Building Regression Models with scikit-learn

UNDERSTANDING LINEAR REGRESSION AS A MACHINE LEARNING PROBLEM

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Linear regression as a machine learning problem

Mean Square Error (MSE) as loss function

Interpreting the results of a regression analysis

R^2 for evaluating regression models

Prerequisites and Course Outline

Prerequisites

Basic Python programming

No prior ML exposure required

High school math

Prerequisite Courses

Building Your First scikit-learn Solution

Course Outline

Understanding the regression problem

Building simple regression models

Building regularized regression models

Advanced regression techniques

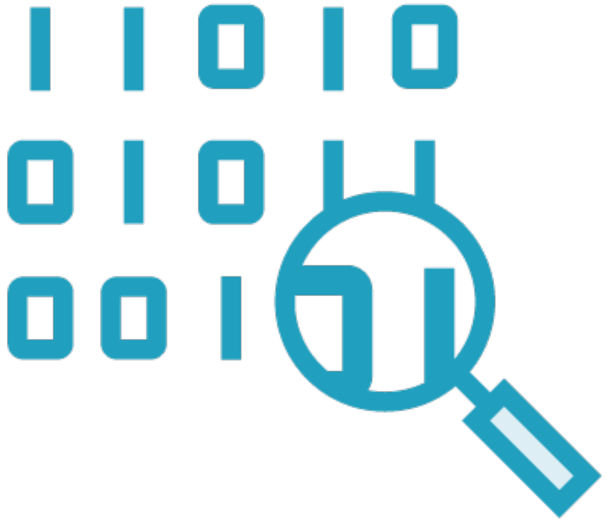
Hyperparameter tuning for regression

Connecting the Dots Using Linear Regression

“My mind is made up. Don’t confuse me with the facts.”

Some powerful person

Thoughtful, Fact-based Point of View



Fact-based

Built with painstakingly collected data

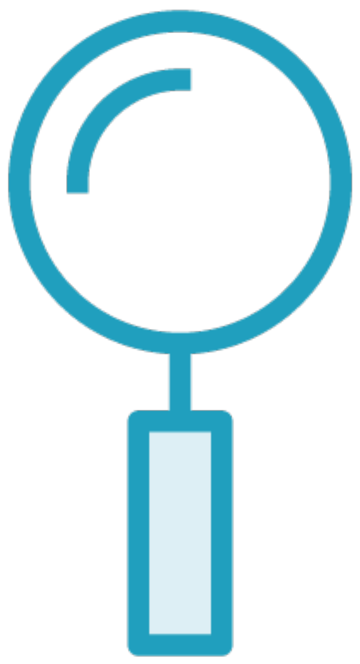
Thoughtful

Balanced, weighing pros and cons

Point of View

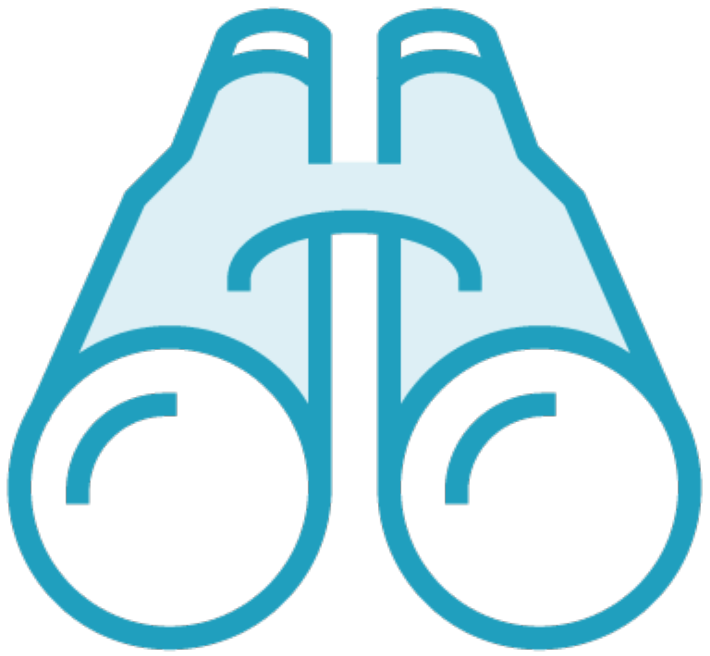
Prediction, recommendation, call to action

Two Sets of Statistical Tools



Descriptive Statistics

Identify important elements in a dataset



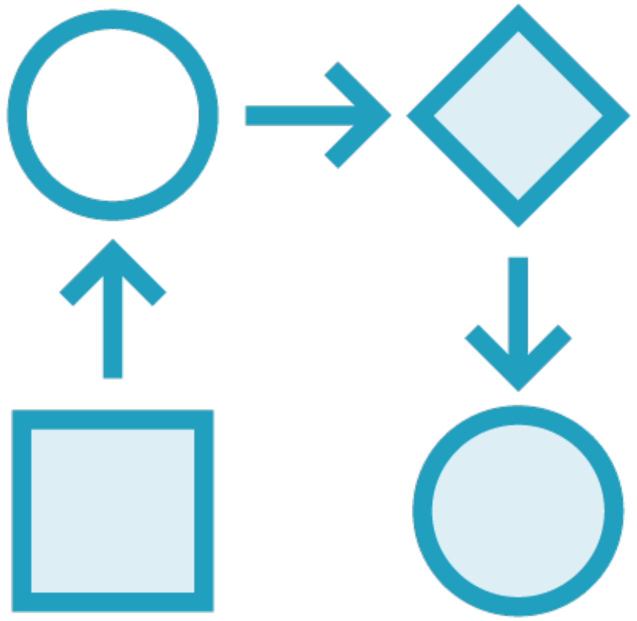
Inferential Statistics

Explain those elements via relationships with other elements

Two Hats of a Data Professional

Find the Dots

Identify important elements in a dataset



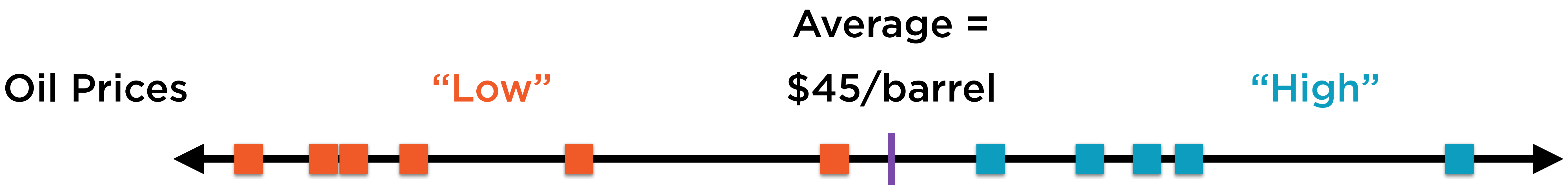
Connect the Dots

Explain those elements via relationships with other elements

Data in One Dimension

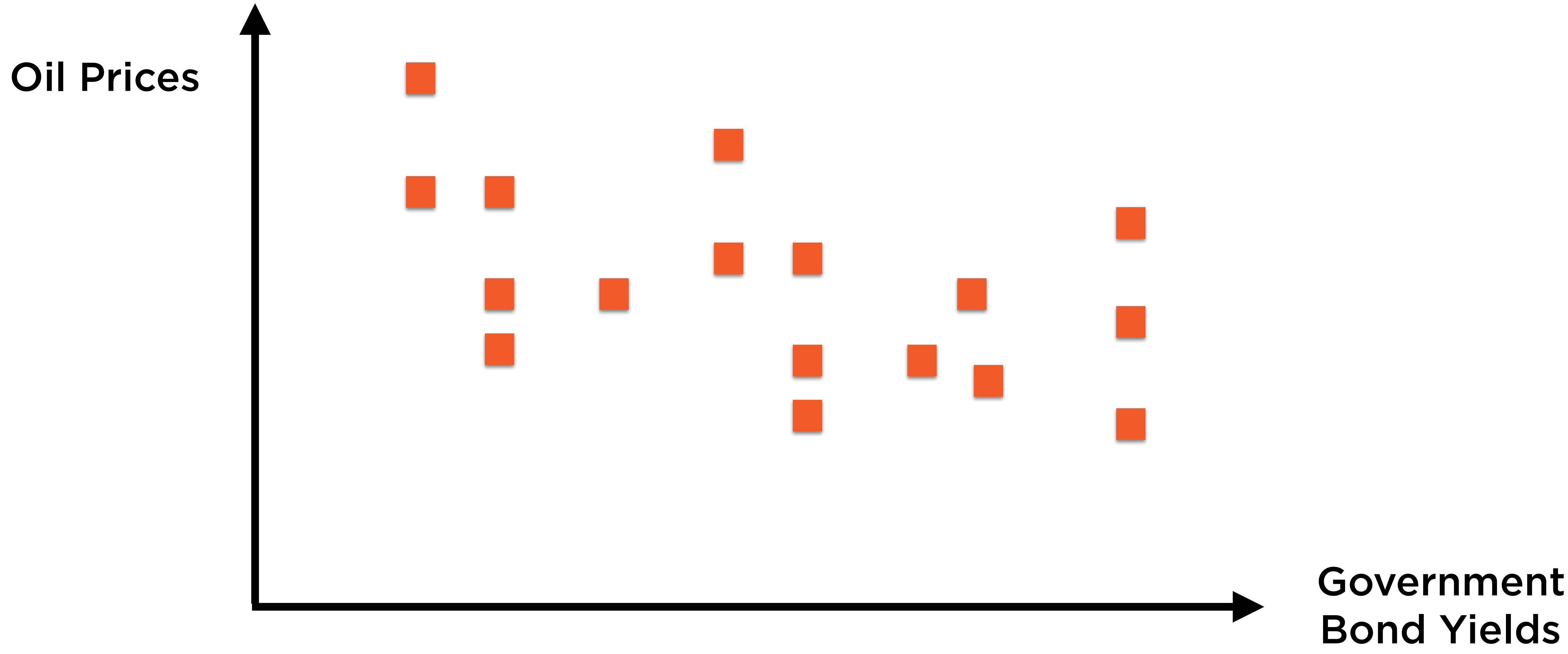
Unidimensional data points can be represented using a line, such as a number line

Data in One Dimension



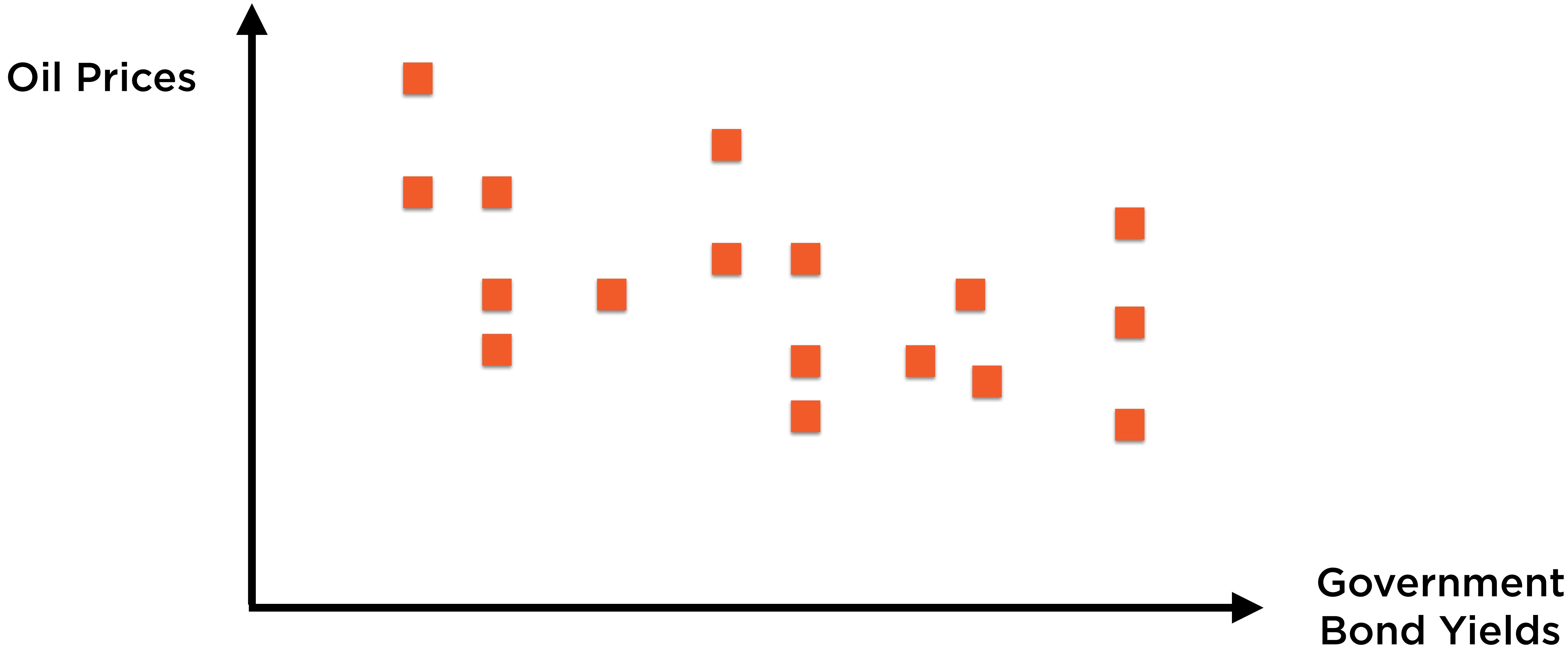
Unidimensional data is analysed using statistics such as mean, median, standard deviation

Data in Two Dimensions



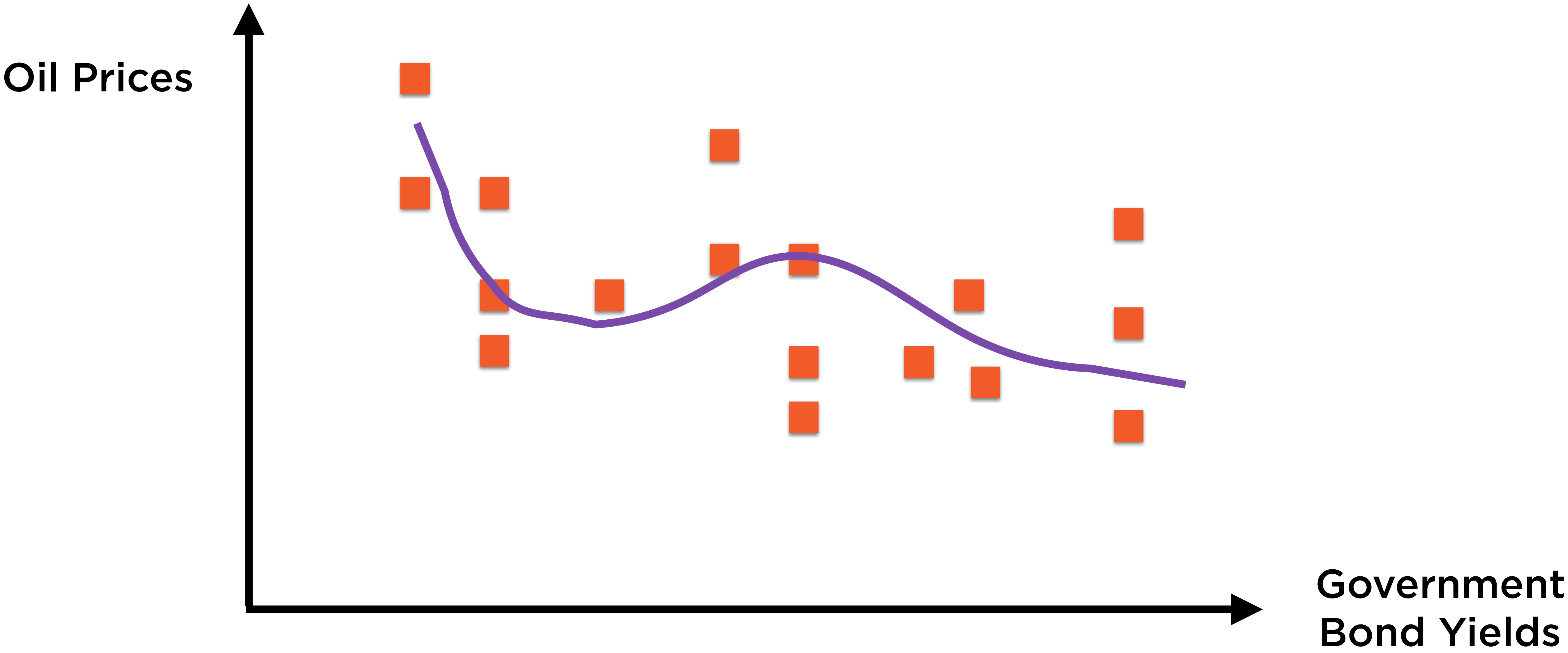
It's often more insightful to view data in relation to some other, related data

Data in Two Dimensions



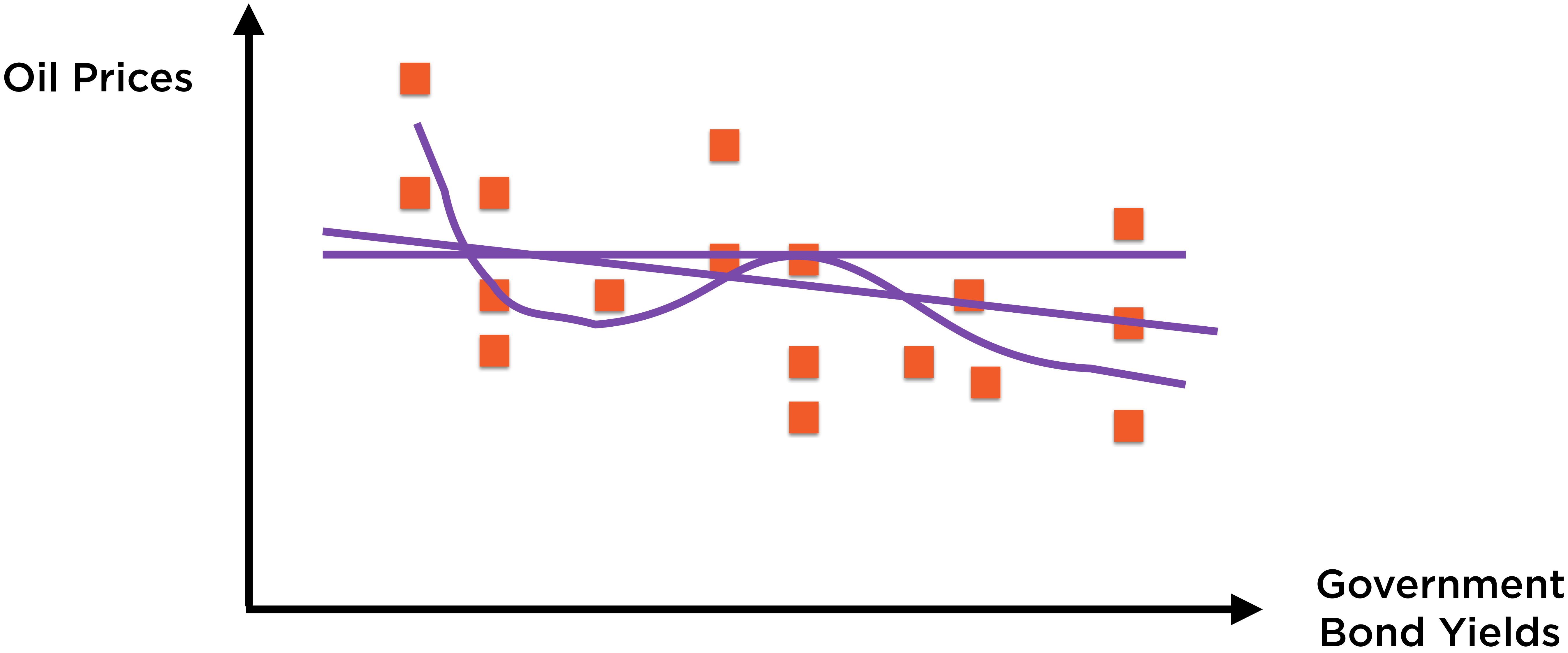
Bidimensional data can be represented in a plane

Data in Two Dimensions



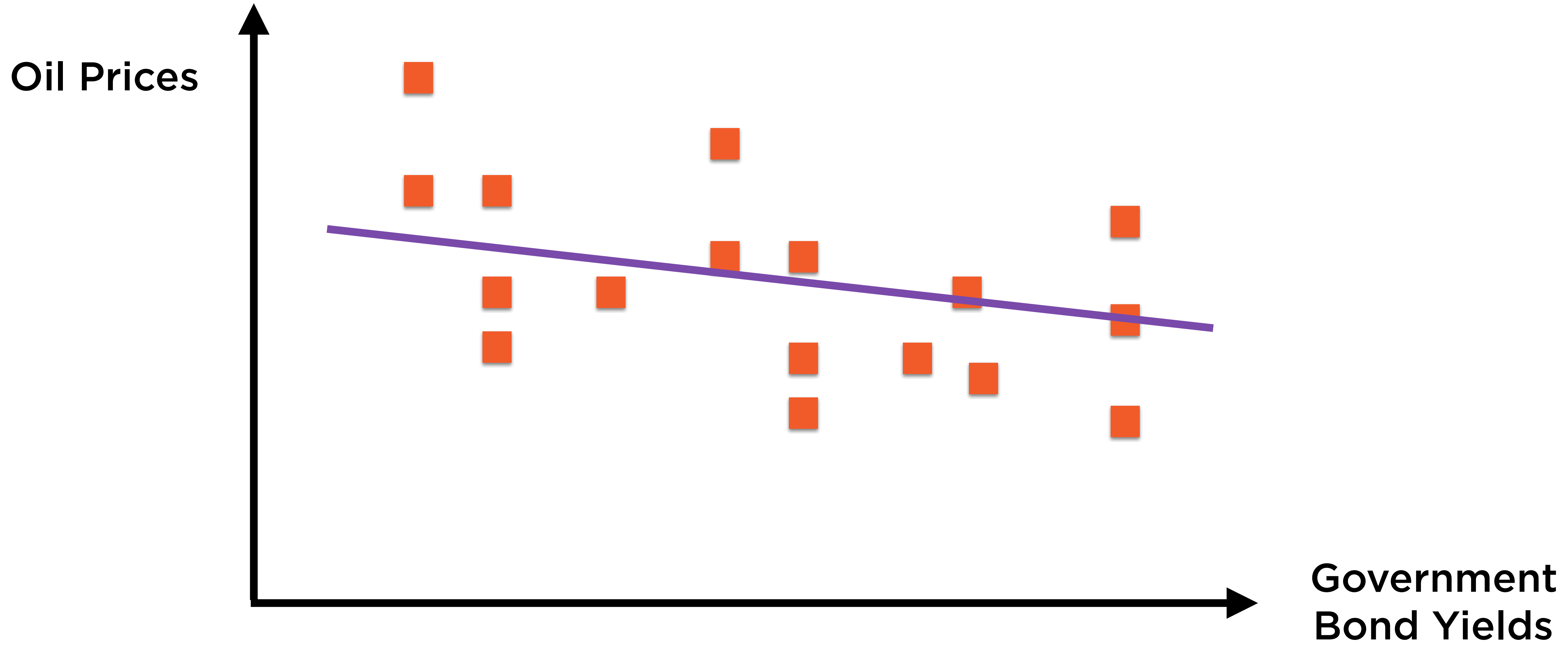
We can draw any number of curves to fit such data

Data in Two Dimensions



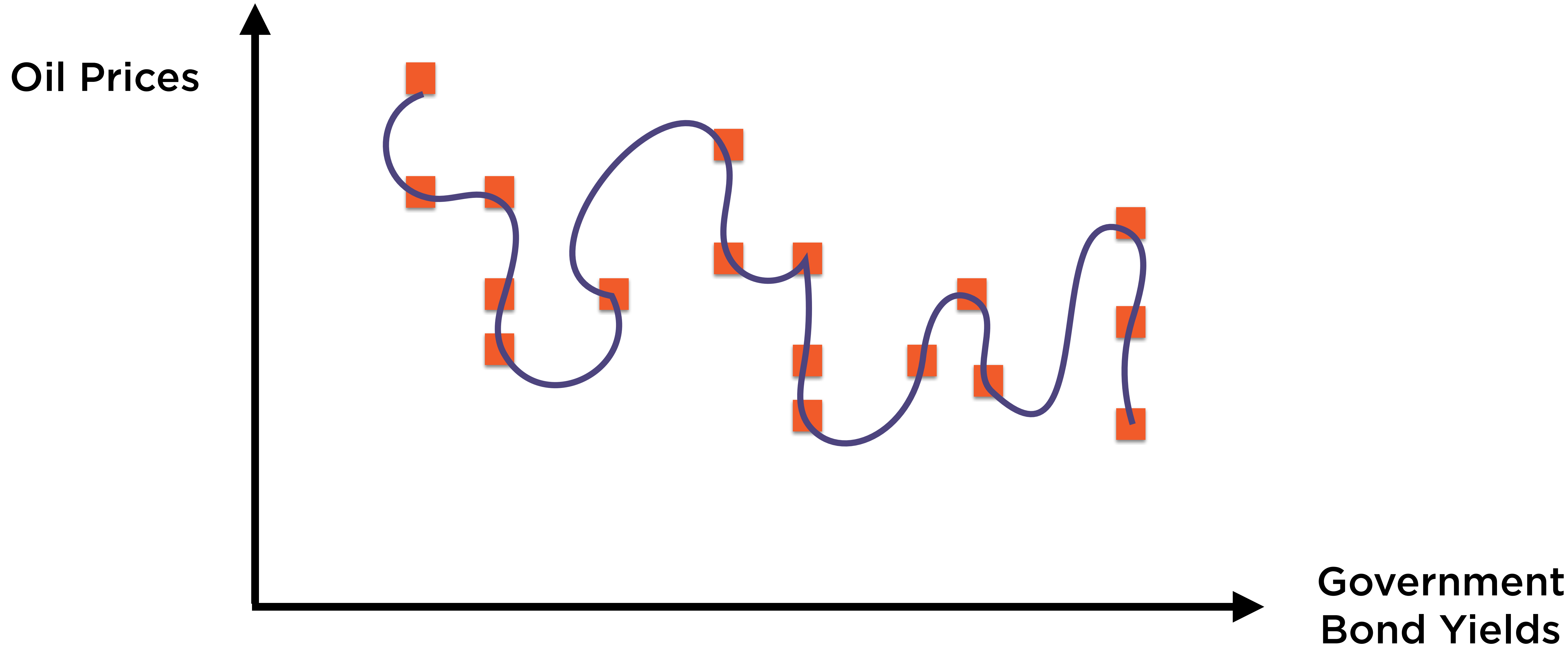
We can draw any number of curves to fit such data

Data in Two Dimensions



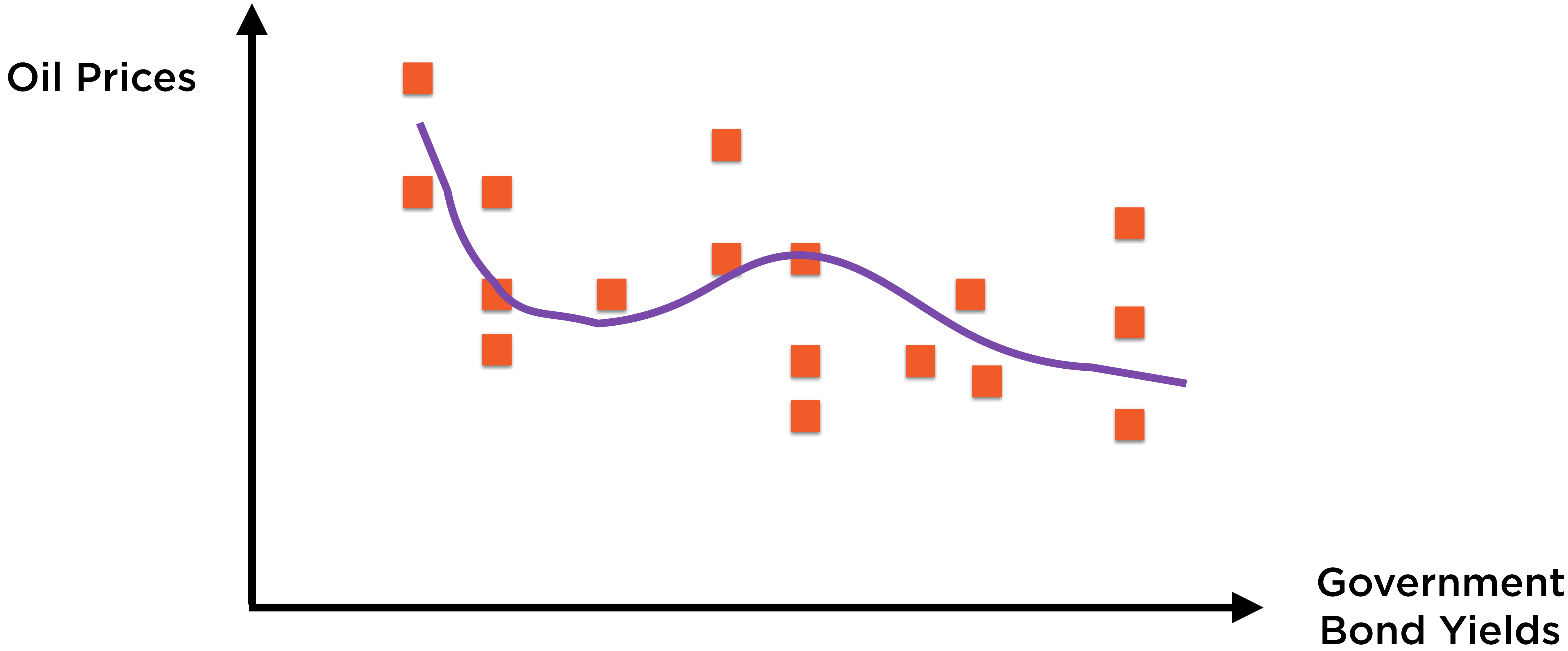
A straight line represents a linear relationship

Data in Two Dimensions



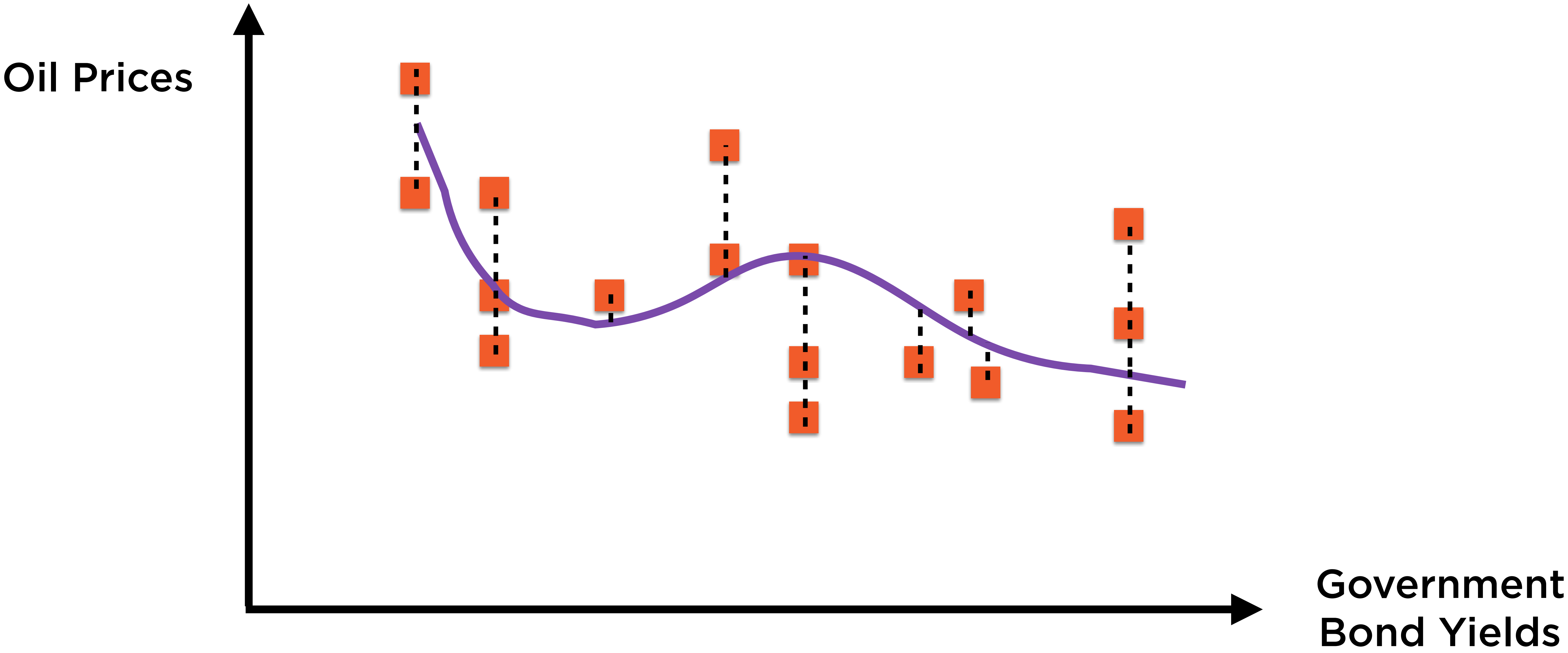
We could either make this curve pass through each point...

Data in Two Dimensions



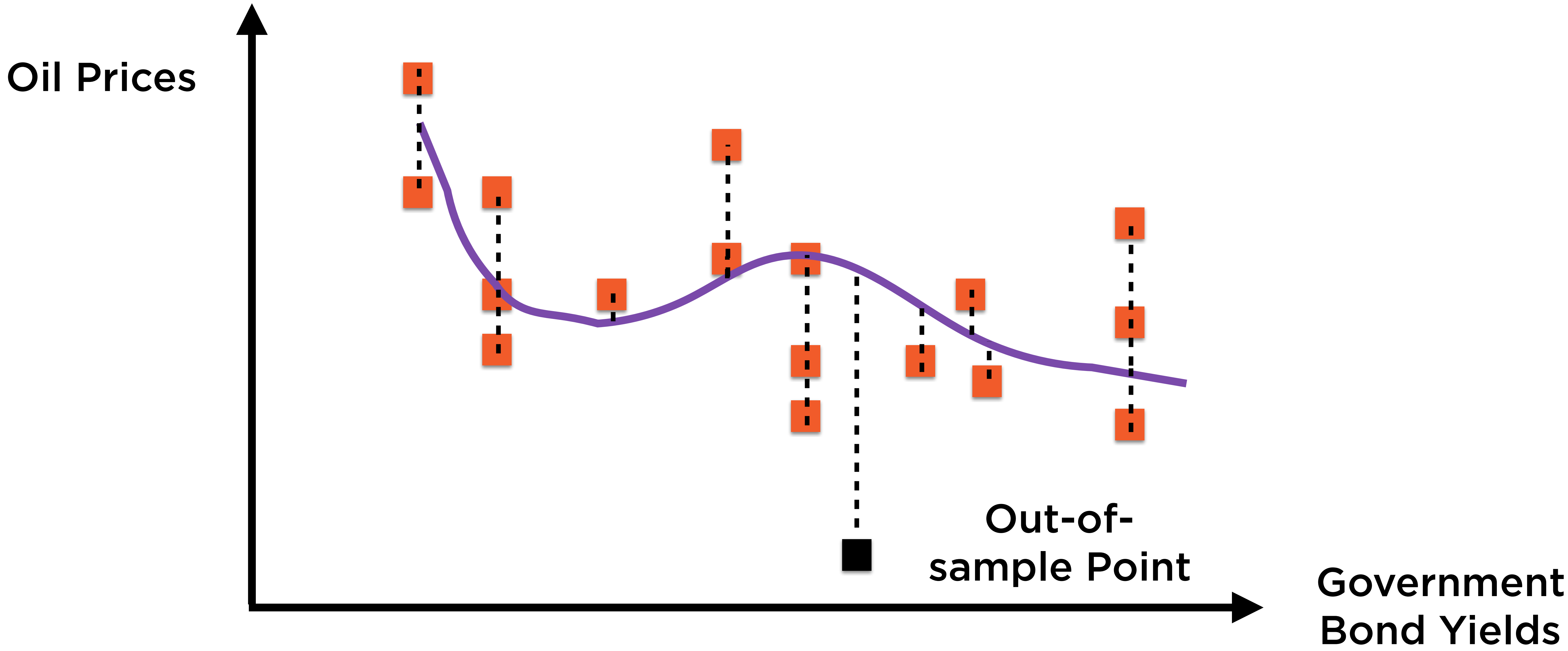
...Or in some sense “fit” the data in aggregate

Data in Two Dimensions



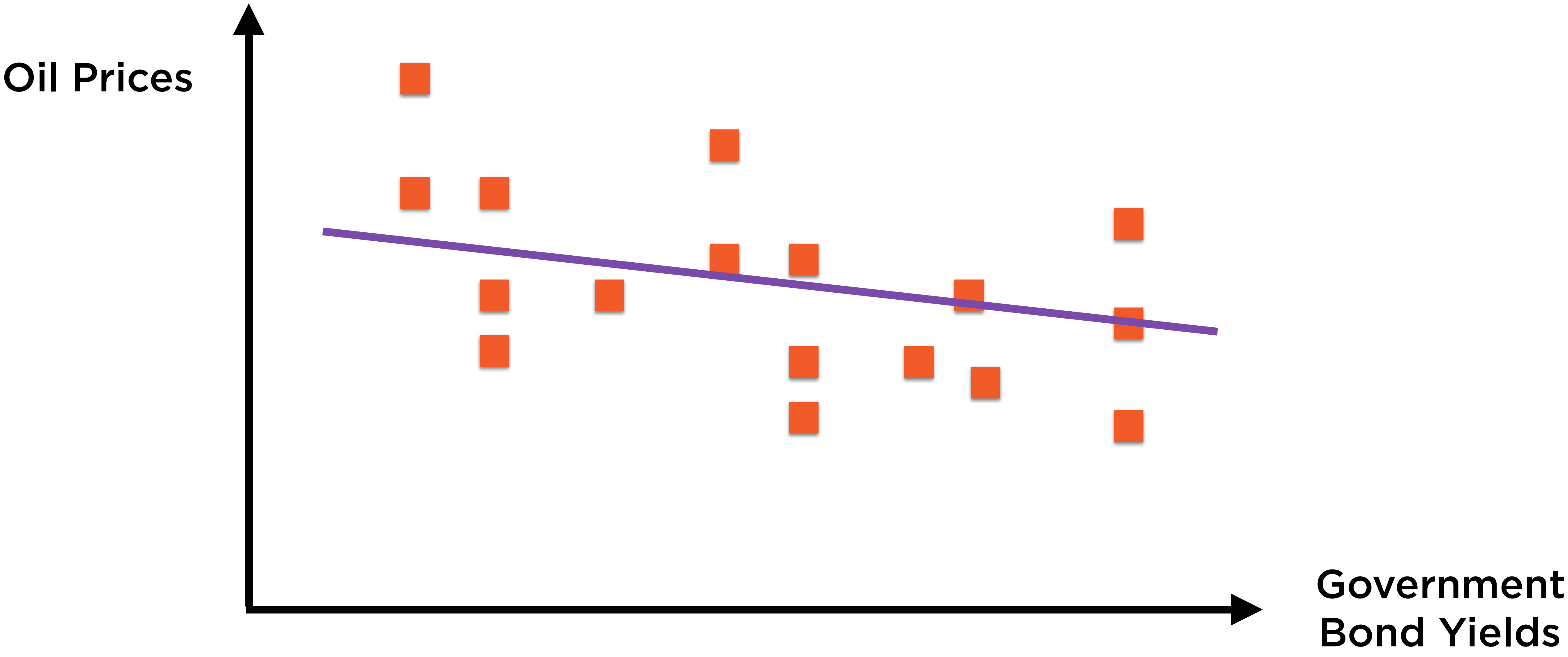
A curve has a "good fit" if the distances of points from the curve are small

Data in Two Dimensions



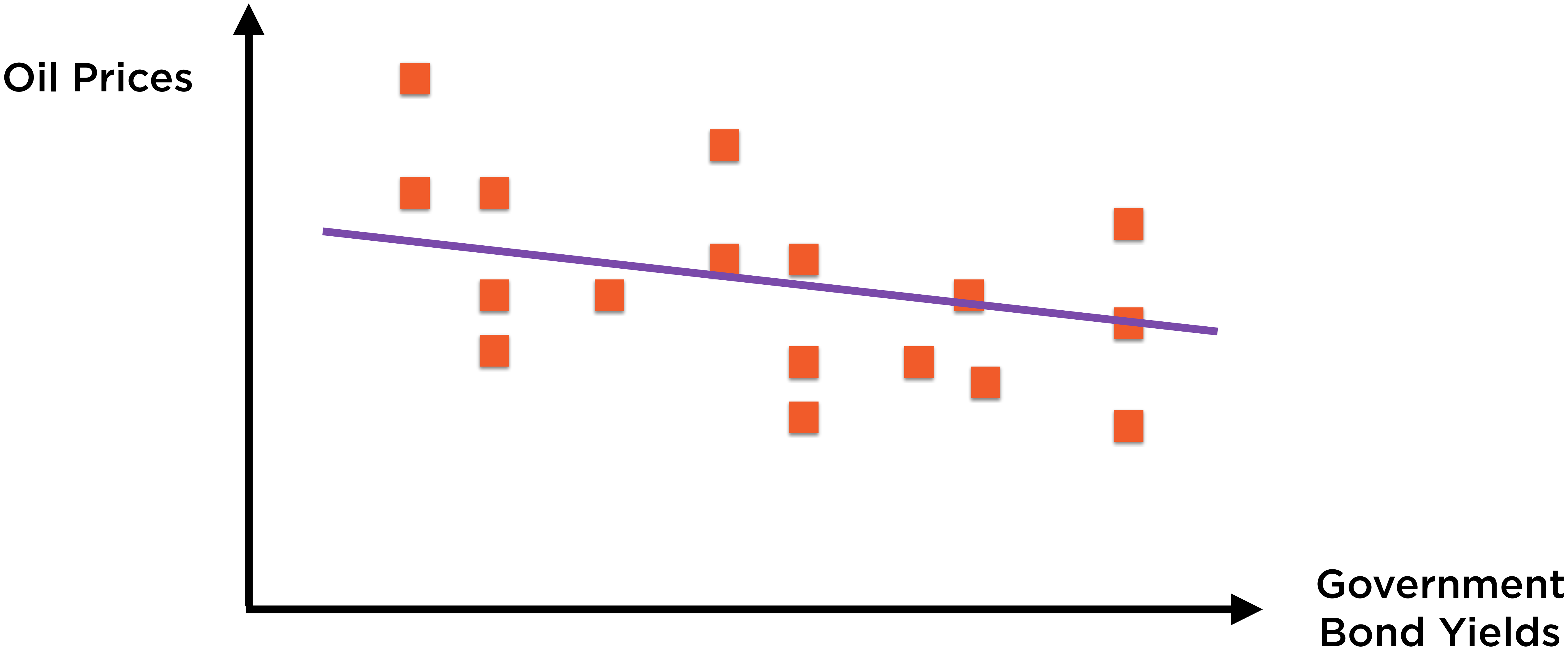
Overfitting by finding a very complicated curve often only hurts predictive accuracy

Data in Two Dimensions



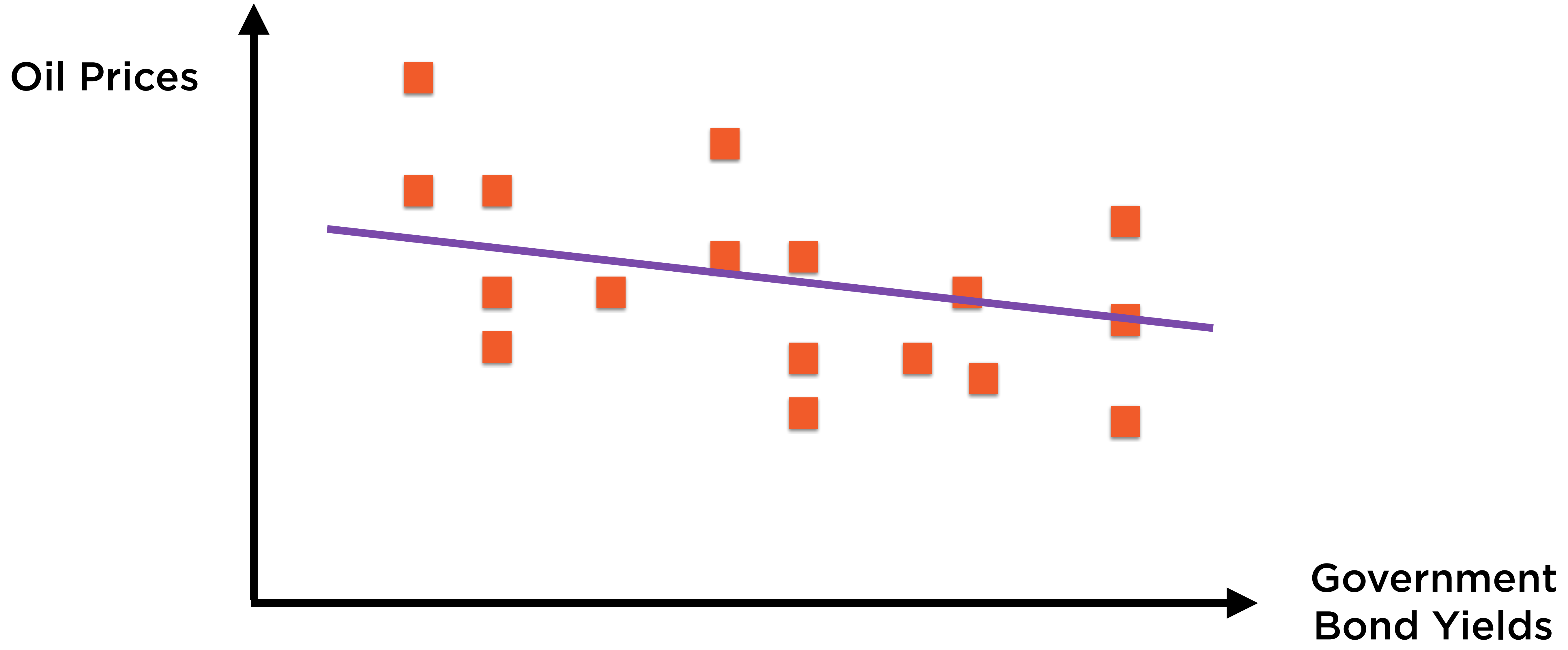
Often, a straight line works just fine

Data in Two Dimensions



Finding the “best” such straight line is called Linear Regression

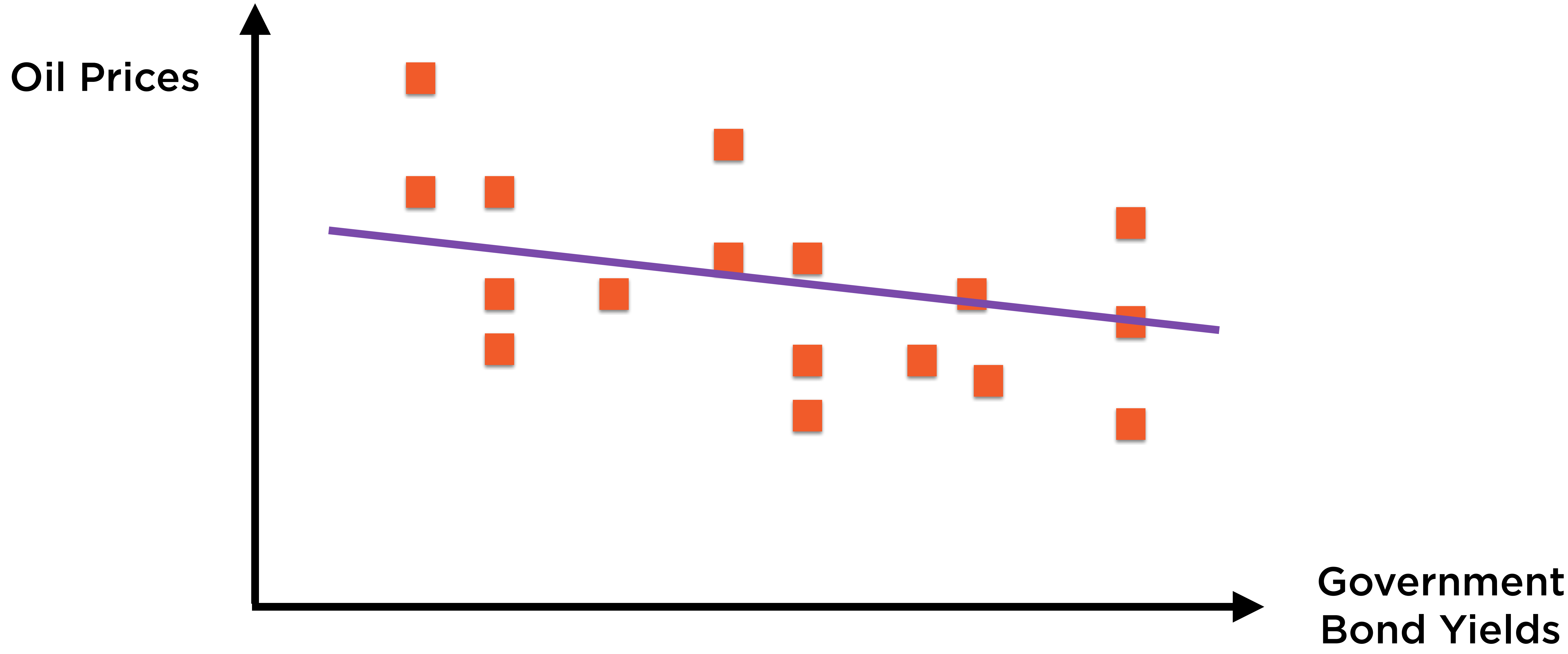
Linear Regression



The linear regression relationship can be expressed as

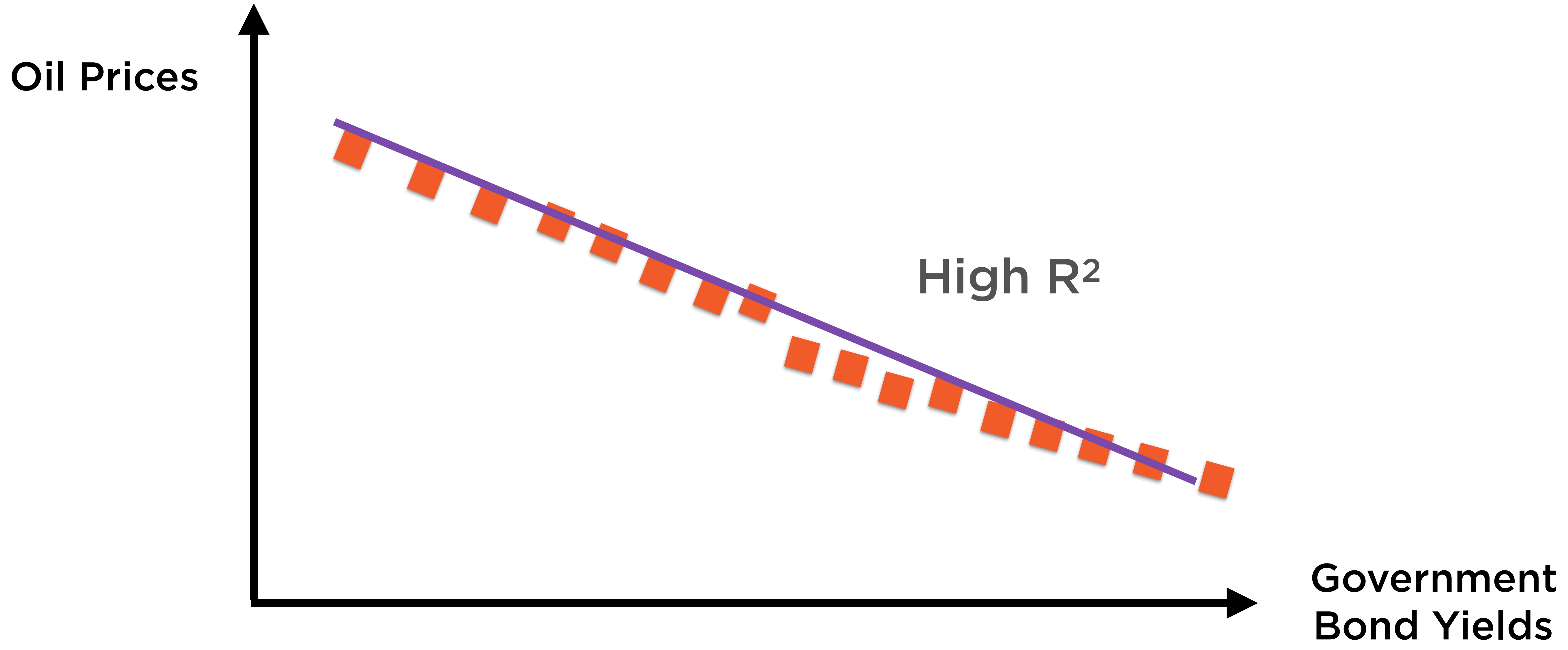
$y = A + Bx$

Linear Regression



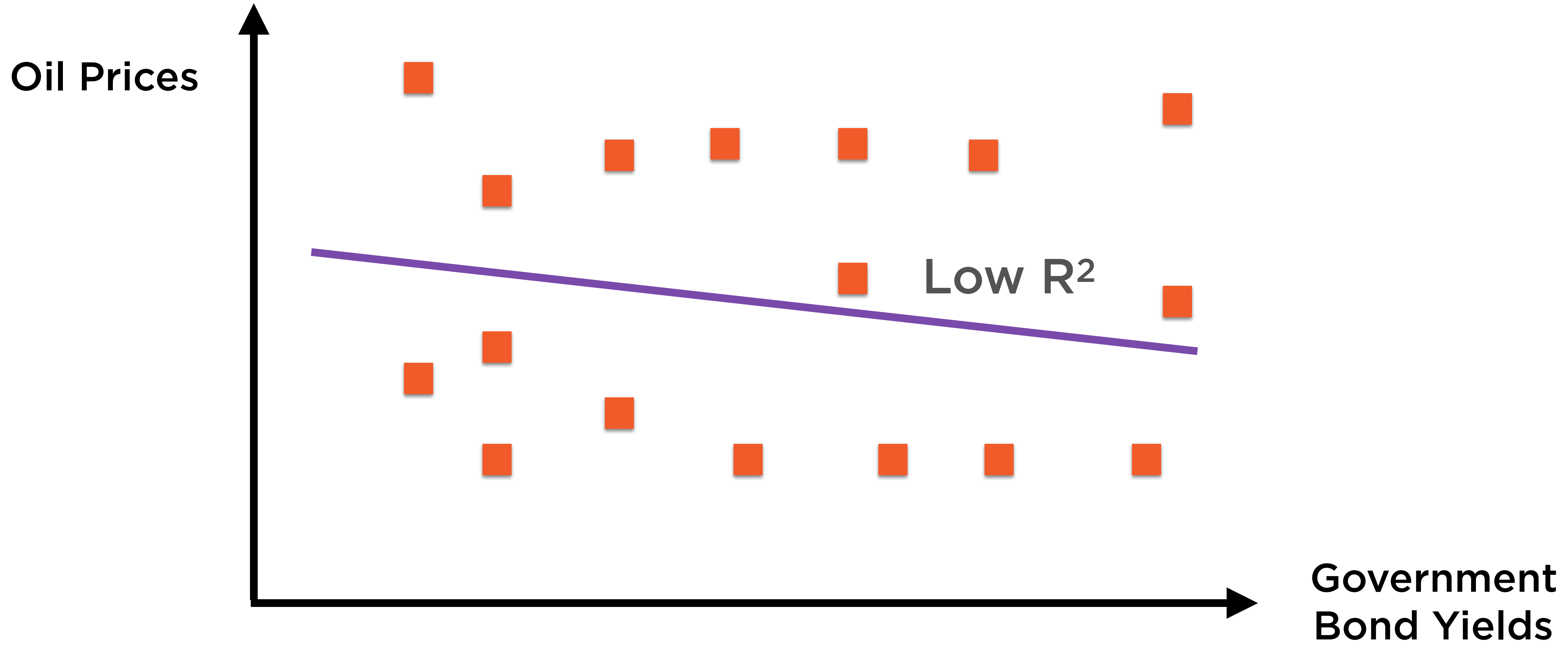
Regression not only gives us the equation of this line, it also signals how reliable the line is

Linear Regression



High quality of fit

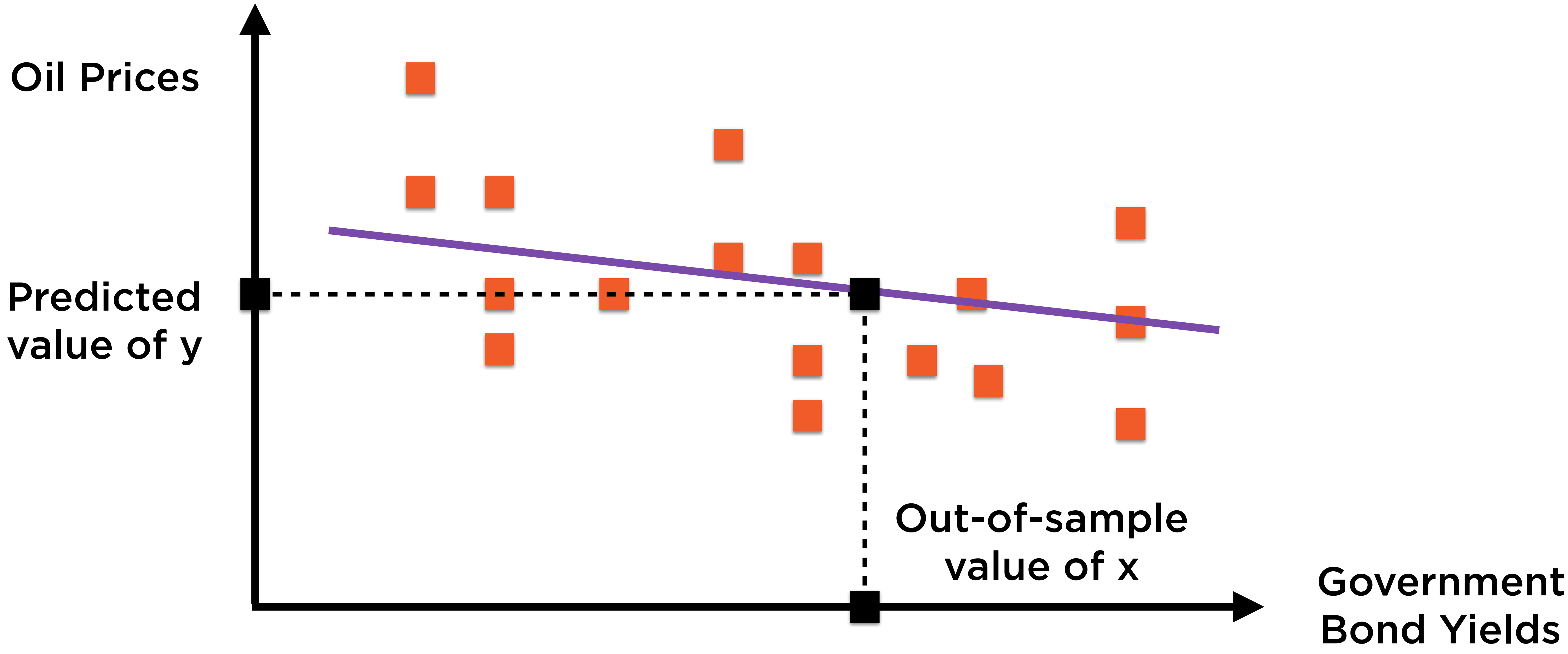
Linear Regression



Low quality of fit

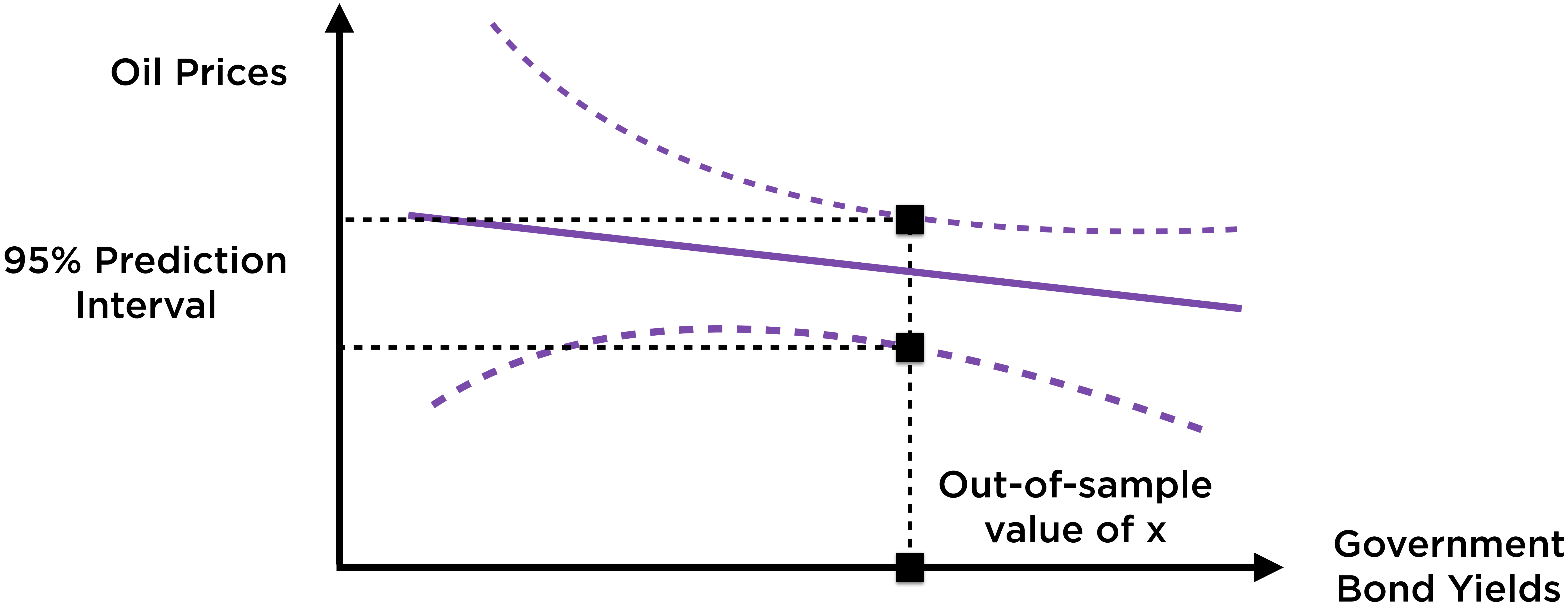
R^2 is a measure of how well the linear regression fits the underlying data

Prediction Using Regression



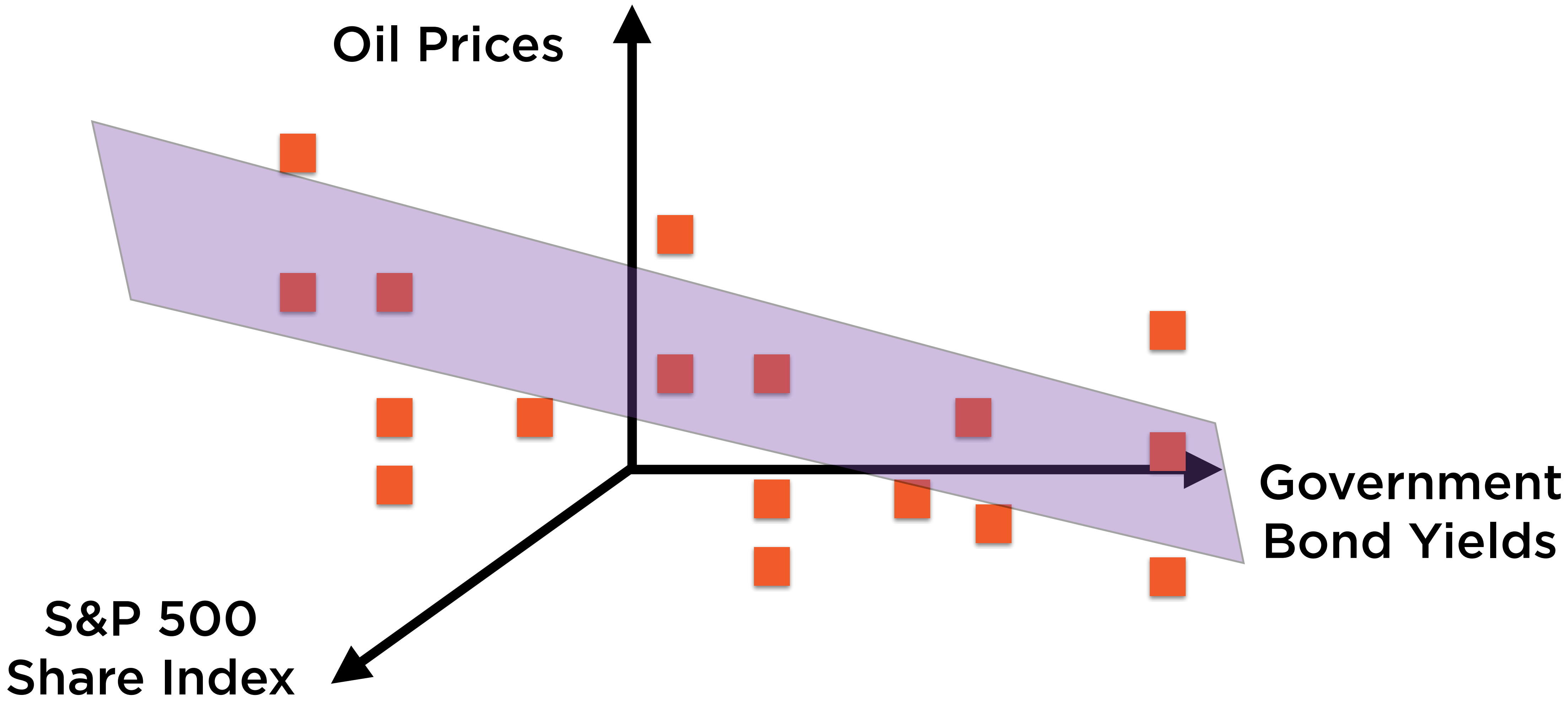
Given a new value of x, use the line to predict the corresponding value of y

Prediction Using Regression



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

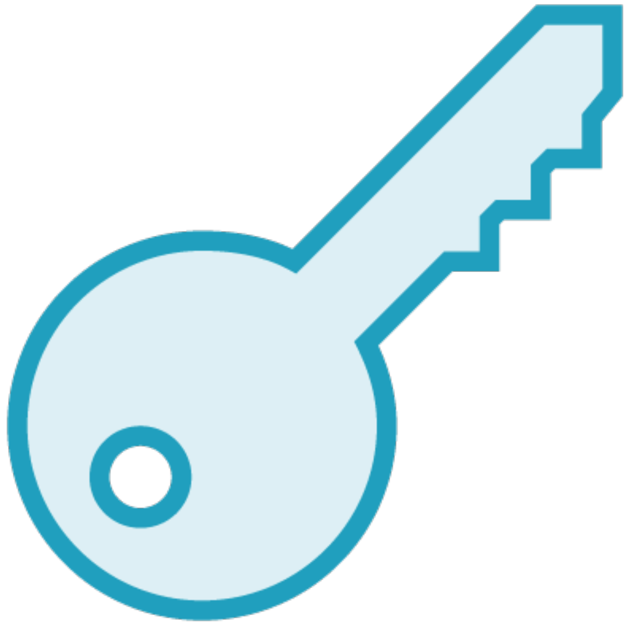
Data in N Dimensions



Linear Regression can easily be extended to n-dimensional data

Setting Up The Regression Problem

X Causes Y



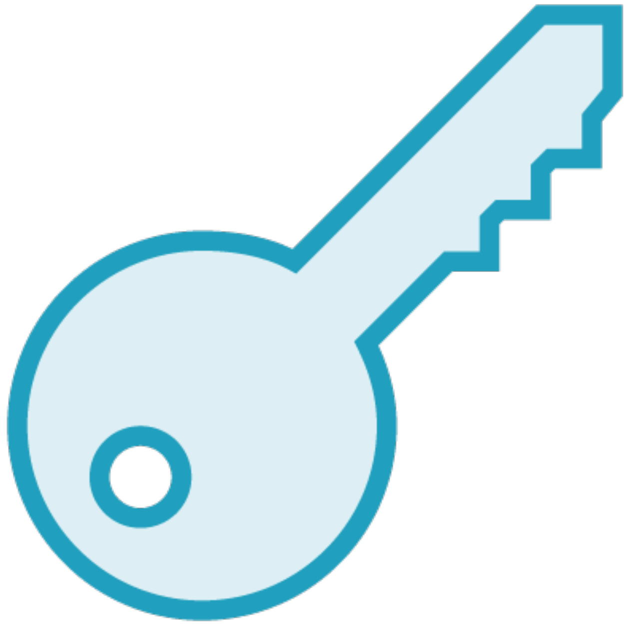
Cause

Independent variable

Effect

Dependent variable

X Causes Y



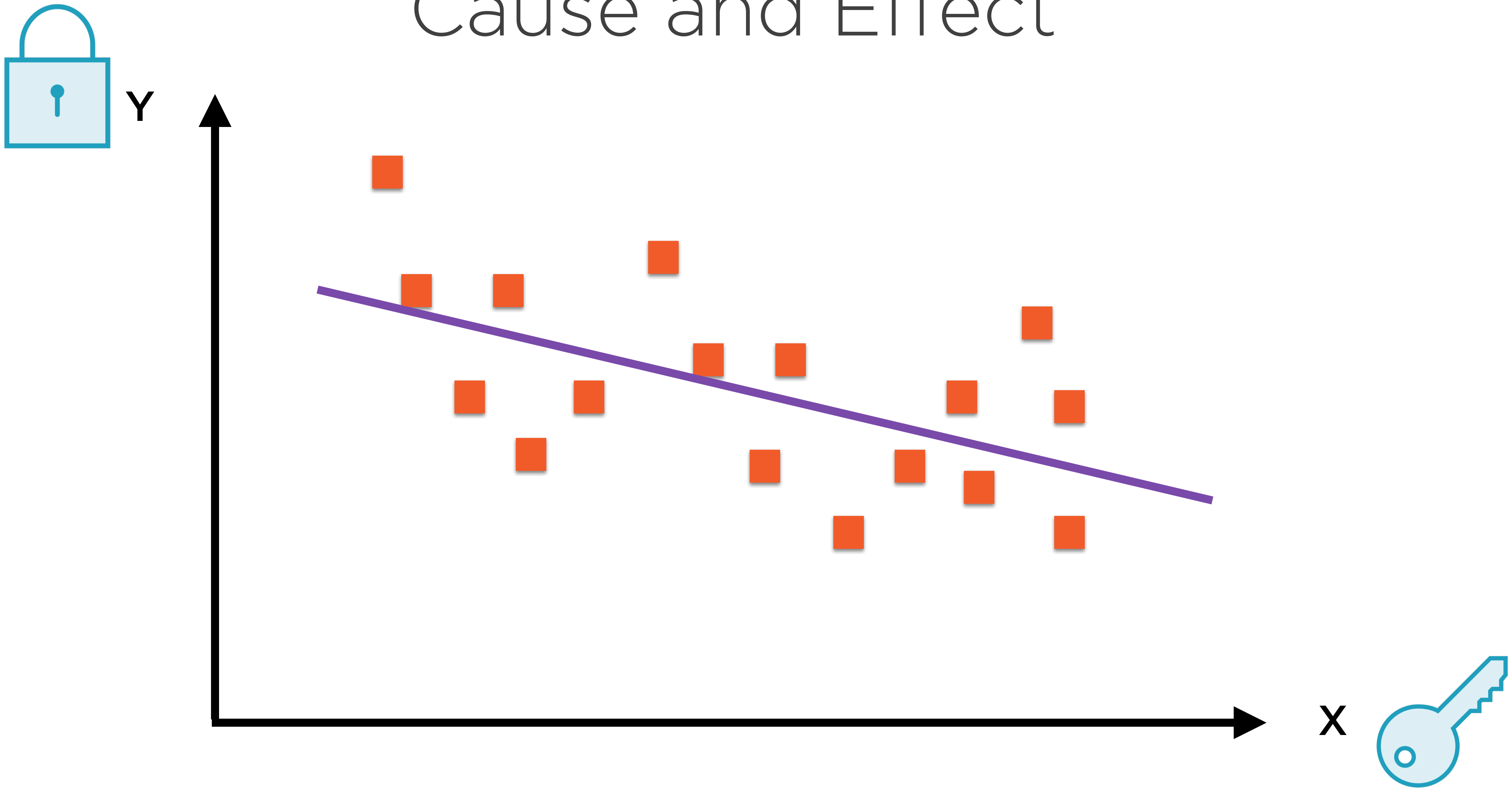
Cause

Explanatory variable

Effect

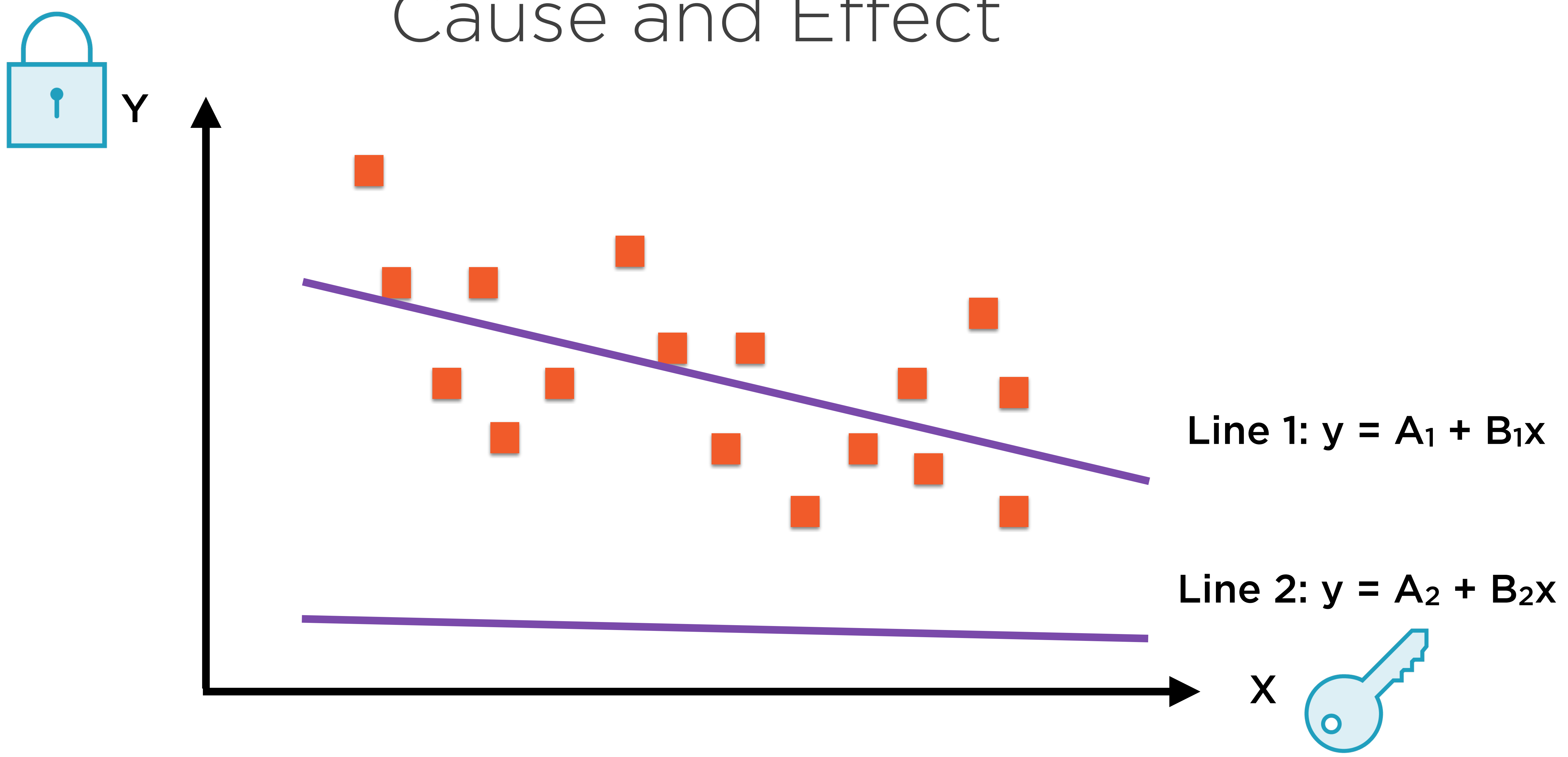
Dependent variable

Cause and Effect



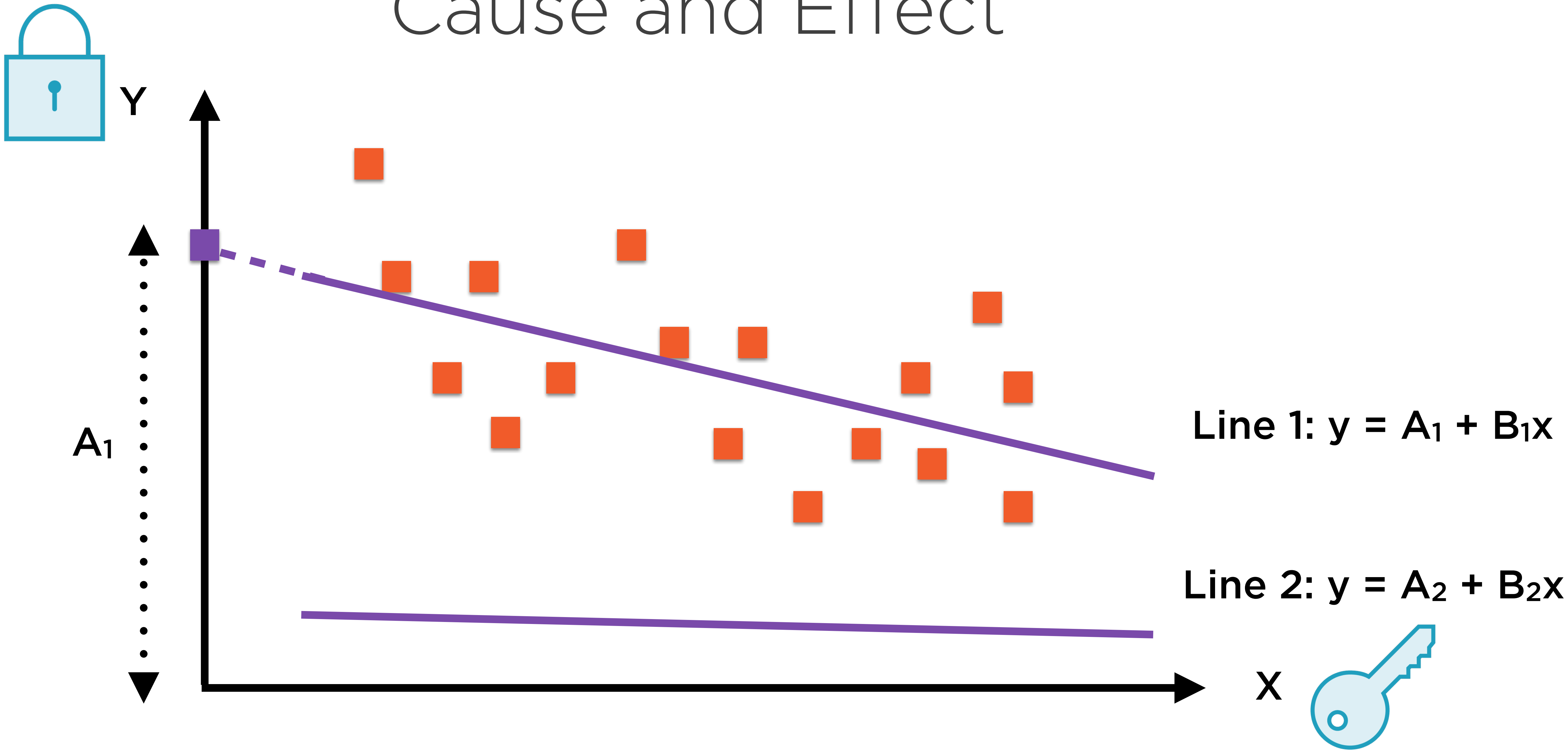
Linear Regression involves finding the “best fit” line

Cause and Effect



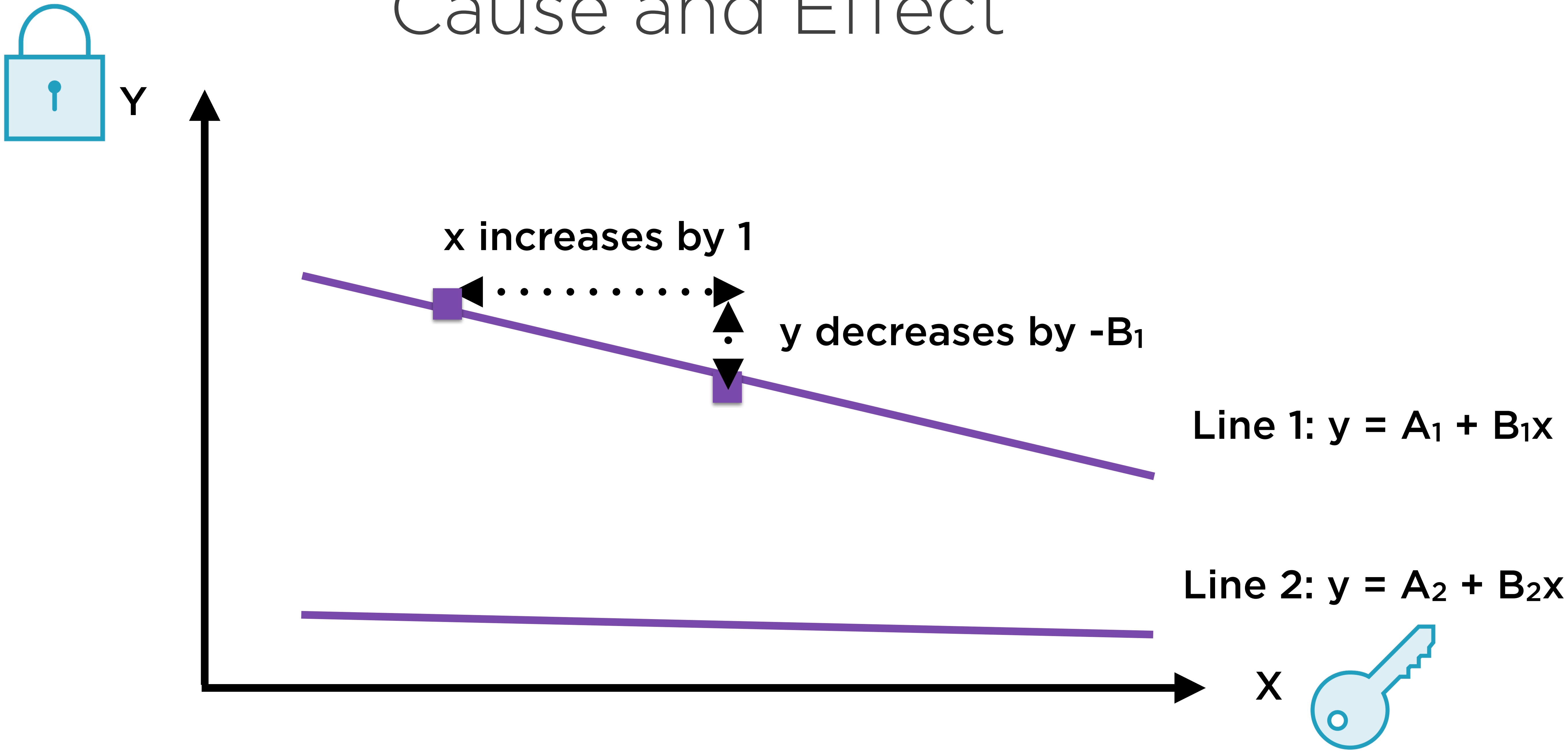
Let’s compare two lines, Line 1 and Line 2

Cause and Effect



The first line has y-intercept A_1

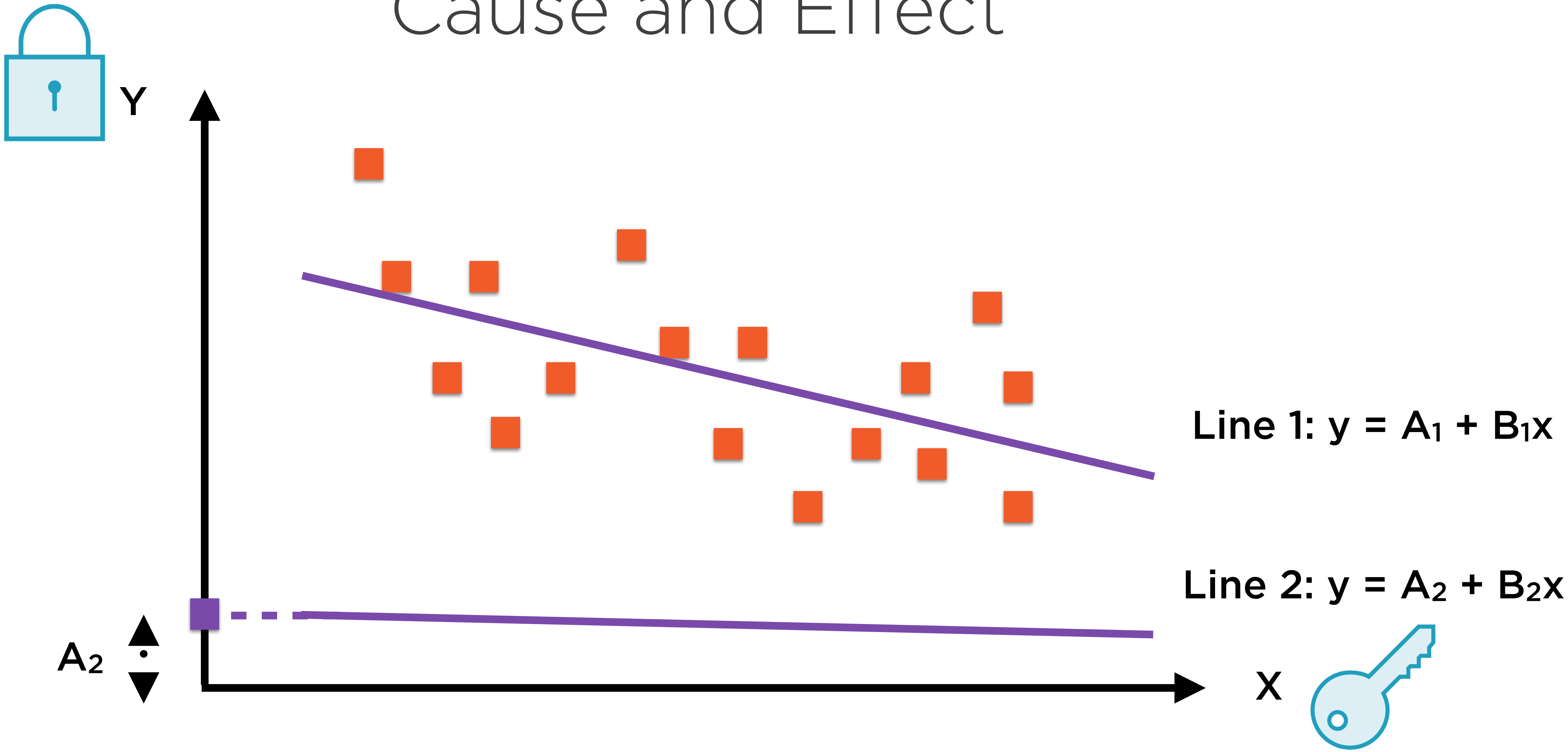
Cause and Effect



In the first line, if x changes by 1 unit, y decreases by $-B_1$ units

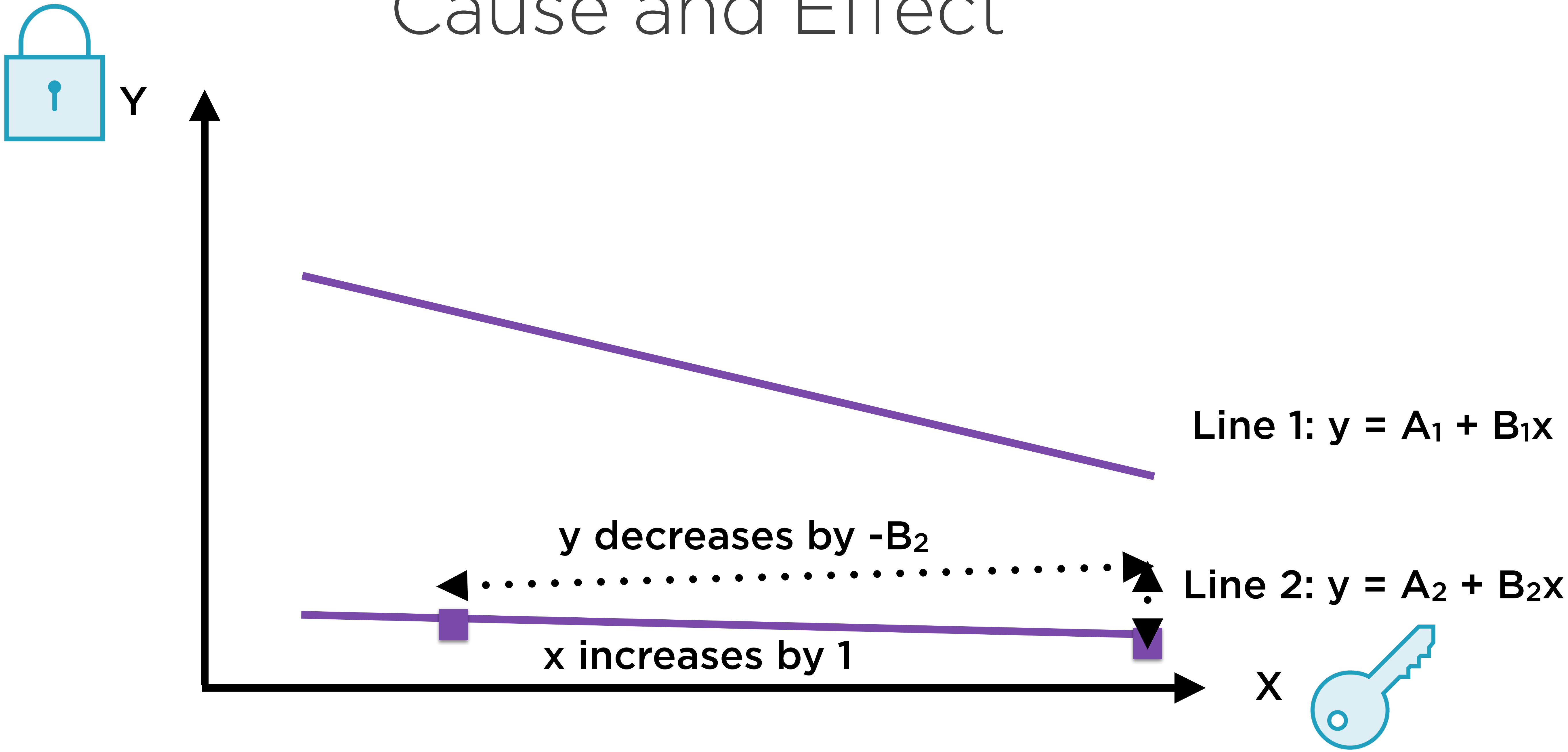
(B_1 is negative because of downward slope, so $-B_1$ is positive)

Cause and Effect



The second line has y-intercept A_2

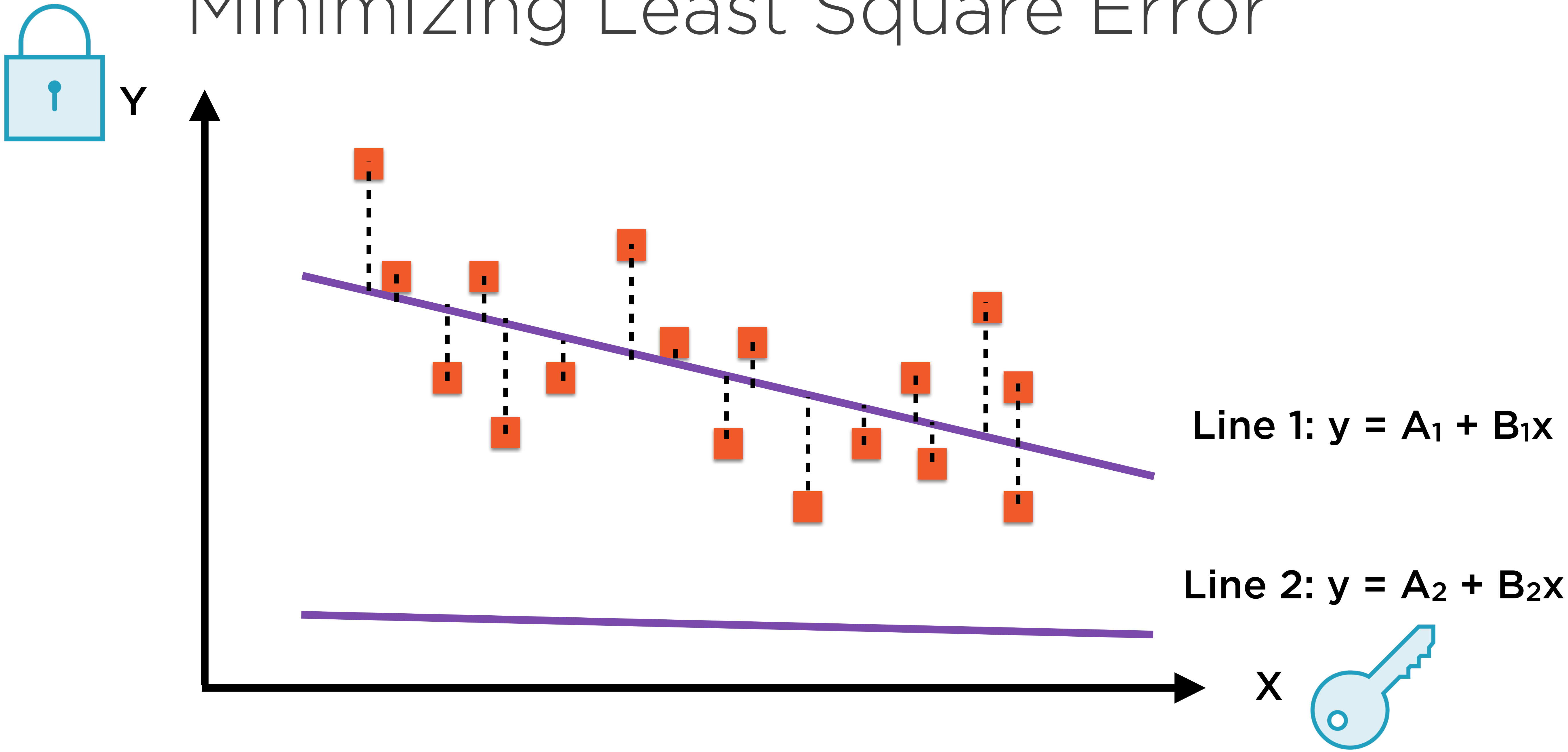
Cause and Effect



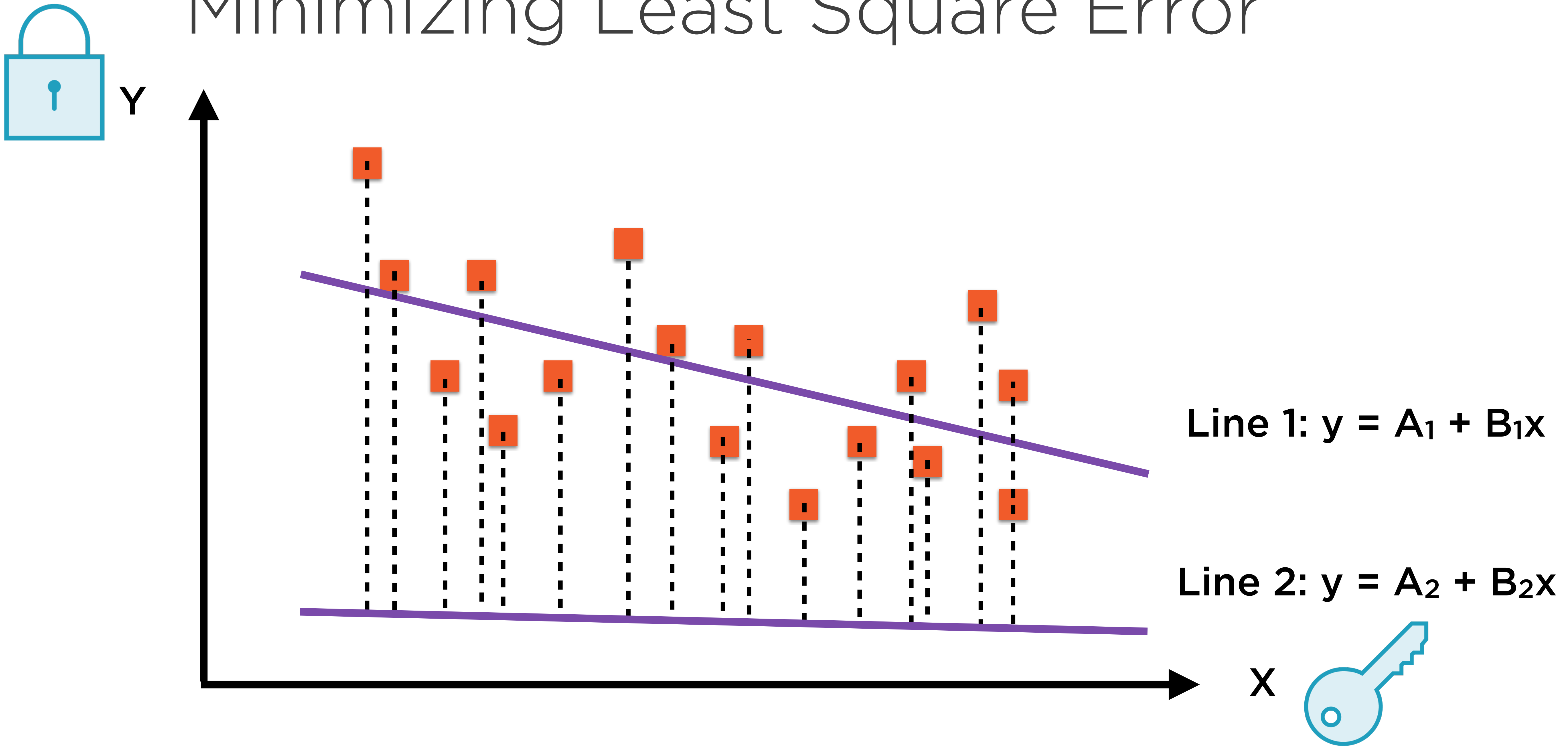
In the second line, if x changes by 1 unit, y decreases by $-B_2$ units

(B_2 is negative because of downward slope, so $-B_2$ is positive)

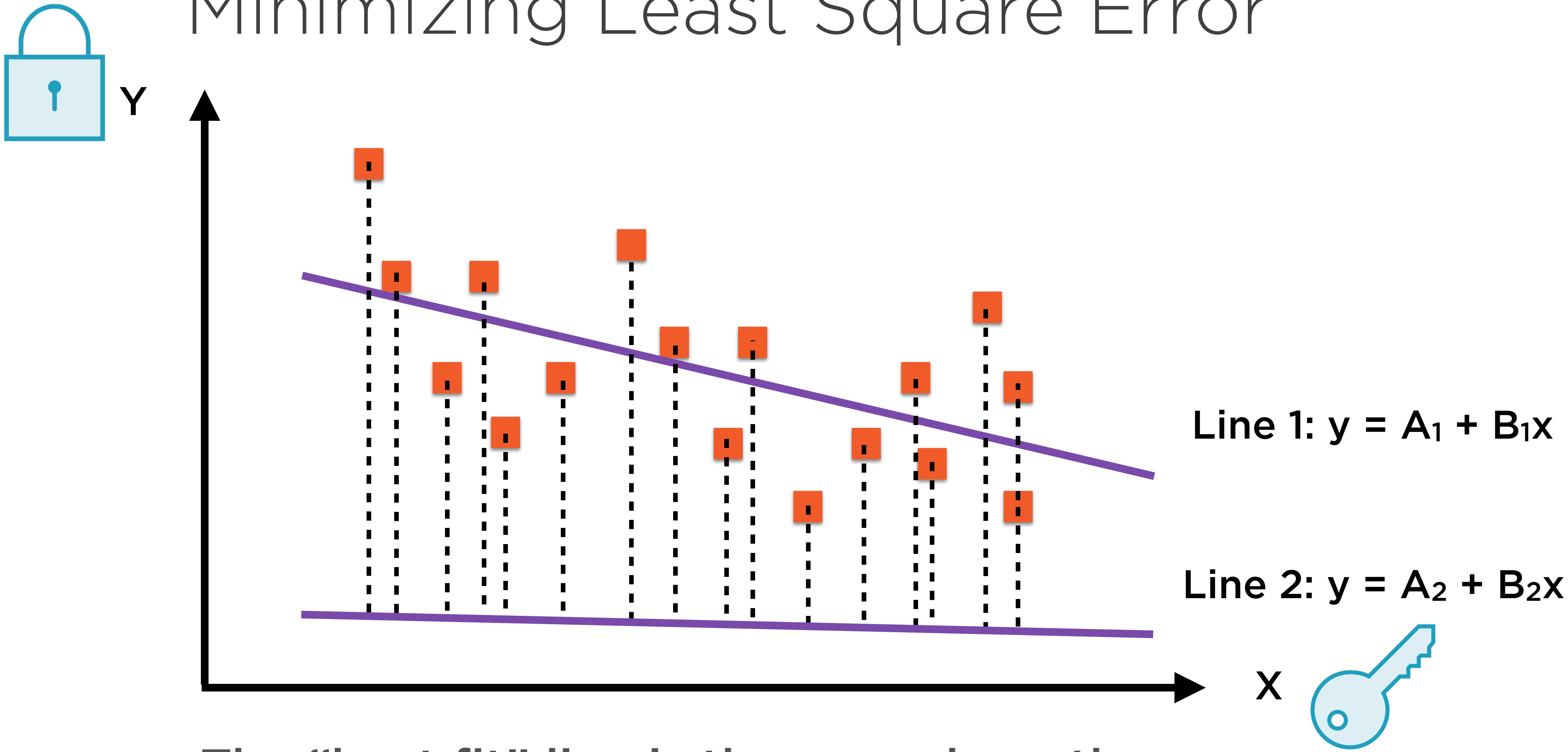
Minimizing Least Square Error



Minimizing Least Square Error

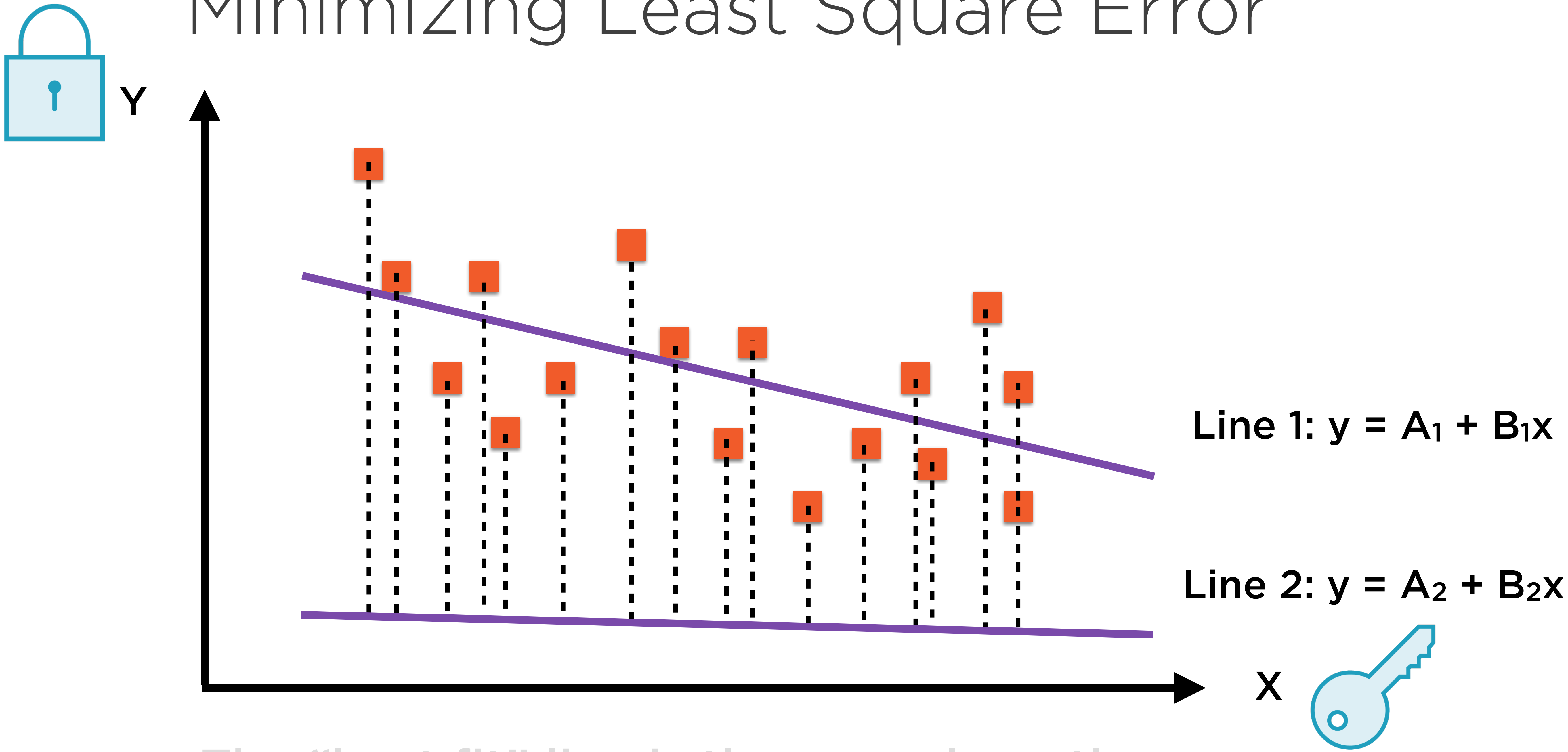


Minimizing Least Square Error



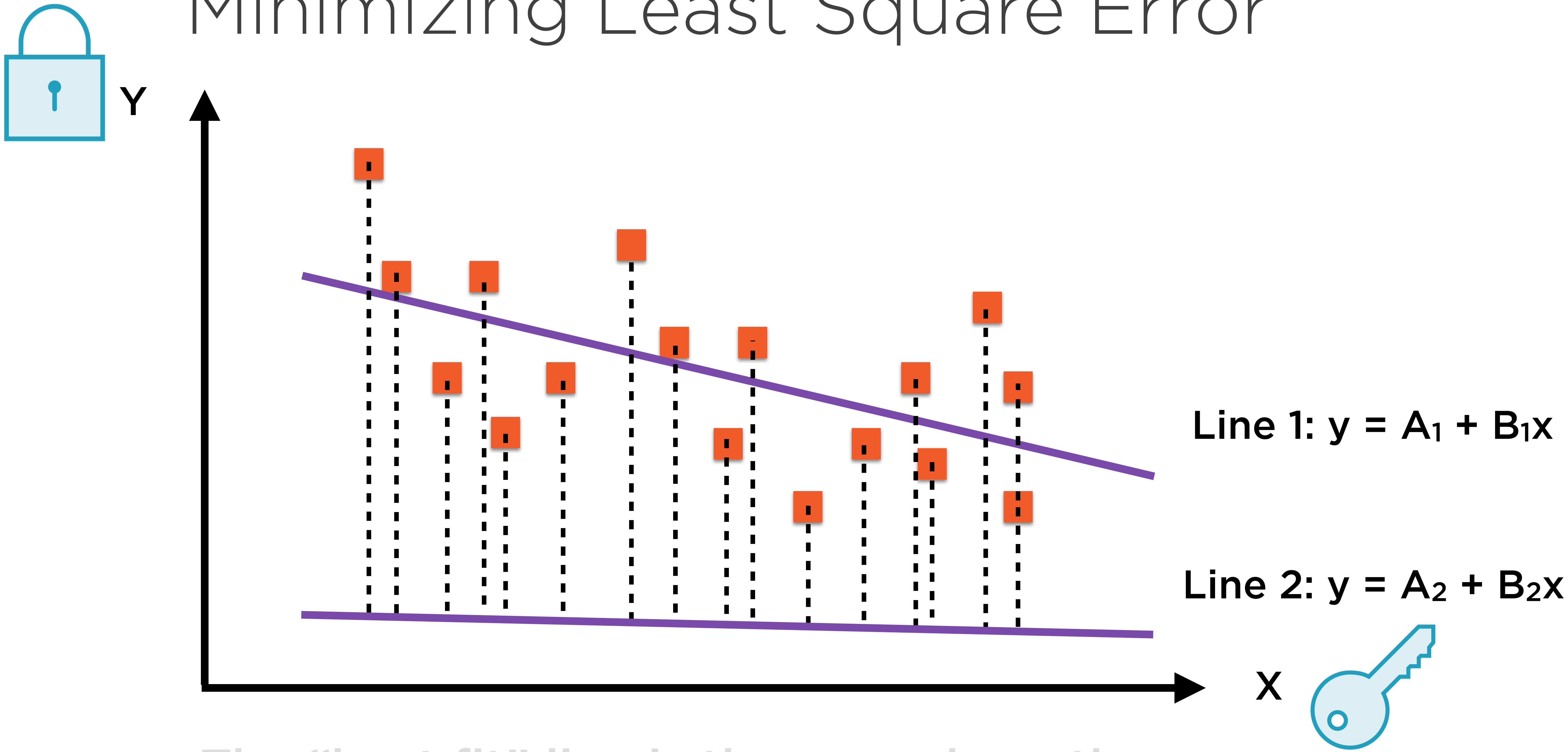
The "best fit" line is the one where the sum of the squares of the lengths of these dotted lines is minimum

Minimizing Least Square Error



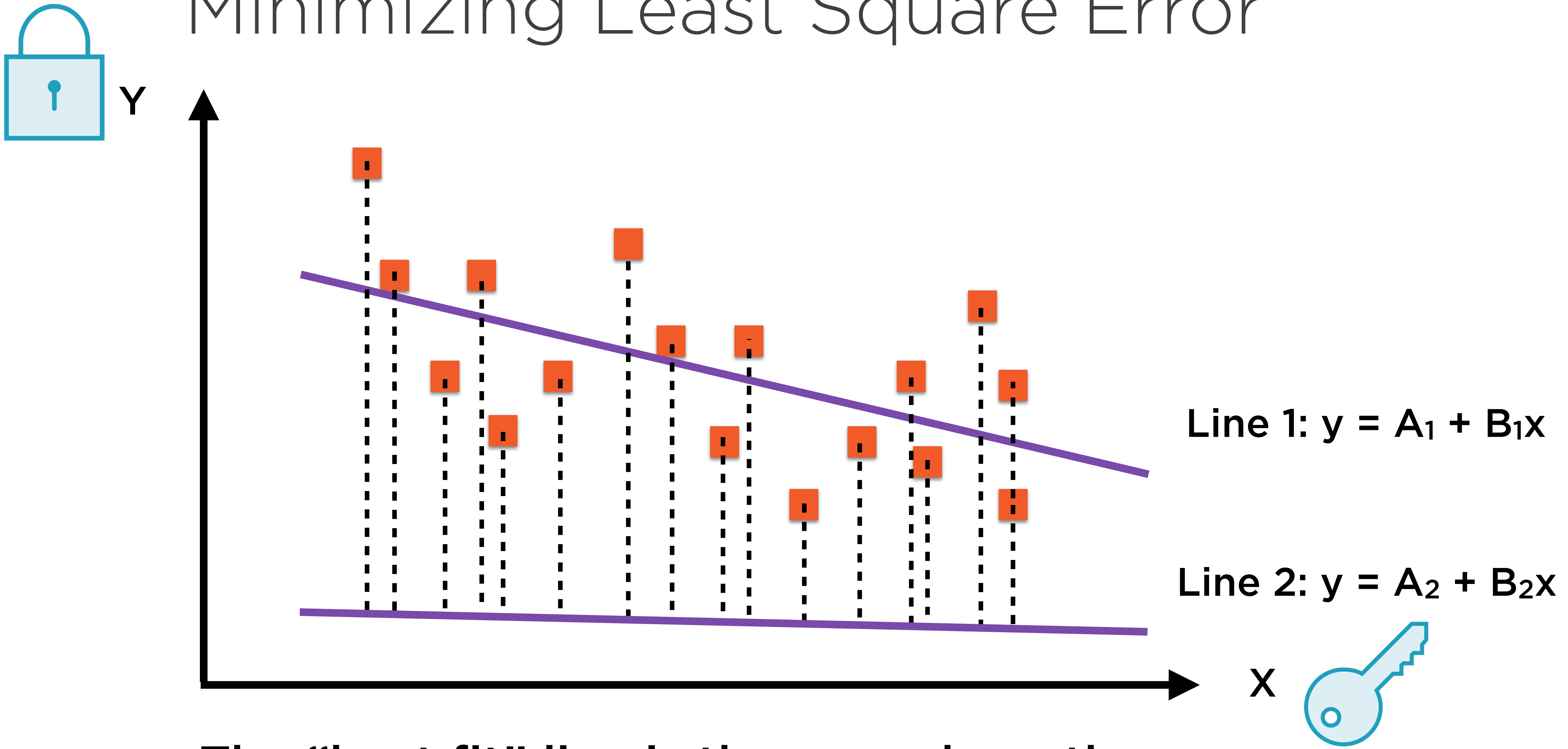
The “best fit” line is the one where the sum of the squares of the lengths of **these dotted lines** is minimum

Minimizing Least Square Error



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

Minimizing Least Square Error

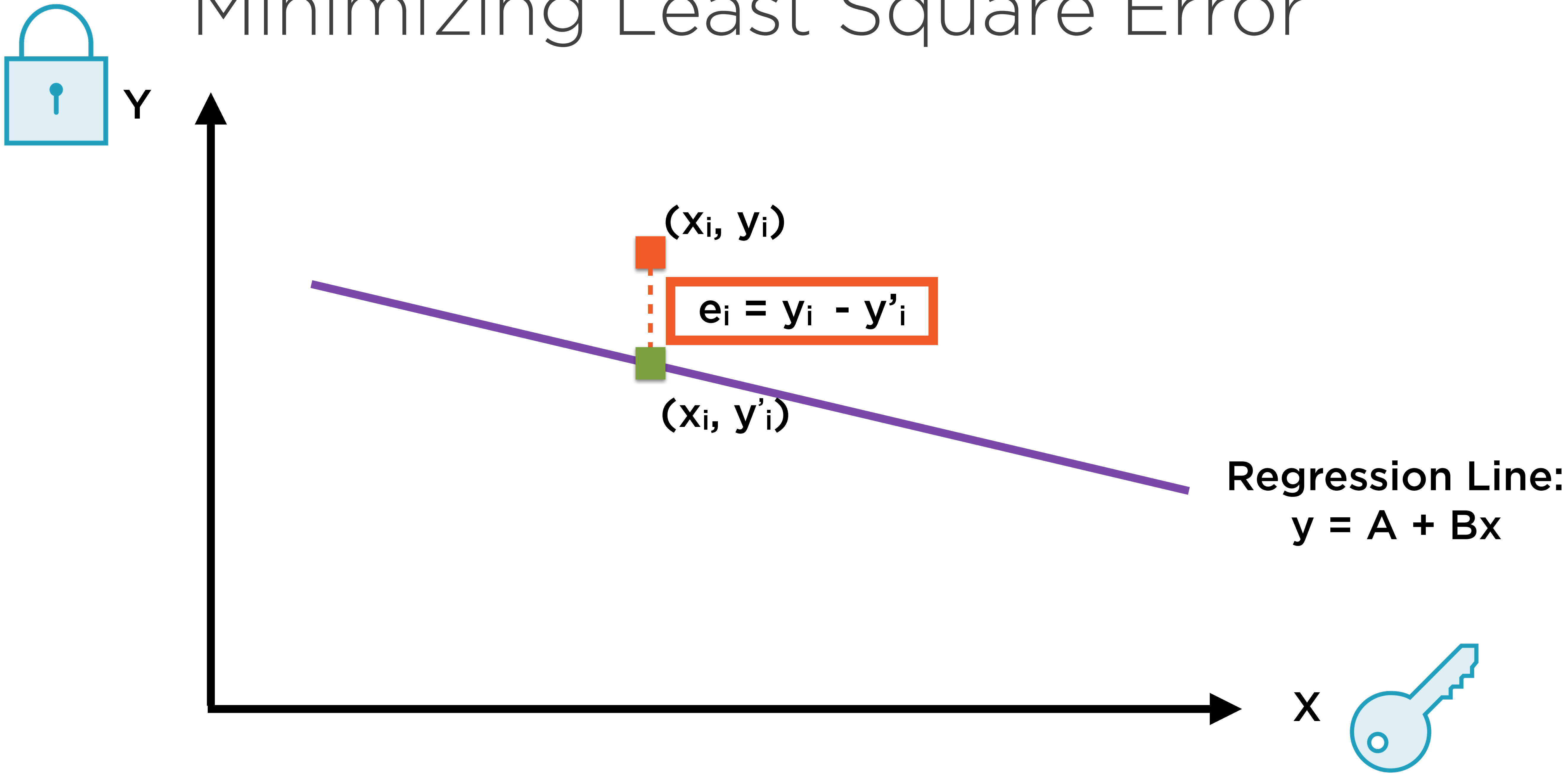


The "best fit" line is the one where the sum of the squares of the lengths of the errors is minimum

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

Finding this line is the objective of the regression problem

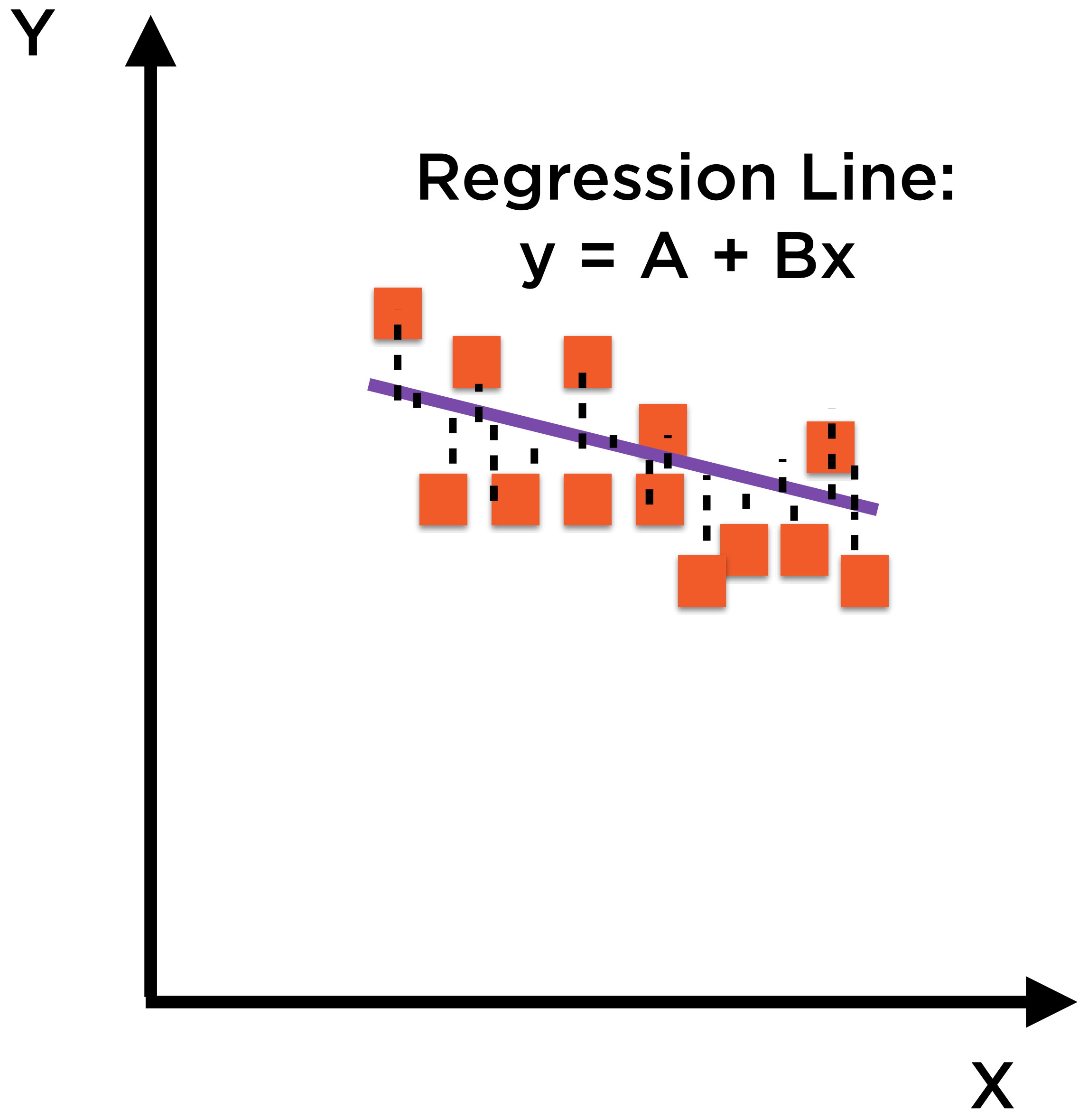
Minimizing Least Square Error



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

To find the “best fit” line we need to make some assumptions about regression error

There is a fine distinction between errors and residuals - but we can ignore it



Ideally, residuals should

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

Demo

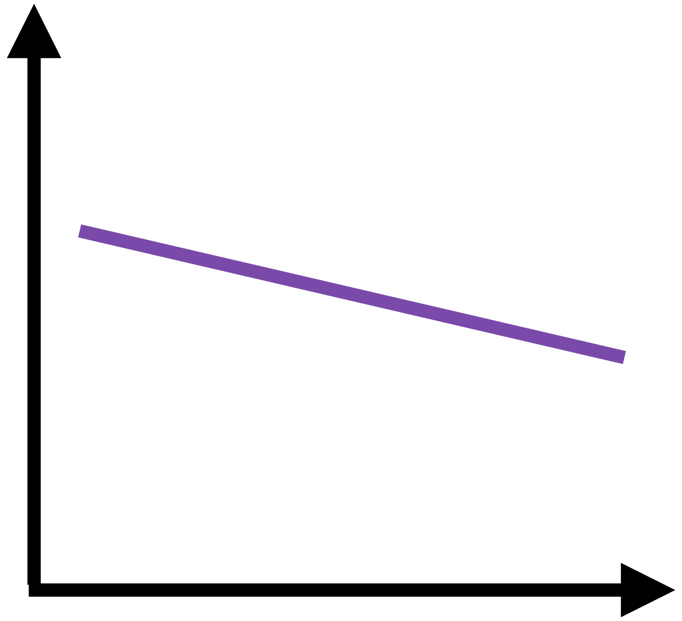
Installing the scikit-learn library

Demo

Exploring and visualizing relationships in data

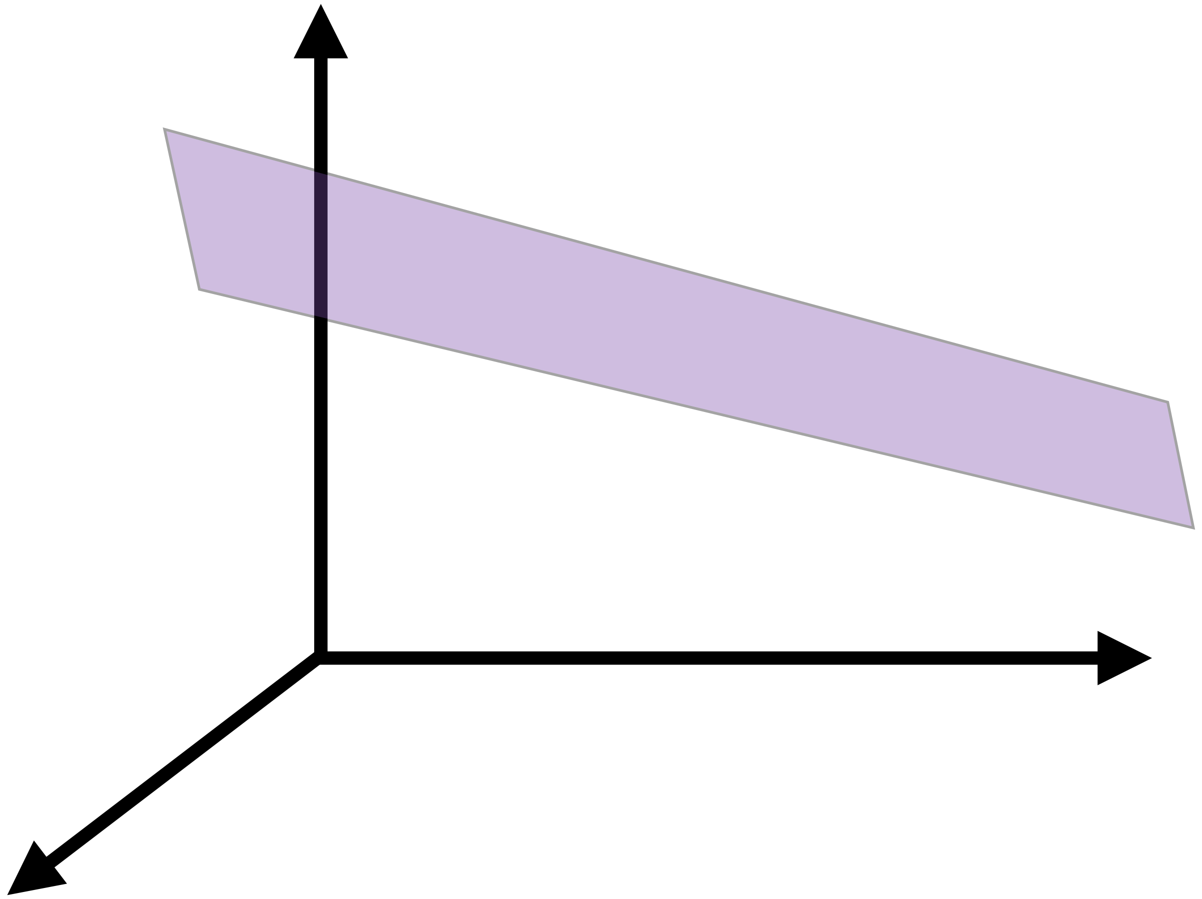
Risks in Multiple Regression

Simple and Multiple Regression



Simple Regression

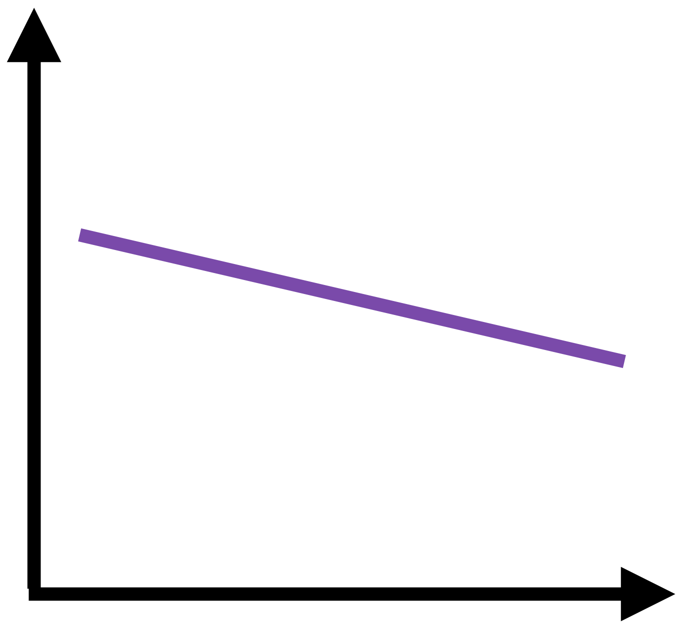
Data in 2 dimensions



Multiple Regression

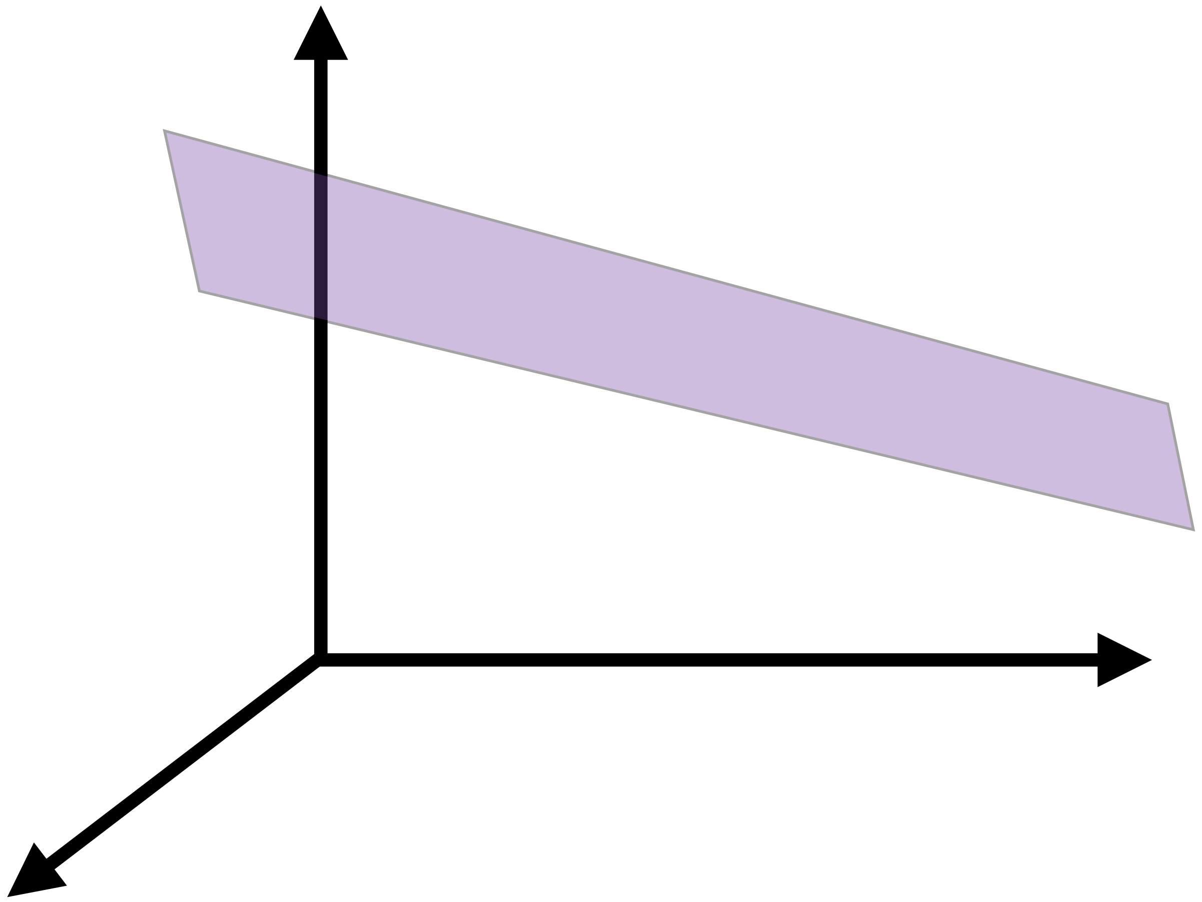
Data in > 2 dimensions

Simple and Multiple Regression



Simple Regression

Risks exist, but can usually be mitigated analysing R^2 and residuals



Multiple Regression

Risks are more complicated, require interpreting regression statistics

Risks in Simple Regression

No cause-effect relationship

Regression on completely unrelated data series

Mis-specified relationship

Non-linear (exponential or polynomial) fit

Incomplete relationship

Multiple causes exist, we have captured just one

Diagnosing Risks in Simple Regression

No cause-effect relationship

low R^2, plot of X ~ Y has no pattern

Mis-specified relationship

high R^2, residuals are not independent of each other

Incomplete relationship

low R^2, residuals are not independent of x

Mitigating Risks in Simple Regression

No cause-effect relationship

Wrong choice of X and Y - back to drawing board

Mis-specified relationship

Transform X and Y - convert to logs or returns

Incomplete relationship

Add X variables (move to multiple regression)

The big new risk with multiple regression is **multicollinearity**: X variables containing the same information

Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \dots + C_kx_{k-1}$$

$$\underset{n \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}} = \underset{n \times k}{\begin{bmatrix} 1 & x_{11} & & x_{1k-1} \\ 1 & x_{21} & & x_{2k-1} \\ 1 & x_{31} & \dots & x_{3k-1} \\ \dots & \dots & & \dots \\ 1 & x_{n1} & & x_{nk-1} \end{bmatrix}} * \underset{k \times 1}{\begin{bmatrix} C_1 \\ C_2 \\ \dots \\ C_k \end{bmatrix}}$$

n Rows,
1 Column

n Rows,
k Columns

k Rows,
1 Column

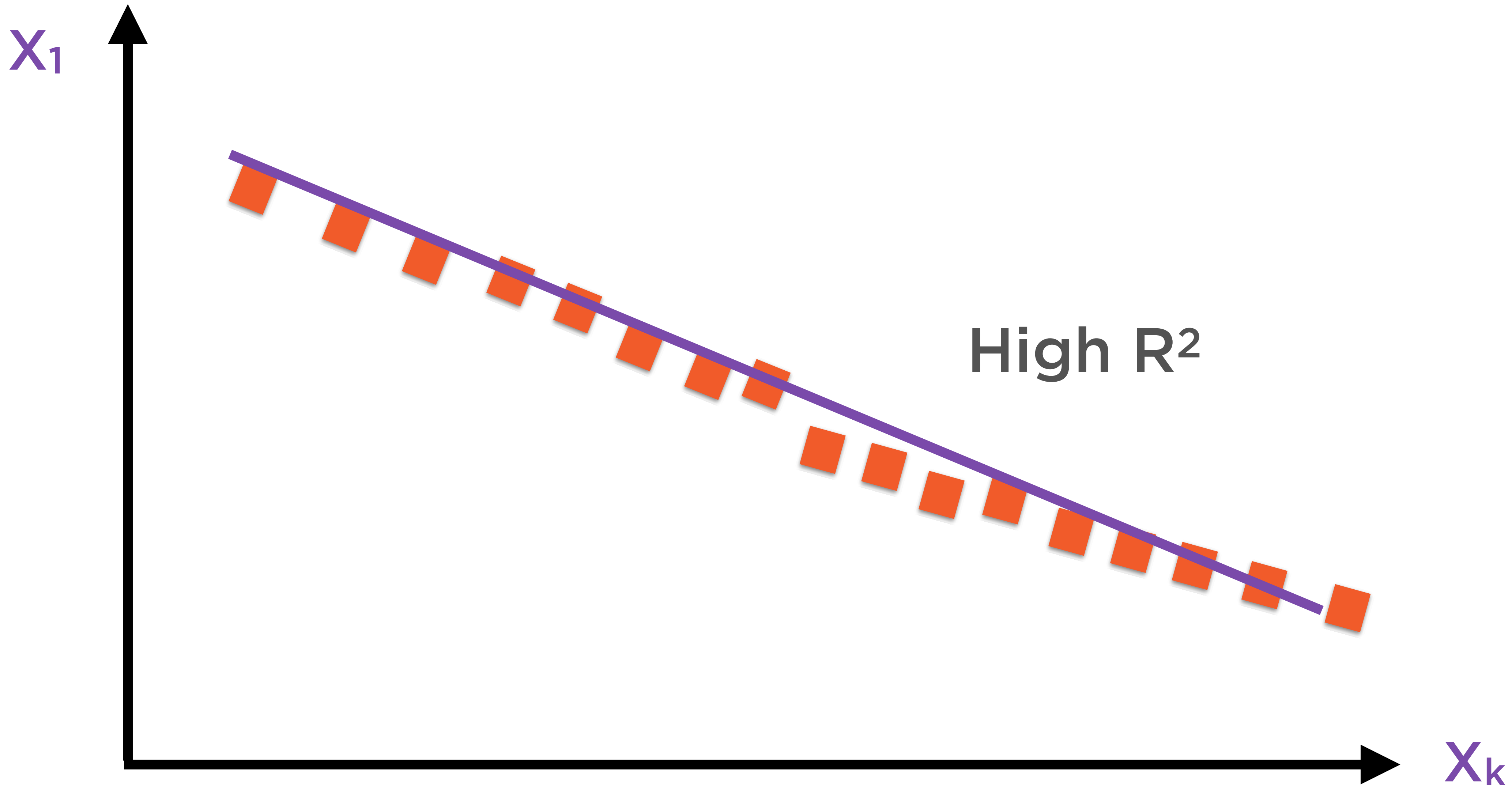
Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \dots + C_kx_{k-1}$$

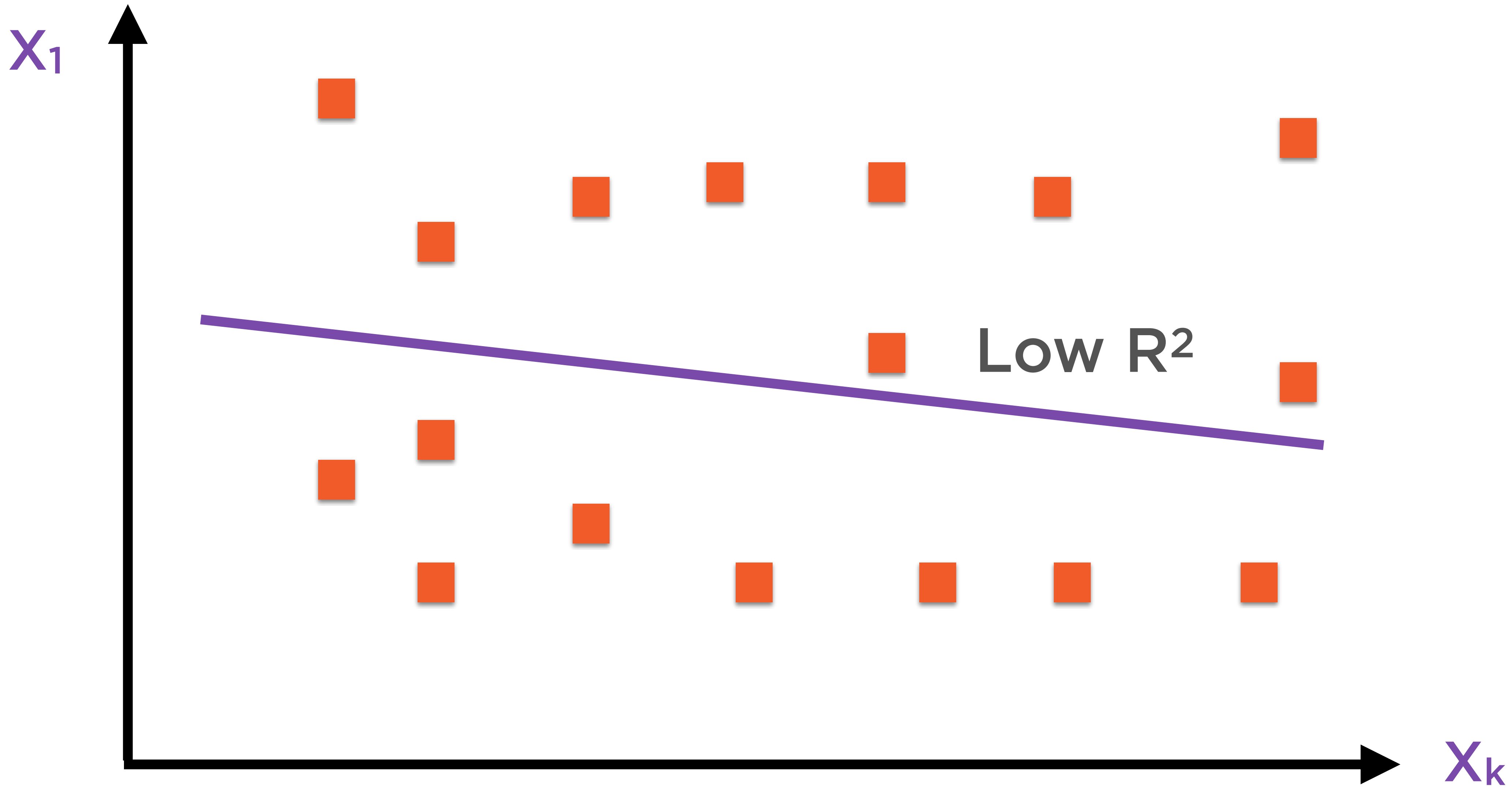
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{array}{c} \begin{bmatrix} 1 & x_{11} & \dots & x_{1k-1} \\ 1 & x_{21} & & x_{2k-1} \\ 1 & x_{31} & & x_{3k-1} \\ \dots & \dots & & \dots \\ 1 & x_{n1} & & x_{nk-1} \end{bmatrix} \\ \begin{array}{cccc} & X_1 & & X_k \end{array} \end{array} * \begin{bmatrix} C_1 \\ C_2 \\ \dots \\ C_k \end{bmatrix}$$

Bad News: Multicollinearity Detected



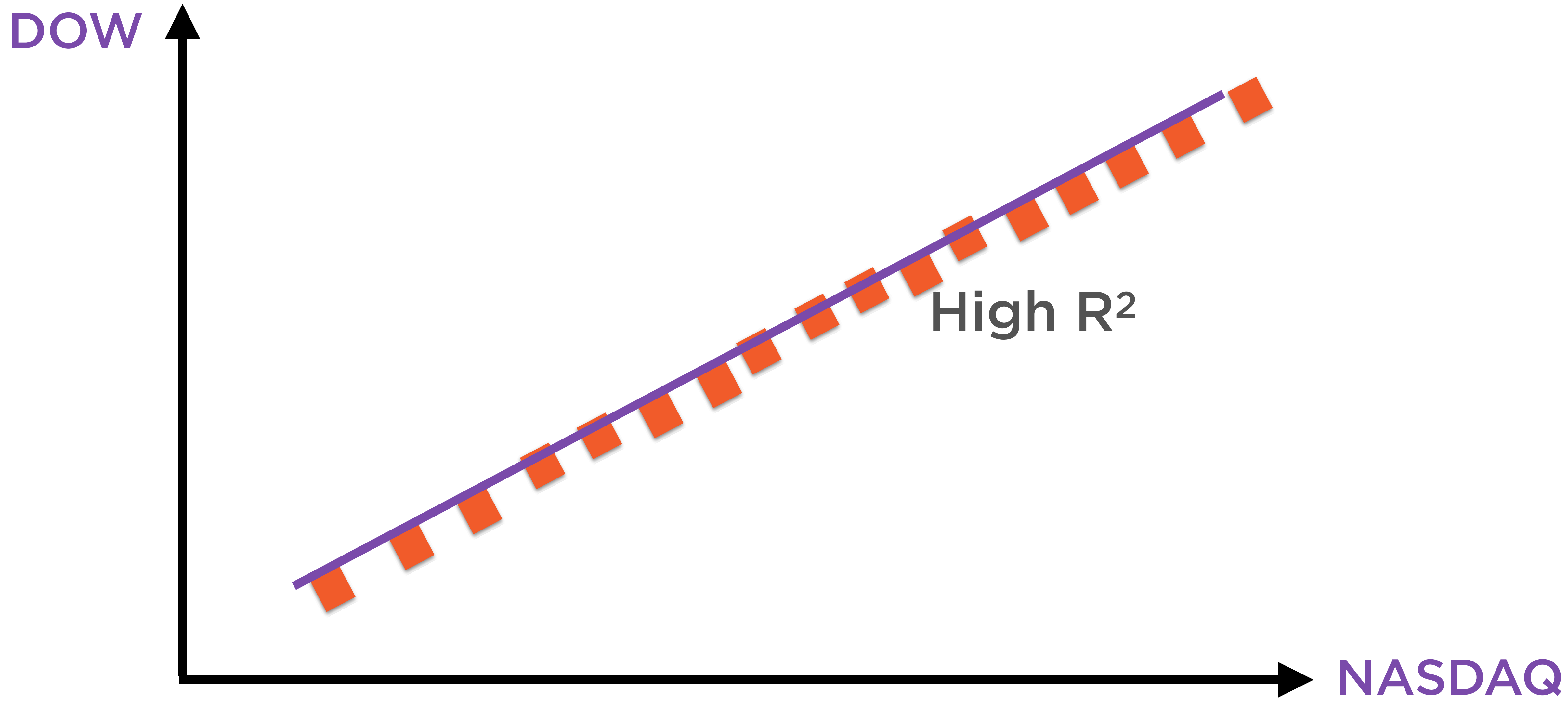
Highly correlated explanatory variables

Good News: No Multicollinearity Detected



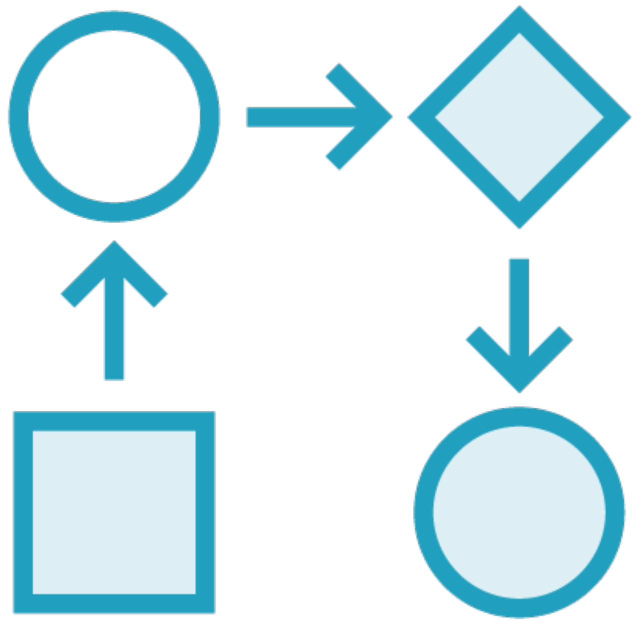
Uncorrelated explanatory variables

Bad News: Multicollinearity Detected



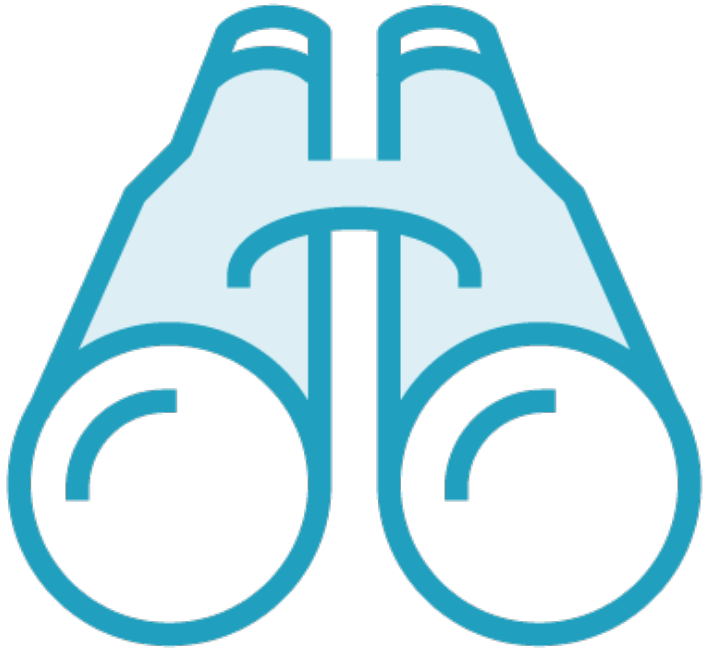
Highly correlated explanatory variables

Multicollinearity Kills Regression's Usefulness



Explaining Variance

The R^2 as well as the regression coefficients are not very reliable



Making Predictions

The regression model will perform poorly with out-of-sample data

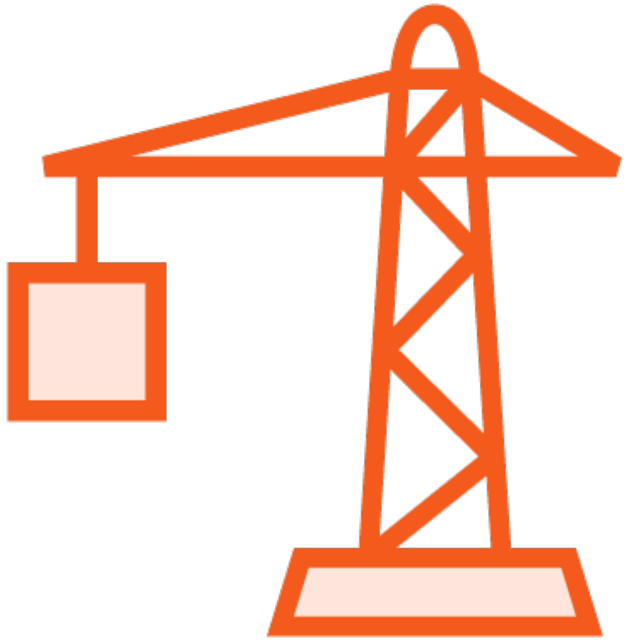
Multicollinearity: Prevention and Cure

Common Sense

Big-picture understanding of the data

Nuts and Bolts

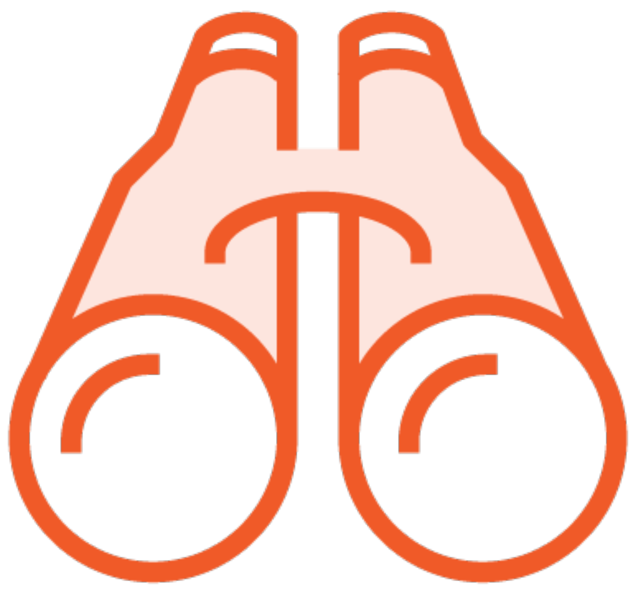
Setting up data right



Heavy Lifting

Factor analysis, principal components analysis (PCA)

Common Sense



Think deeply about each x variable

Eliminate closely related ones

Dig down to underlying causes

Nuts and Bolts

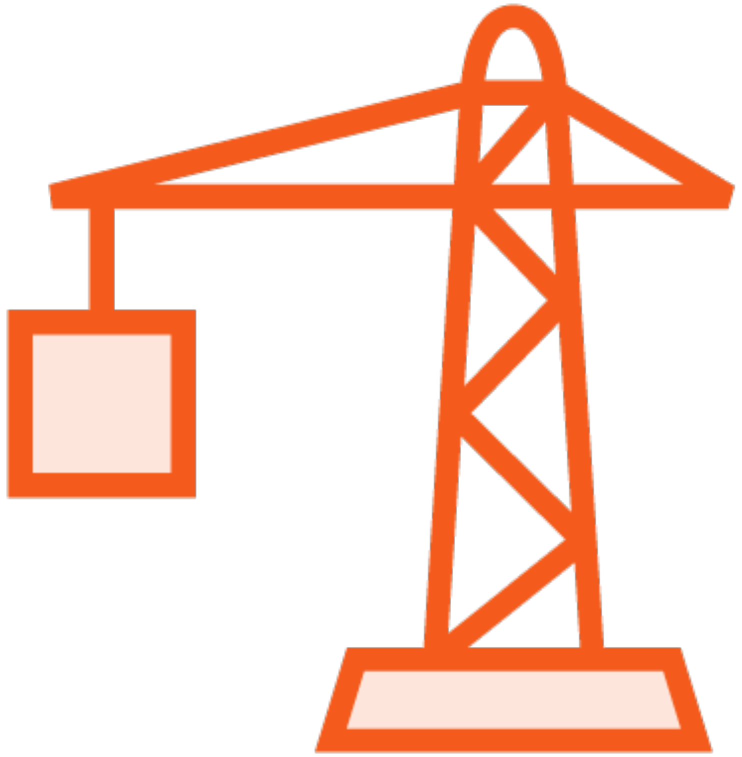
‘Standardize’ the variables

Rely on adjusted-R^2, not plain R^2

Set up dummy variables right

Distribute lags

Heavy Lifting



Find underlying factors that drive the correlated x variables

Principal Component Analysis (PCA) is a great tool

Interpreting the Results of a Regression Analysis

R^2

The most common and popular metric for evaluating regression

Between 0 and 100%

Unfortunately, always increases by adding new x variables

Can lead to overfitting

Adjusted R^2 preferred for evaluating multiple regression

$$\text{Adjusted-}R^2 = R^2 \times \text{(Penalty for adding irrelevant variables)}$$

Adjusted-R^2

Increases if irrelevant* variables are deleted

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

Regression with Categorical Variables

A Simple Regression

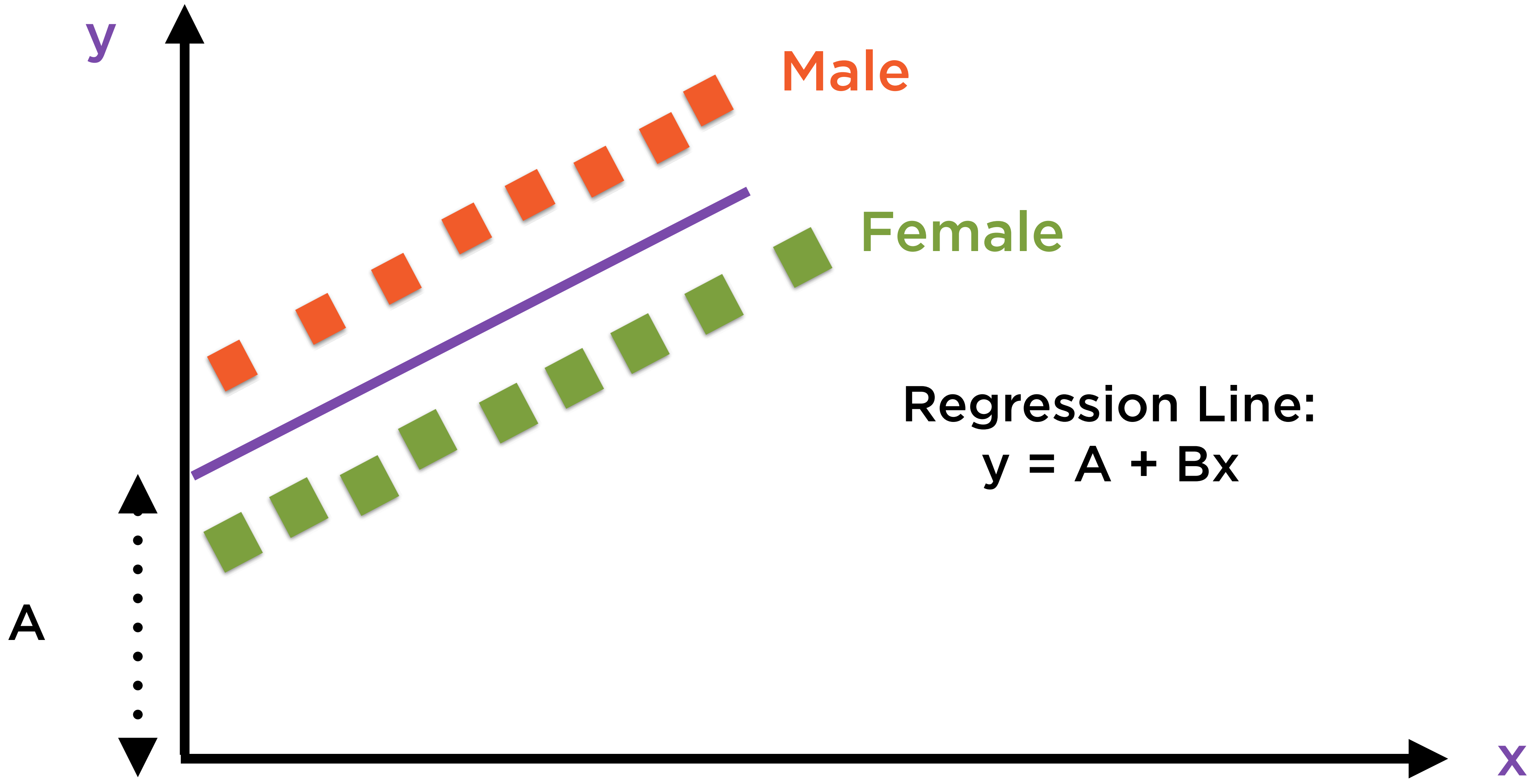
Proposed Regression Equation:

$$y = A + Bx$$

Height of individual

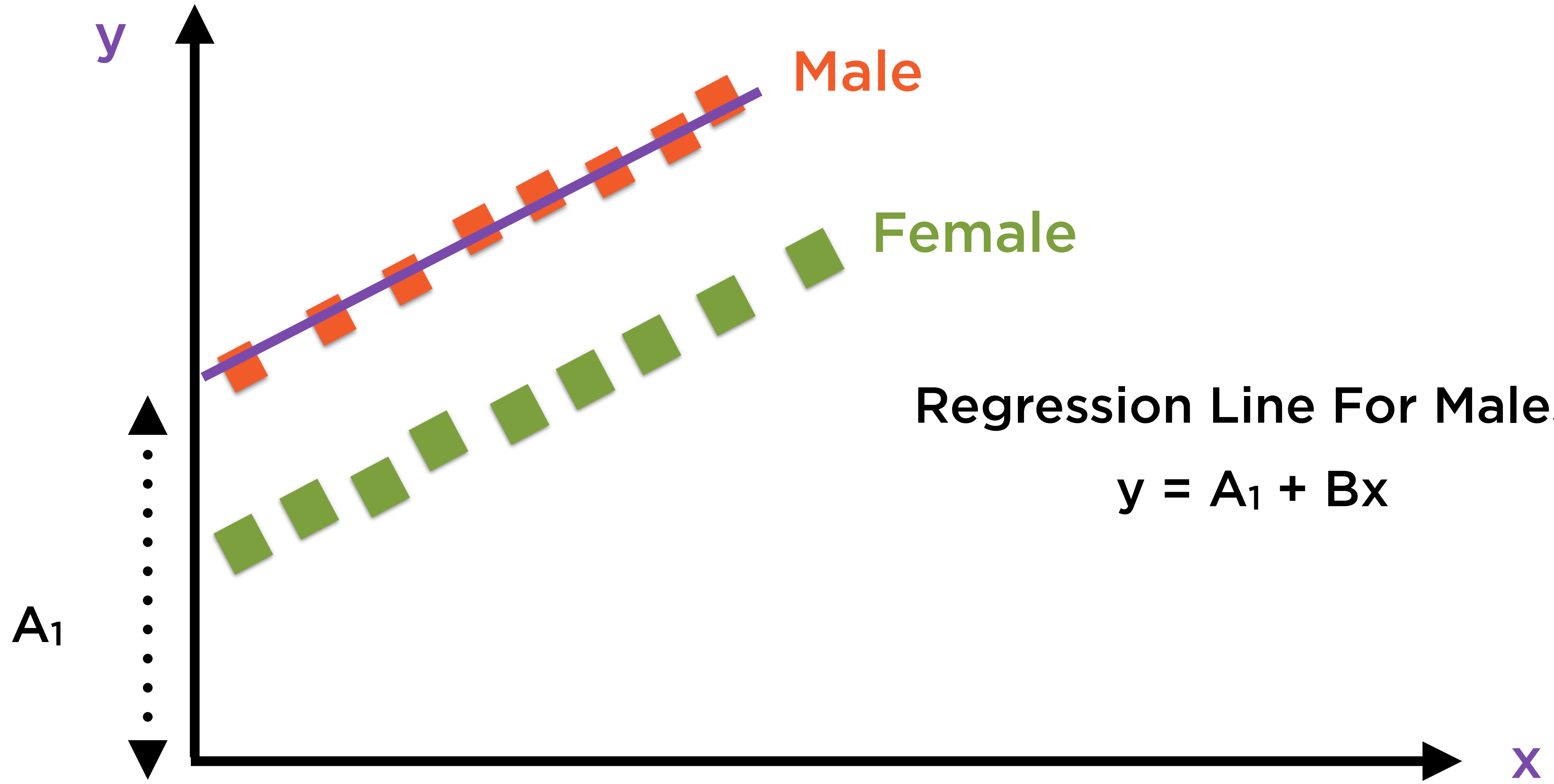
Average height of parents

A Simple Regression



Not a great fit - regression line is far from all points!

A Simple Regression

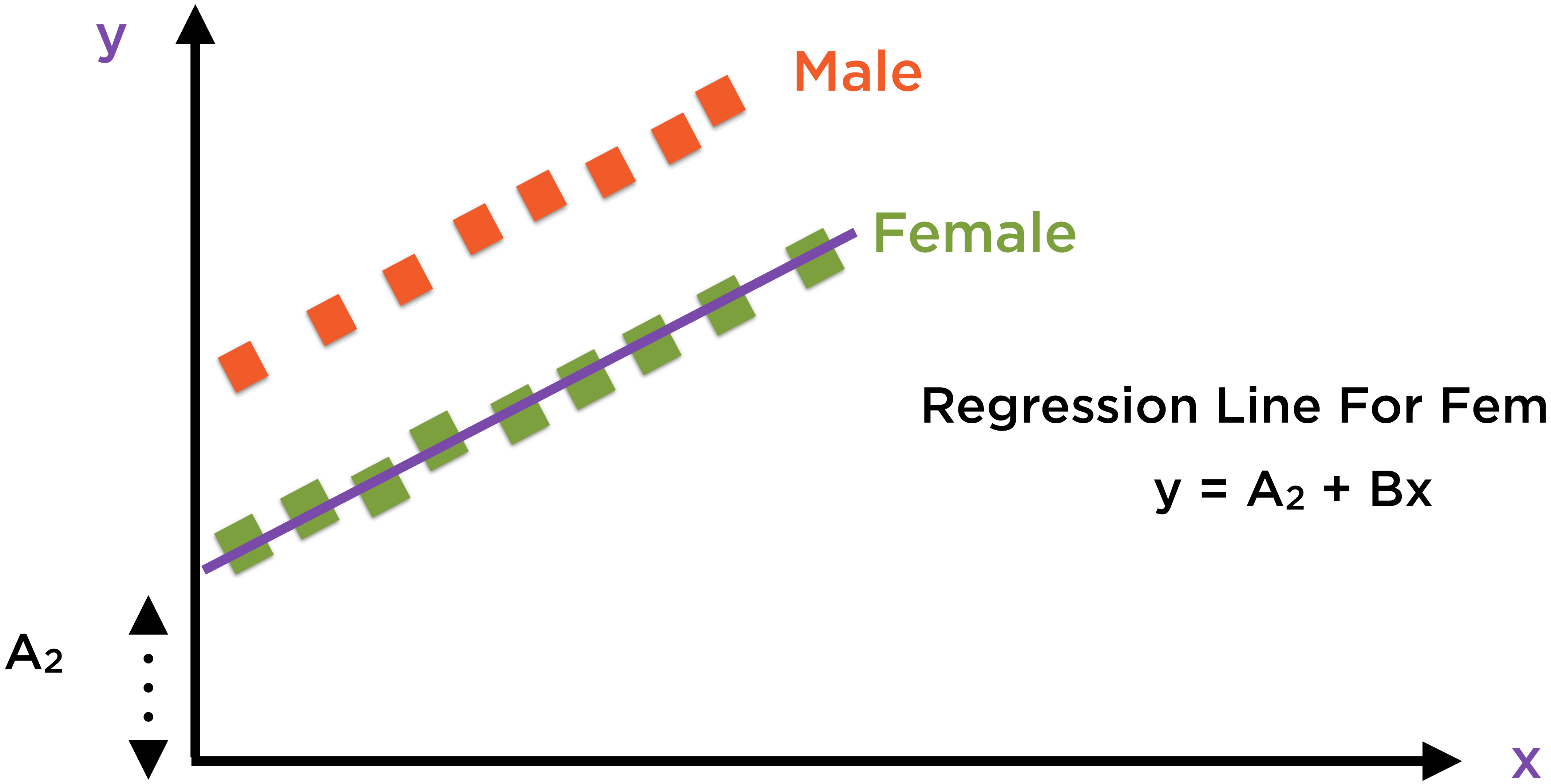


Regression Line For Males:

$y = A_1 + Bx$

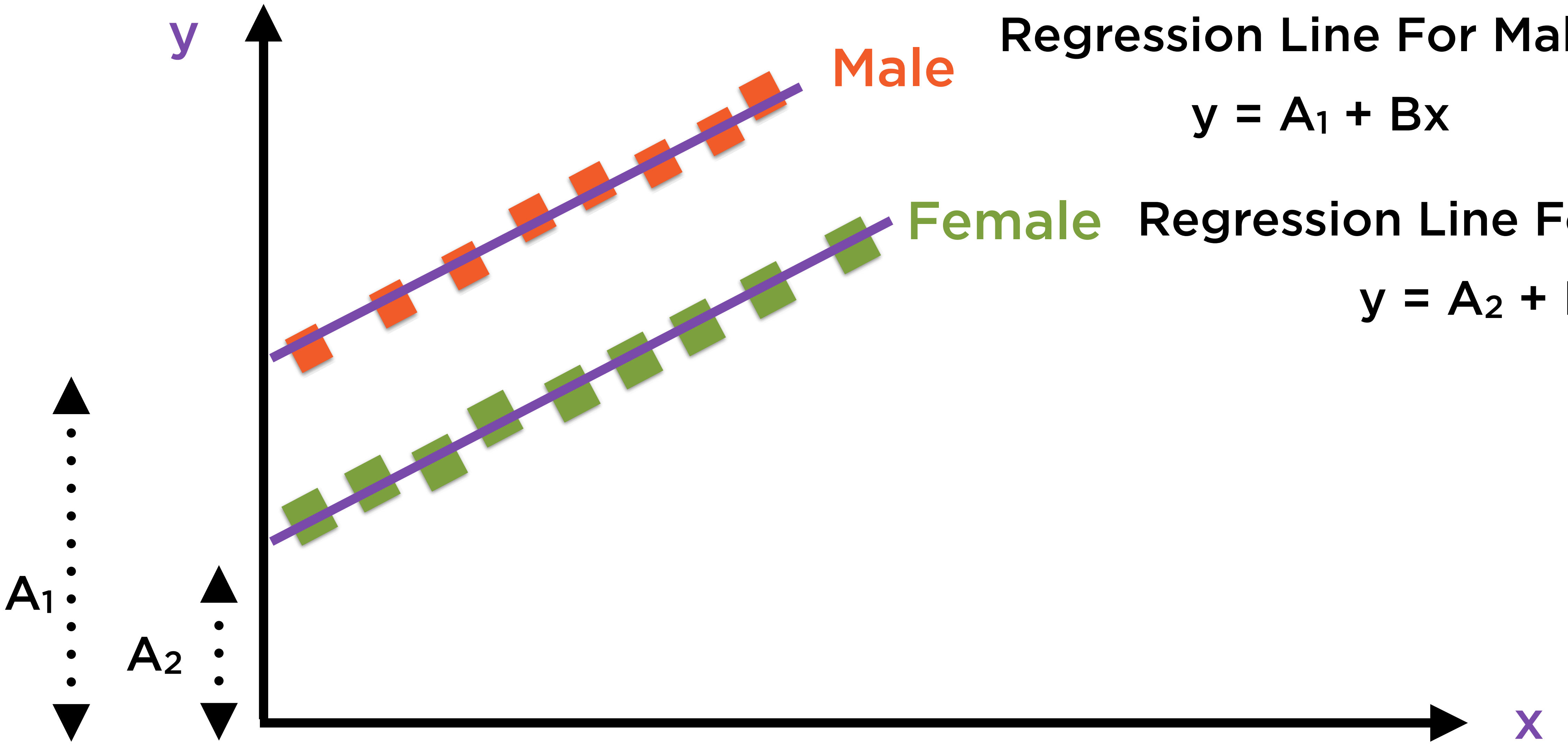
We can easily plot a great fit for males...

A Simple Regression



...and another great fit for females

A Simple Regression



Regression Line For Males:

$$y = A_1 + Bx$$

Regression Line For Females:

$$y = A_2 + Bx$$

Two lines - same slope, different intercepts

Adding A Dummy Variable

Regression Line For Males:

$$y = A_1 + Bx$$

Regression Line For Females:

$$y = A_2 + Bx$$

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$= 1$ for females

Adding A Dummy Variable

Regression Line For Males:

$y = A_1 + Bx$

Regression Line For Females:

$y = A_2 + Bx$

Combined Regression Line:

$y = A_1 + (A_2 - A_1)D + Bx$

$D = 0$ for males

$$y = A_1 + \cancel{(A_2 - A_1)D} + Bx$$

$$= A_1 + Bx$$

Adding A Dummy Variable

Regression Line For Males:

$$y = A_1 + Bx$$

Regression Line For Females:

$$y = A_2 + Bx$$

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 1$ for females

$$y = \cancel{A_1} + (A_2 \cancel{- A_1}) + Bx$$

$$= A_2 + Bx$$

Adding A Dummy Variable

Original Regression Equation:

$$y = A + Bx$$

Height of individual

Average height of parents

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$= 1$ for females

Adding A Dummy Variable

Combined Regression Line:

$$y = A_1 + (A_2 - A_1)D + Bx$$

$D = 0$ for males

$= 1$ for females

The data contained 2 groups, so we added 1 dummy variable

Given data with k groups, set up k-1 dummy variables, else multicollinearity occurs

Dummy and Other Categorical Variables

Dummy Variables

Binary - 0 or 1

Categorical Variables

Finite set of values - e.g. days of week, months of year...

To include non-binary categorical variables, simply add more dummies

Testing for Seasonality

Proposed Regression Equation:

$$y = A + BQ_1 + CQ_2 + DQ_3$$

Average stock returns

Quarter of the year

The data contains 4 groups, so we added 3 dummy variables

Testing for Seasonality

$$y = A + BQ_1 + CQ_2 + DQ_3$$

The data contains 4 groups, so we added 3 dummy variables

Q_1	$= 1$	for Jan, Feb, Mar
	$= 0$	for other quarters
Q_2	$= 1$	for Apr, May, Jun
	$= 0$	for other quarters
Q_3	$= 1$	for July, Aug, Sep
	$= 0$	for other quarters

Summary

Linear regression as a machine learning problem

Mean Square Error (MSE) as loss function

Interpreting the results of a regression analysis

R^2 for evaluating regression models