Understanding and Applying Linear Regression

MODELING RELATIONSHIPS BETWEEN VARIABLES USING REGRESSION

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Introduce regression models as a way to connect the dots

Set up the regression problem

Understand why regression is such a popular tool

See how regression is an example of Machine Learning

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

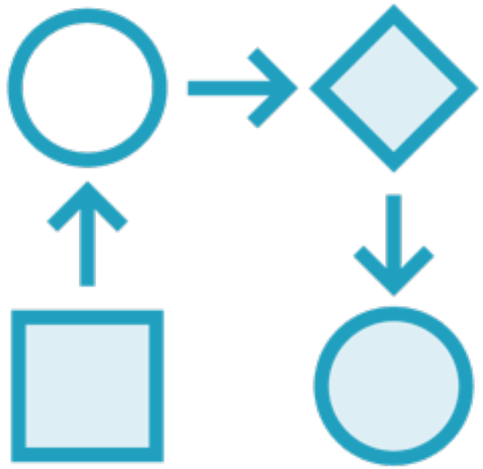
Inferential Statistics

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

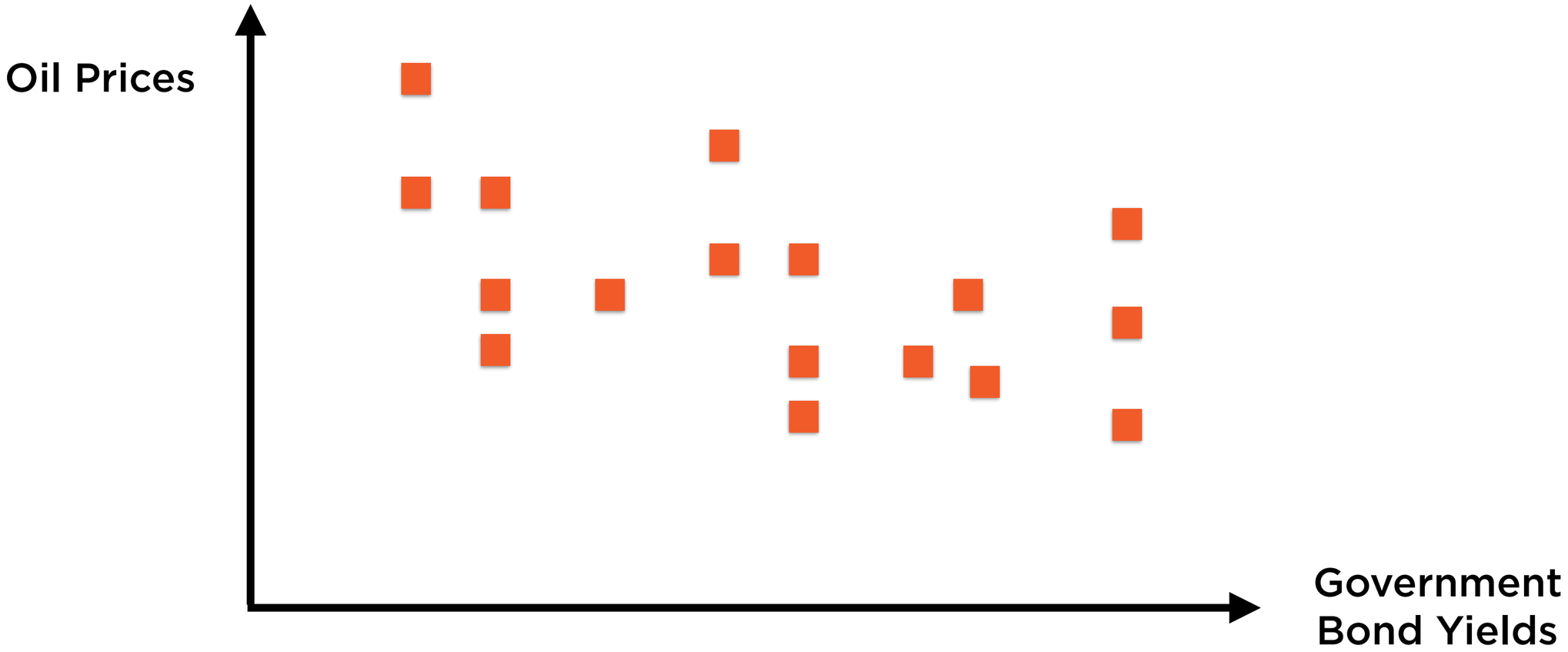
Data in One Dimension

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

Data in One Dimension

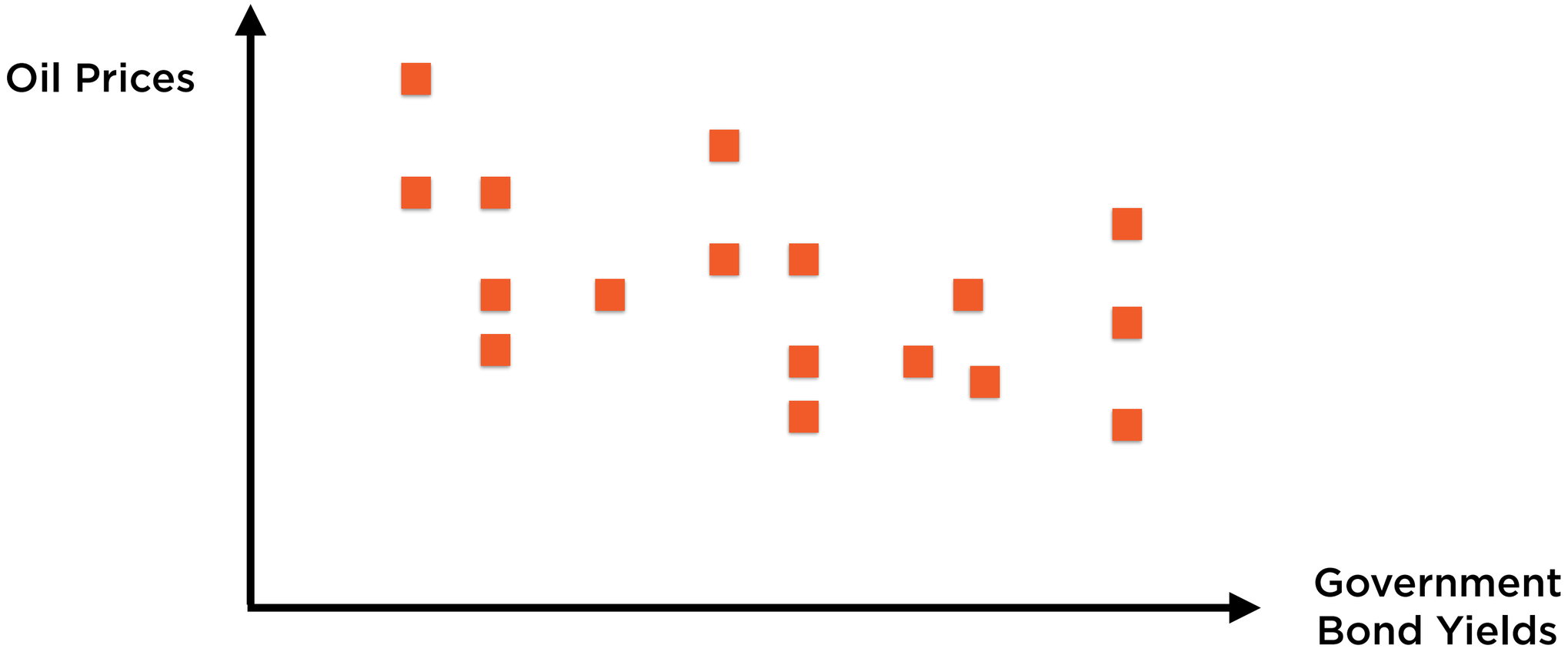
Unidimensional data is analysed using statistics such as mean, median, standard deviation

Data in Two Dimensions



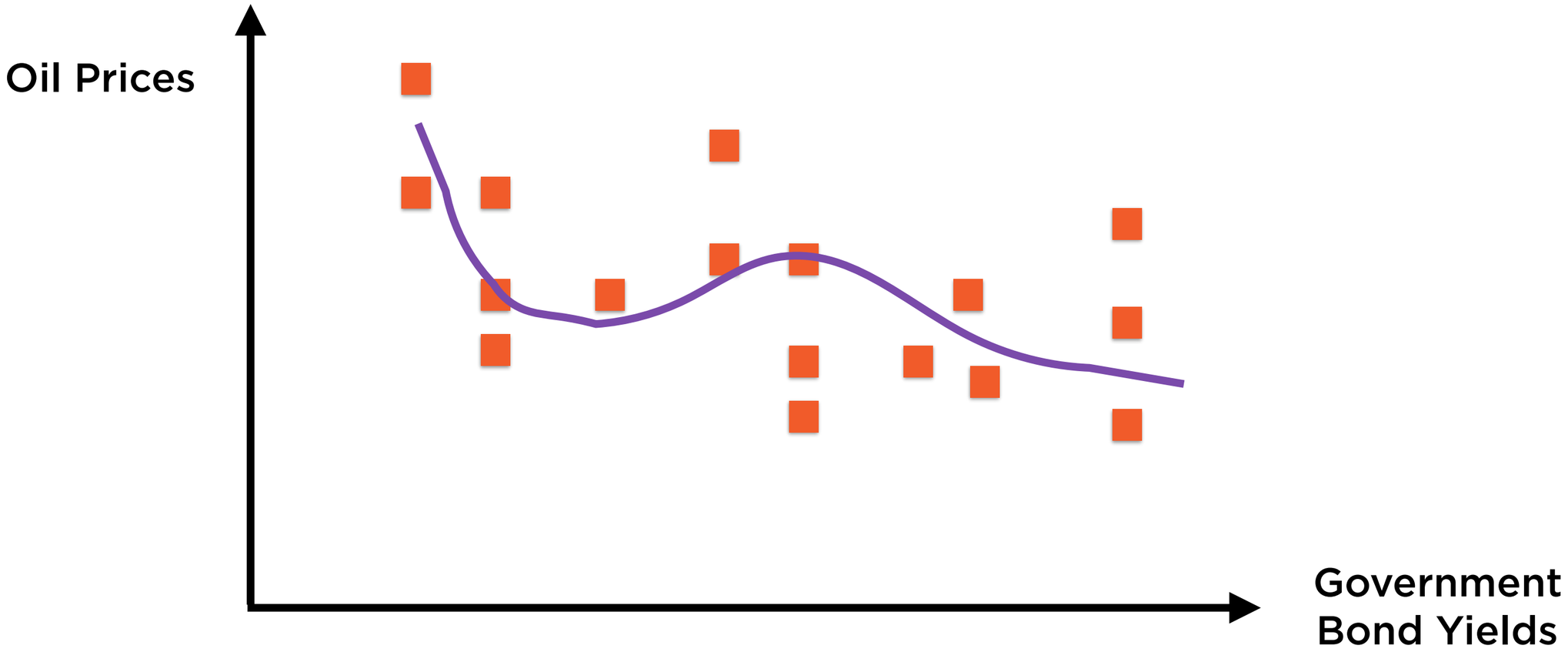
Its often more insightful to view data in relation to some other, related data

Data in Two Dimensions



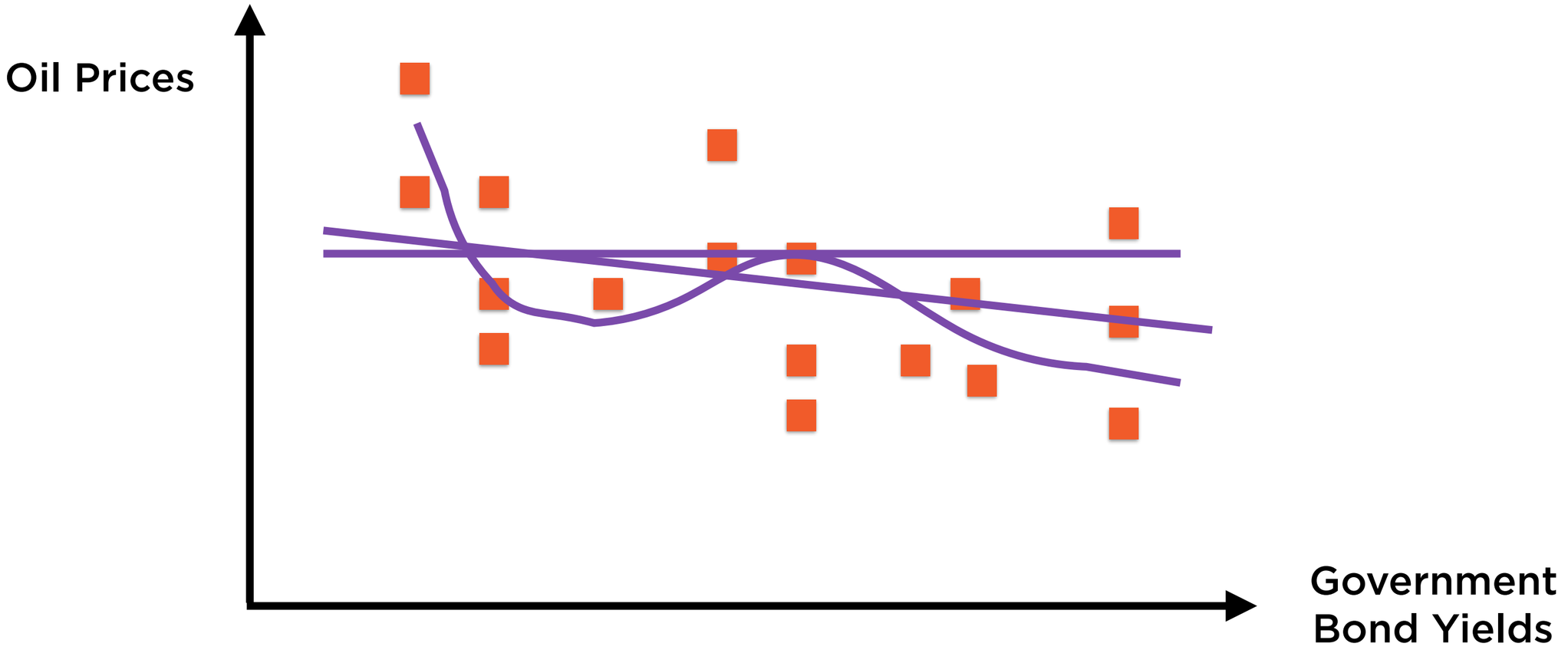
Bidimensional data can be represented in a plane

Data in Two Dimensions



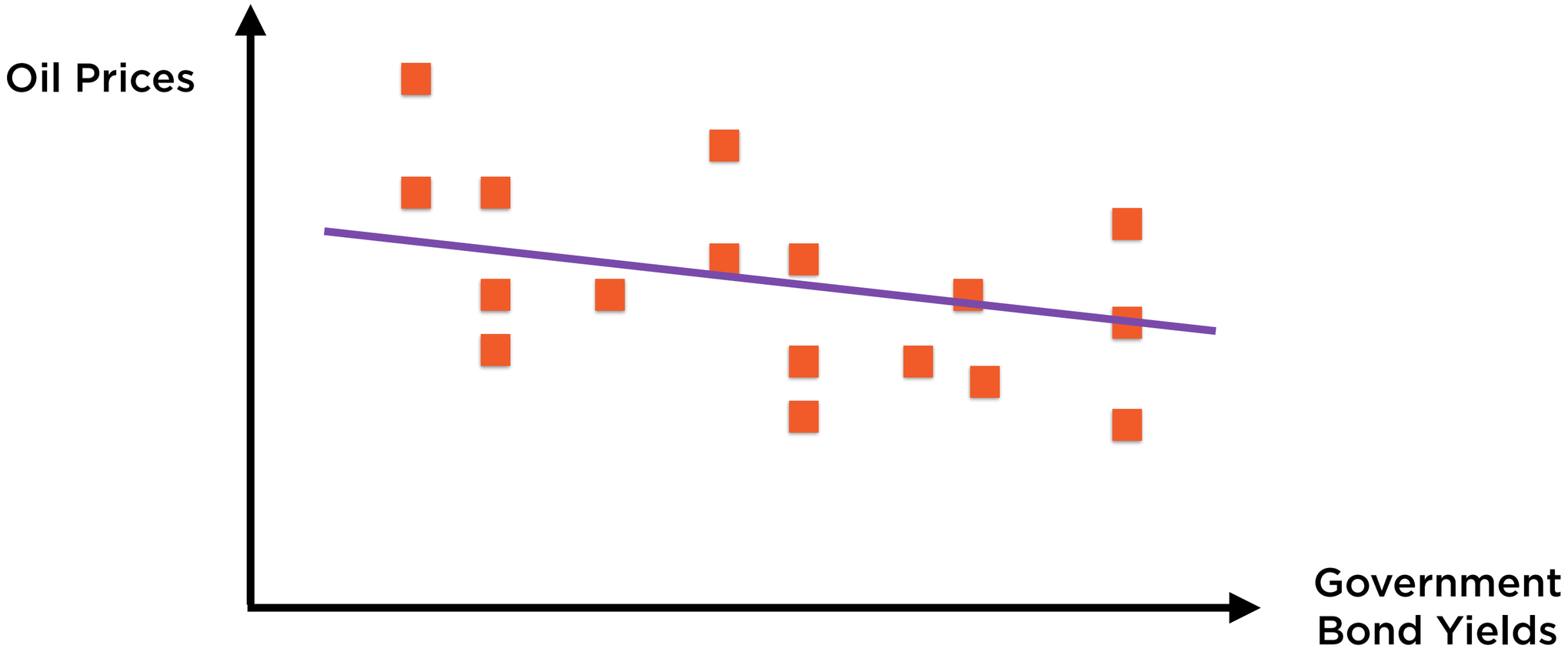
We can draw any number of curves to fit such data

Data in Two Dimensions



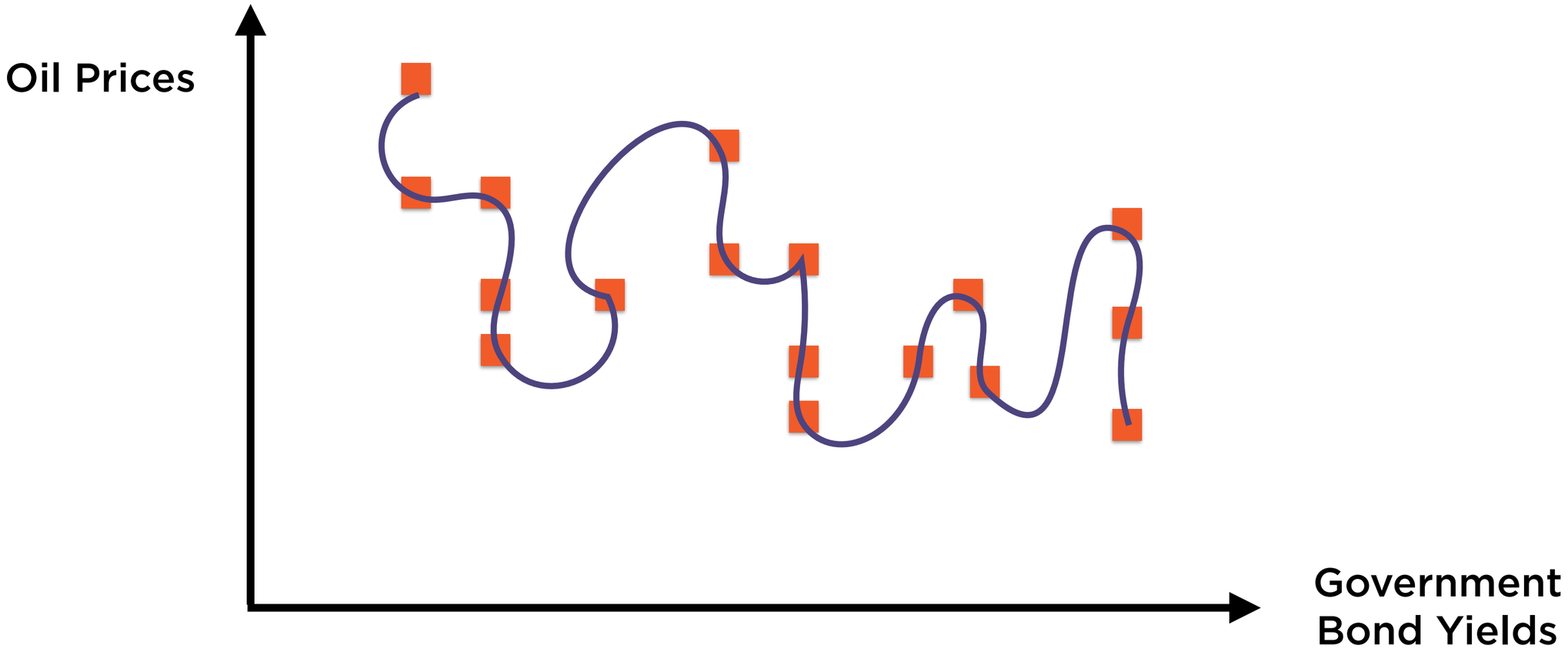
We can draw any number of curves to fit such data

Data in Two Dimensions



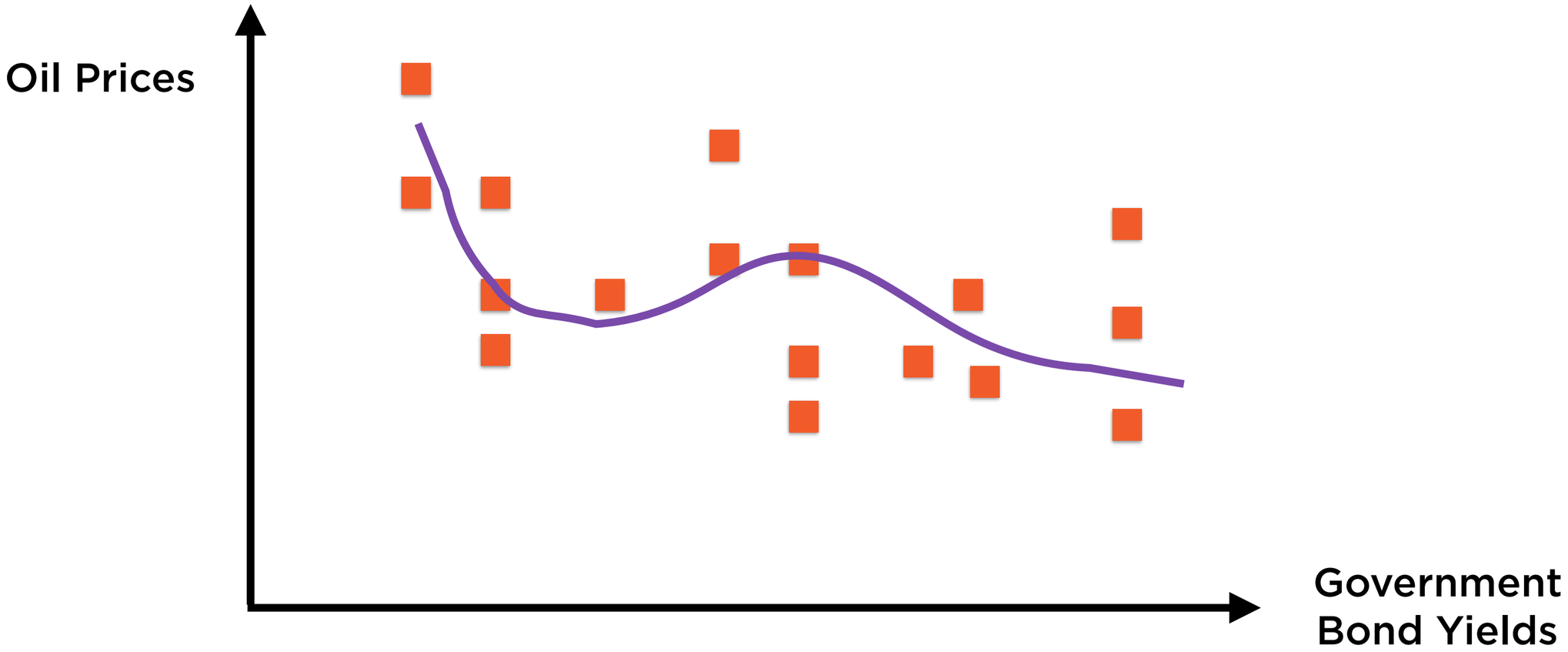
A straight line represents a linear relationship

Data in Two Dimensions



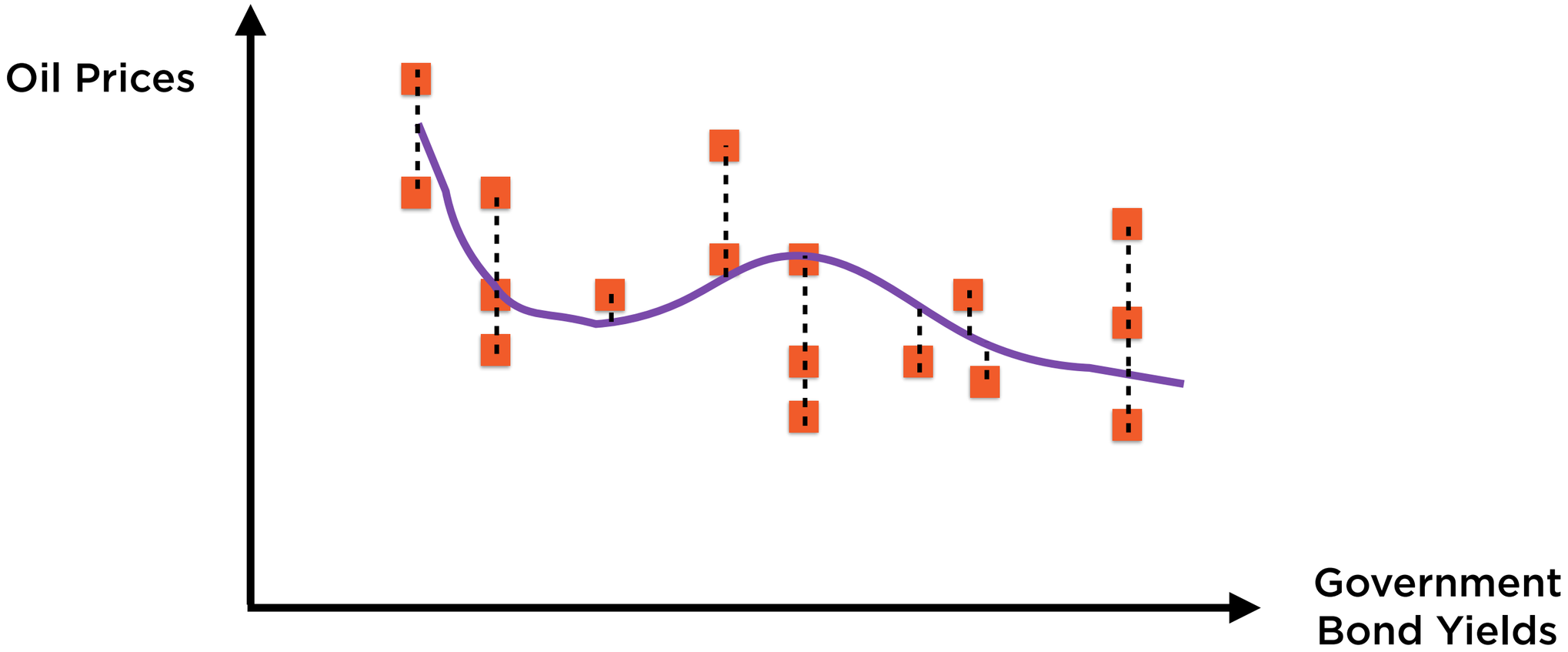
We could either make this curve pass through each point...

Data in Two Dimensions



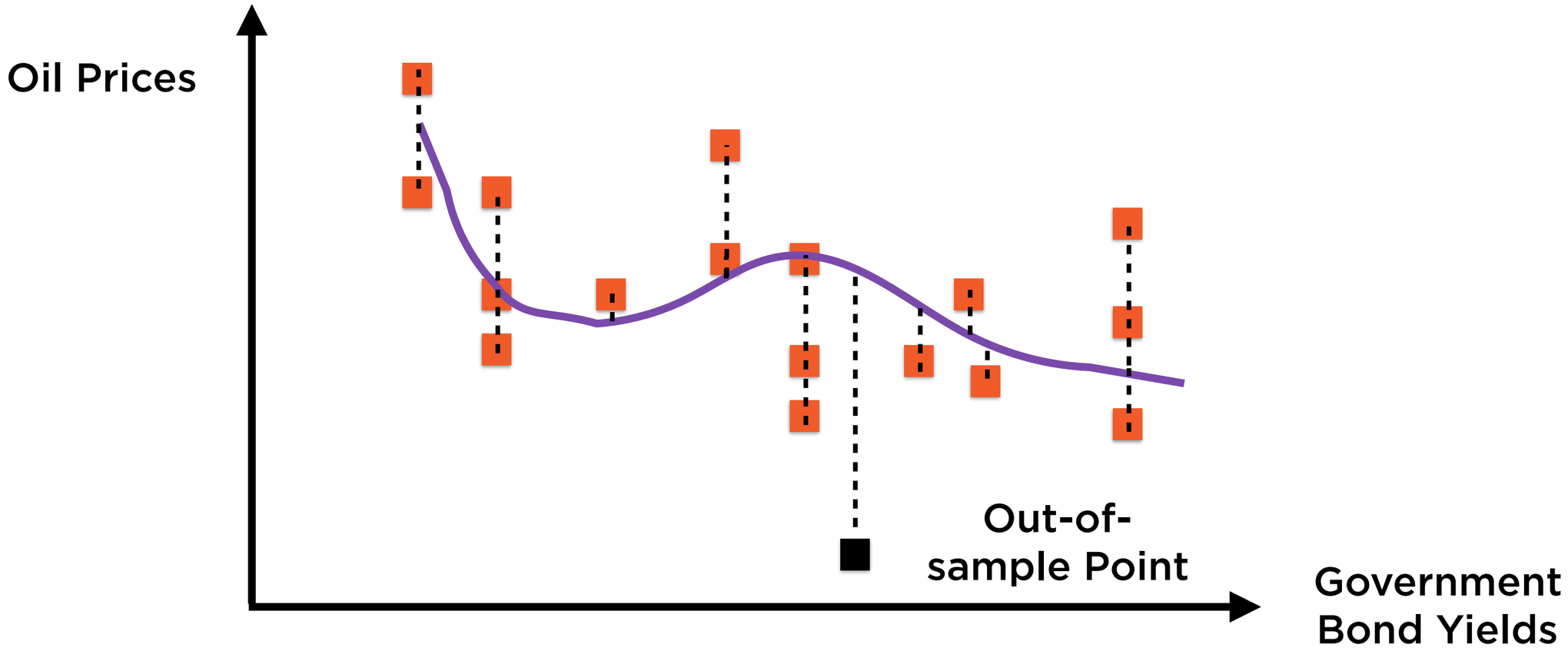
...Or in some sense “fit” the data in aggregate

Data in Two Dimensions



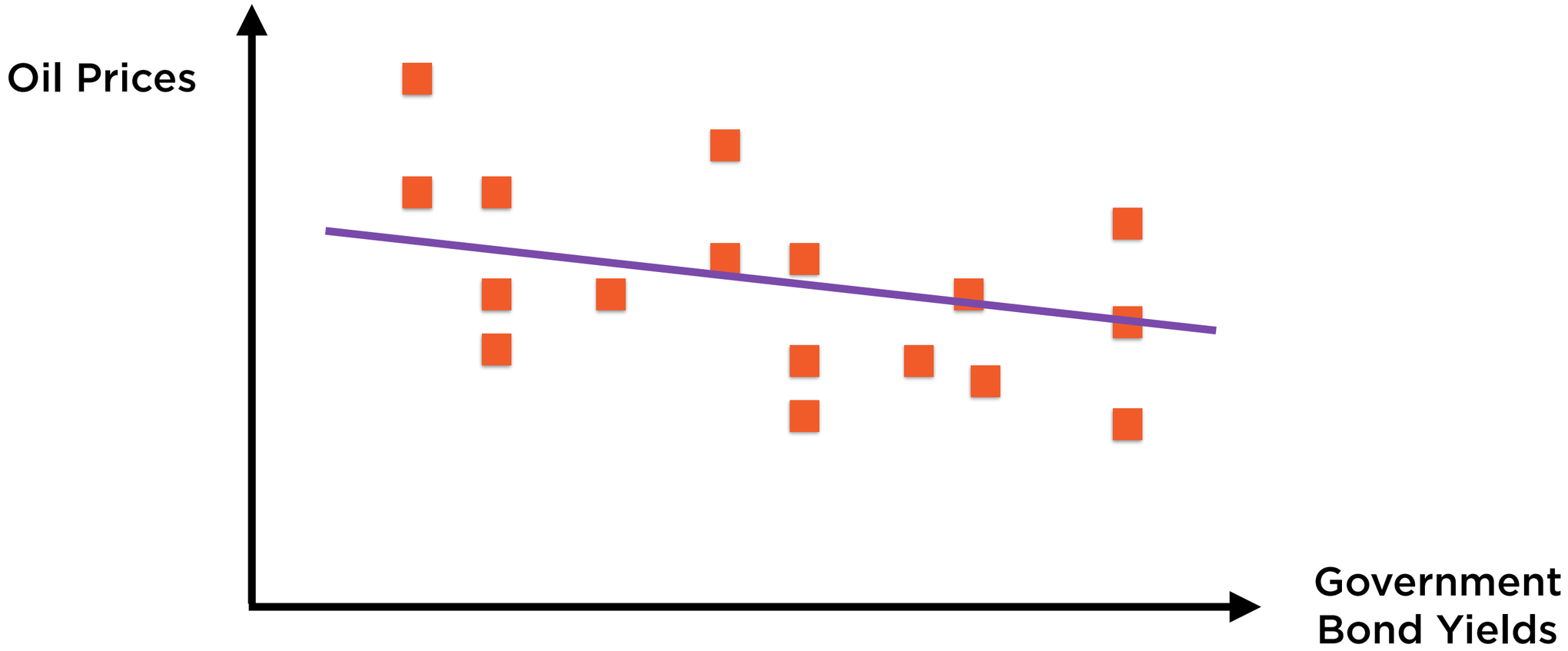
A curve has a “good fit” if the distances of points from the curve are small

Data in Two Dimensions



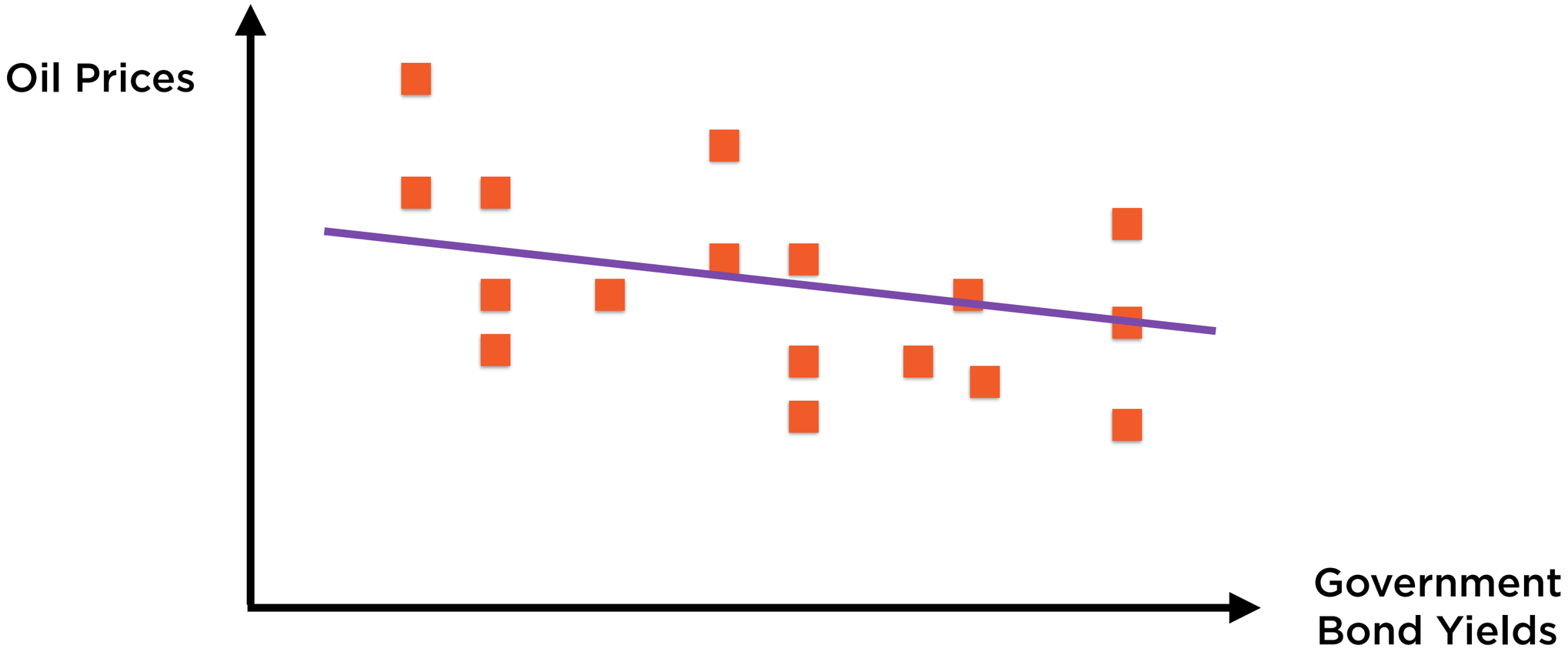
Overfitting by finding a very complicated curve often only hurts predictive accuracy

Data in Two Dimensions



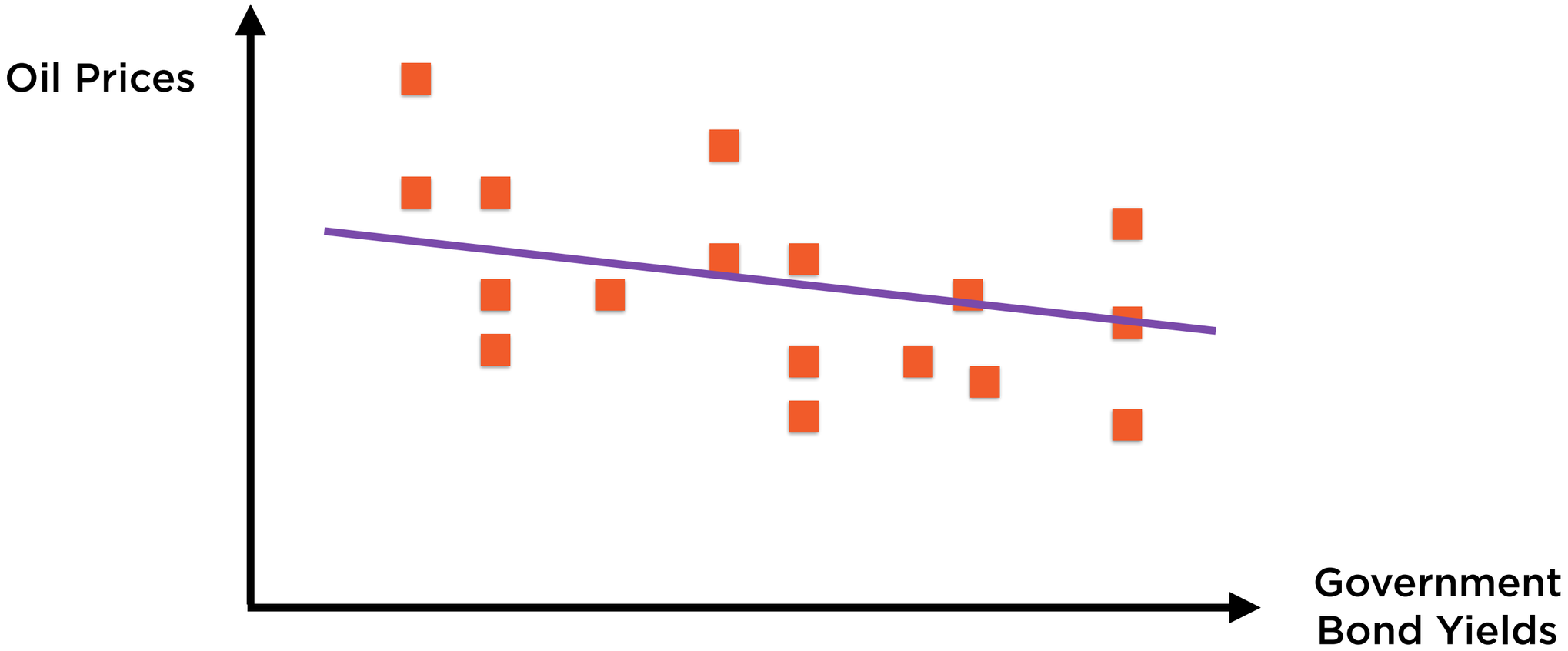
Often, a straight line works just fine

Data in Two Dimensions



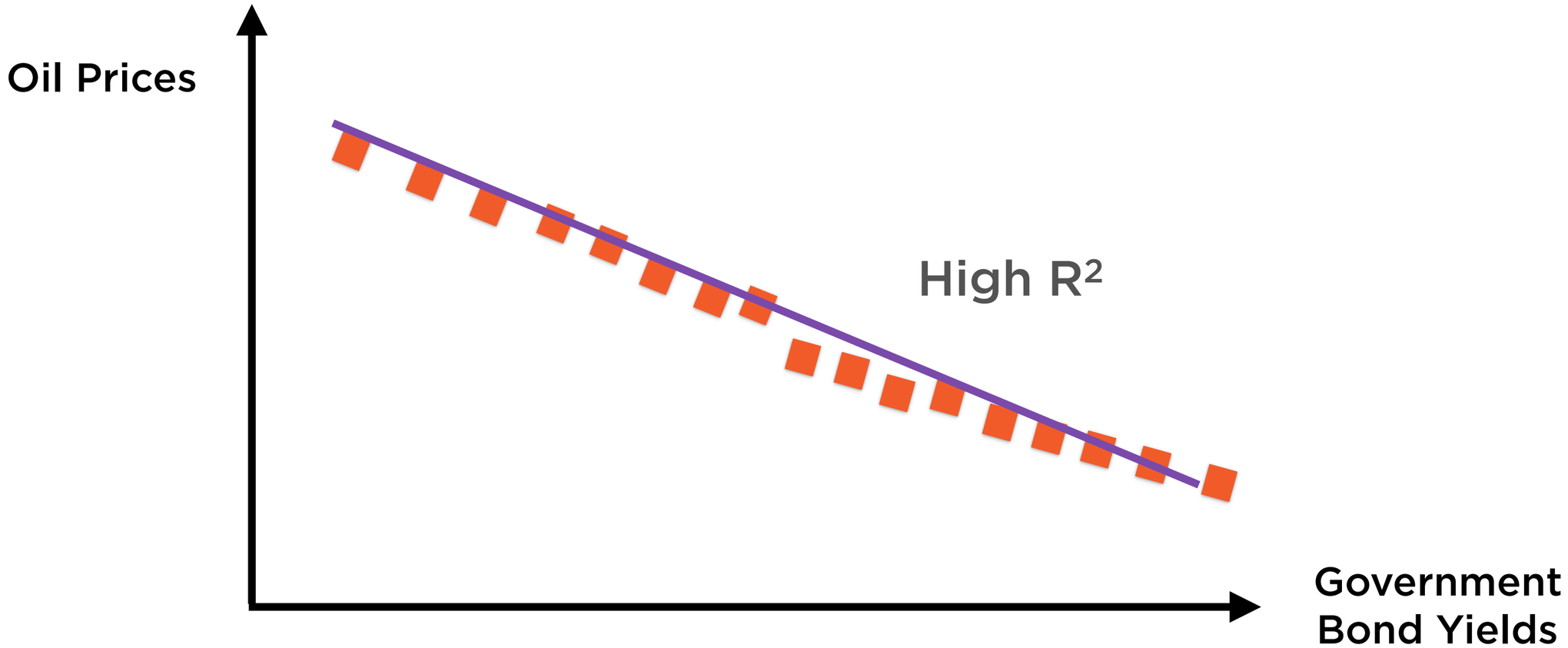
Finding the “best” such straight line is called Linear Regression

Data in Two Dimensions



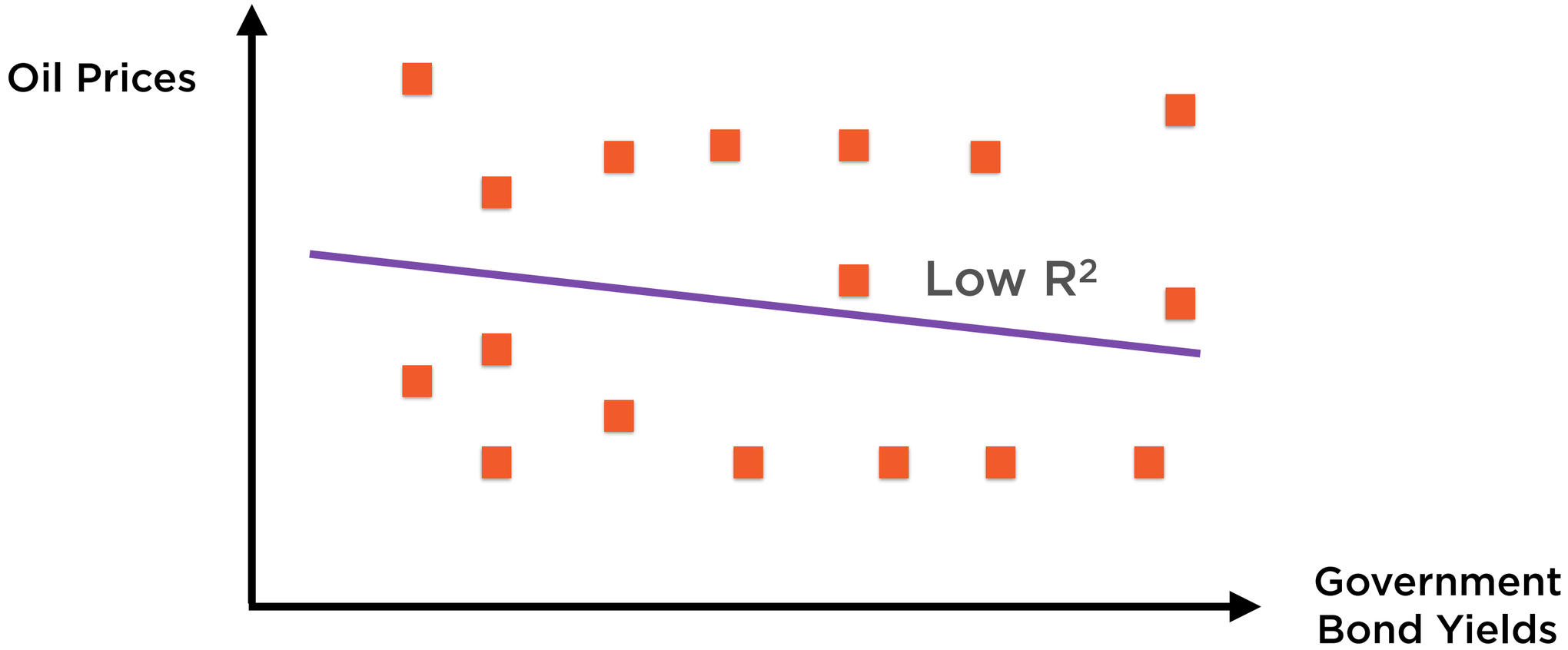
Regression not only gives us the equation of this line, it also signals how reliable the line is

Data in Two Dimensions



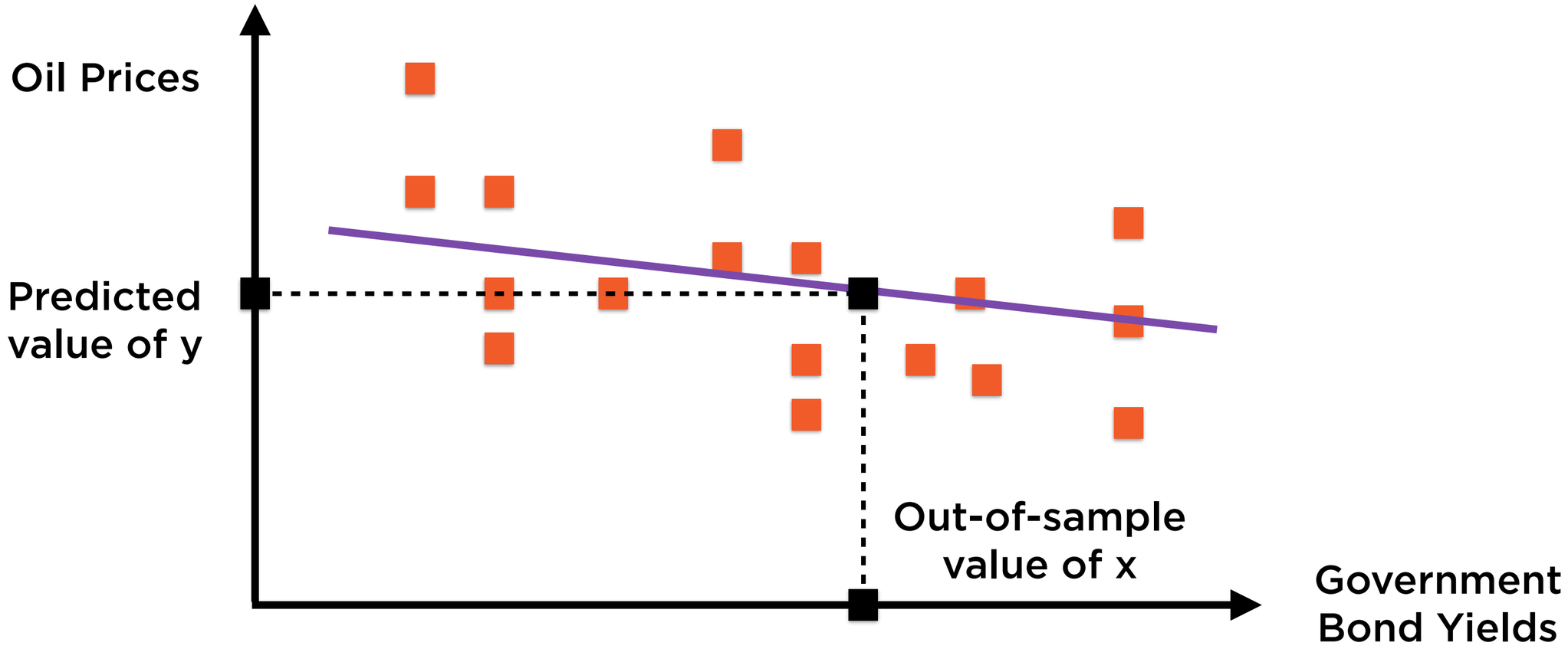
High quality of fit

Data in Two Dimensions



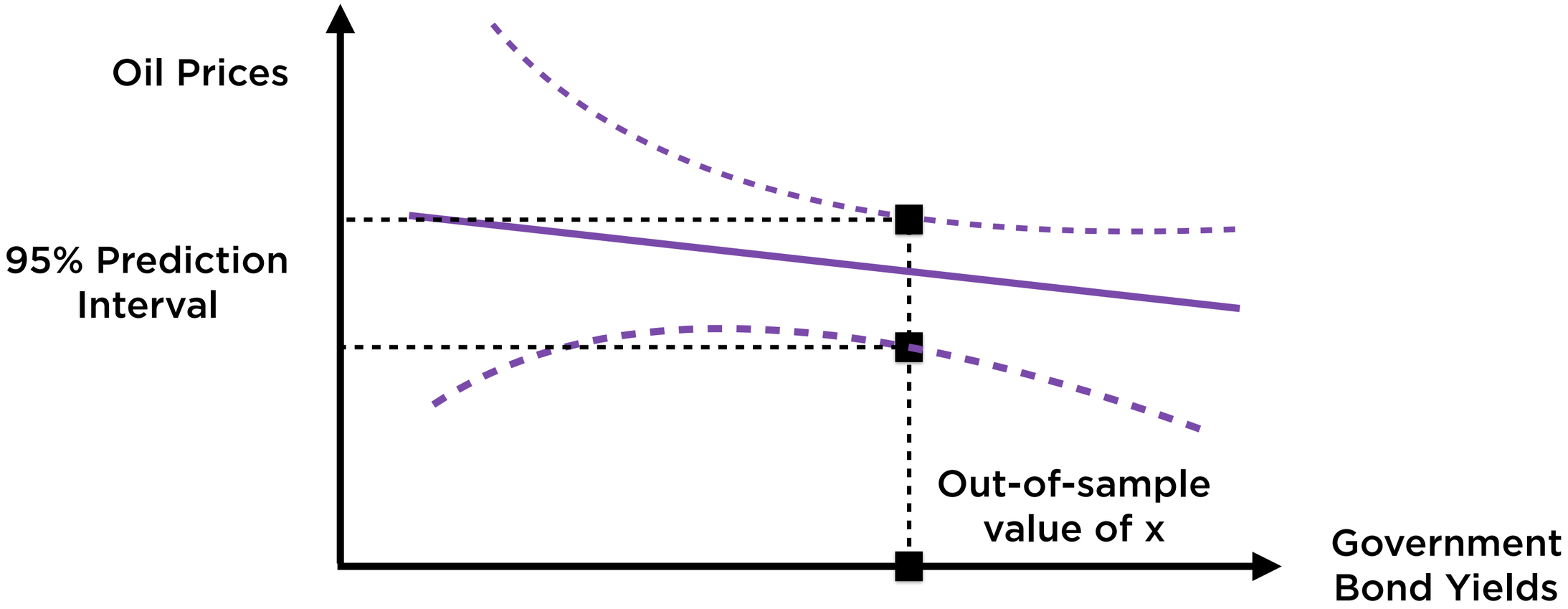
Low quality of fit

Prediction Using Regression



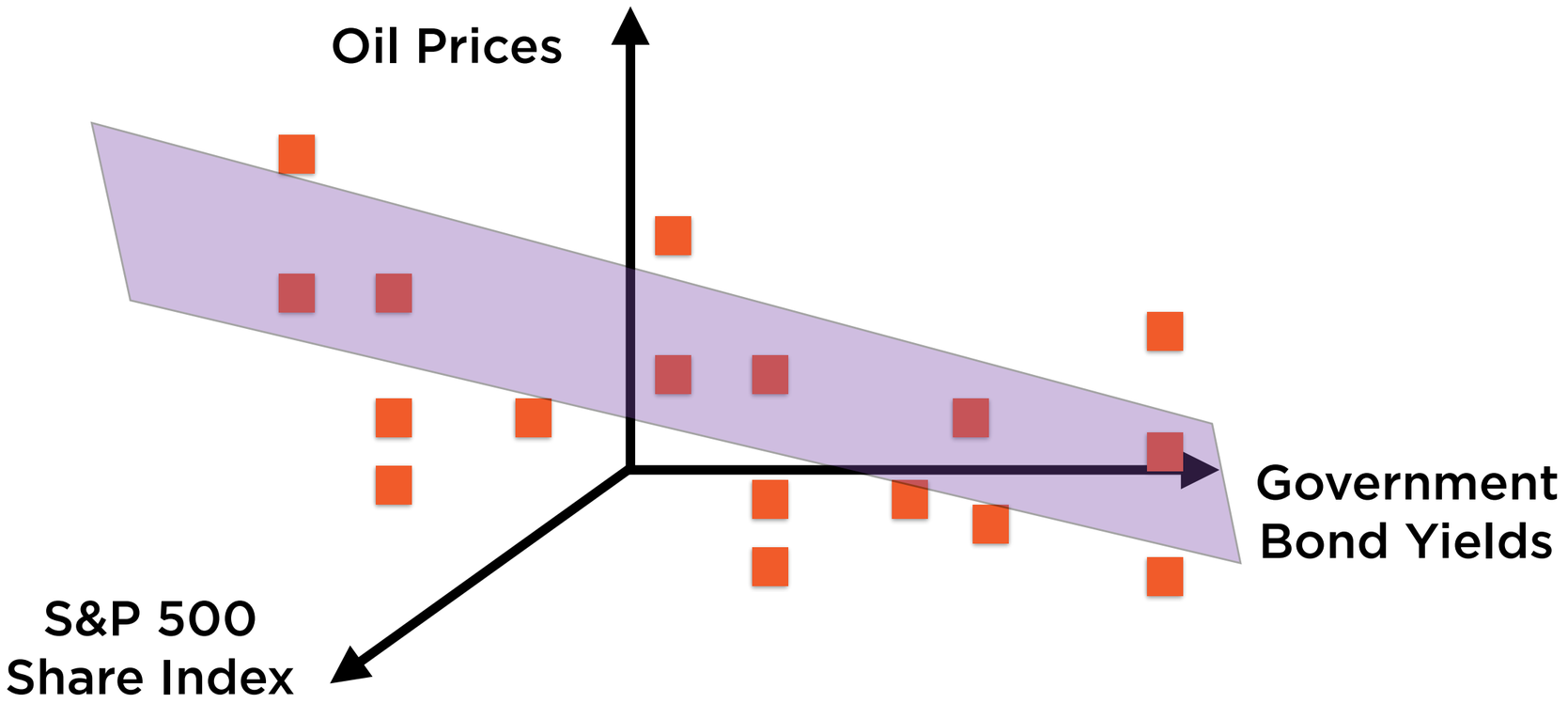
Given a new value of x, use the line to predict the corresponding value of y

Prediction Using Regression



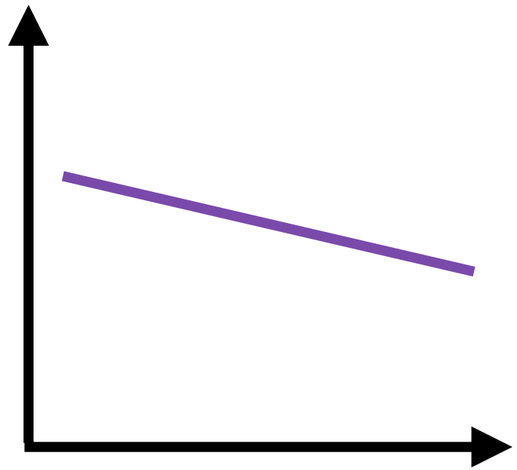
Regression also allows to specify prediction intervals (similar to confidence intervals) around this point estimate

Data in N Dimensions



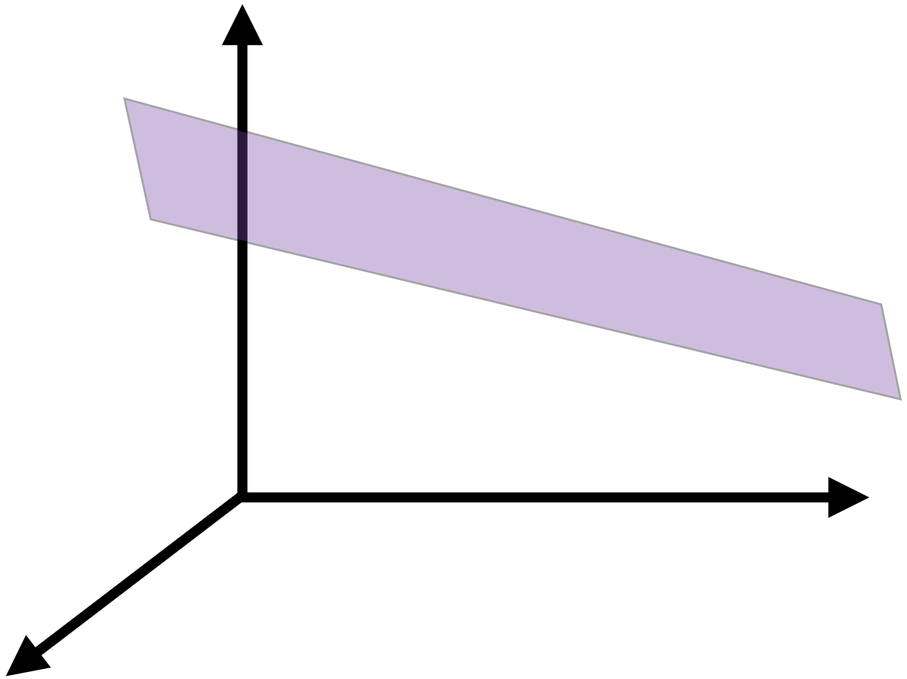
Linear Regression can easily be extended to n-dimensional data

Simple and Multiple Regression



Simple Regression

Data in 2 dimensions



Multiple Regression

Data in > 2 dimensions

Reasons for Using Regression

Regression Is a Great Tool

Powerful

Perfectly suited to two common use-cases

Versatile

Easily extended to non-linear relationships

Deep

The first “crossover hit” from Machine Learning

Regression Is a Great Tool

Powerful

Perfectly suited to two common use-cases

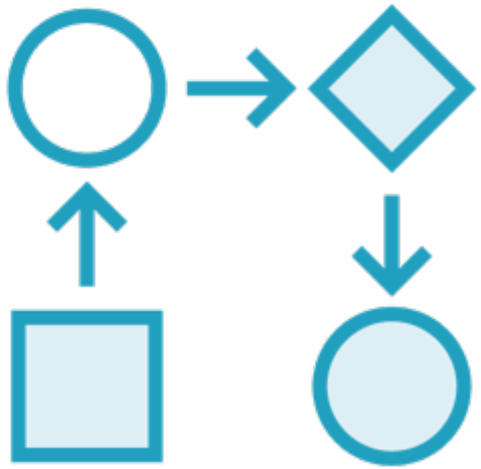
Versatile

Easily extended to non-linear relationships

Deep

The first “crossover hit” from Machine Learning

Two Common Applications of Regression



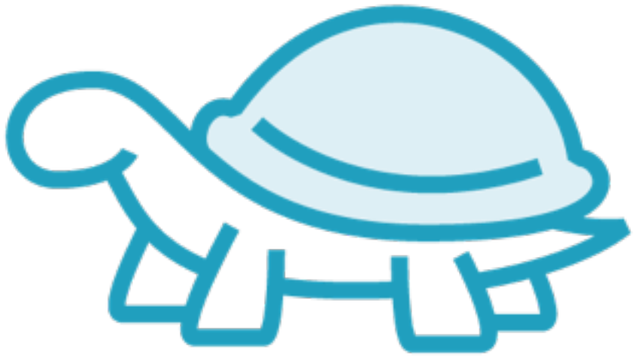
Explaining Variance

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

Making Predictions

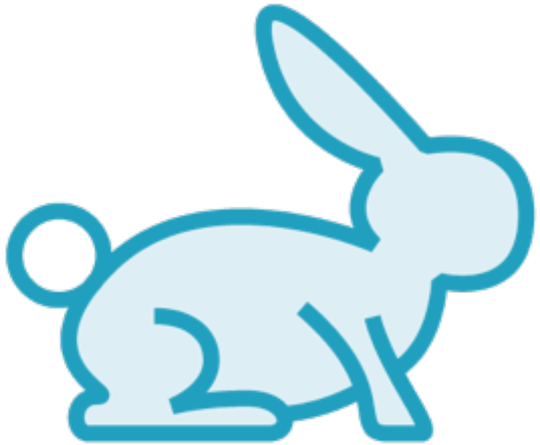
How much does a move in one series impact another?

Rising Stock: Alpha or Beta?



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

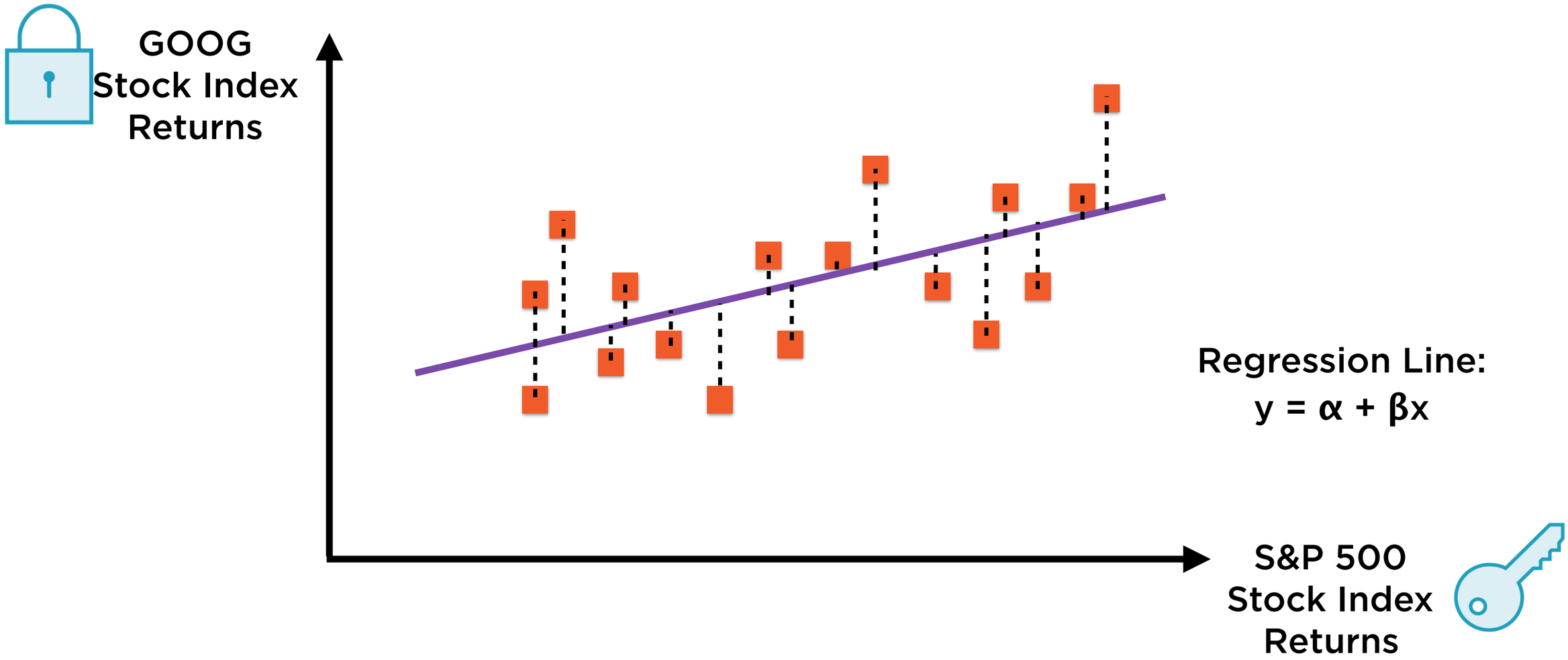
Cause

Explanatory variable: Changes in level of the market as a whole

Effect

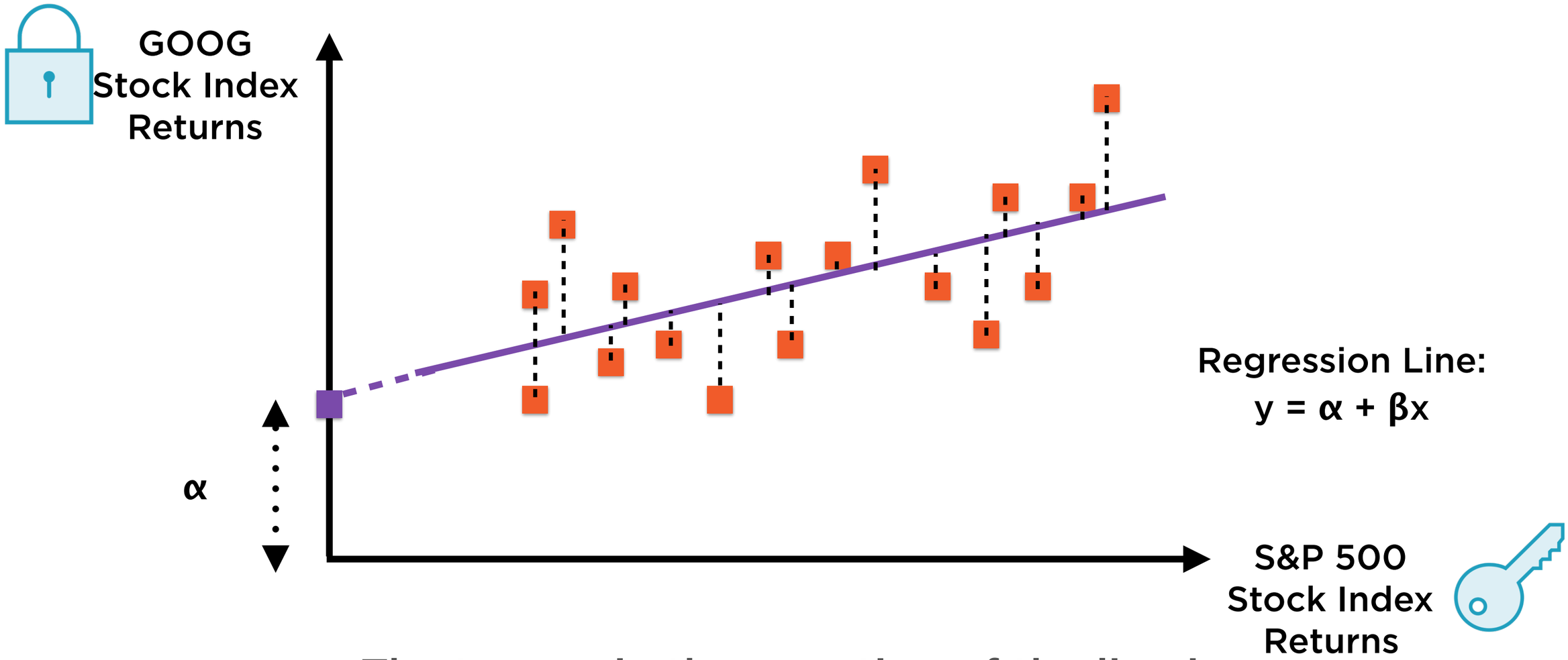
Dependent variable: Changes in the level of one particular stock

Minimising Least Square Error



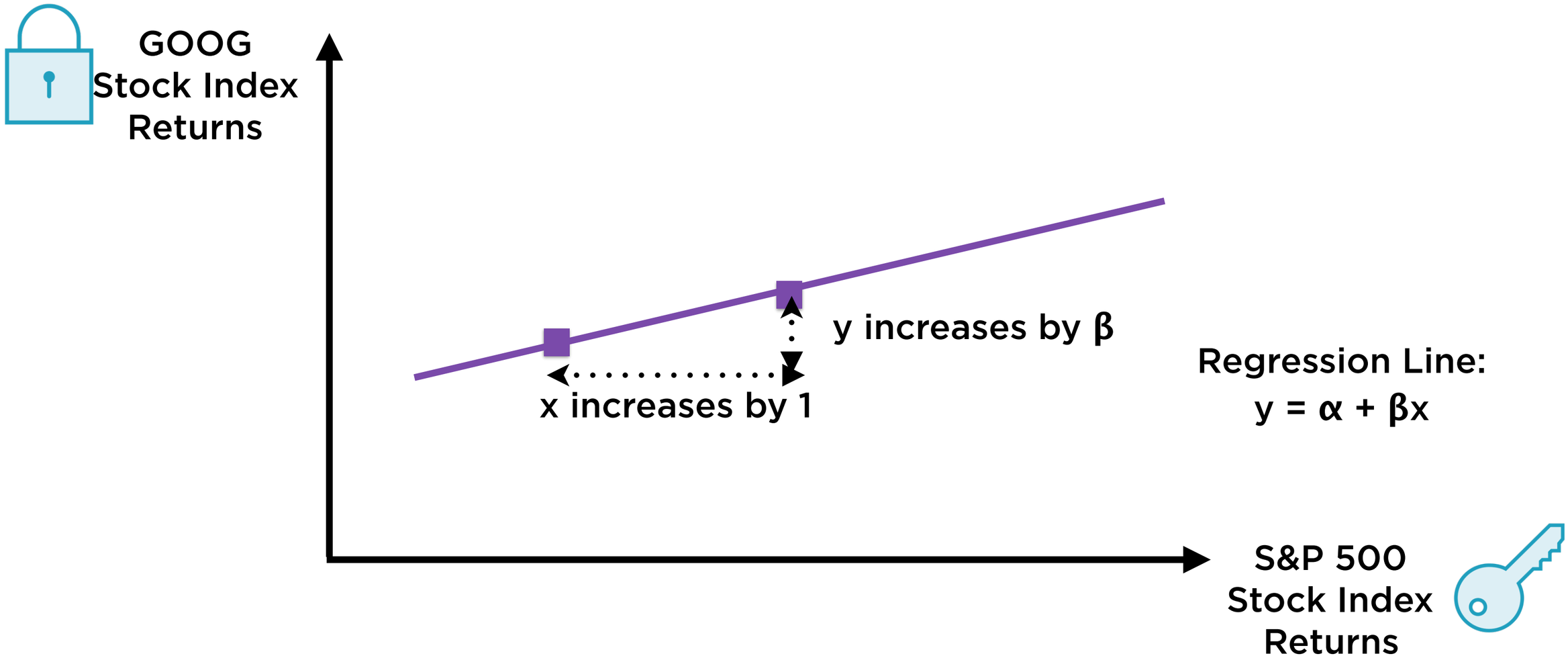
The axes are usually calculated as “excess returns” over bonds, but that’s not important here

Minimising Least Square Error



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

Minimising Least Square Error



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

Regression Models in Commodity Trading

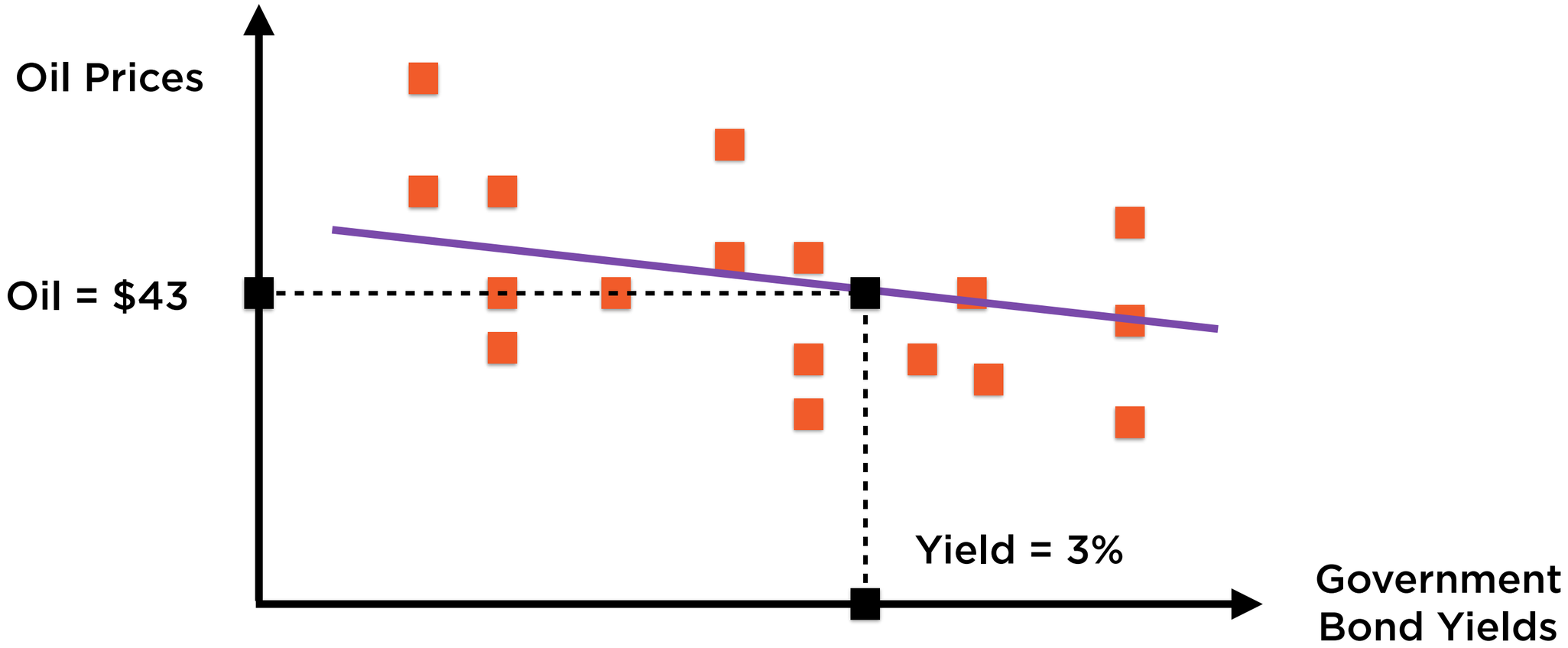
Interest Rates are Rising

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



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

Regression Is a Great Tool

Powerful

Perfectly suited to two common use-cases

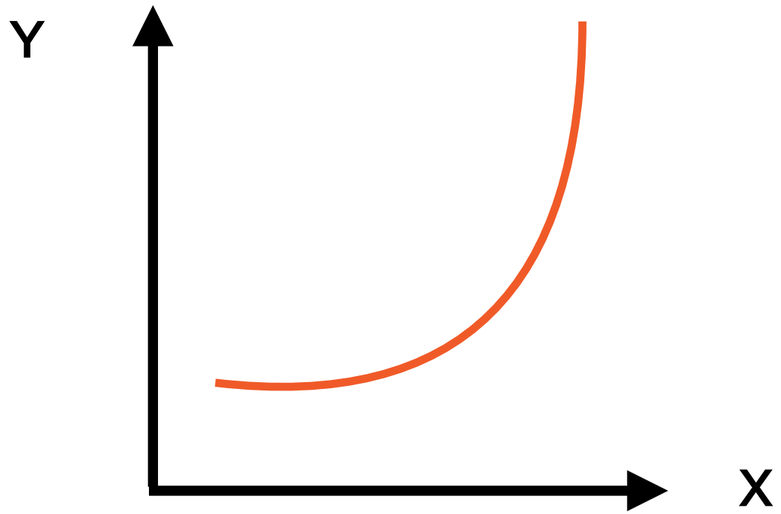
Versatile

Easily extended to non-linear relationships

Deep

The first “crossover hit” from Machine Learning

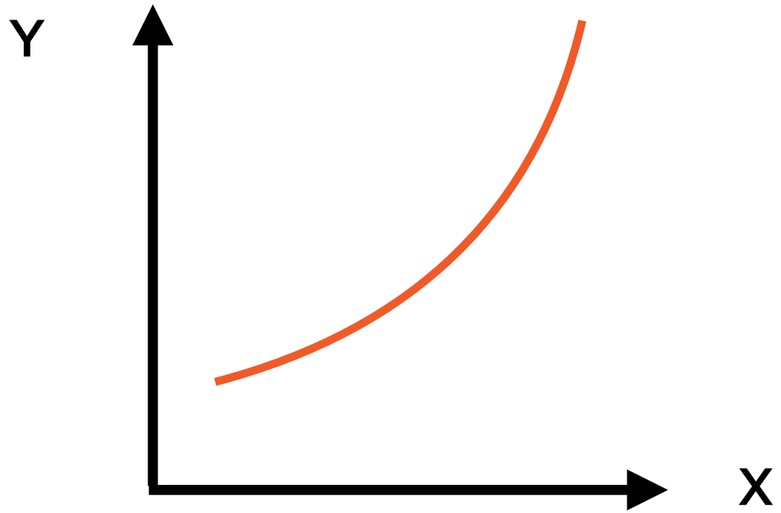
Transform Non-linear Data



Exponential

$y = A + Be^x$

Transform using logarithms

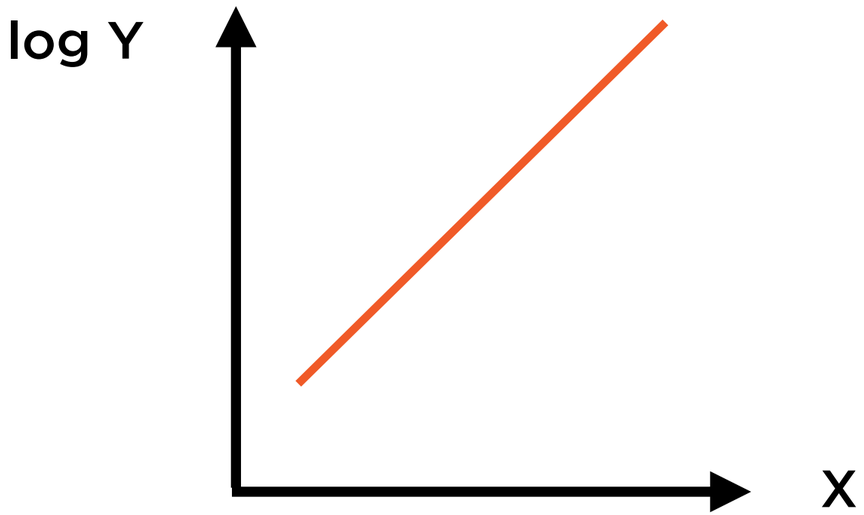


Polynomial

$y = A + Cx^2$

Transform using logarithms or simply regress on x^2

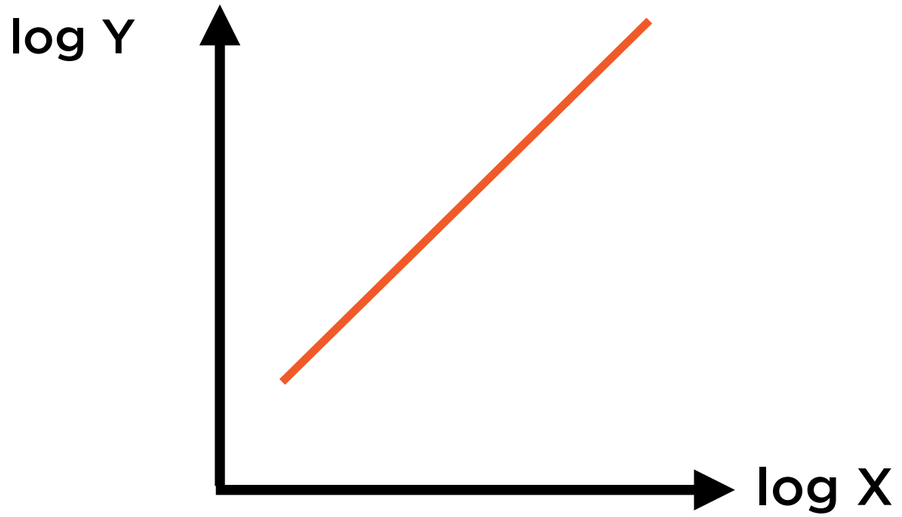
Transform Non-linear Data



Exponential

$$\log y = C + Dx$$

Now regress log y on x



Polynomial

$$\log y = C + D \log x$$

or simply regress y on x^2

Regression Is a Great Tool

Powerful

Perfectly suited to two common use-cases

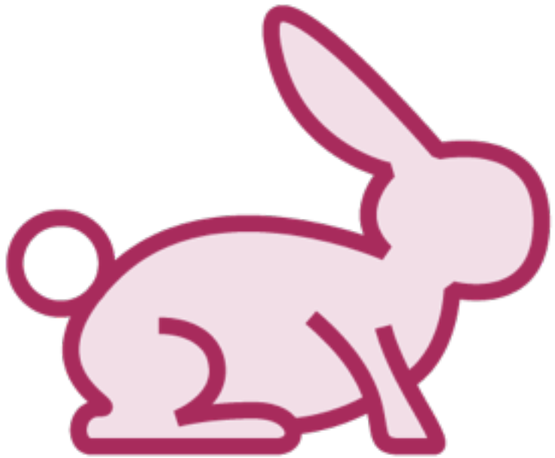
Versatile

Easily extended to non-linear relationships

Deep

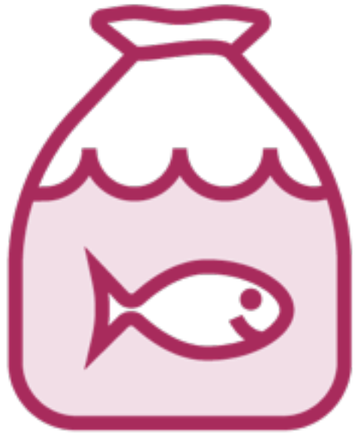
The first “crossover hit” from Machine Learning

Whales: Fish or Mammals?



Mammals

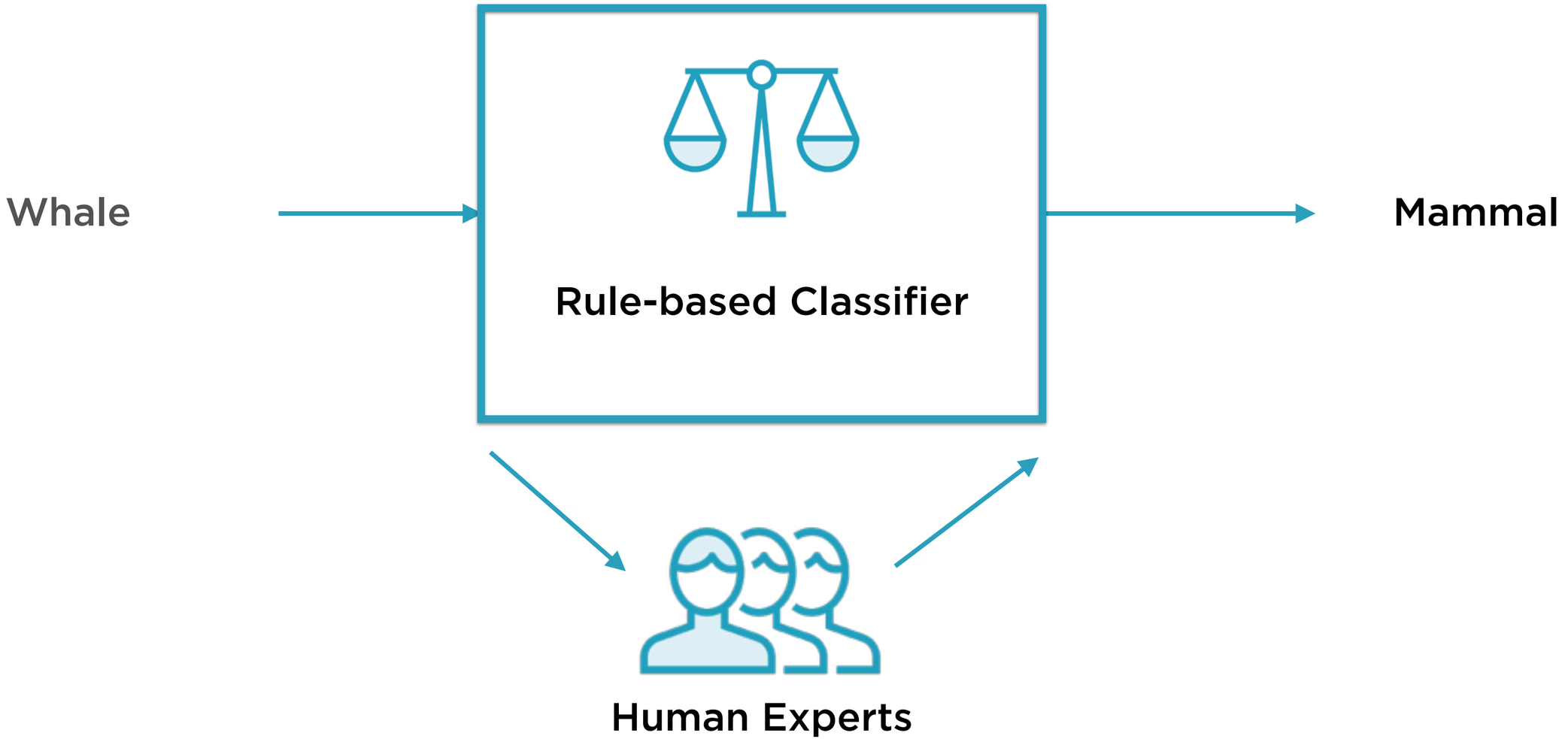
Members of the infraorder *Cetacea*



Fish

Look like fish, swim like fish, move like fish

Rule-based Binary Classifier



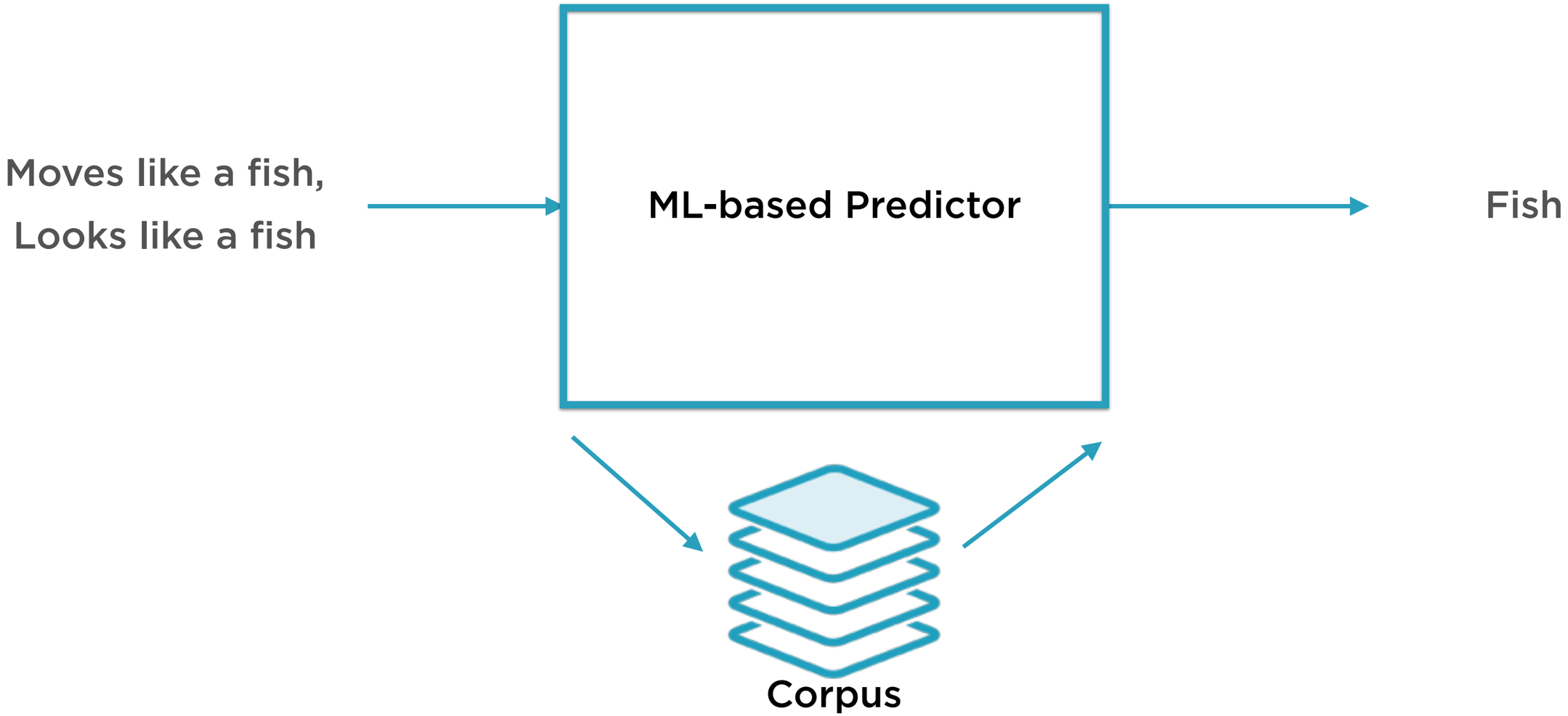
ML-based Binary Classifier

Corpus

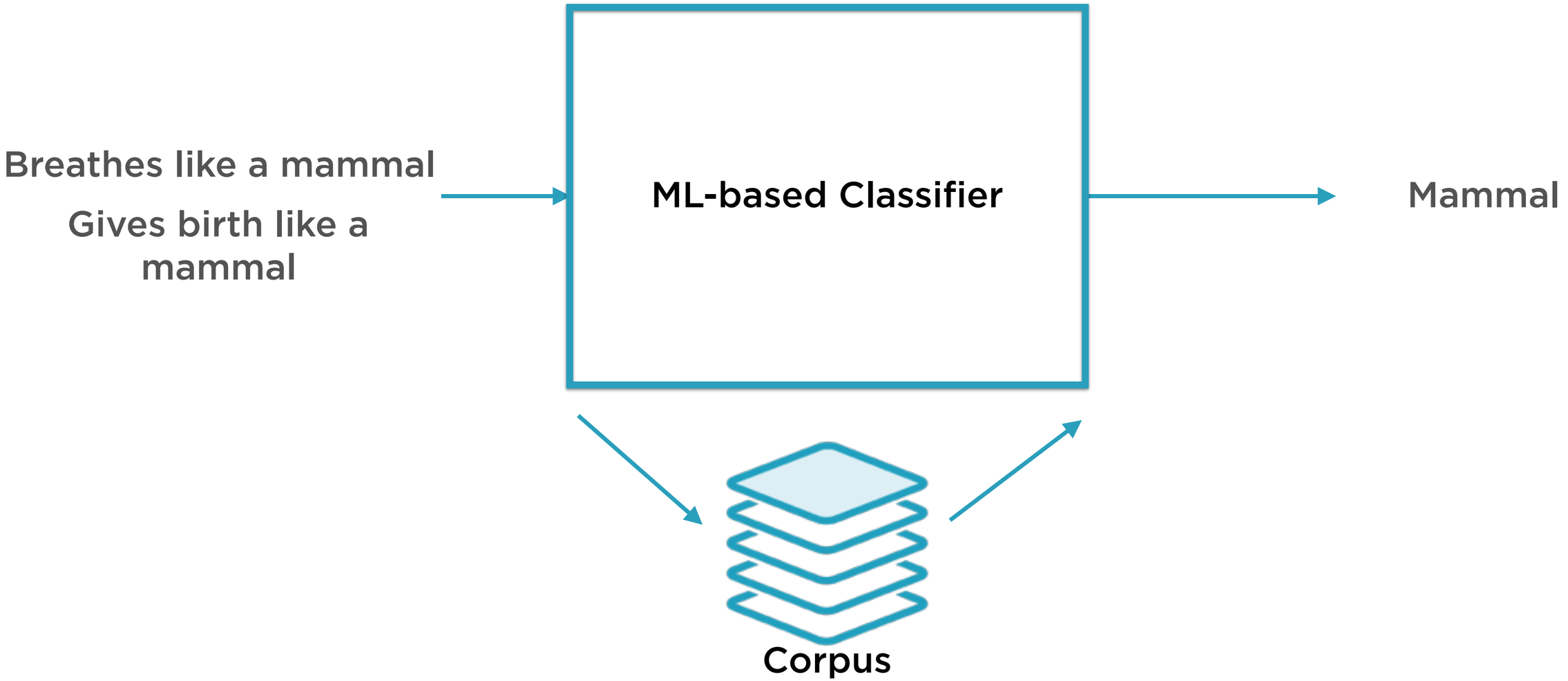
Classification Algorithm

ML-based Classifier

ML-based Binary Classifier



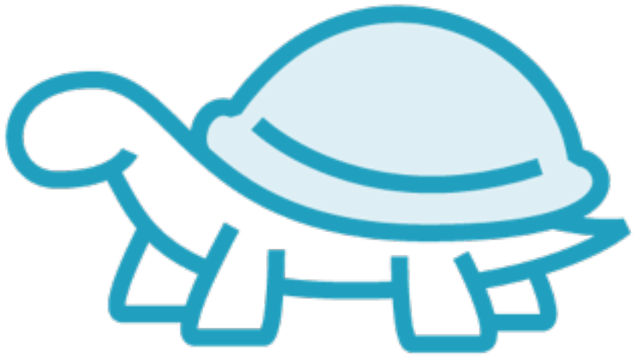
ML-based Binary Classifier



Rule-based or ML-based?

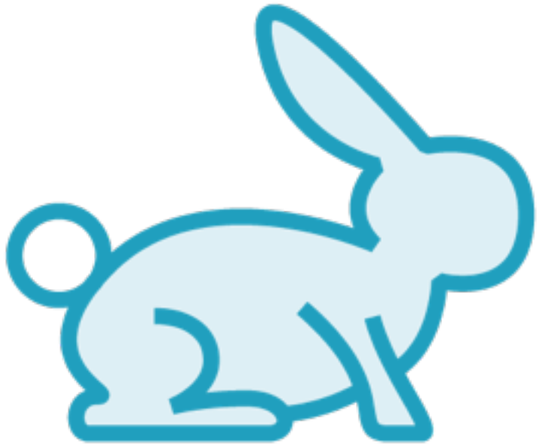
ML-based	Rule-based
Dynamic	Static
Experts optional	Experts required
Corpus required	Corpus optional
Training step	No training step

GOOG: Buy or Sell?



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

ML-based Predictor

Corpus

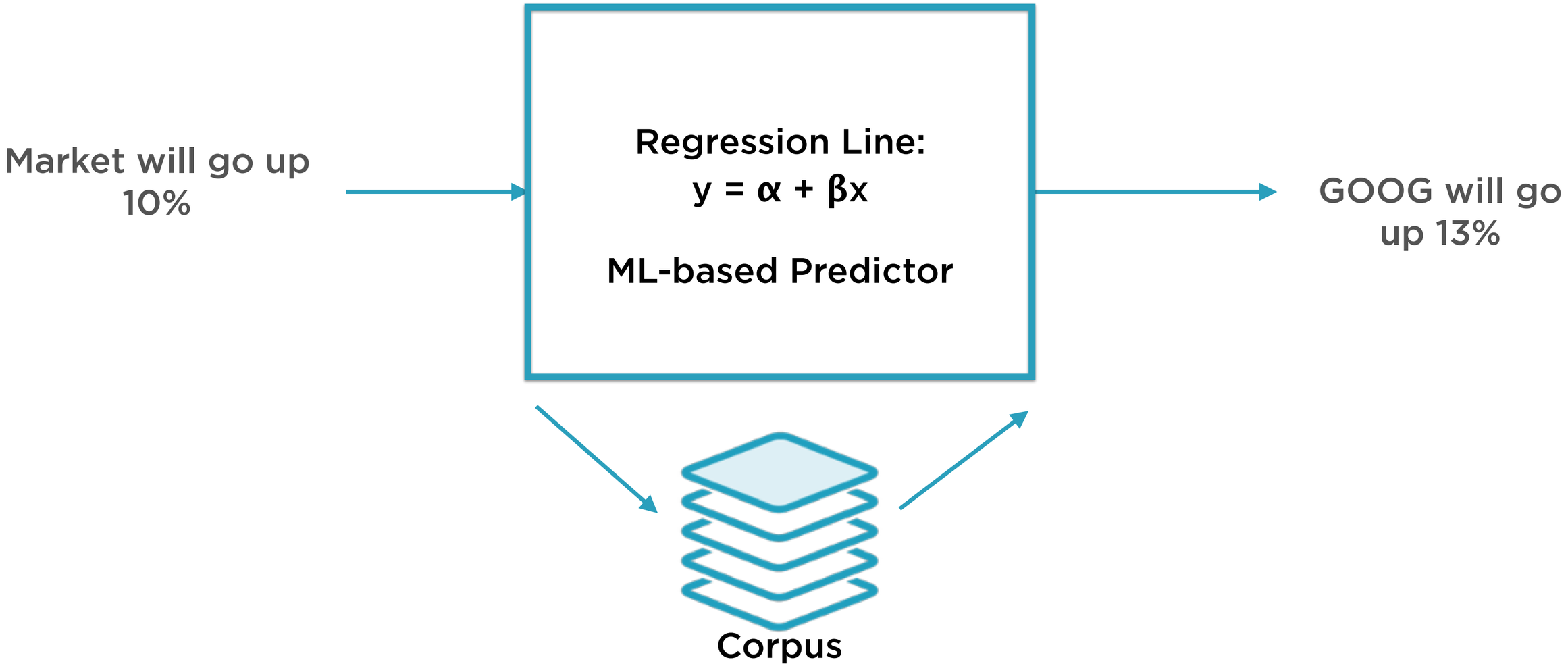
Regression Algorithm

ML-based Predictor

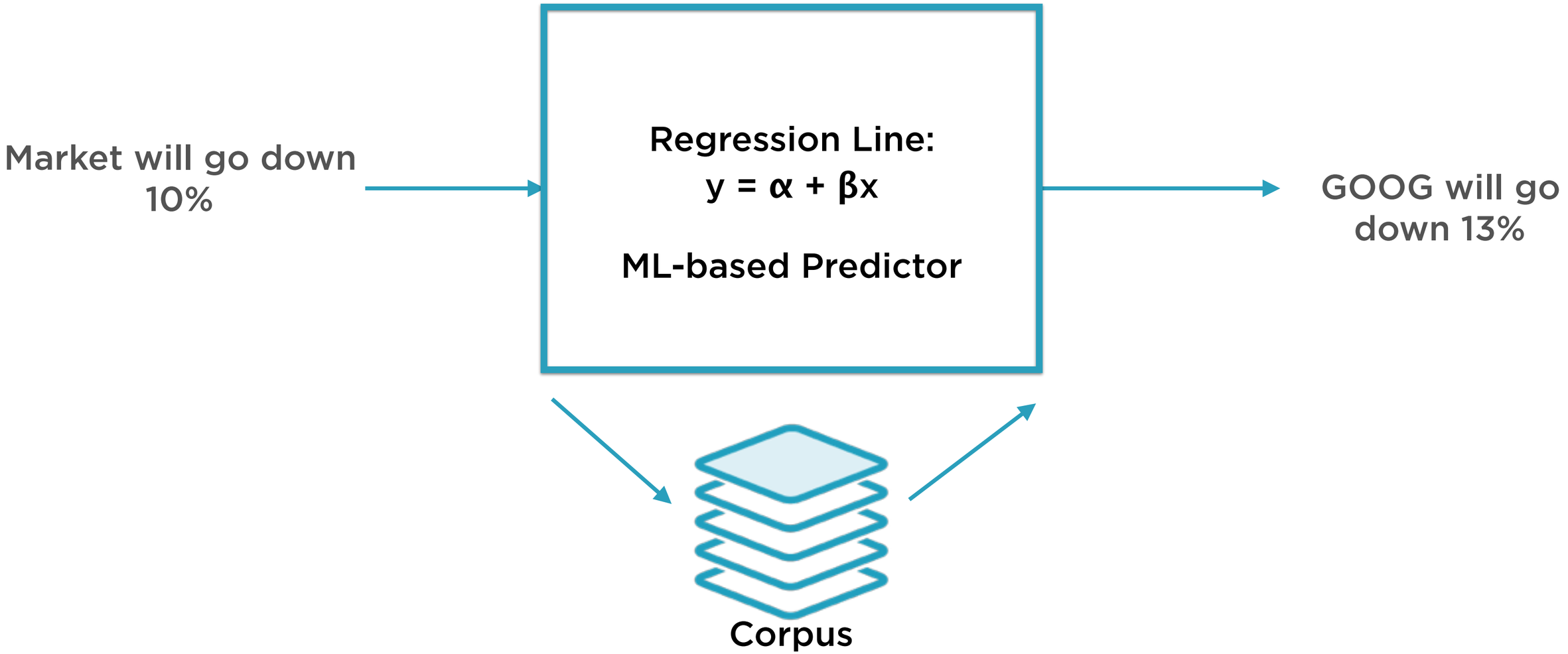
Regression Line:

$$y = \alpha + \beta x$$

ML-based Predictor



ML-based Predictor



Mean and Variance

Data in One Dimension

Pop quiz: Your thoughtful, fact-based point-of-view on these numbers, please

Data in One Dimension

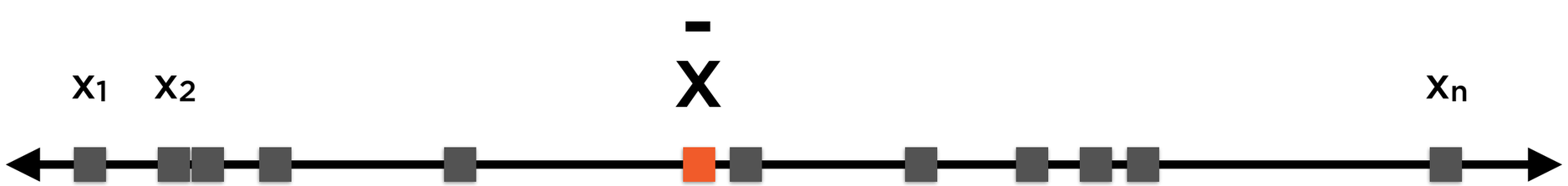
Boss

5-second attention span

Go-to Colleague

10-second attention span

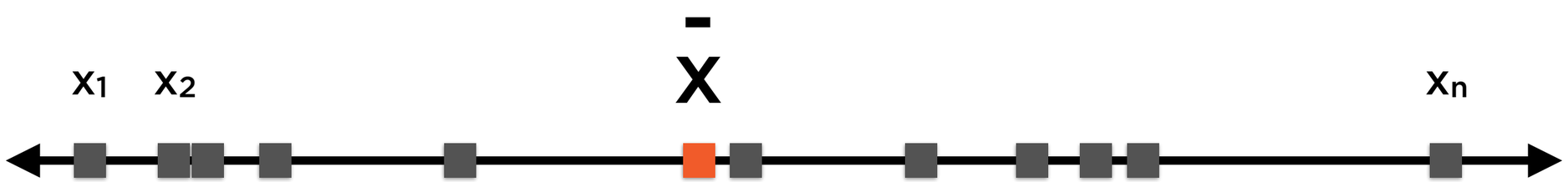
Mean as Headline



The mean, or average, is the one number that best represents all of these data points

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Variation Is Important Too

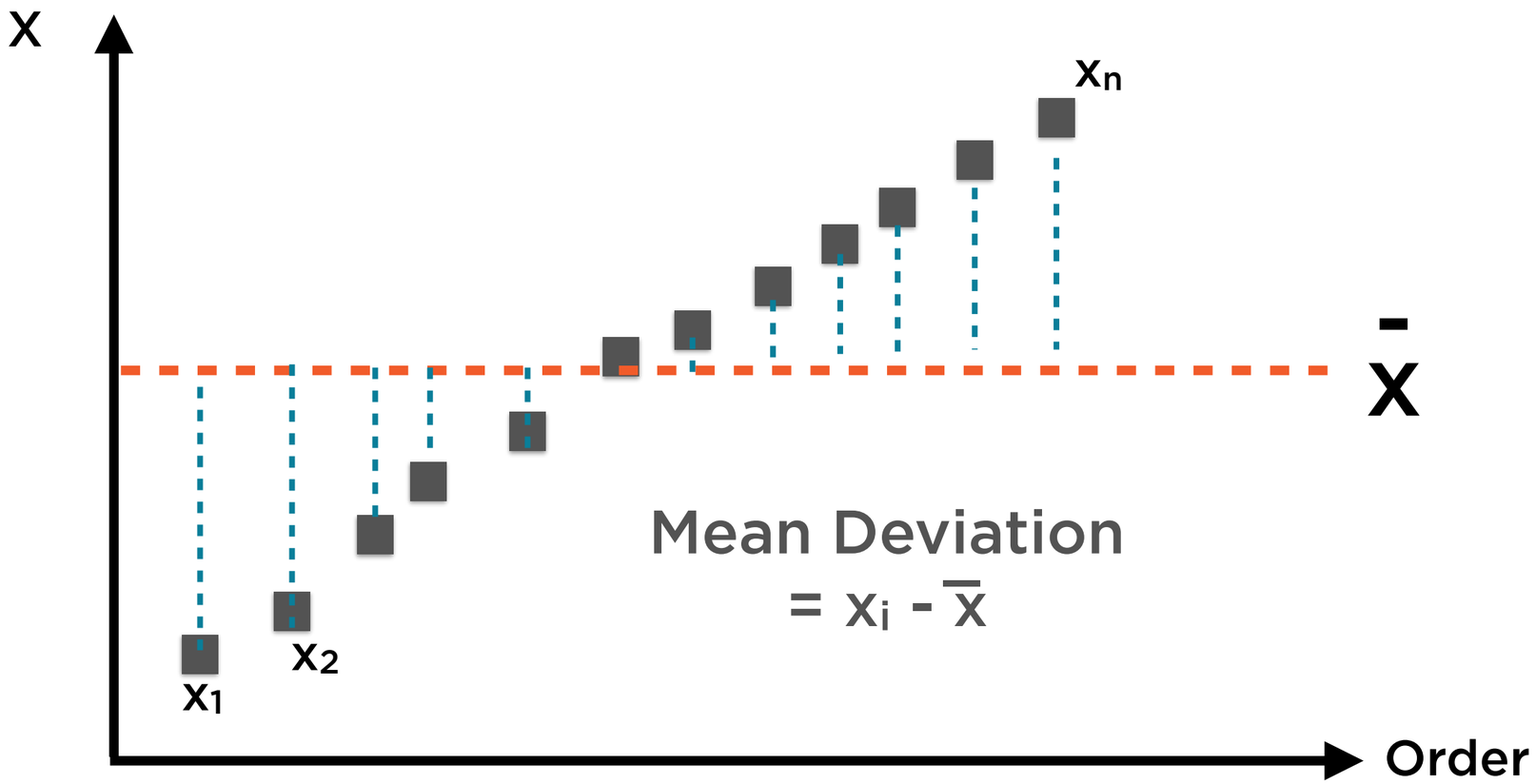


"Do the numbers jump around?"

$$\text{Range} = X_{max} - X_{min}$$

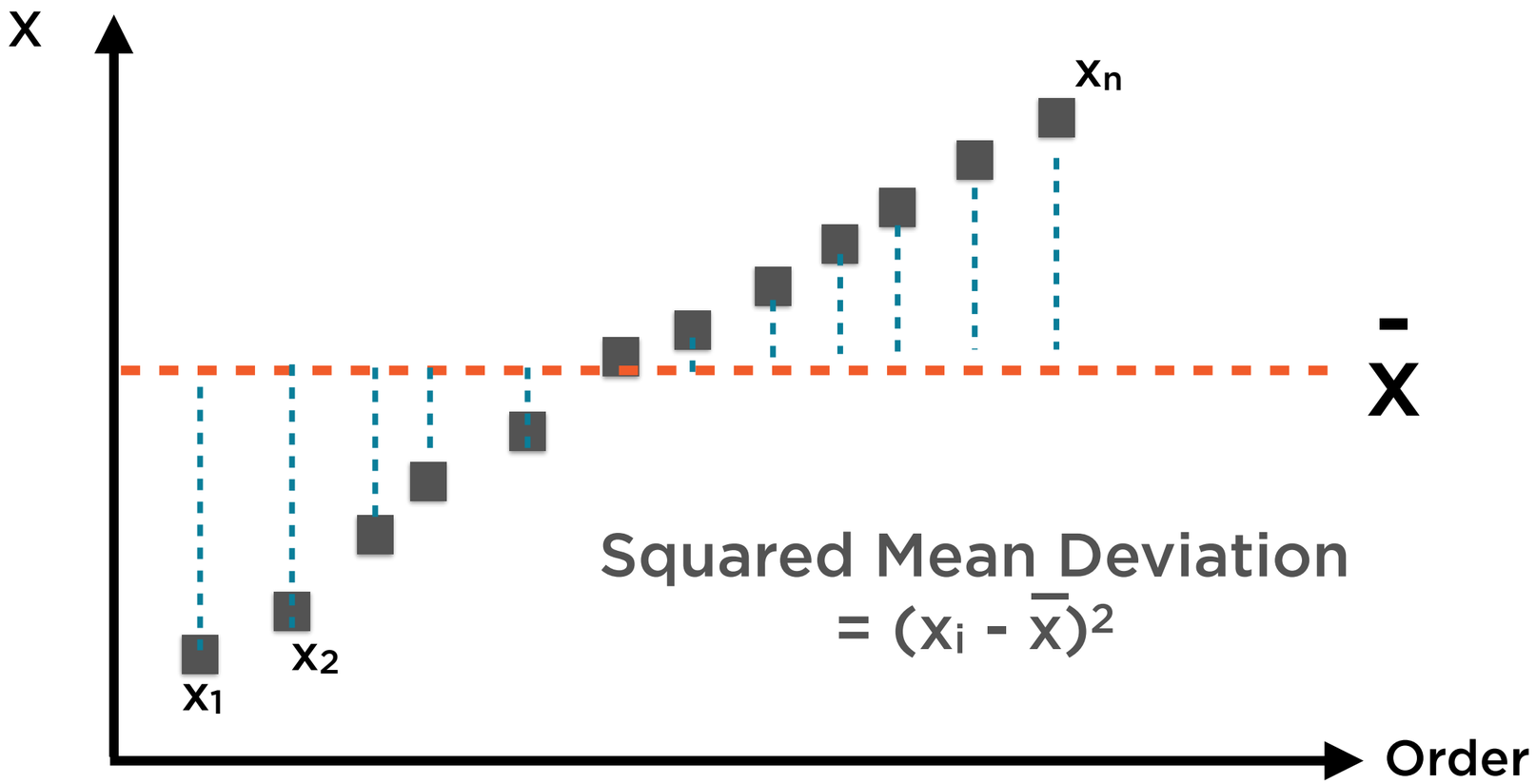
The range ignores the mean, and is swayed by outliers - that's where variance comes in

Variance as Asterisk



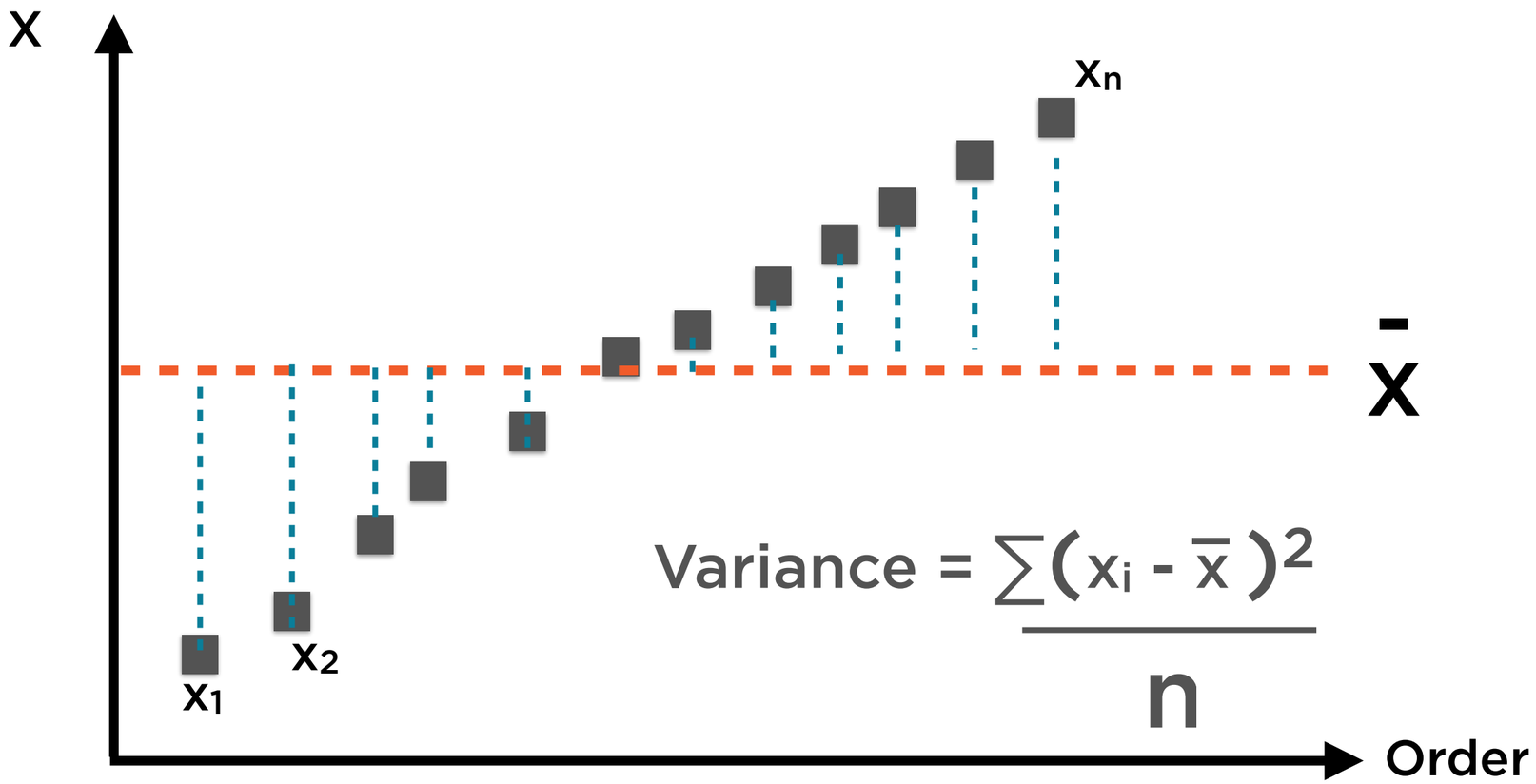
Variance is the second-most important number to summarise this set of data points

Variance as Asterisk



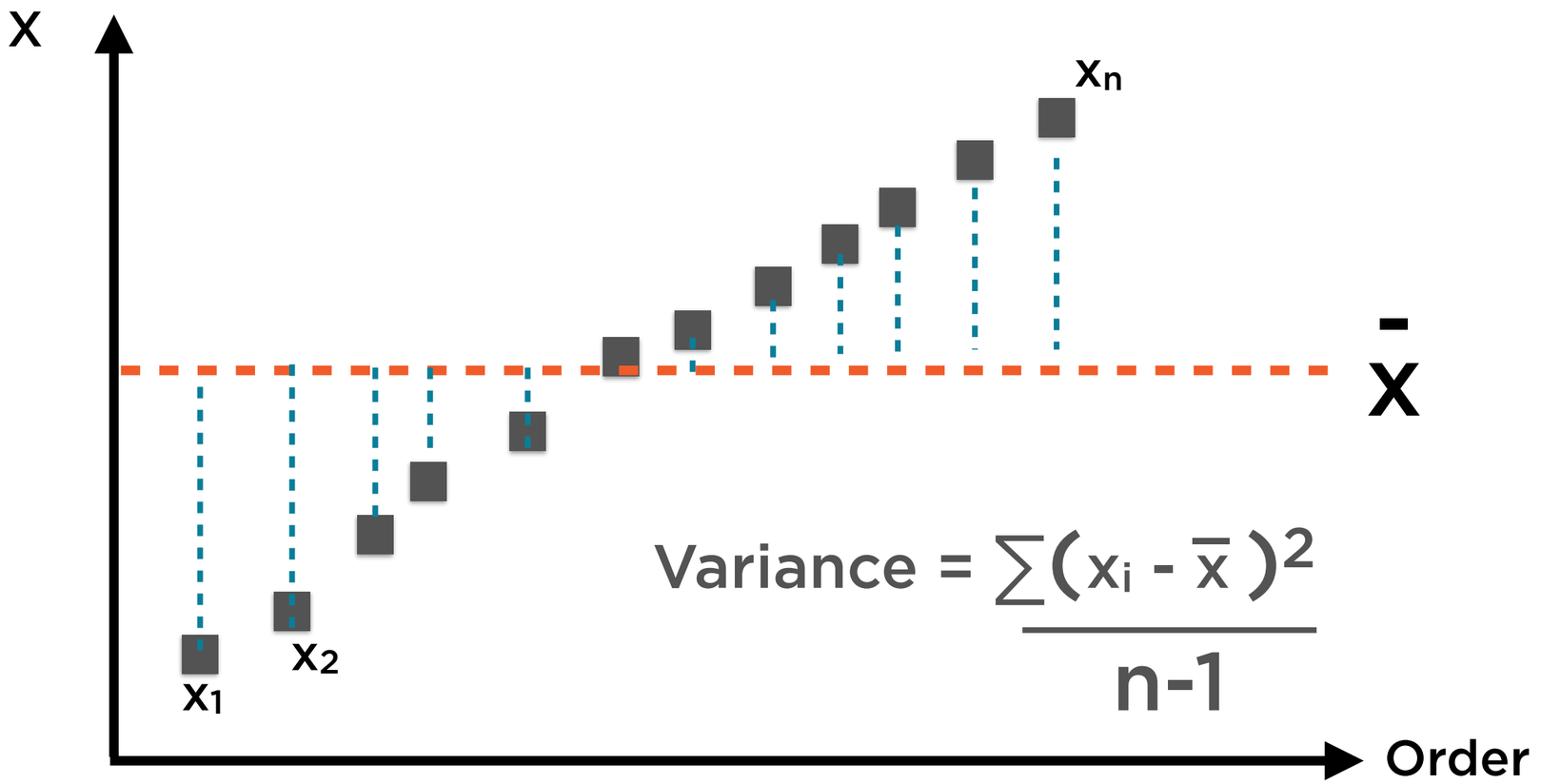
Variance is the second-most important number to summarise this set of data points

Variance as Asterisk



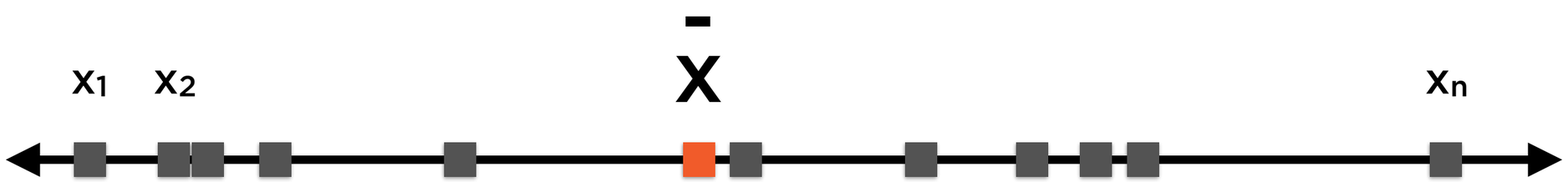
Variance is the second-most important number to summarise this set of data points

Variance as Asterisk



We can improve our estimate of the variance by tweaking the denominator - this is called Bessel's Correction

Mean and Variance

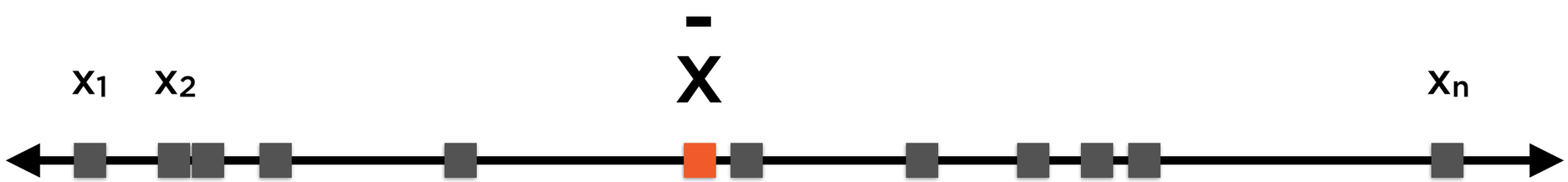


Mean and variance succinctly summarise a set of numbers

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Variance and Standard Deviation



Standard deviation is the square root of variance

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{Std Dev} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Mean and Variance

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

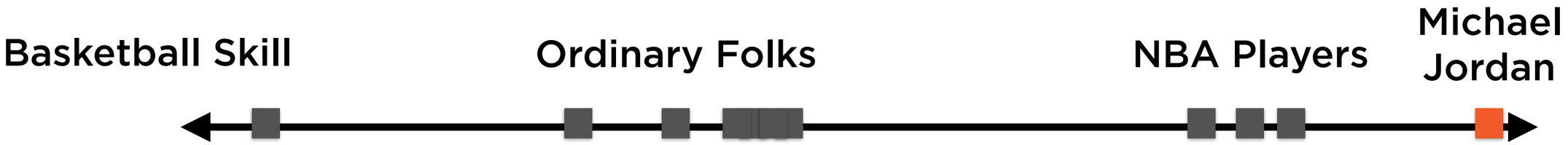
These statistics only apply to the sample of data, and so are known as sample statistics

The corresponding figures for all possible data points out there are called population statistics

Probability Distributions and the Bell Curve

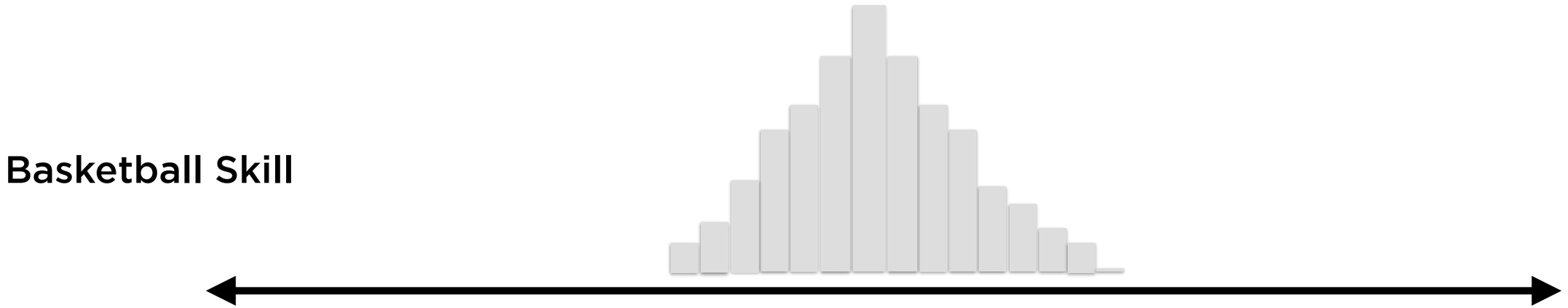
“Michael Jordan is a once-in-a-lifetime player”

Outliers



A once-in-a-lifetime player is an outlier, a point far from the pack

Outliers

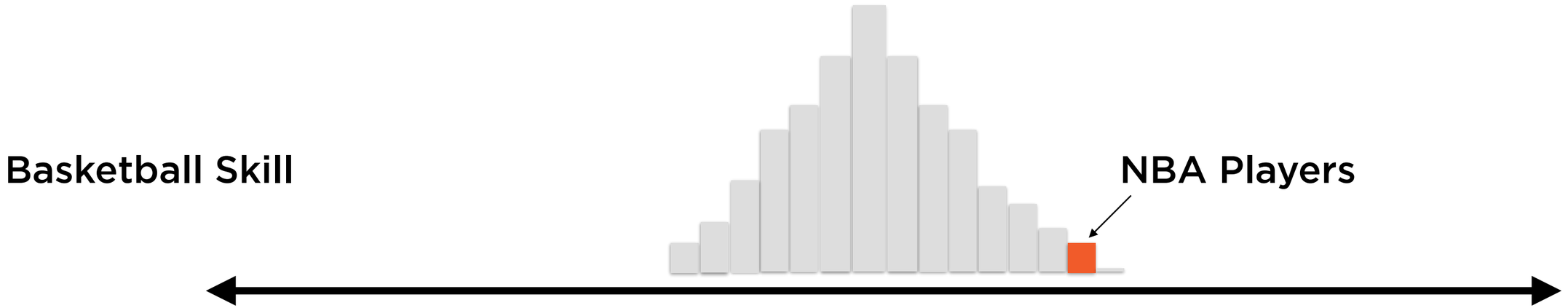


In reality, most ordinary folks would be clustered around an average level of skill

The NBA players would be outliers

Michael Jordan would be an even greater outlier

Outliers

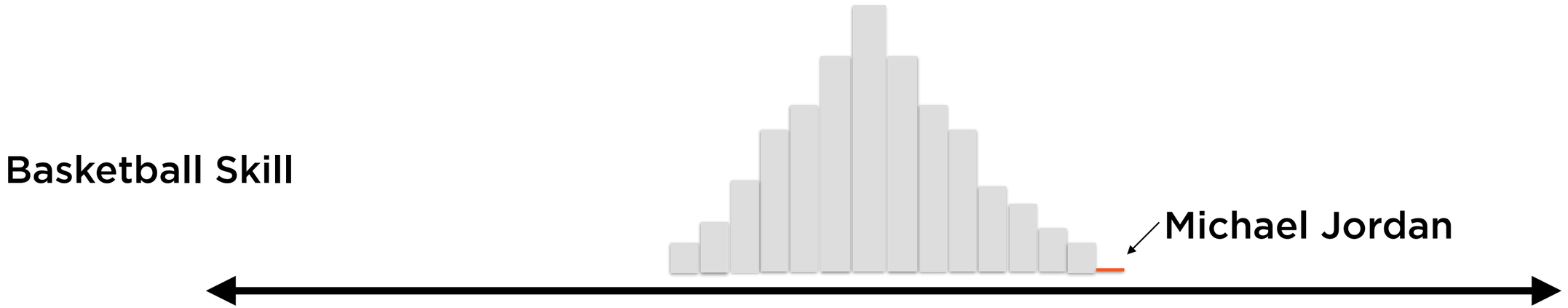


In reality, most ordinary folks would be clustered around an average level of skill

The NBA players would be outliers

Michael Jordan would be an even greater outlier

Outliers

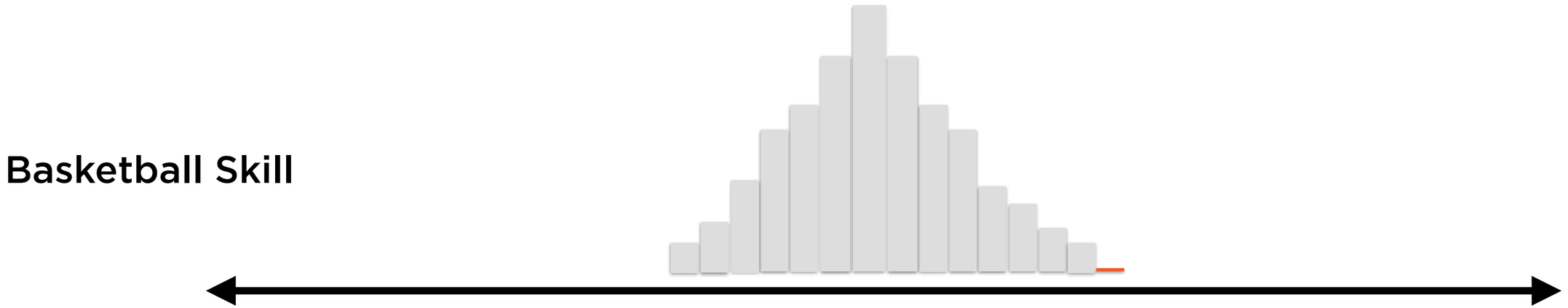


In reality, most ordinary folks would be clustered around an average level of skill

The NBA players would be outliers

Michael Jordan would be an even greater outlier

Outliers

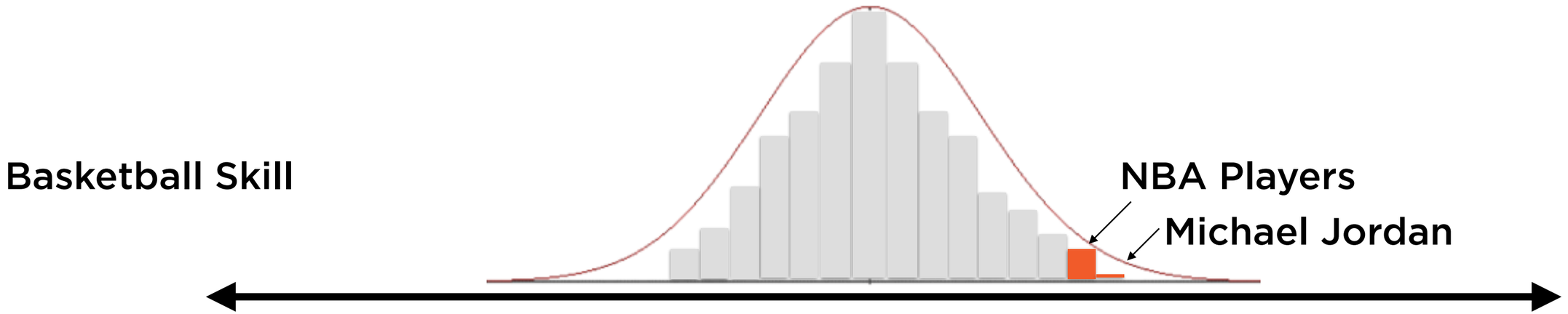


This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

Outliers

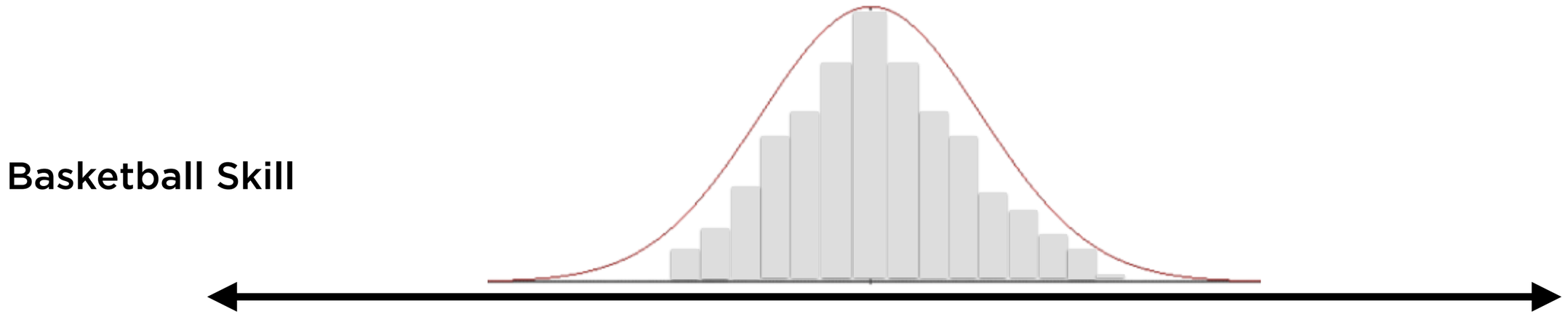


This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

Outliers



This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

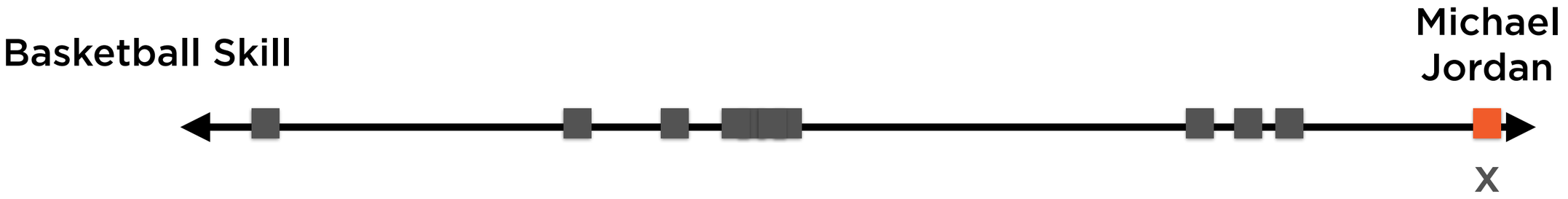
Outliers

Average is common

Very high and very low are both unusual

The bell curve occurs everywhere in nature

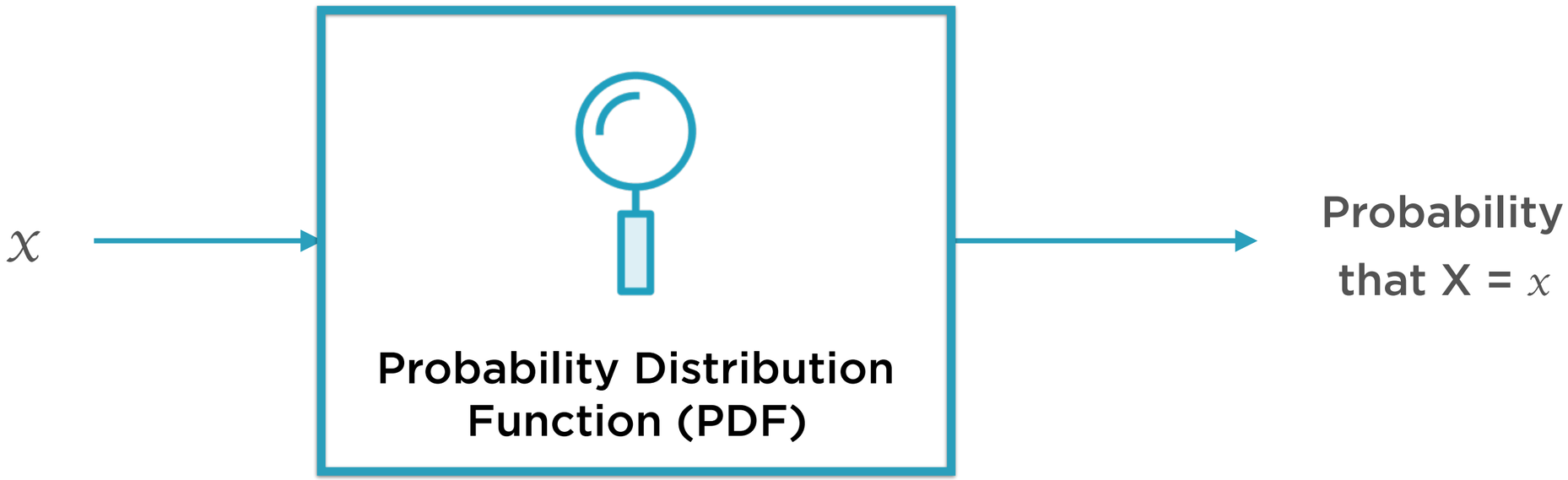
Outliers



What is the probability of any specific value x occurring in the data?

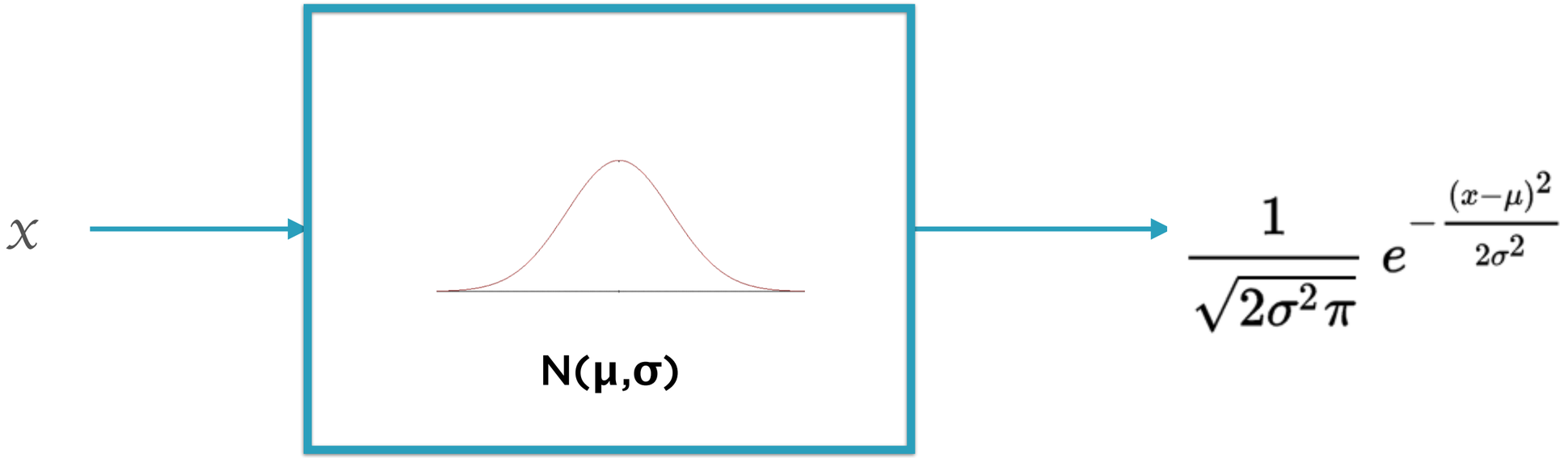
The answer lies in a probability distribution function

Probability Distribution Function



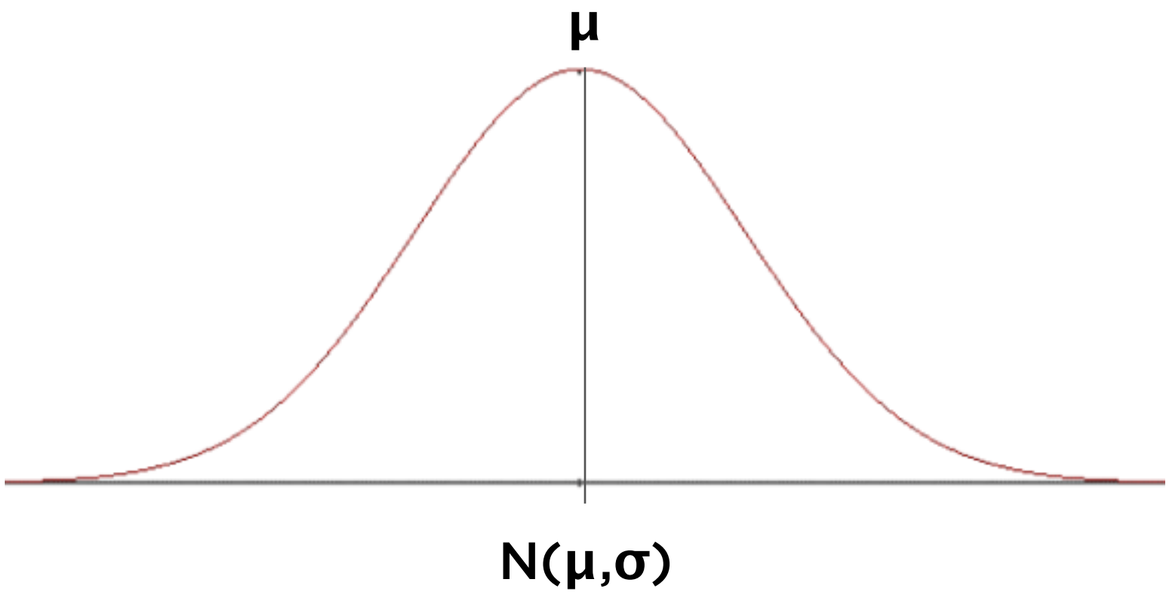
Given any value x, how likely is that value to be found in the data?

Probability Distribution Function



A Normal Distribution is a probability distribution that occurs ubiquitously in nature

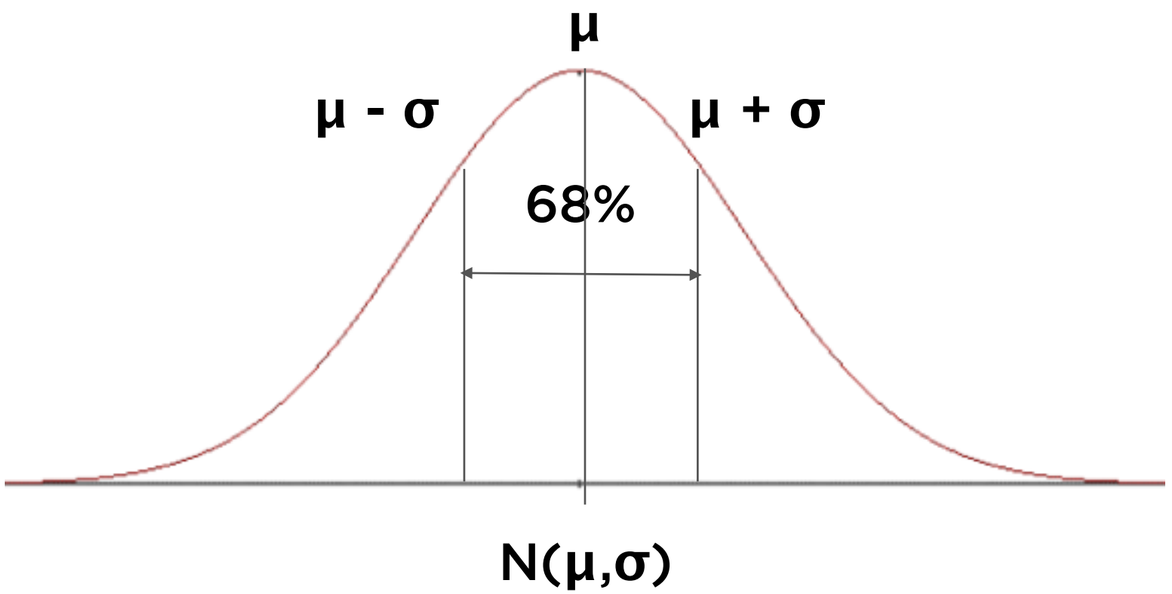
Normal Distribution



Average (mean) is μ

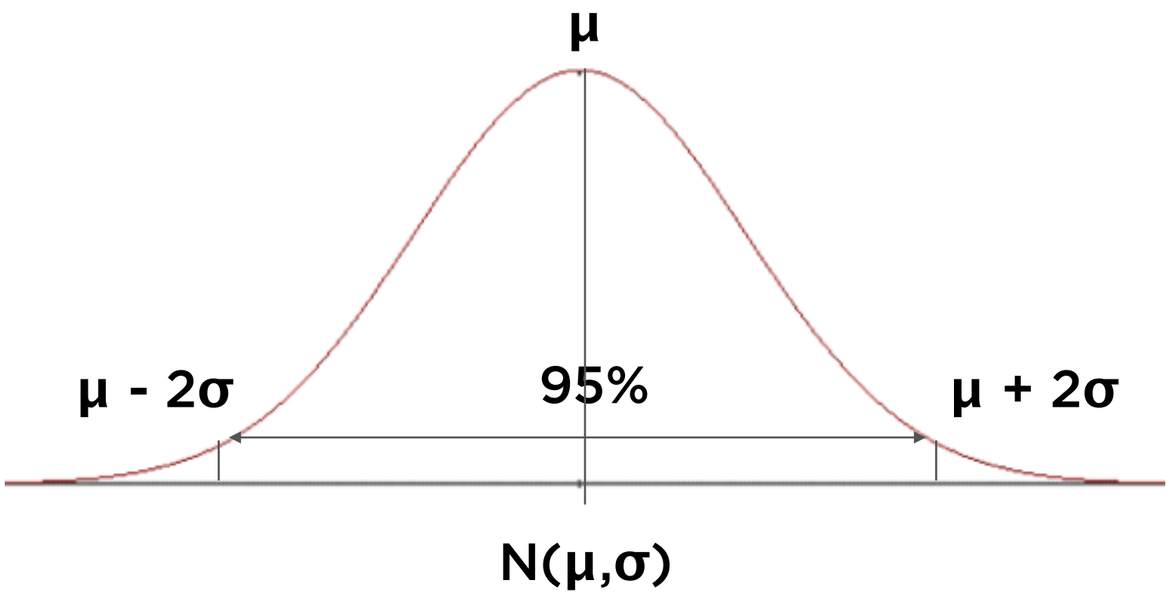
Standard deviation is σ

Normal Distribution



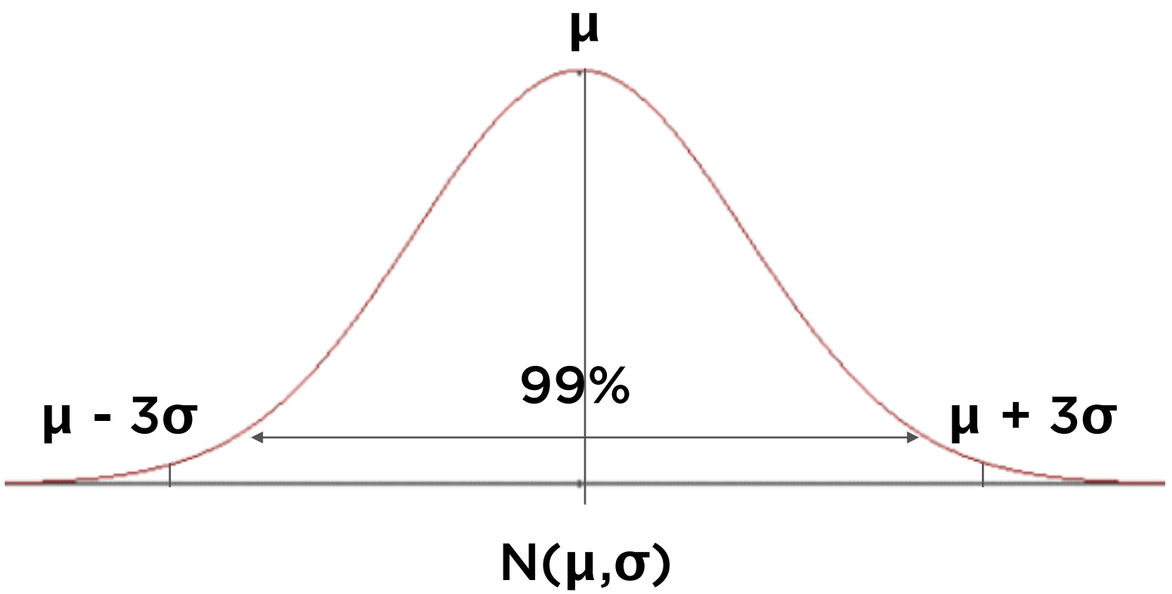
68% within 1 standard deviation of mean

Normal Distribution



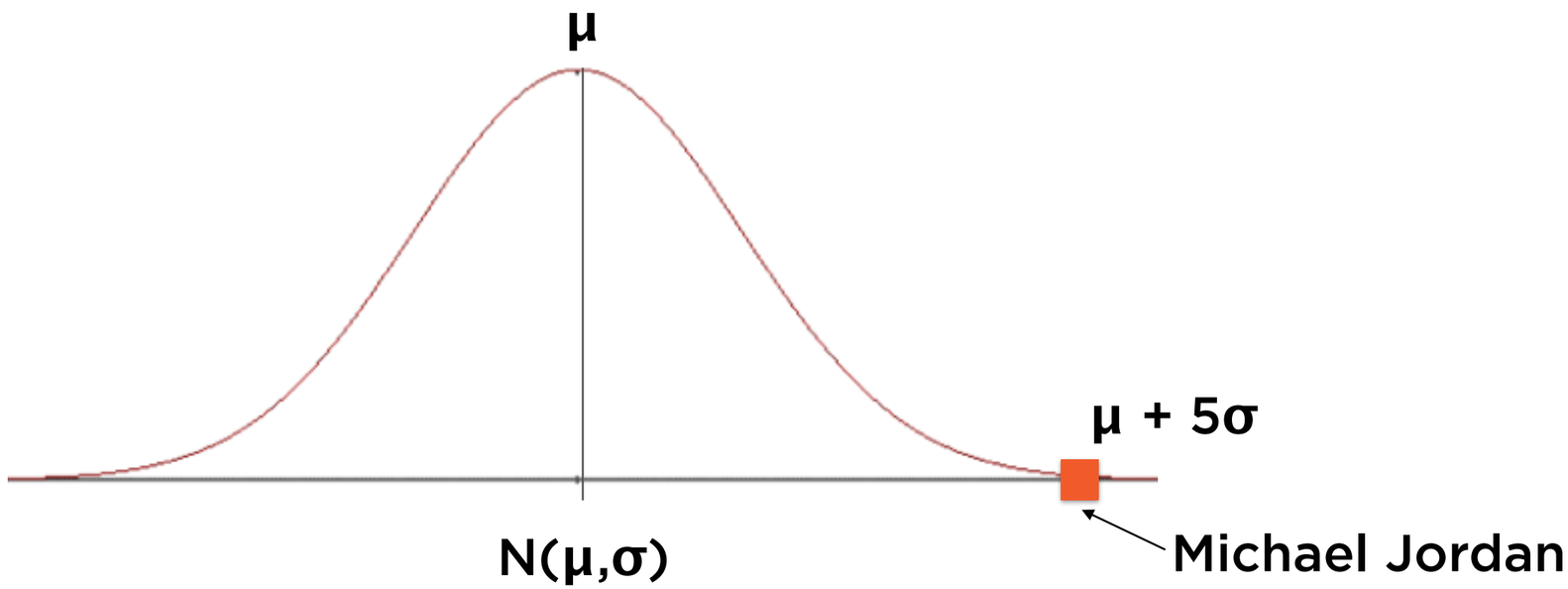
95% within 2 standard deviations of mean

Normal Distribution



99% within 3 standard deviations of mean

Normal Distribution



“Michael Jordan is a once-in-a-lifetime player”

Connecting the Dots with Regression

Regression Equation:

$$y = A + Bx$$

$$\begin{aligned} y_1 &= A + Bx_1 \\ y_2 &= A + Bx_2 \\ y_3 &= A + Bx_3 \\ \dots &\quad \dots \\ y_n &= A + Bx_n \end{aligned}$$

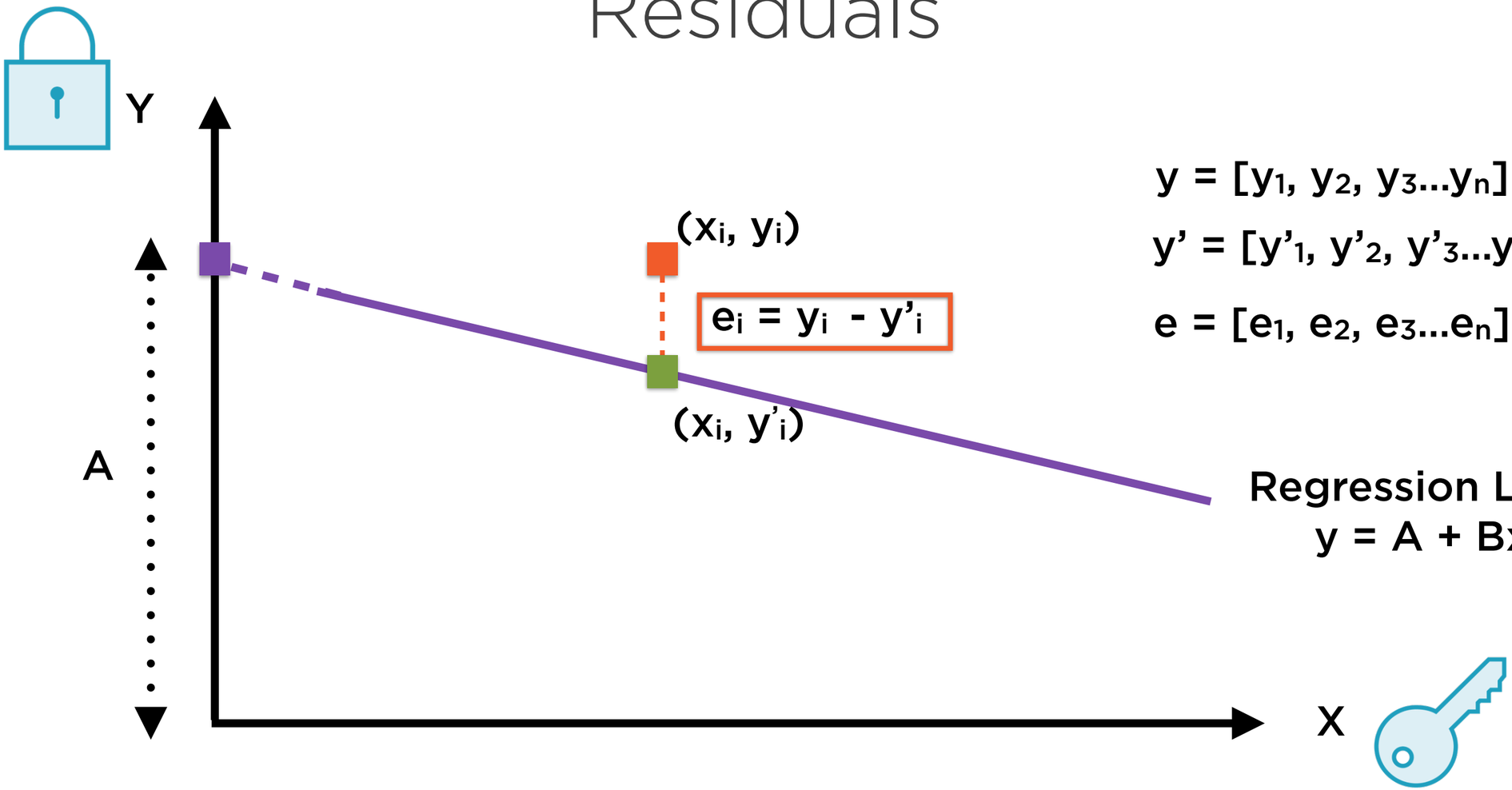
Connecting the Dots with Regression

Regression Equation:

$$y = A + Bx$$

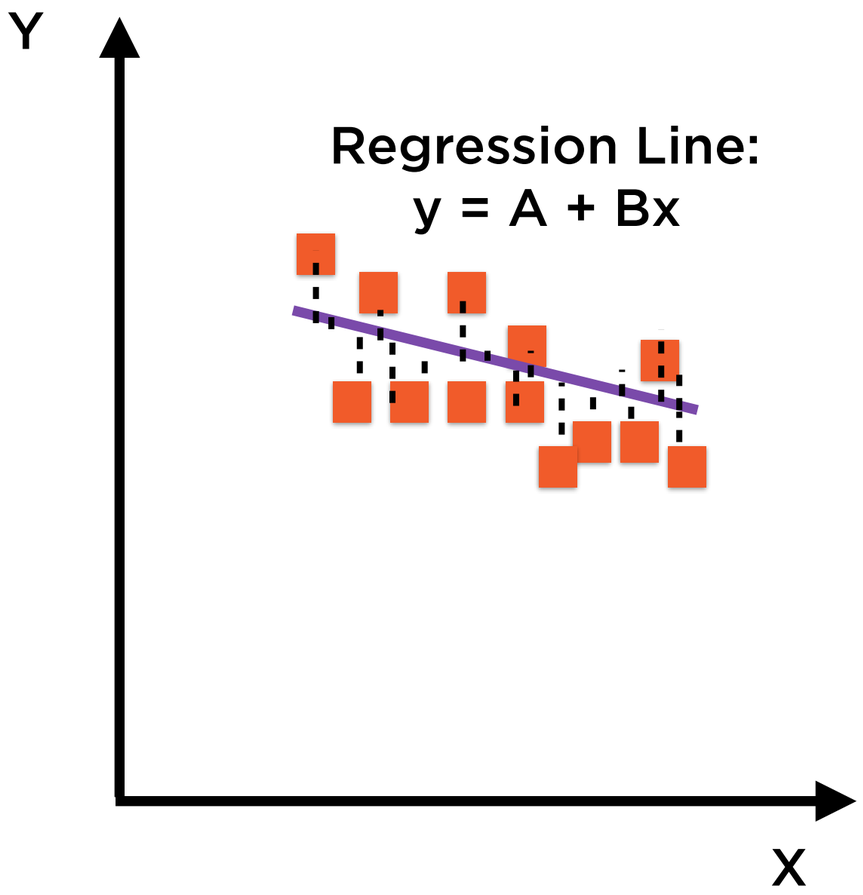
$$
\begin{aligned}
y_1 &= A + Bx_1 + e_1 \\
y_2 &= A + Bx_2 + e_2 \\
y_3 &= A + Bx_3 + e_3 \\
\dots &\quad \dots \\
y_n &= A + Bx_n + e_n
\end{aligned}
$$

Residuals



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 residuals



Ideally, residuals should

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

Summary

Regression is a way to fit a curve through a set of points

It is widely used in quantifying cause-effect relationships and in forecasting

Regression is powerful, versatile and deep

Prediction using regression is an application of Machine Learning