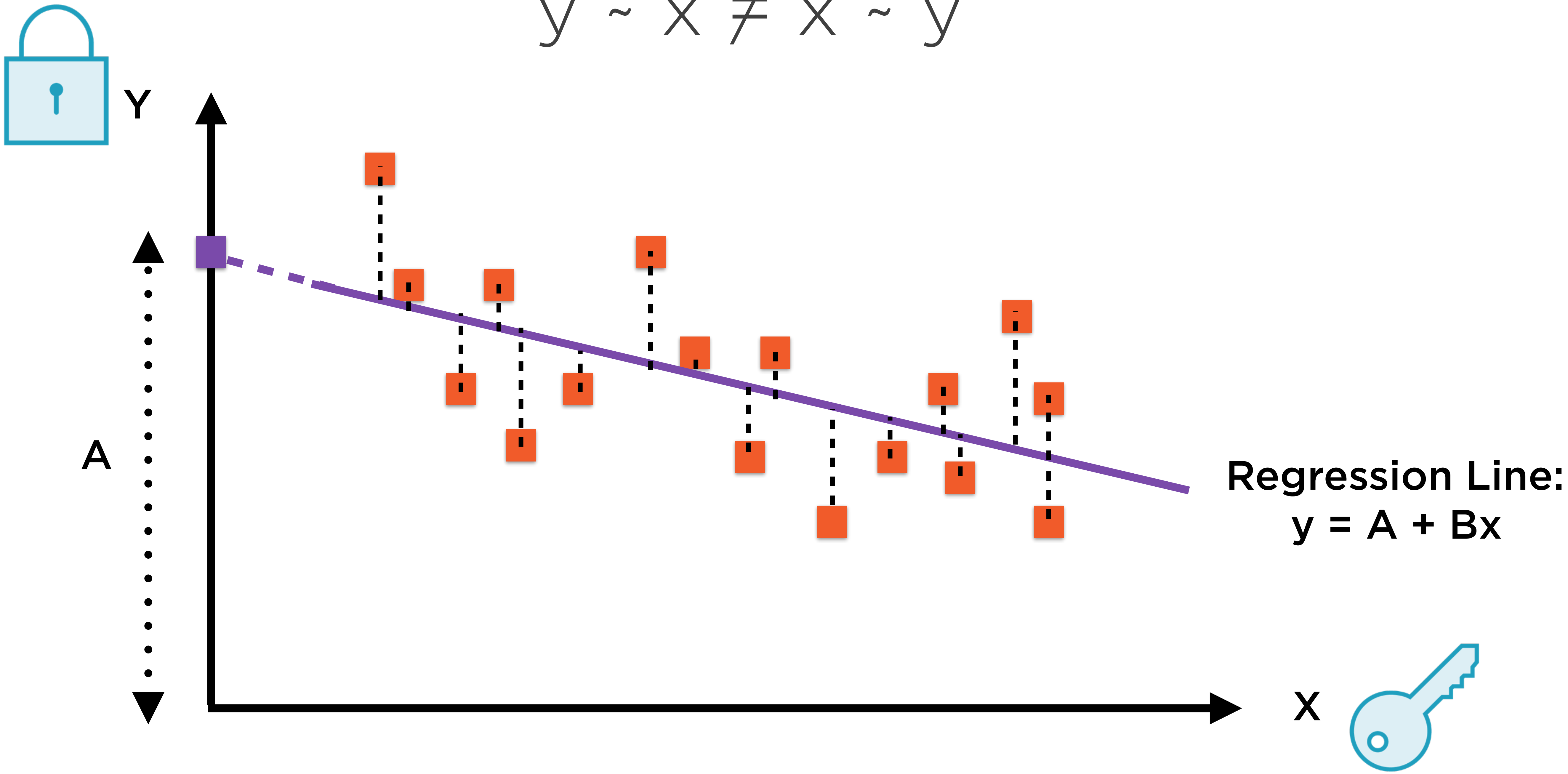
Residuals of a regression are the difference between actual and fitted values of the dependent variable

Minimising Least Square Error



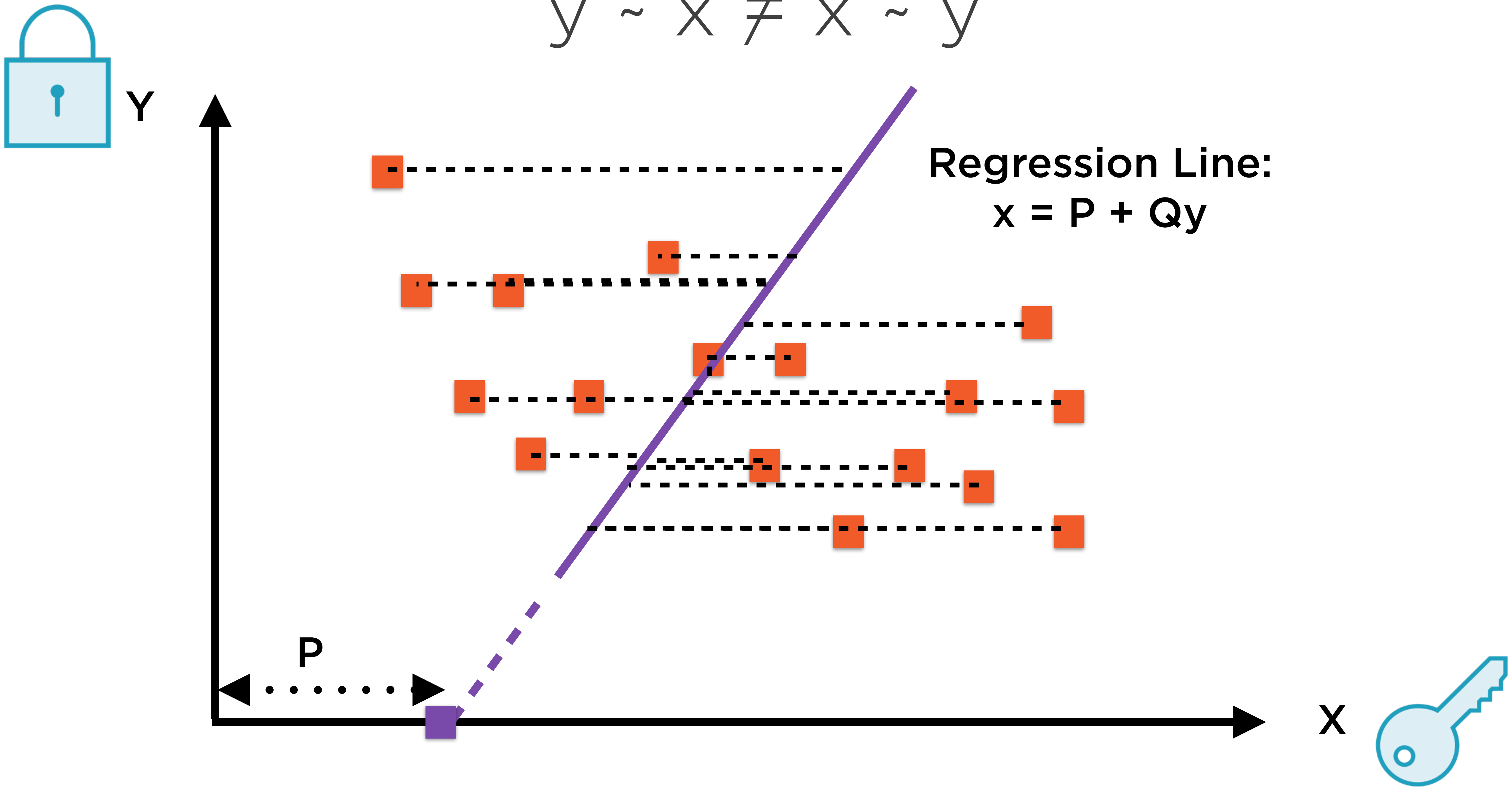
For each point, the difference between y_i and y'_i is called e_i, the residual or the error

$$y \sim x \neq x \sim y$$



Regressing y on x - minimise sum of square of vertical errors

$$y \sim x \neq x \sim y$$

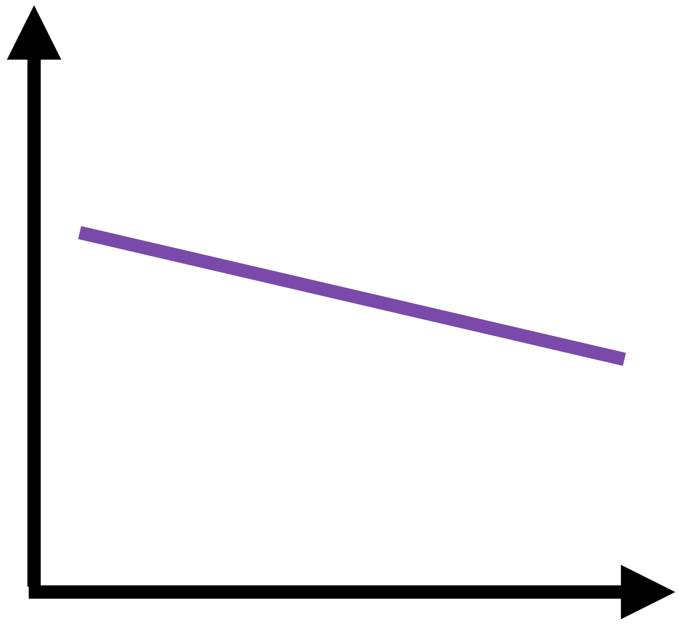


Regressing y on x - minimise sum of square of vertical errors

Demo

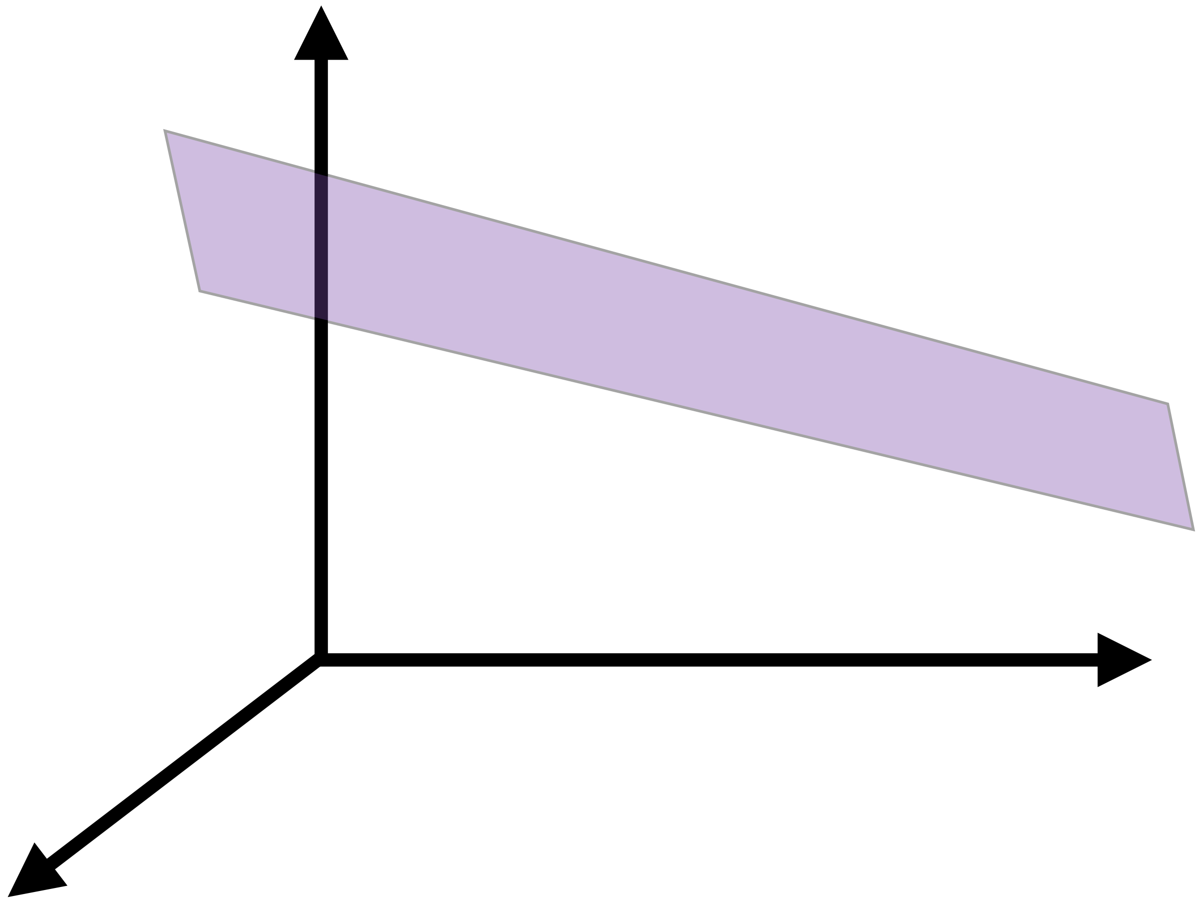
Perform a simple regression in Excel

Simple and Multiple Regression



Simple Regression

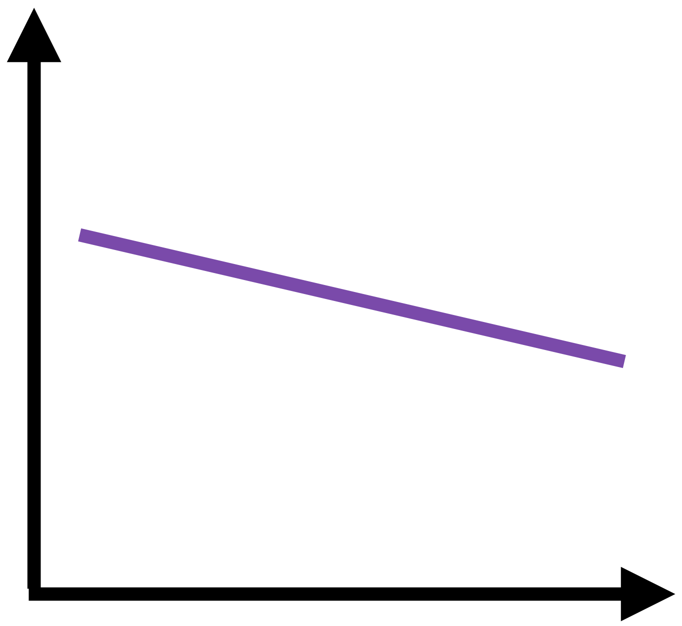
Data in 2 dimensions



Multiple Regression

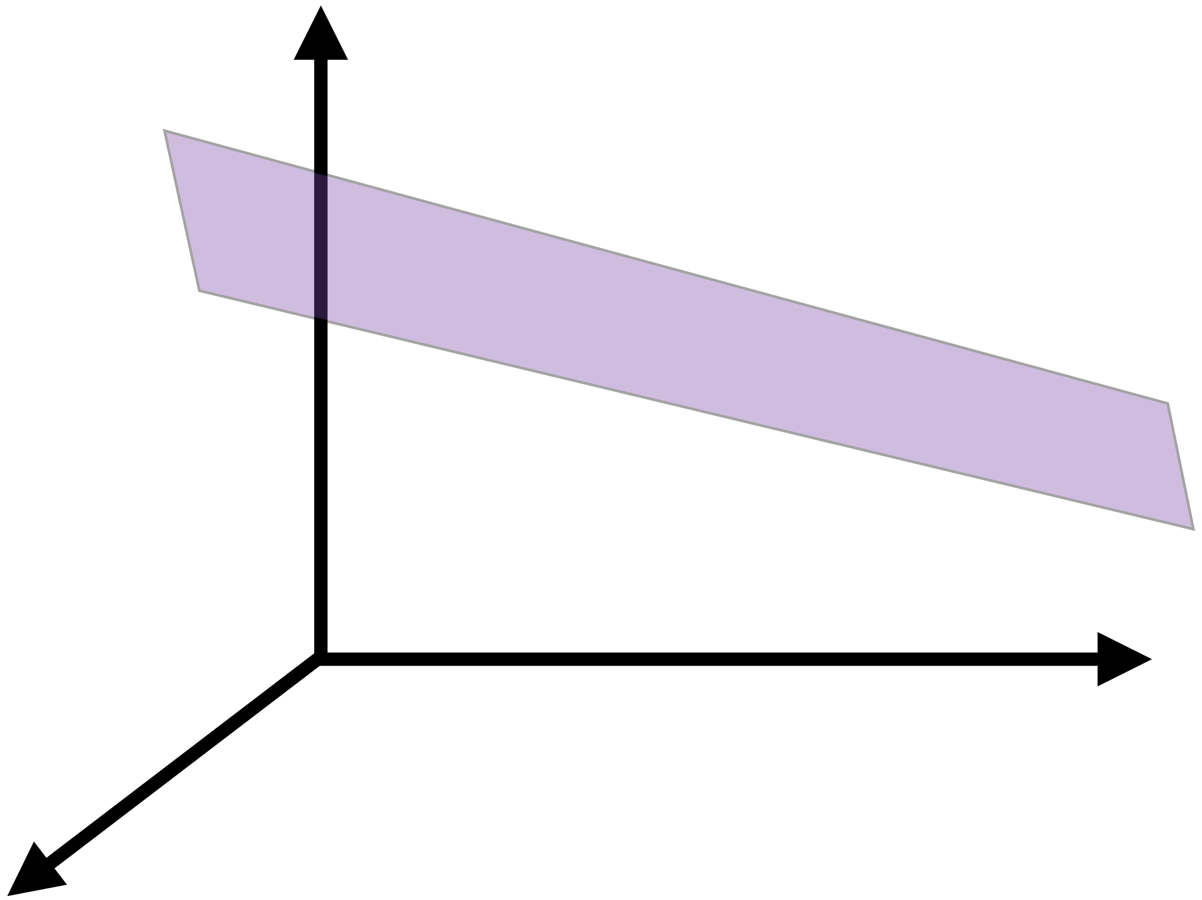
Data in > 2 dimensions

Simple and Multiple Regression



Simple Regression

One independent variable



Multiple Regression

Multiple independent variables

Three Estimation Methods

Method of moments

Method of least squares

Maximum likelihood estimation

Cookie cutter techniques to determine the values of A and B (regression coefficients)

"Best Linear Unbiased Estimator" (BLUE)

"Best"

Coefficients have minimum variance, i.e. are estimated with relatively high certainty

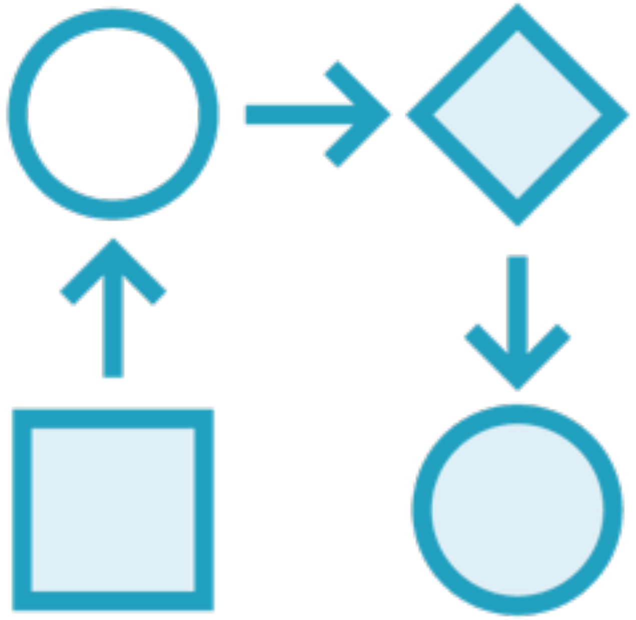
"Unbiased"

Residuals have zero mean, are uncorrelated to each other and have equal variance

Solving the regression problem with the method of least squares gives a BLUE solution

Explaining Variance Using Simple Regression

Two Common Applications of Regression



Explaining Variance

How much variation in one data series is caused by another?

Making Predictions

How much does a move in one series impact another?

Rising Stock: Alpha or Beta?

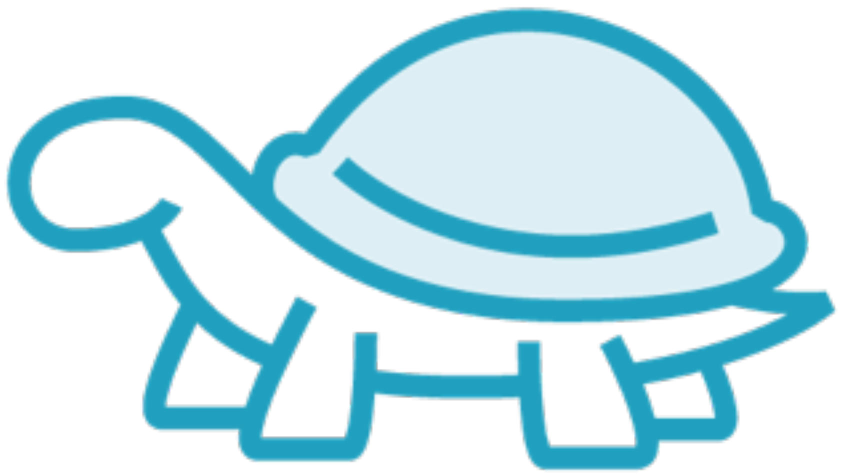
Company X's Stock Is Rising

The stock has risen 10% this year; the market is up 8% in the same period

Financial Analysts are Divided

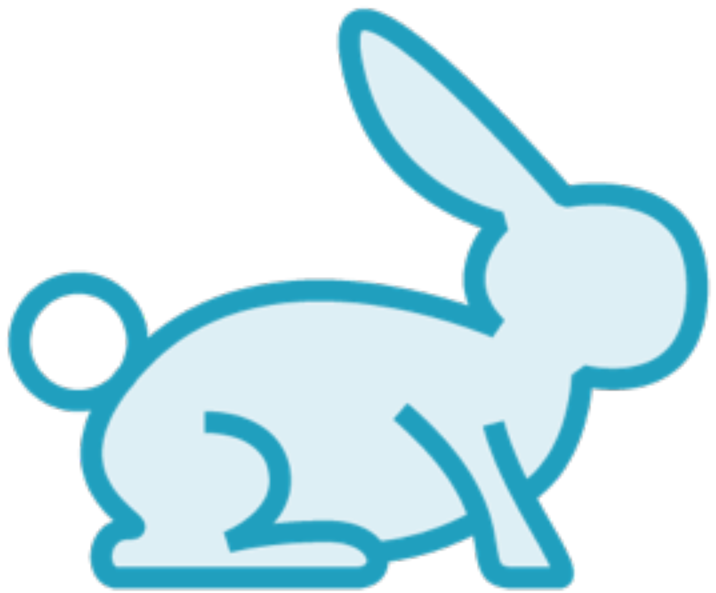
How much of the increase is explained by the market rise?

Rising Stock: Alpha or Beta?



Explanation #1: Beta

Price rise driven by beta, i.e. explained by market rise



Explanation #2: Alpha

Price rise can not be explained by market rise - company really has done something right

X Causes Y

Cause

Independent variable

Effect

Dependent variable

X Causes Y

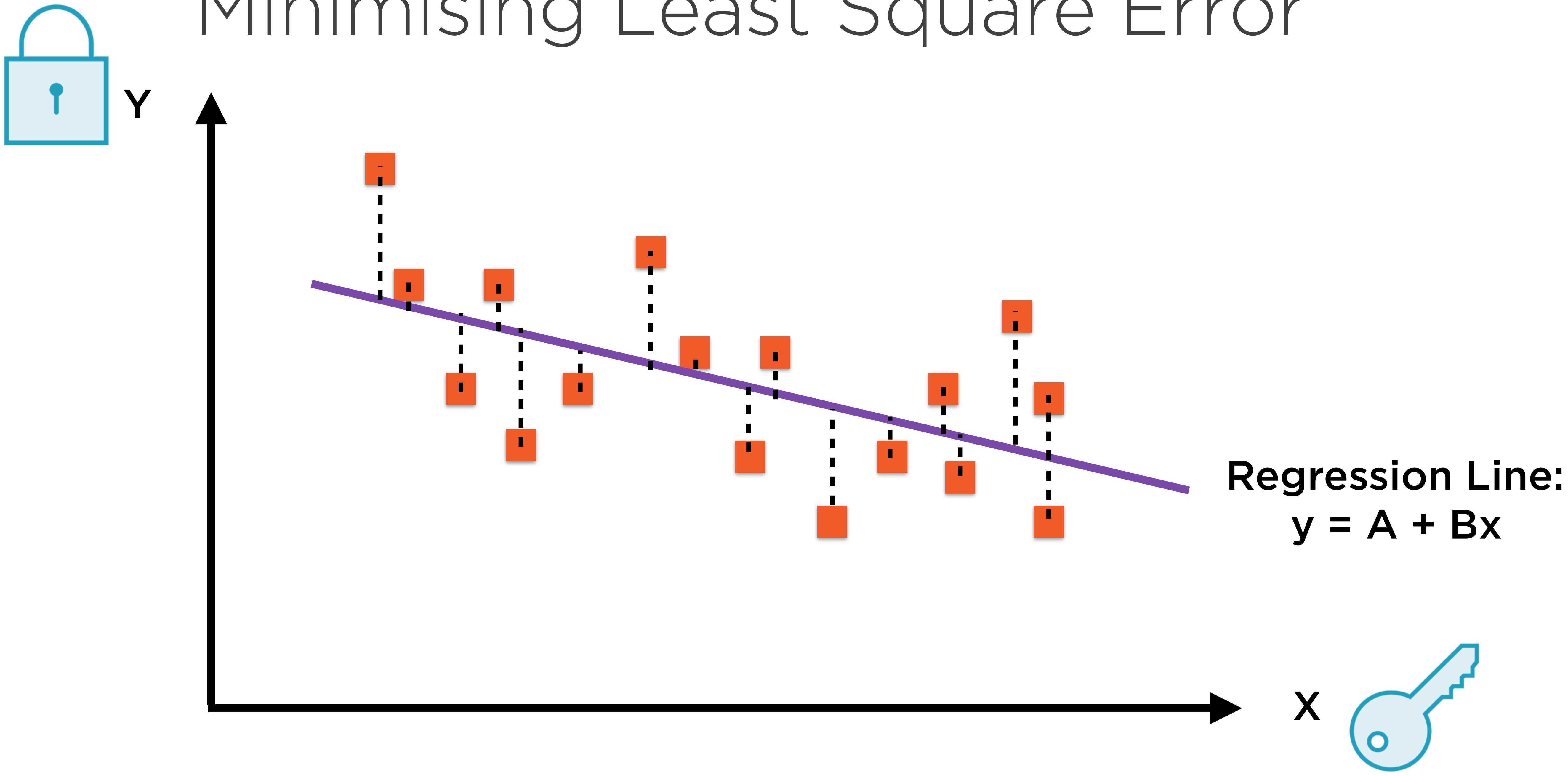
Cause

Explanatory variable

Effect

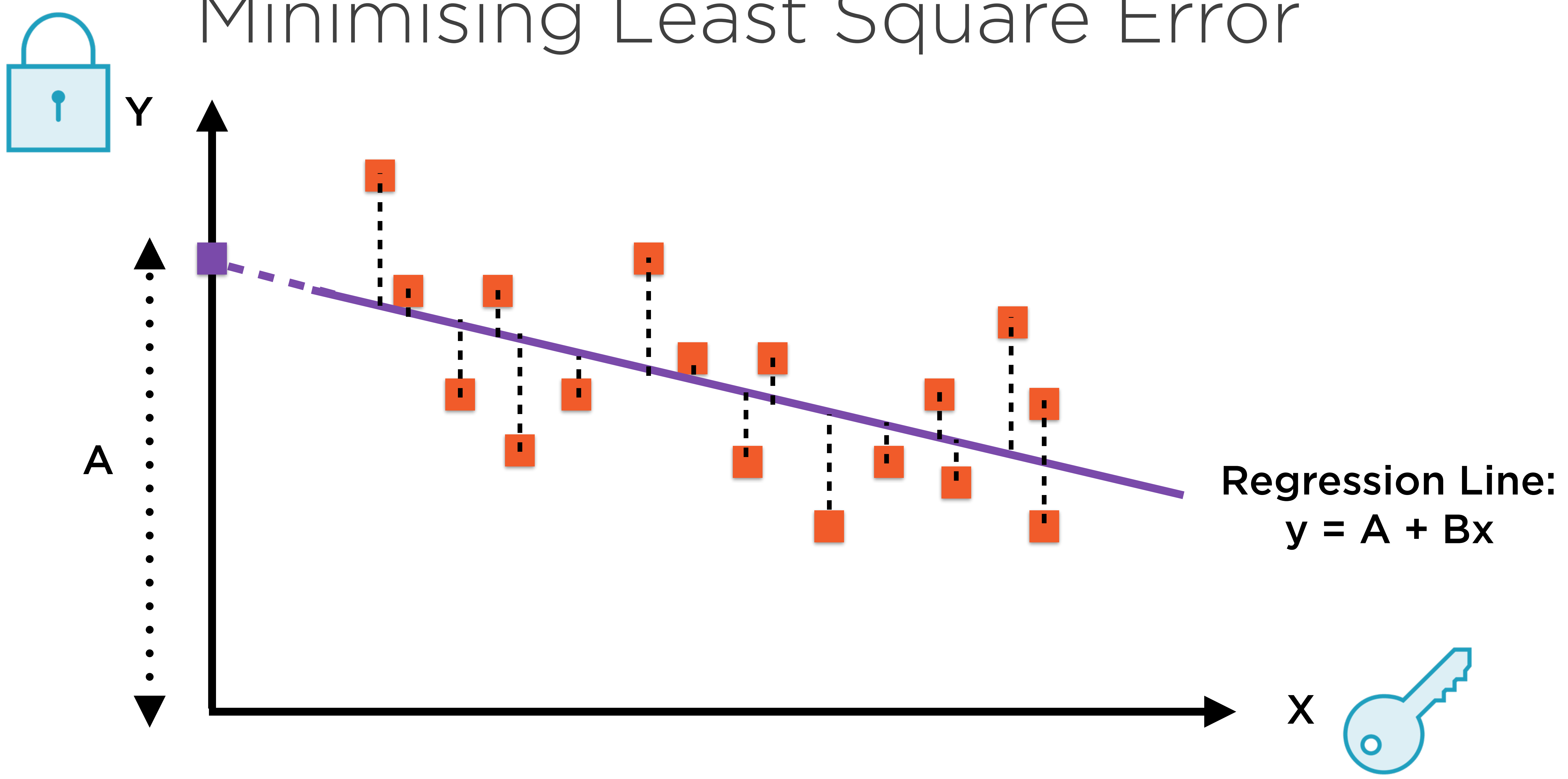
Dependent variable

Minimising Least Square Error



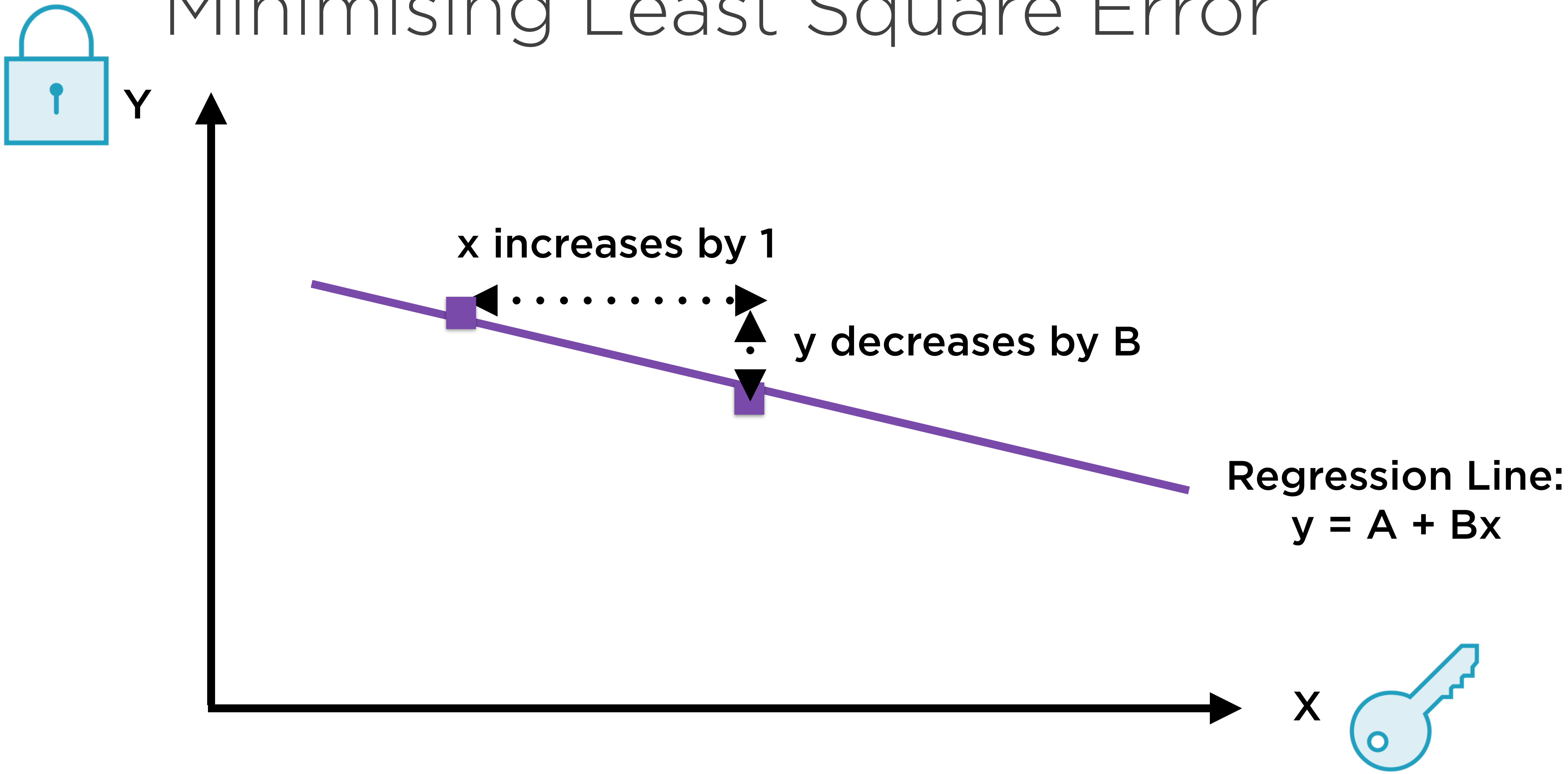
The “best fit” line is called the regression line

Minimising Least Square Error



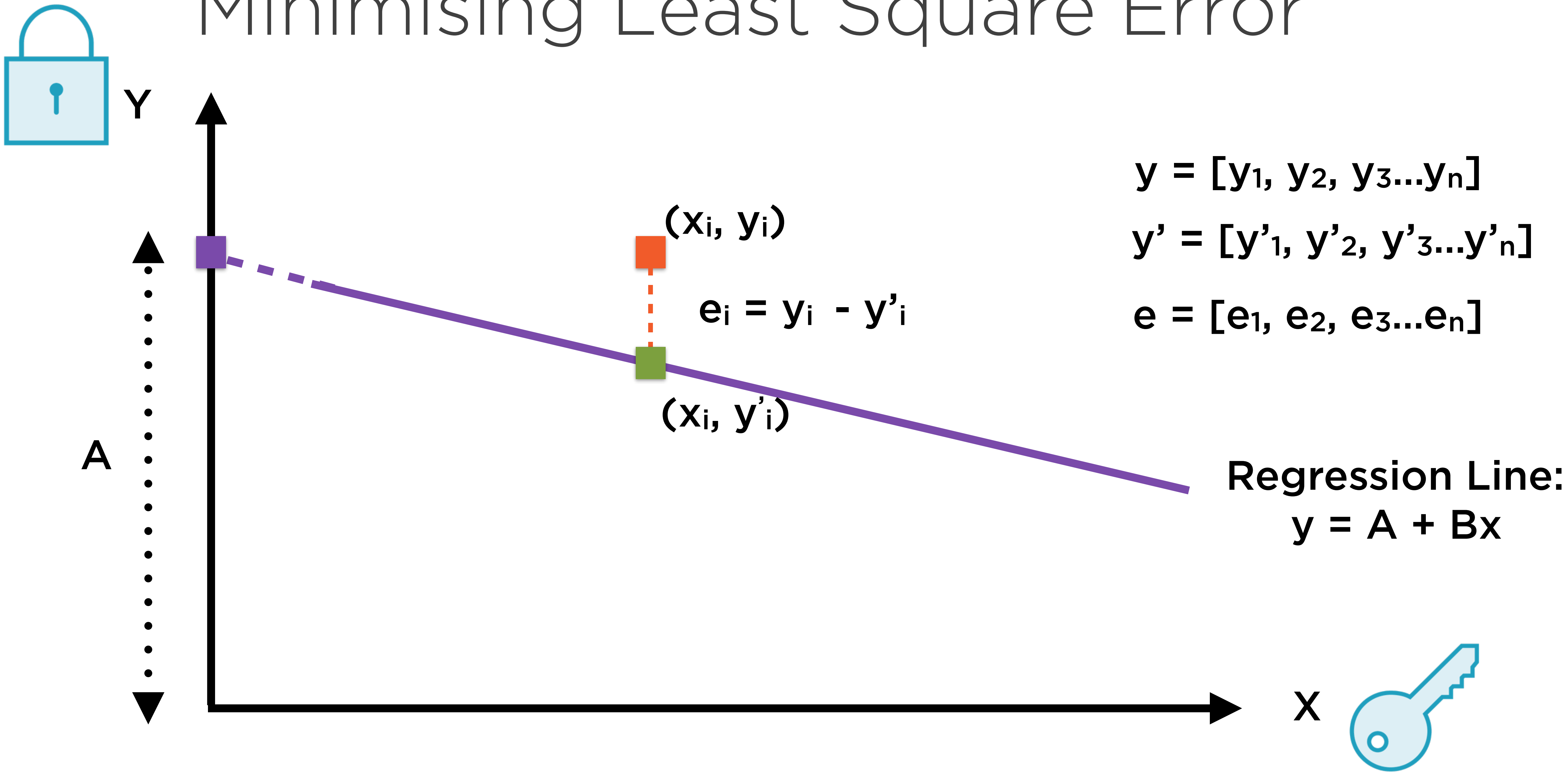
The term A in the equation of the line is the y-intercept

Minimising Least Square Error



The term B is the slope, and gives the sensitivity of y to a change of 1 unit in x

Minimising Least Square Error



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

Post Hoc Fallacy

New Roommate Moves In

Weather Turns Gloomy

Just because X happened before Y, it does not mean that X caused Y

Correlation Is Not Causation

Economy is Booming

Banks are Lending Freely

Not even Nobel Prize-winning economists can agree on this one!

Cause and Effect

Precedes

X happens before Y

Accompanies

X and Y happen together

Causes

X causes Y

Cause and Effect

Post hoc fallacy

X happens before Y, so we conclude that X causes Y

Correlation is causation fallacy

X and Y happen together so we conclude that X causes Y

Genuine Causation

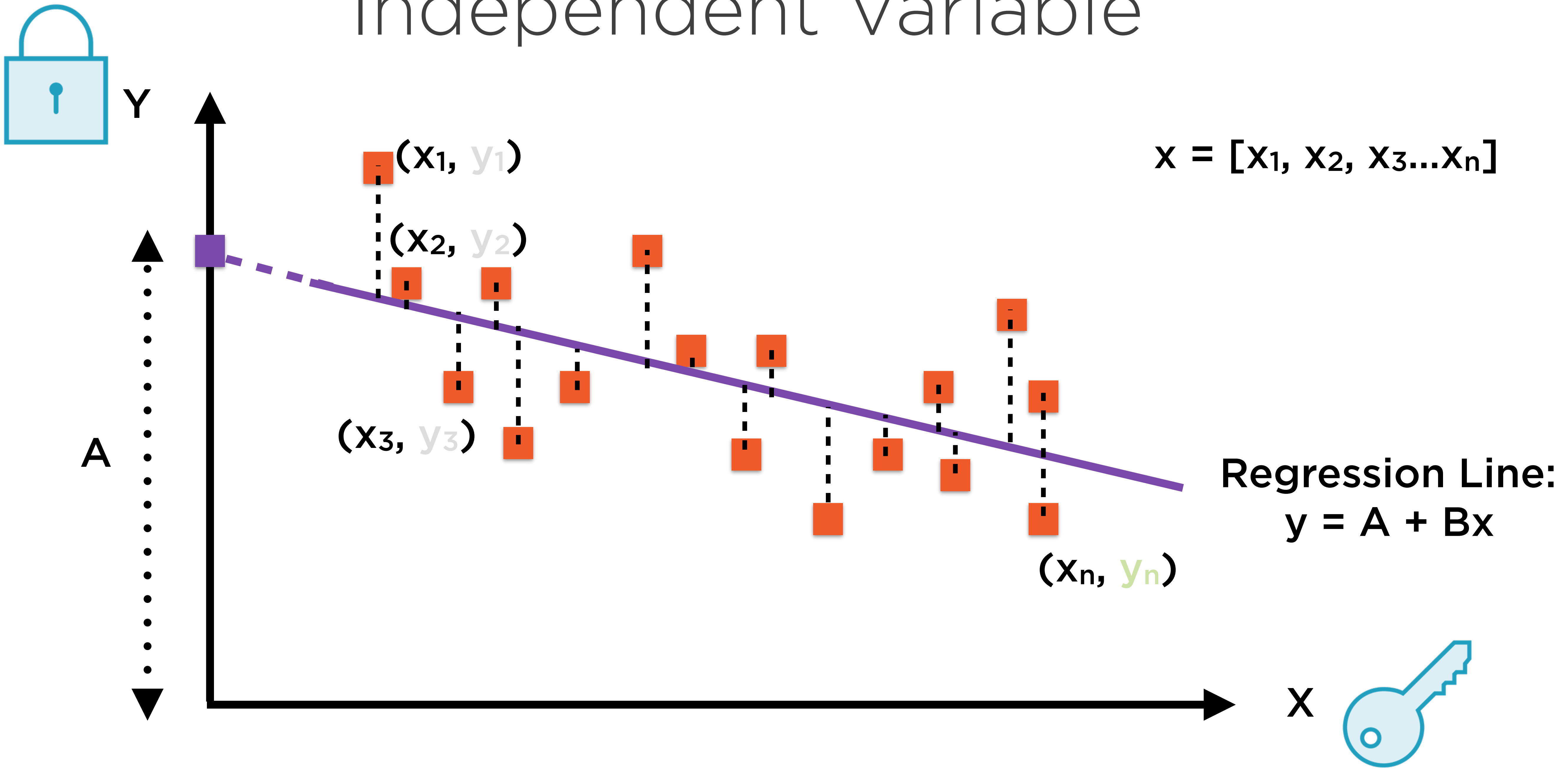
X actually causes Y; we can use regression to quantify causation

$x = [x_1, x_2, x_3 \ldots x_n]$

Independent Variable

If X causes Y, then values of x form a vector, called the independent variable or explanatory variable

Independent Variable



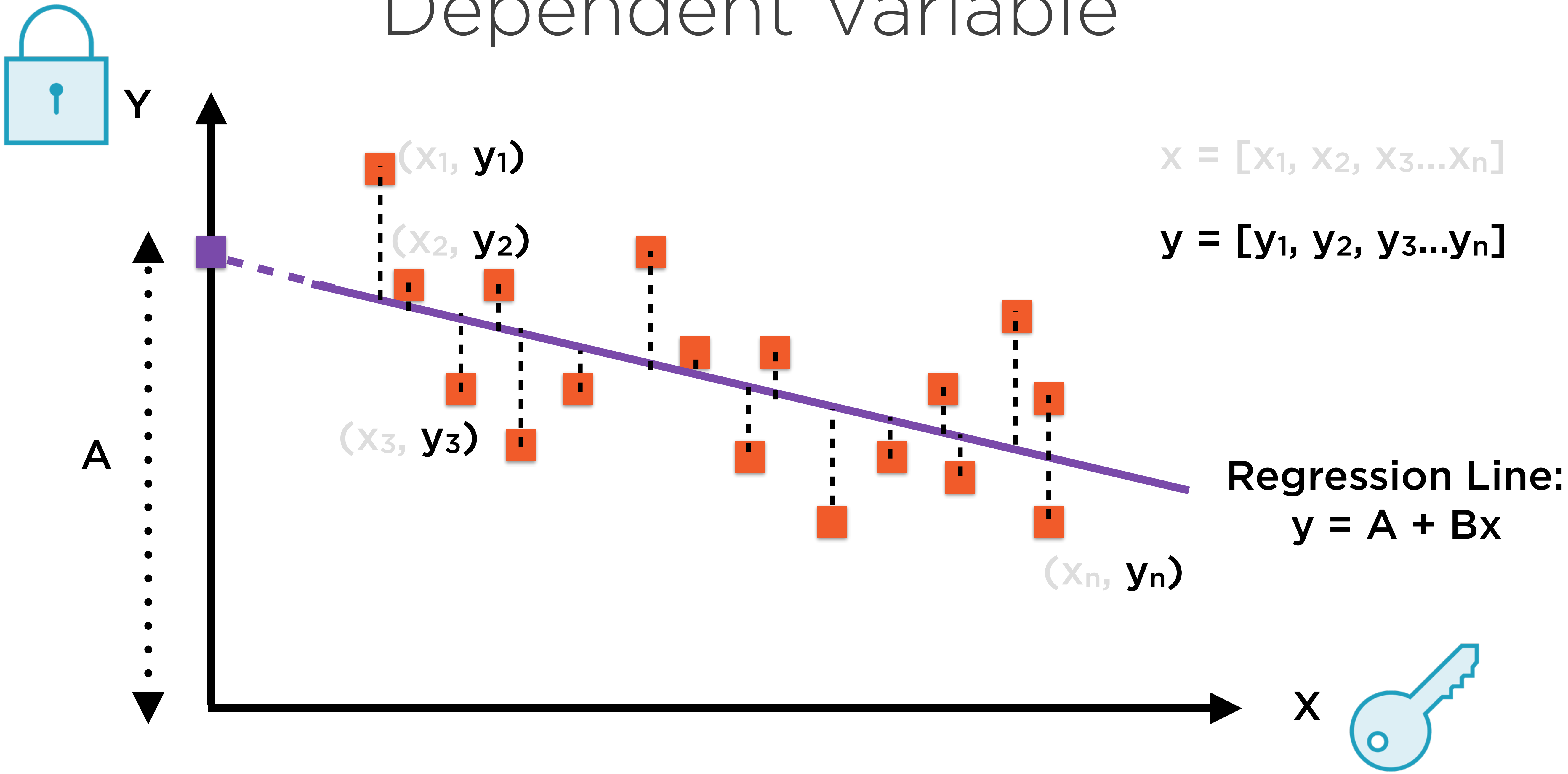
x in the regression line refers to the vector of all x coordinates

$y = [y_1, y_2, y_3 \ldots y_n]$

Dependent Variable

If X causes Y, then values of y form a vector, called the dependent variable or explained variable

Dependent Variable



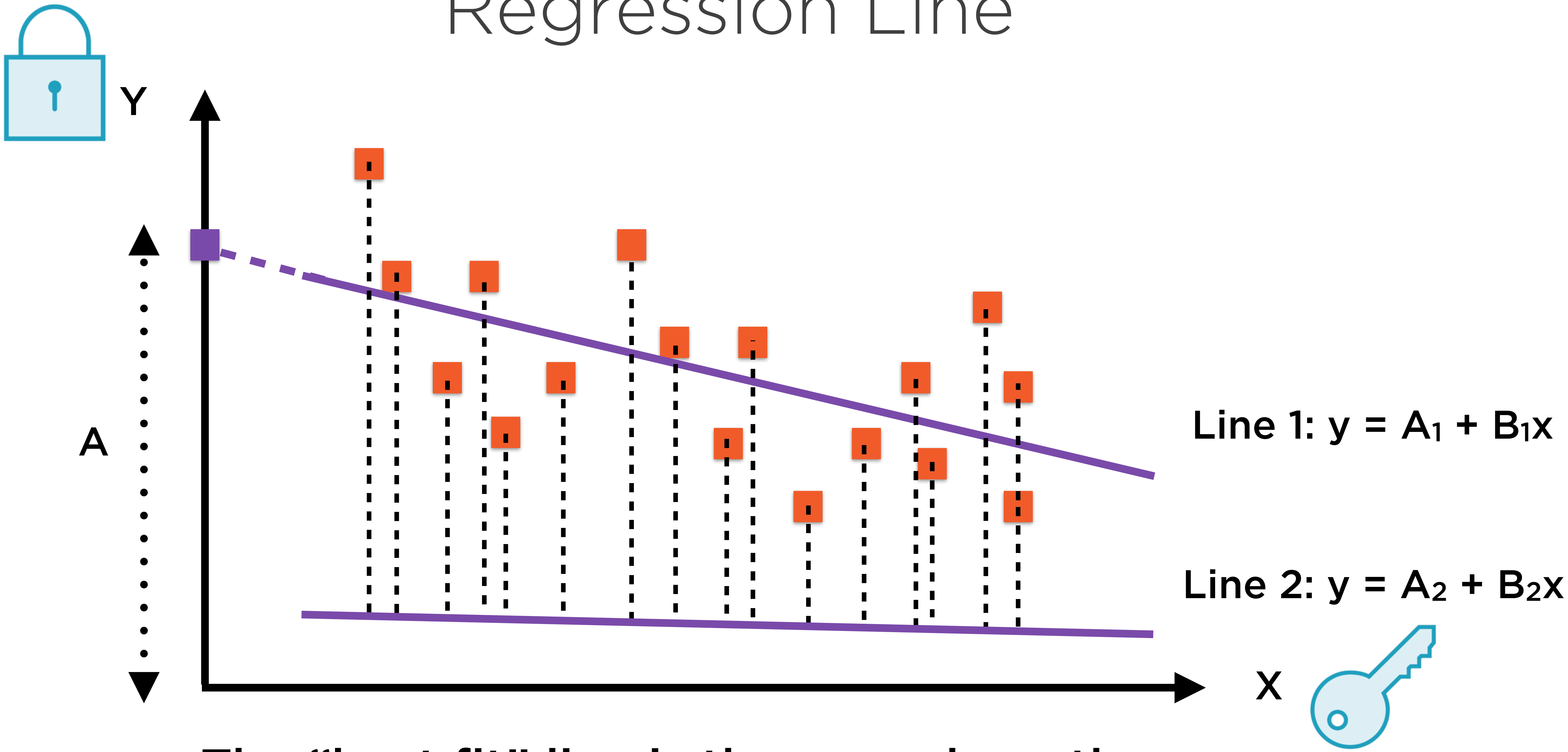
y in the regression line refers to the vector of all y coordinates

$y = A + Bx$

Regression Line

The “best fit” line which minimises the sum of the squares of the errors

Regression Line



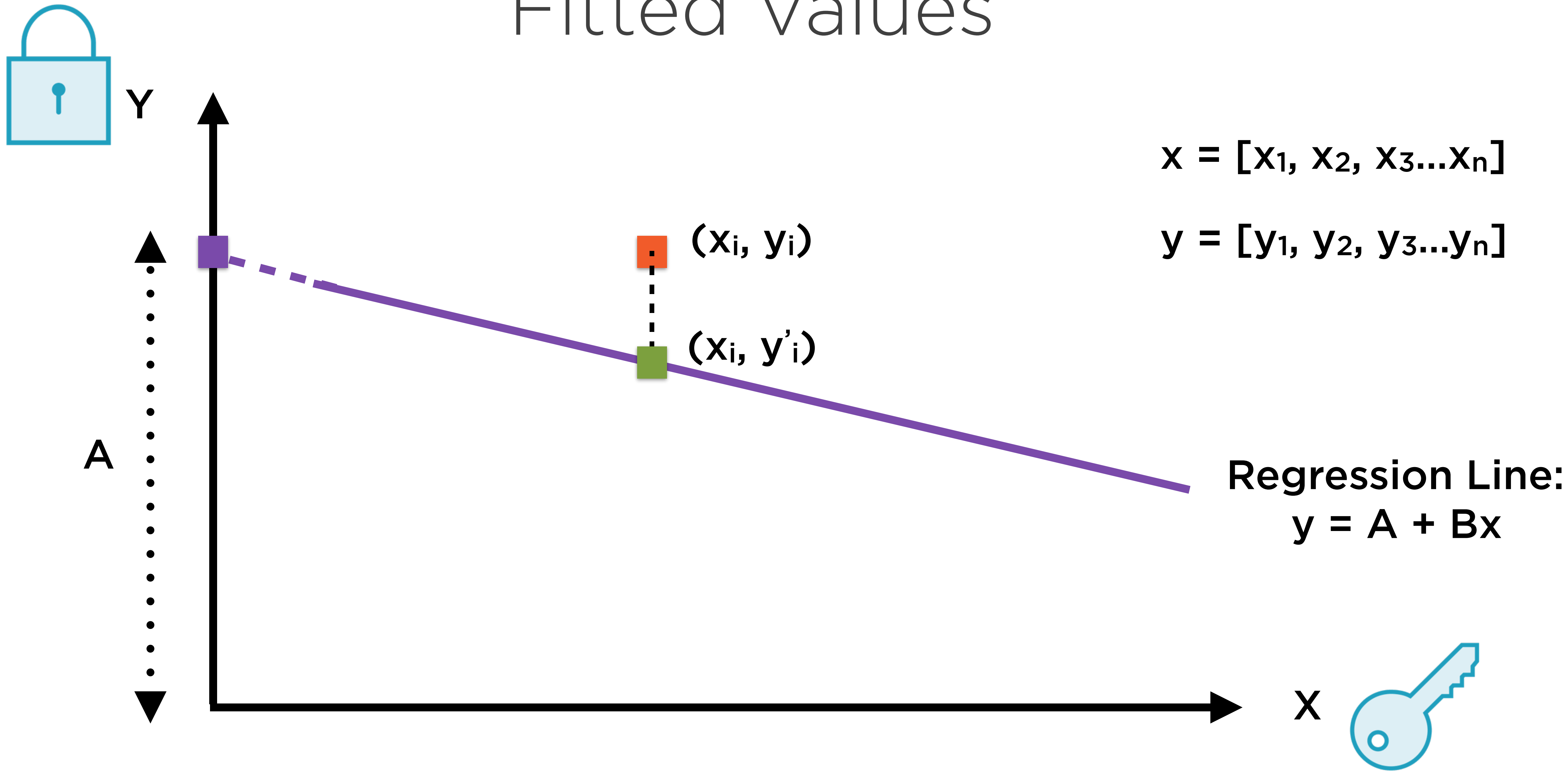
The "best fit" line is the one where the sum of the squares of the lengths of the errors is minimum

$$y' = [y'_1, y'_2, y'_3 \ldots y'_n] = A + Bx$$

Fitted Values of Dependent Variable

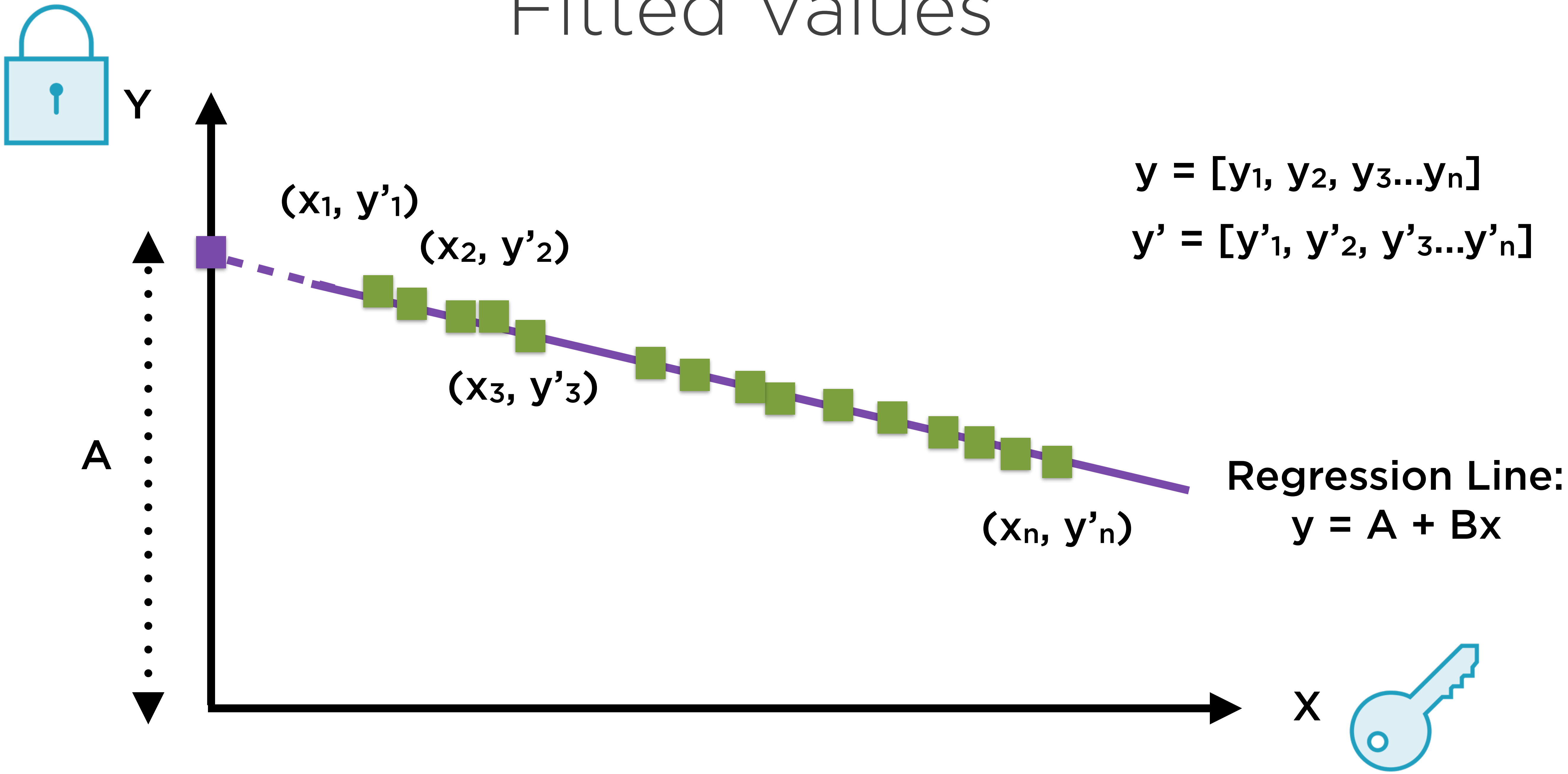
The fitted line $y = A + Bx$ will yield a different set of values, called the fitted values

Fitted Values



Each point (x_i, y_i) has a corresponding point (x_i, y'_i) on the regression line

Fitted Values



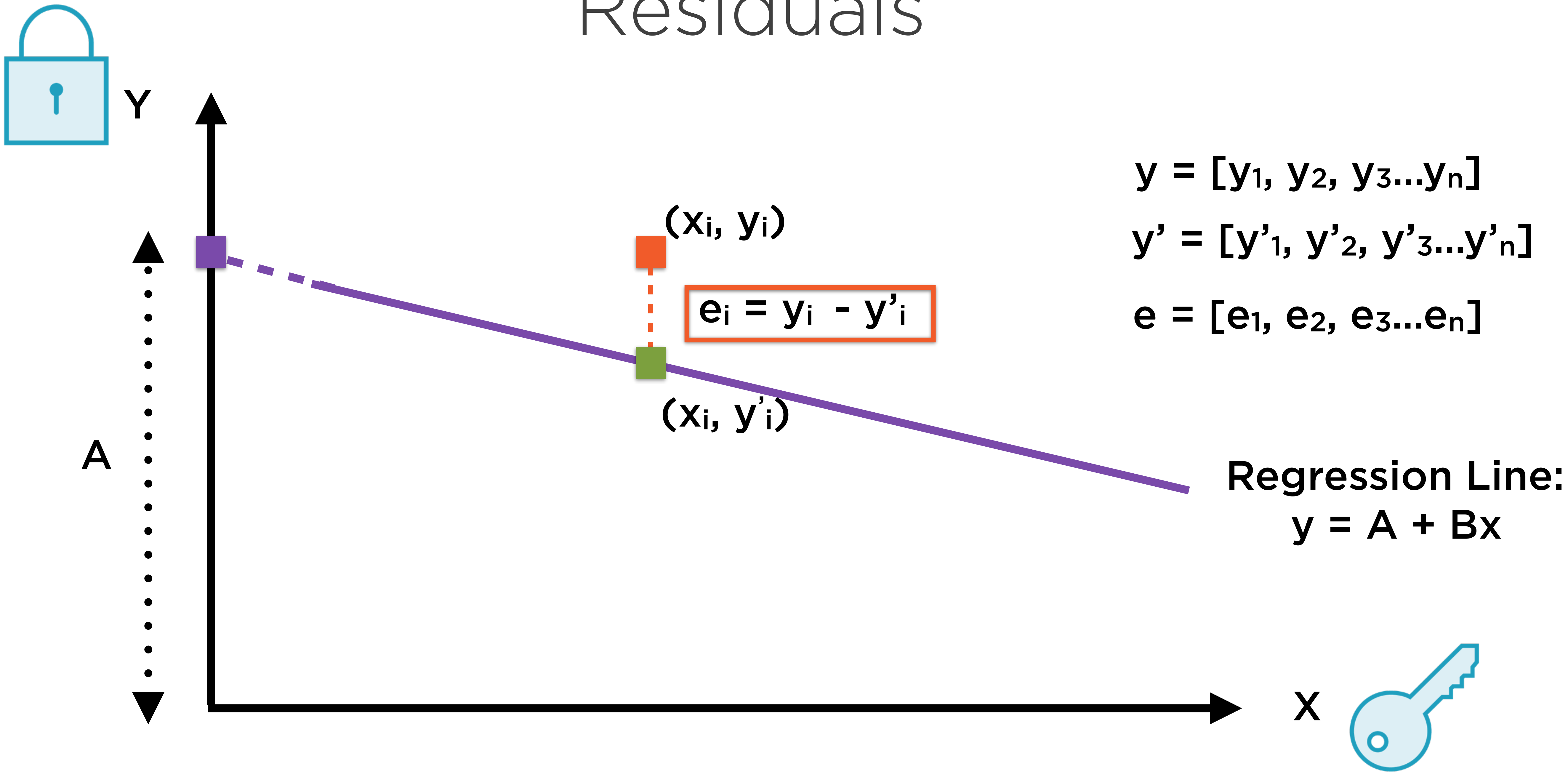
The corresponding values of y'_i are called the fitted values

$e = y - y'$

Residuals

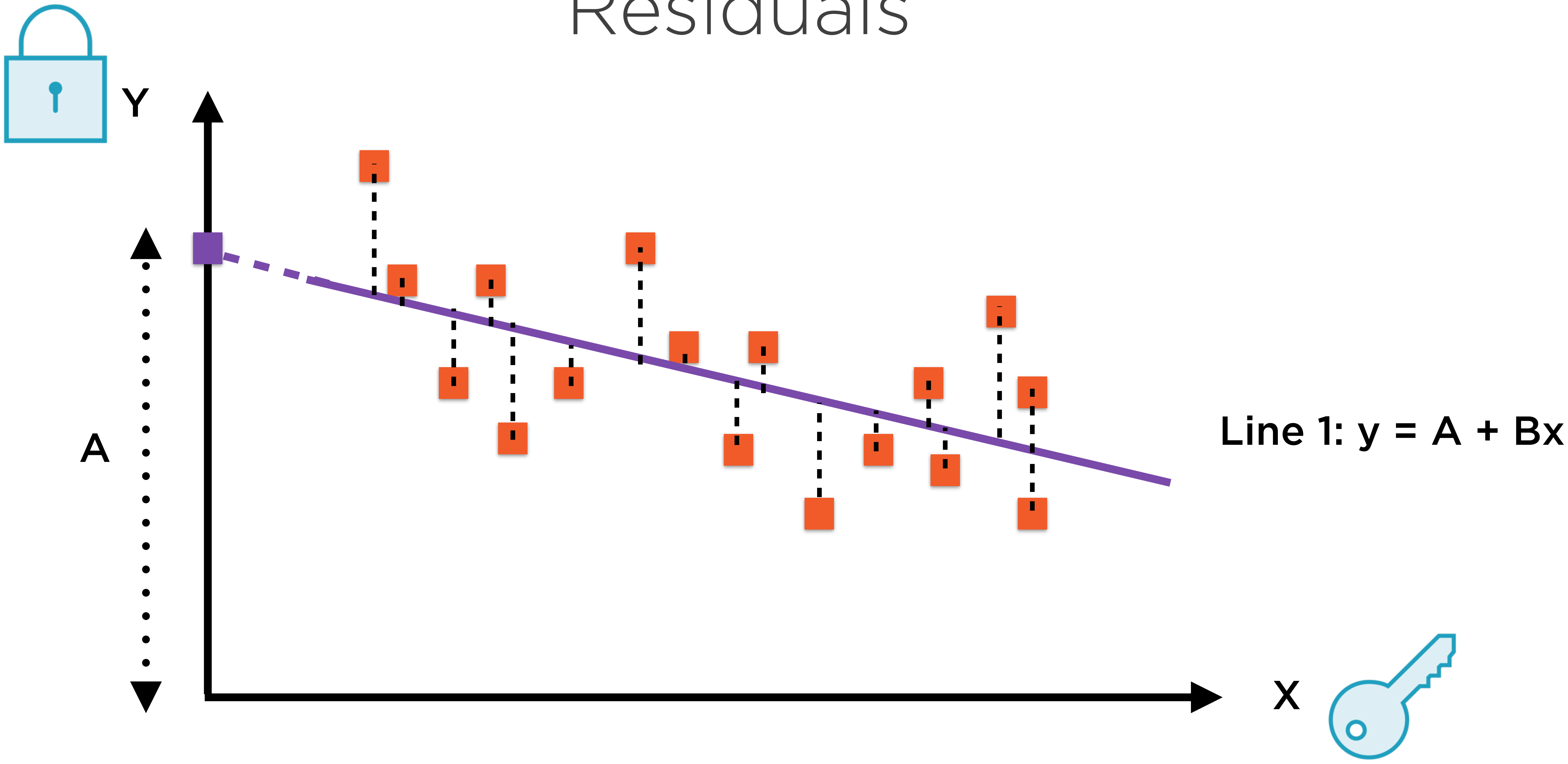
The residuals, or errors, are the differences between the actual and fitted values of the dependent variable

Residuals



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

Residuals



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

e = y - y’

=> y = y’ + e

=> Variance(y) = Variance(y’ + e)

=> Variance(y) = Variance(y’) + Variance(e) + Covariance(y’,e)

A Not-Very-Important Intermediate Step

Variance of the dependent variable can be decomposed into variance of the regression fitted values, and that of the residuals

$e = y' - y$

$\Rightarrow \quad y = y' + e$

$\Rightarrow \quad \text{Variance}(y) = \text{Variance}(y' + e)$

Always = 0

$\Rightarrow \quad \text{Variance}(y) = \text{Variance}(y') + \text{Variance}(e) + \boxed{\text{Covariance}(y',e)}$

Covariance: Only a Passing Mention

This is the only time in the course we will allude to covariance

Variance(y) = Variance(y’) + Variance(e)

Variance Explained

Variance of the dependent variable can be decomposed into variance of the regression fitted values, and that of the residuals

$\boxed{\text{Variance(y)}}$ = Variance(y') + Variance(e)

Total Variance (*TSS*)

A measure of how volatile the dependent variable is, and of much it moves around

TSS = $\boxed{\text{Variance(y')}}$ + Variance(e)

Explained Variance *(ESS)*

A measure of how volatile the fitted values are - these come from the regression line

TSS = Variance(y)

TSS = ESS + $\boxed{\text{Variance(e)}}$

Residual Variance *(RSS)*

This the variance in the dependent variable that can not be explained by the regression

TSS = Variance(y)   ESS = Variance(y')

TSS = ESS + RSS

Variance Explained

Variance of the dependent variable can be decomposed into variance of the regression fitted values, and that of the residuals

TSS = Variance(y)   ESS = Variance(y')   RSS = Variance(e)

$R^2 = ESS / TSS$

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%)

TSS = Variance(y)   ESS = Variance(y') RSS = Variance(e)

Rising Stock: Alpha or Beta?

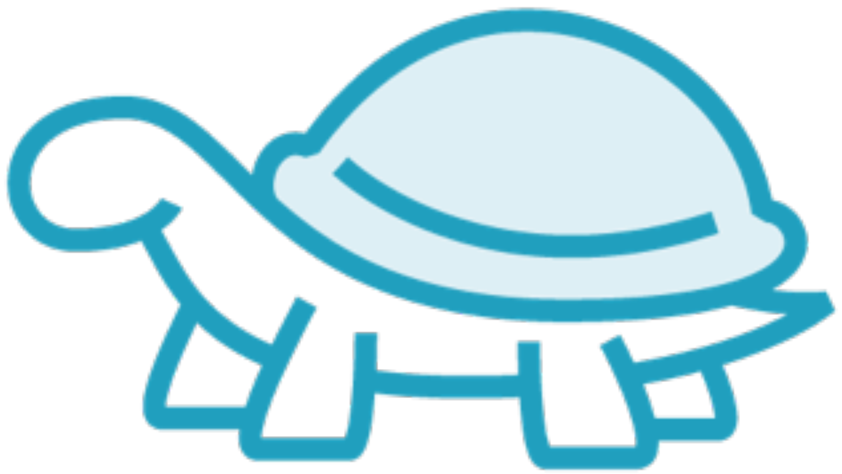
Company X's Stock Is Rising

The stock has risen 10% this year; the market is up 8% in the same period

Financial Analysts are Divided

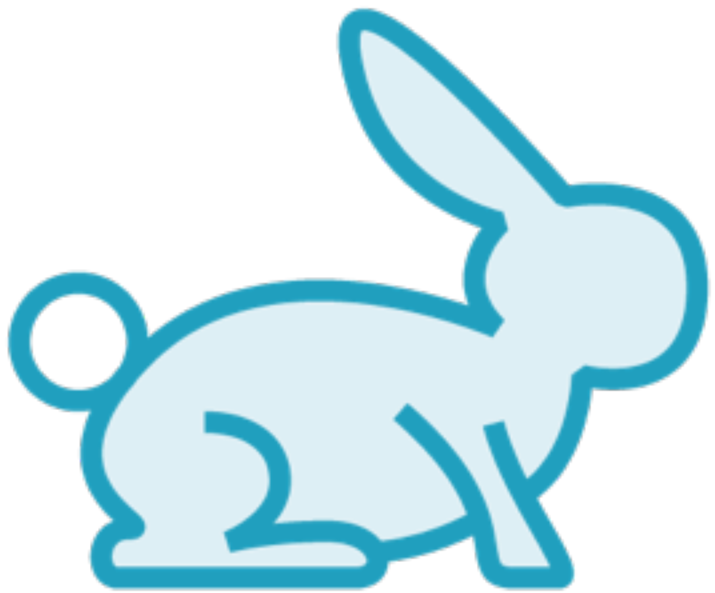
How much of the increase is explained by the market rise?

Rising Stock: Alpha or Beta?



Explanation #1: Beta

Price rise driven by beta, i.e. explained by market rise

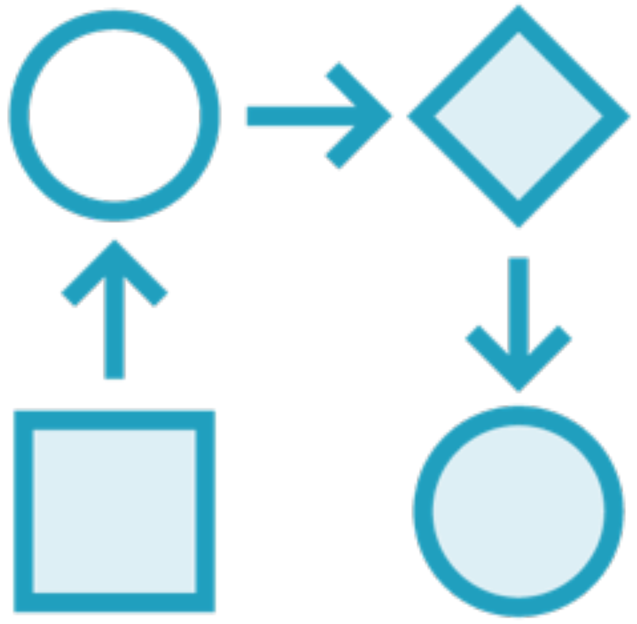


Explanation #2: Alpha

Price rise can not be explained by market rise - company really has done something right

Prediction Using Simple Regression

Two Common Applications of Regression



Explaining Variance

How much variation in one data series is caused by another?

Making Predictions

How much does a move in one series impact another?

Predictions Using Regression

Connect sets of dots

Express relationships between data series

Avoid jumping to conclusions

Measure how strong those relationships are

Predict where new dots will be

Make forecasts, recommendations

Predictions Using Regression

Input Data Series

Two columns, x and y

From underlying database

Specify Functional Form

y = A + Bx

Values of A and B yet to be determined

Find Model Parameters

Find values of A and B

Excel, R and most tools

Check Model Quality

Residuals, R^2

Also in Excel, R...

Predict

Given new x, what is y?

Answer using y = A + Bx

Act

Forewarned is forearmed

Based on possible outcomes

Regression Models in Commodity Trading

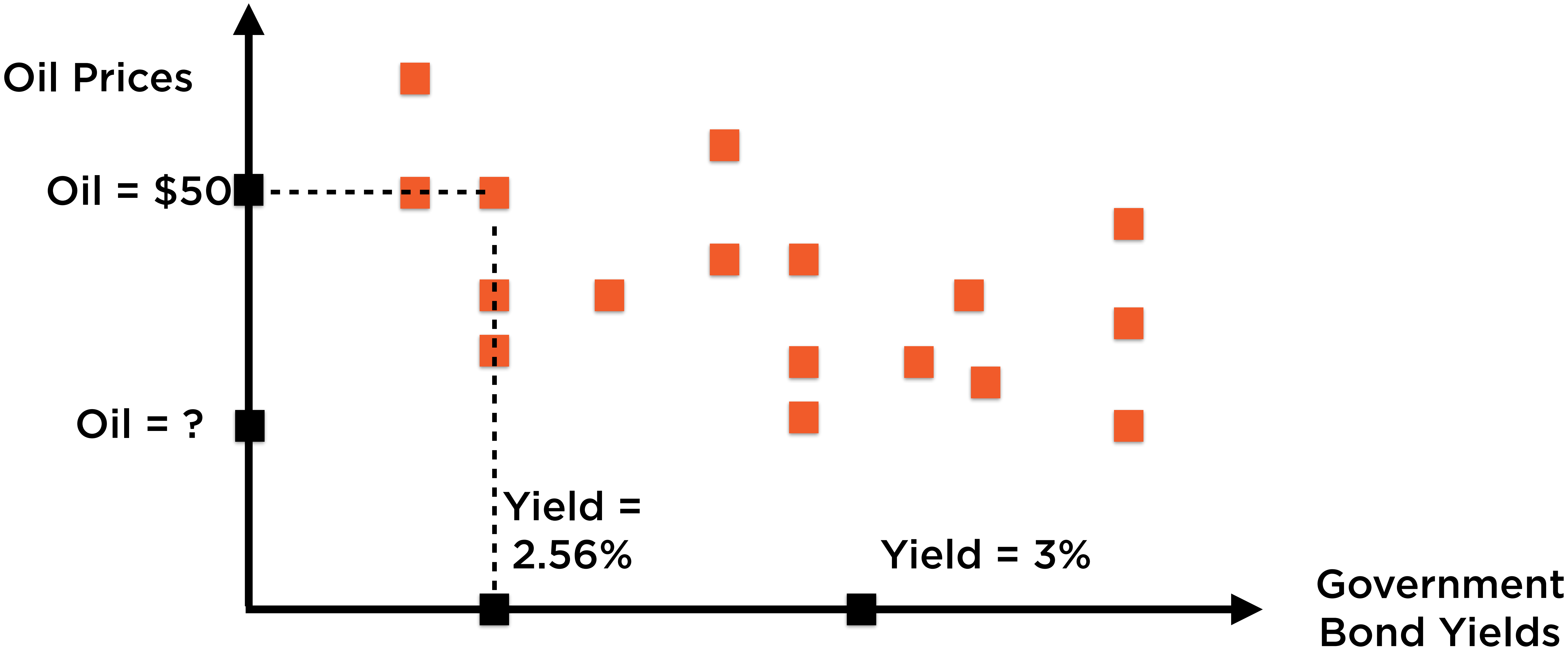
Interest Rates are Rising

US government bond yields are now at 2.56%, but could go to 3%

Commodity Traders are Worried

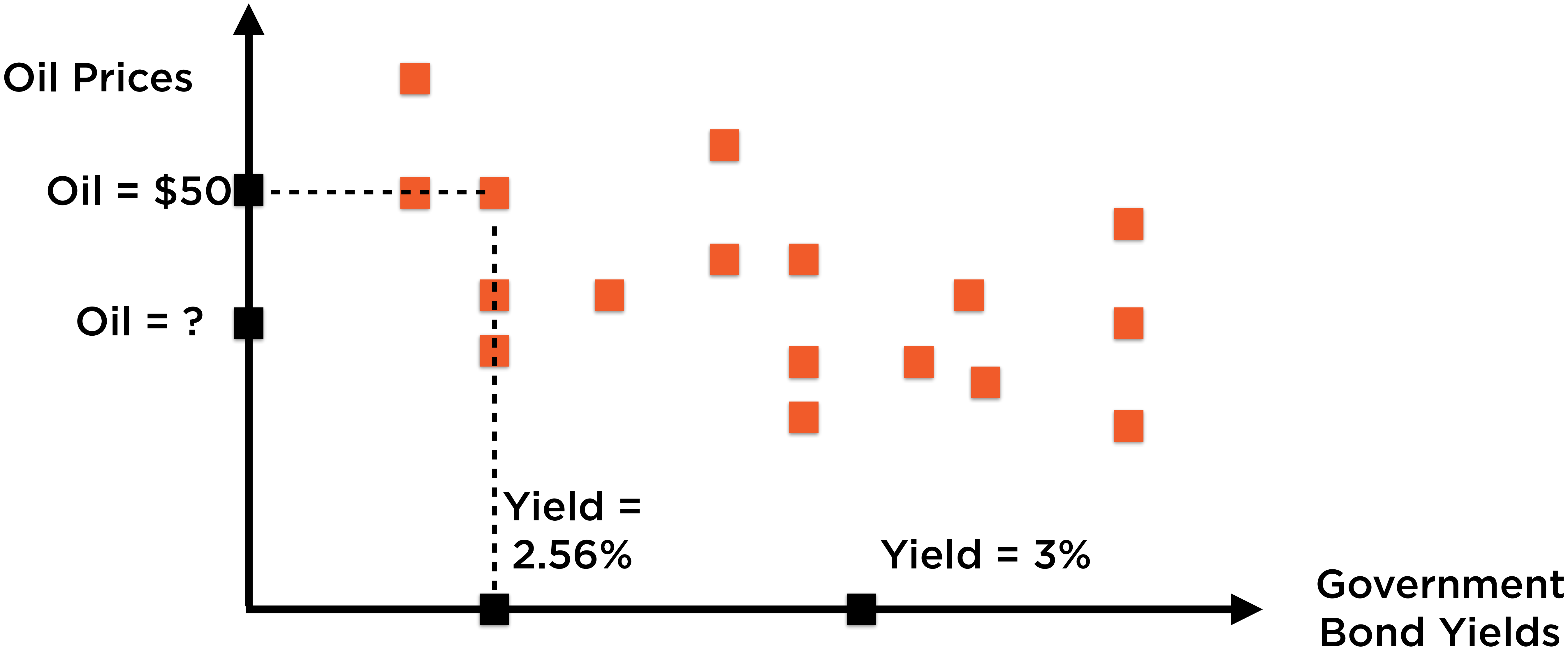
Oil is currently trading at $50/ barrel - buy or sell?

Prediction Using Regression



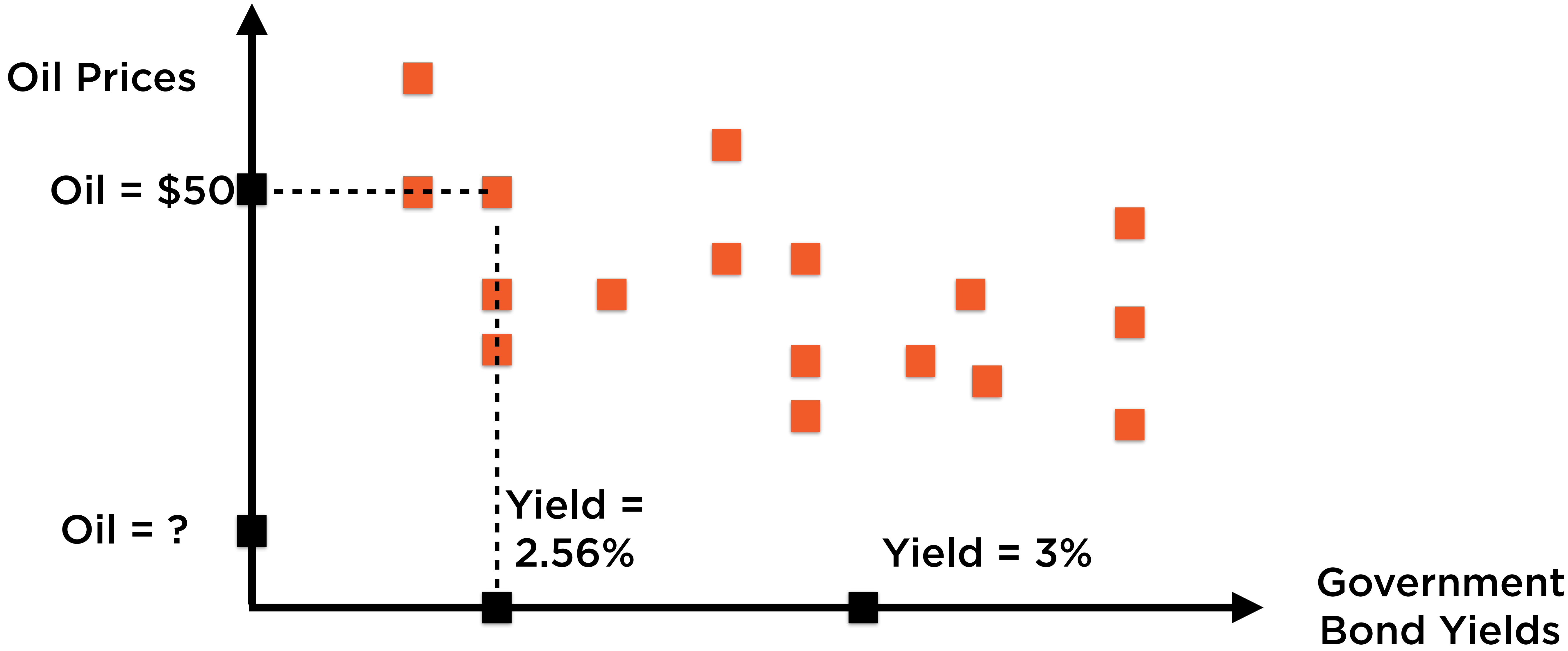
Today, 10-year yield = 2.56%, oil price = $50
Tomorrow, if 10-year yield at 3%, oil price = ?

Prediction Using Regression



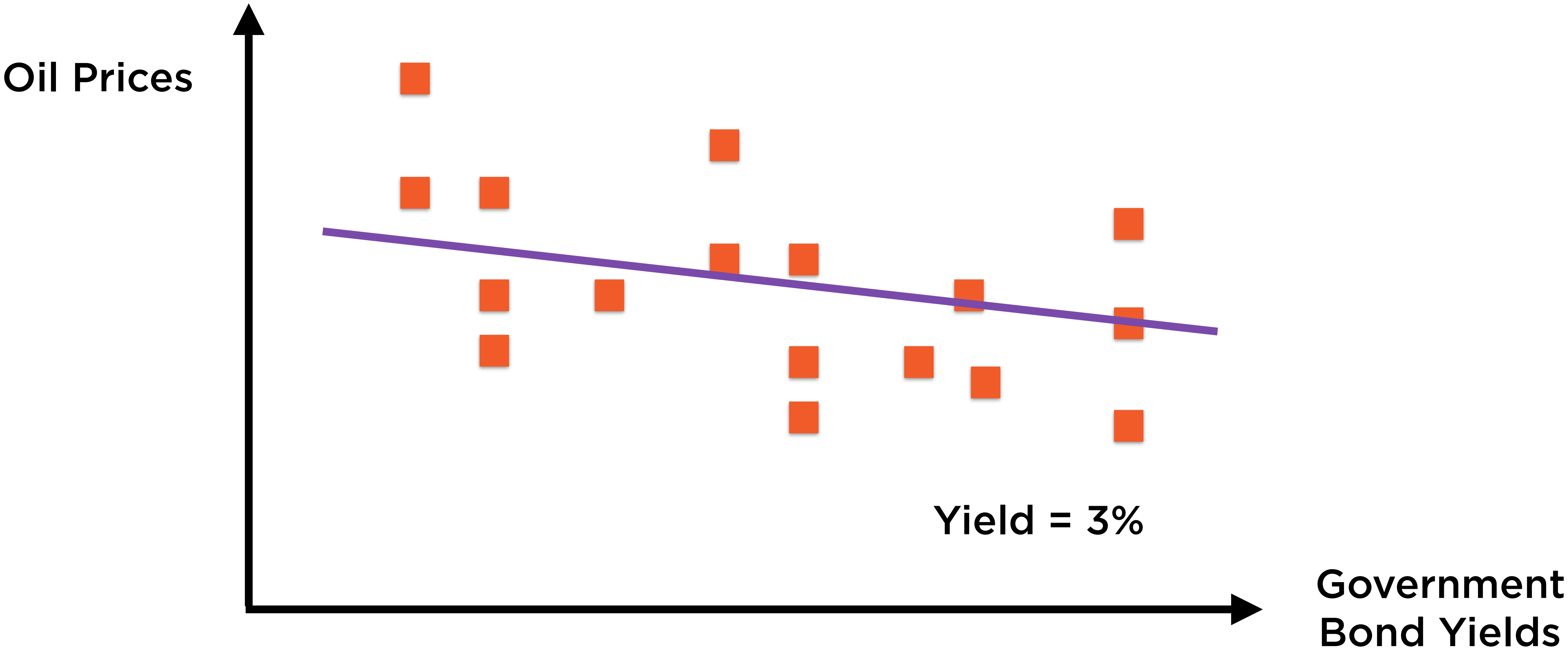
Today, 10-year yield = 2.56%, oil price = $50
Tomorrow, if 10-year yield at 3%, oil price = ?

Prediction Using Regression



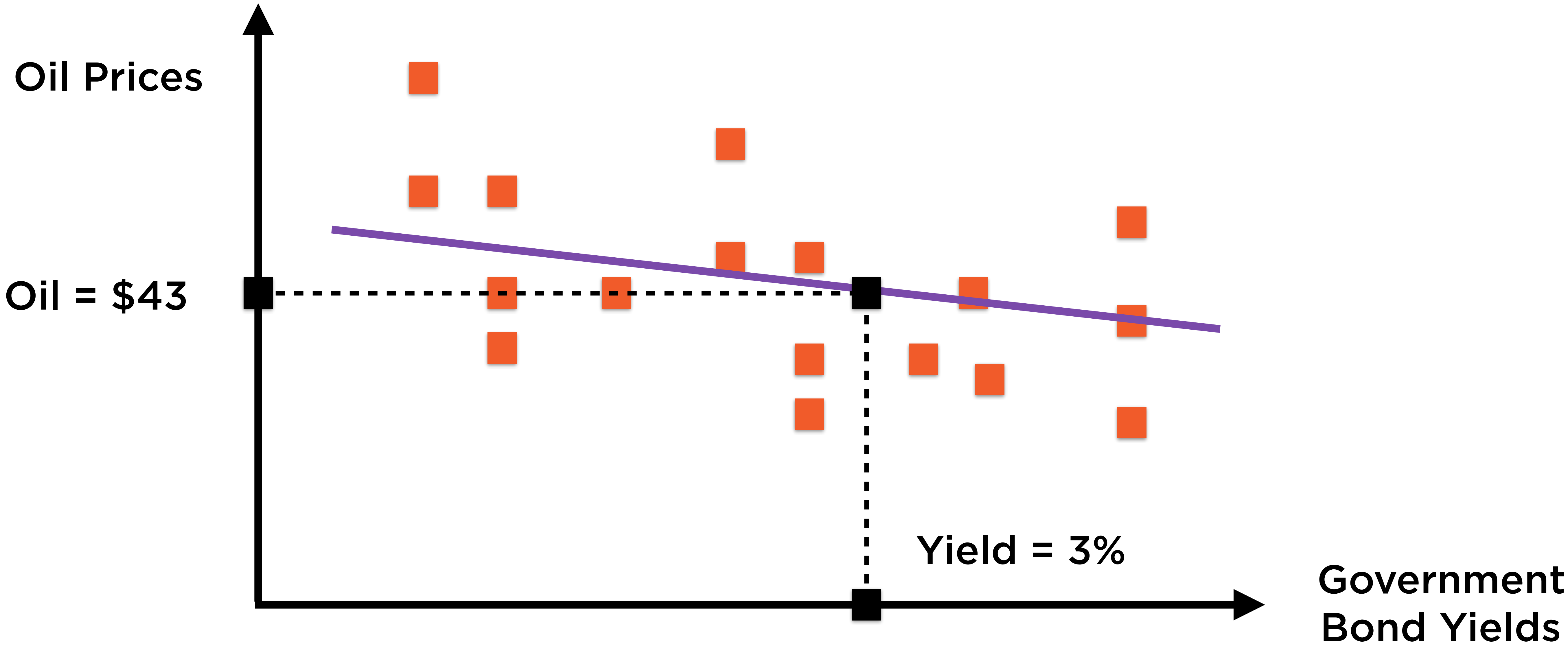
Today, 10-year yield = 2.56%, oil price = $50
Tomorrow, if 10-year yield at 3%, oil price = ?

Prediction Using Regression



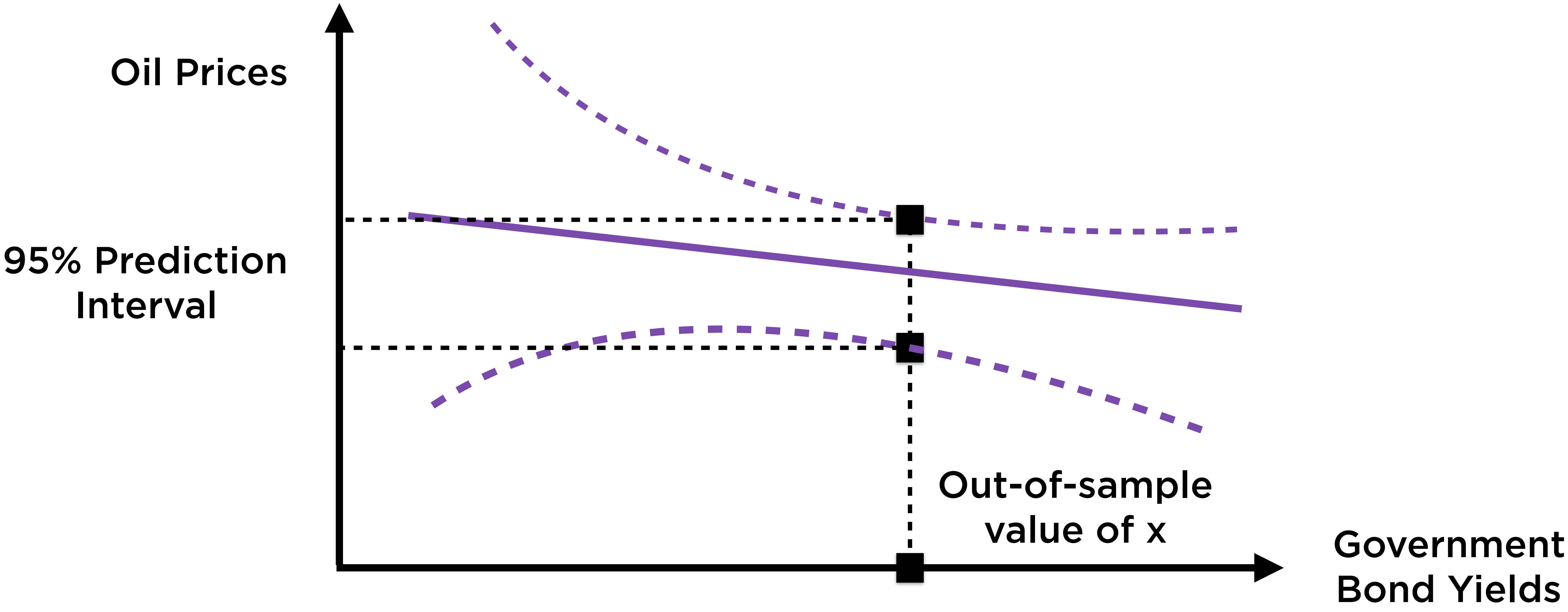
Find the regression line - the line with the "best fit"

Prediction Using Regression



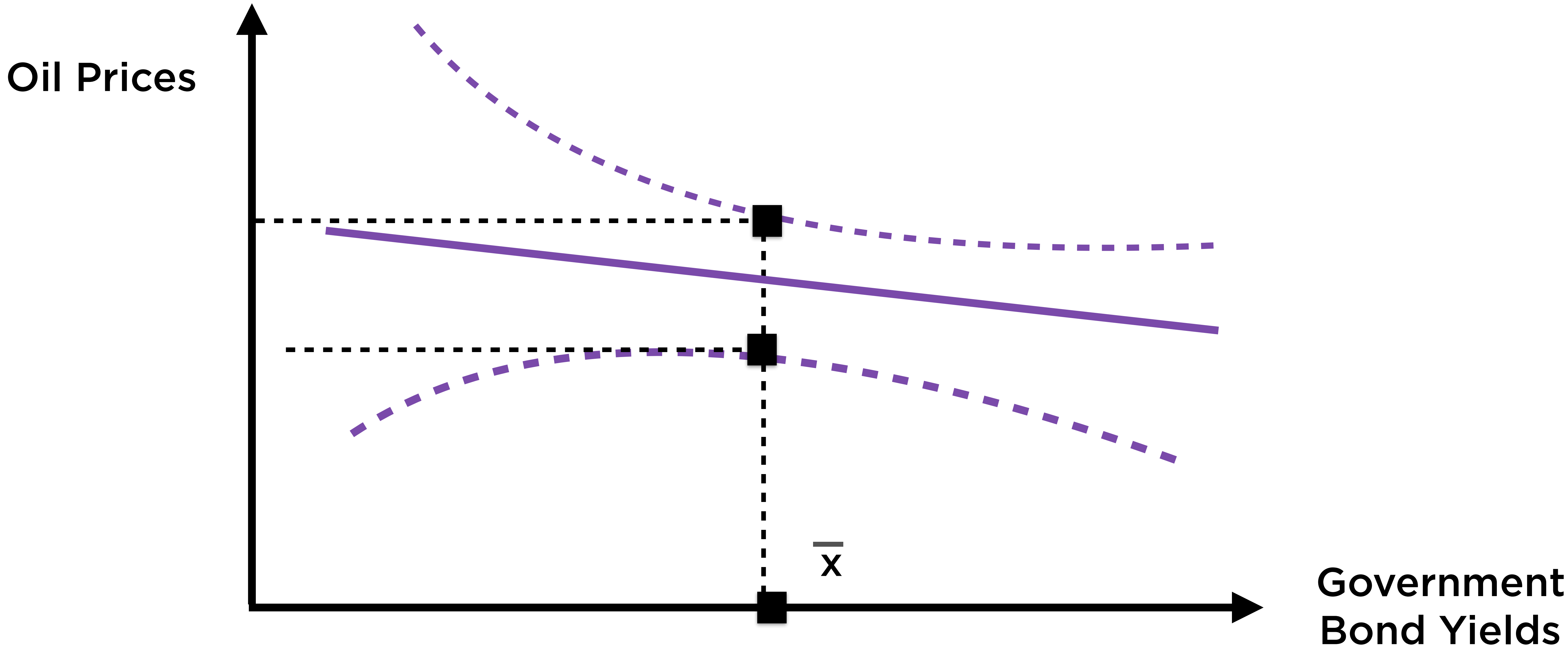
Given a new value of x, use the line to predict the corresponding value of y

Prediction Using Regression



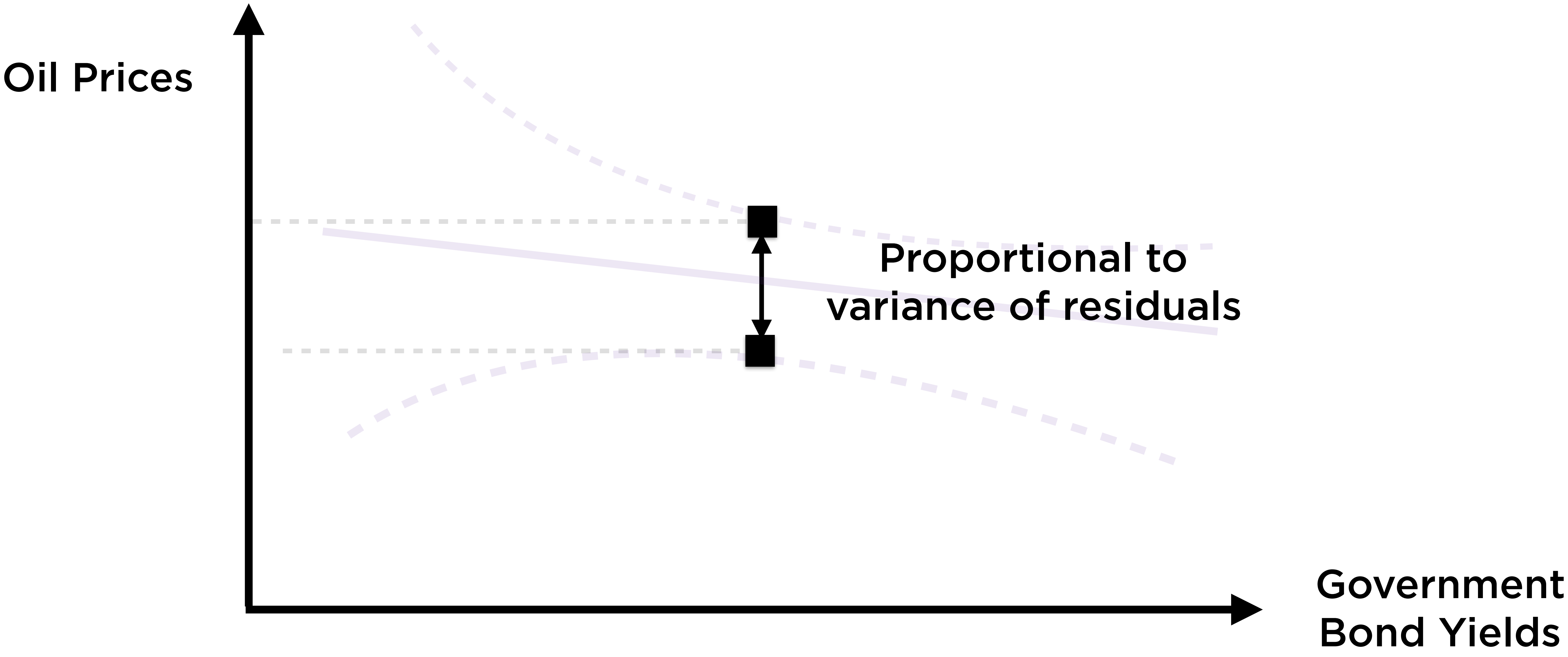
Regression also allows to specify **prediction intervals** (similar to confidence intervals) around this point estimate

Prediction Using Regression



This error is least at $x = \overline{x}$

Prediction Using Regression



The less the variance of the residuals, the more precise the prediction

Summary

Set up the regression problem

Understood the least-squares estimator and its BLUE property

Applied regression to forecasting and explaining variance

Discussed the assumptions about residuals that underpin regression