Understanding and Applying Factor Analysis and PCA

INTRODUCING FACTOR ANALYSIS AND PCA

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

Overview

Introduce factor analysis and PCA and their link to linear regression

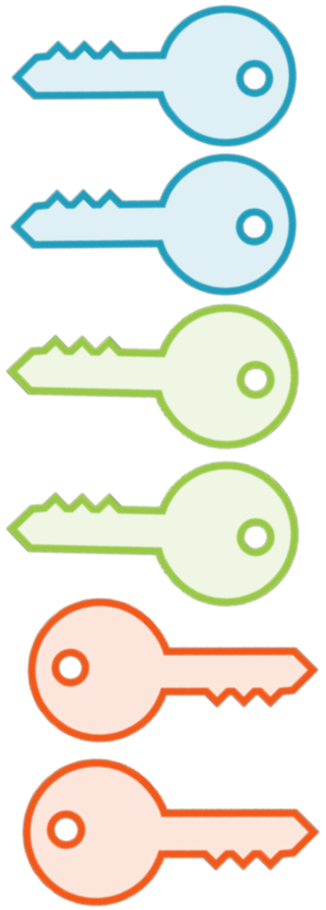
Learn when to use factor analysis and PCA

Understand just enough linear algebra and statistics to do so

Cutting Through Clutter with Factor Analysis

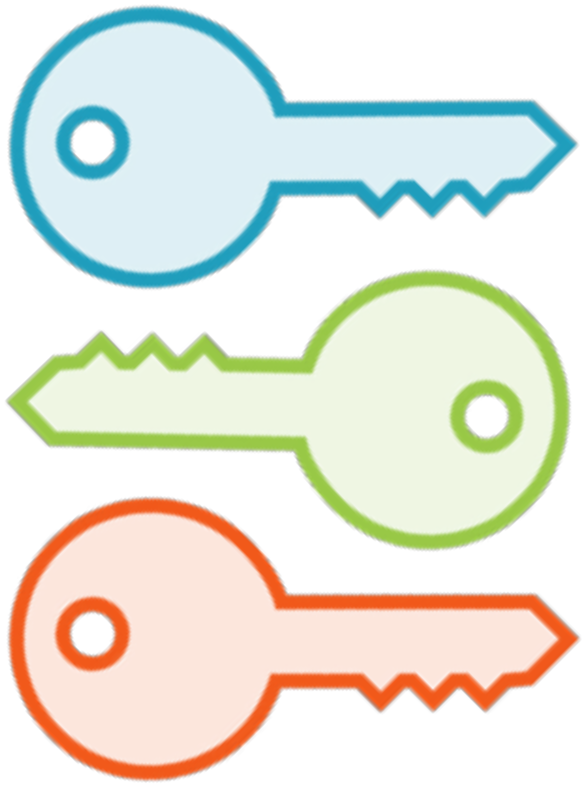
Keeping things simple is quite complicated

Similar, yet Different



Regression

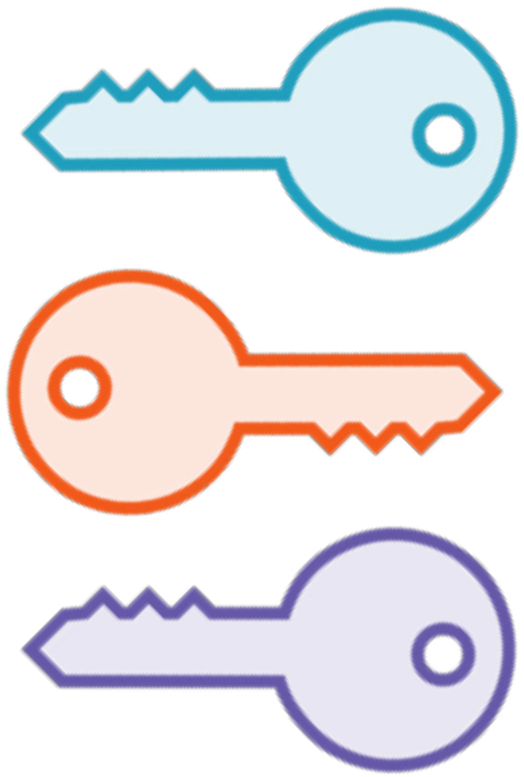
Connect the dots



Factor Analysis

Cut through the clutter

Regression



Causes

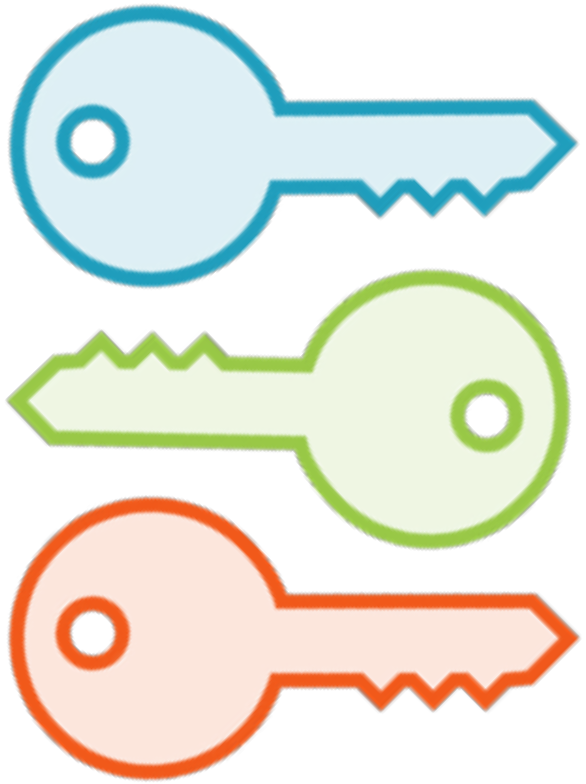
Independent variables

Effect

Dependent variable

Factor Analysis

Many Observed Causes



Few Underlying Causes

One Effect

Simplistic

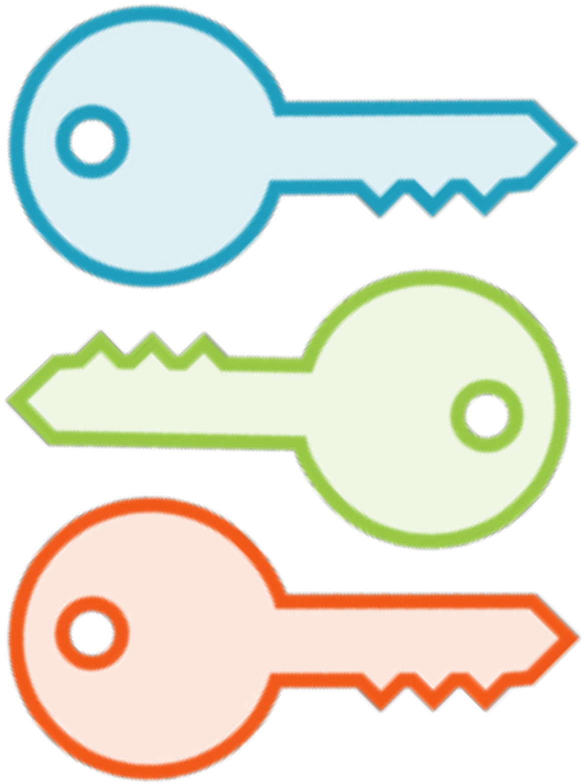
Causes

Independent variables

Effect

Dependent variable

Simple



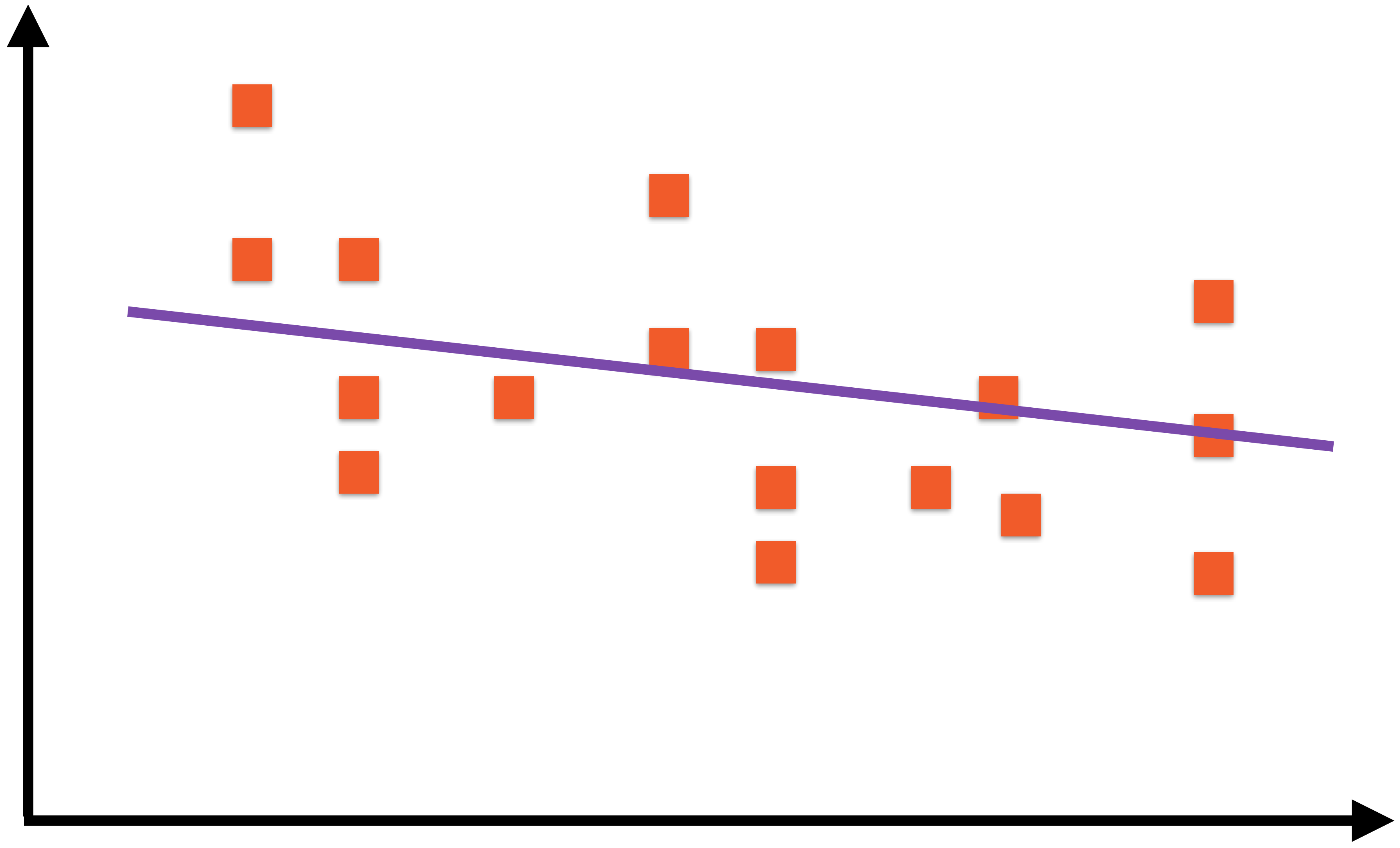
Causes

Independent variables

Effect

Dependent variable

Connecting the Dots with Regression



Regression is a technique to find the “best” line through a set of dots

Connecting the Dots with Regression

Cause

Independent variable

Effect

Dependent variable

Success as a Salesperson

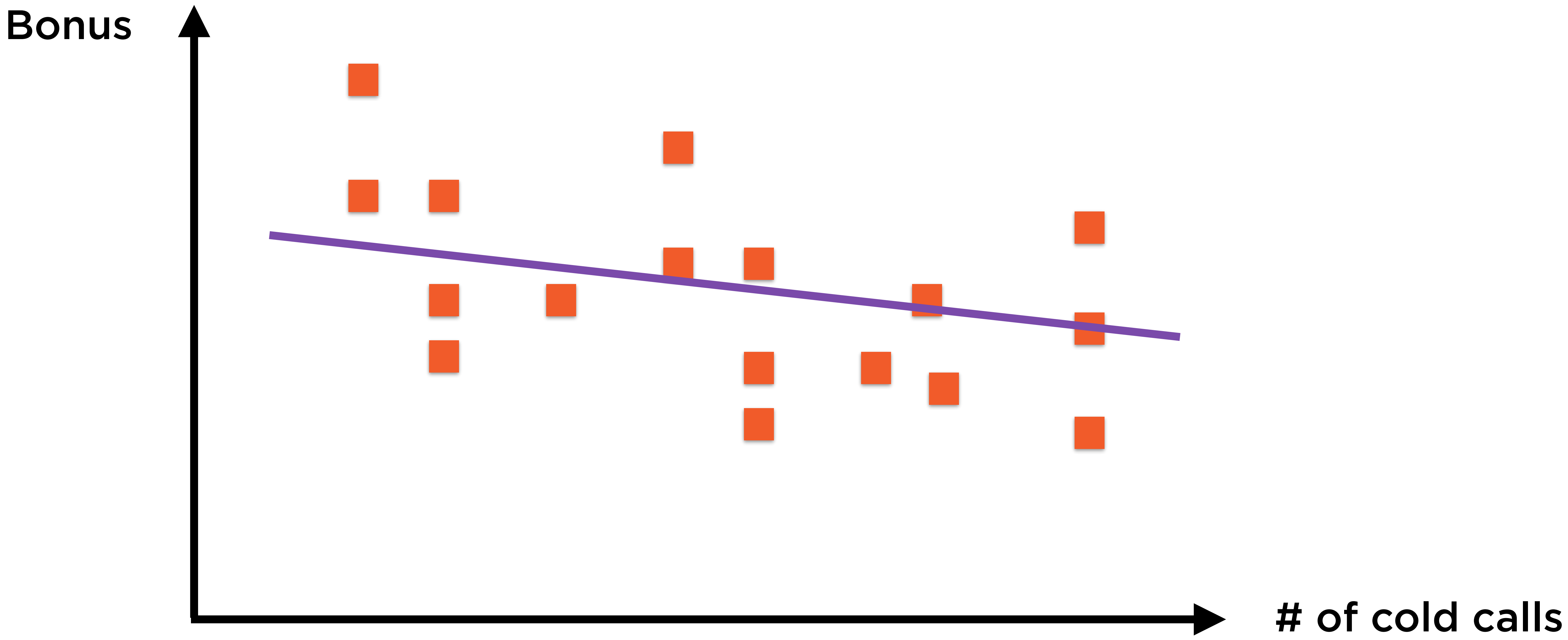
Cause

Number of cold calls initiated

Effect

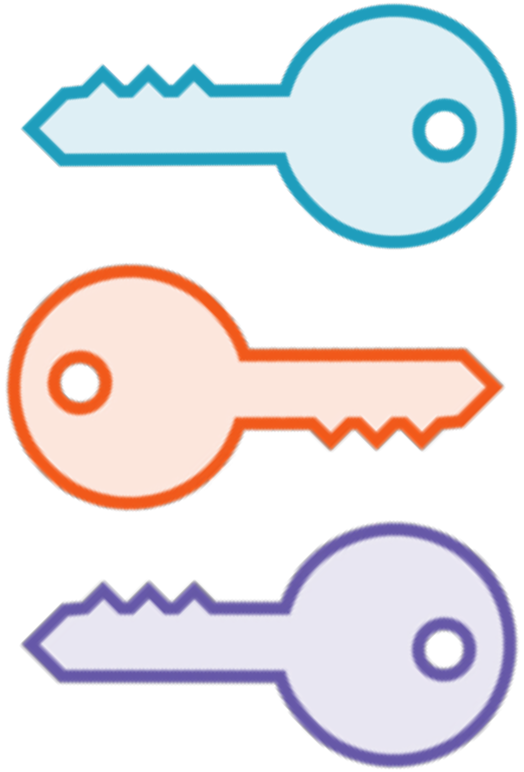
Bonus as member of sales team

Simple Regression



One cause, one effect

Multiple Regression



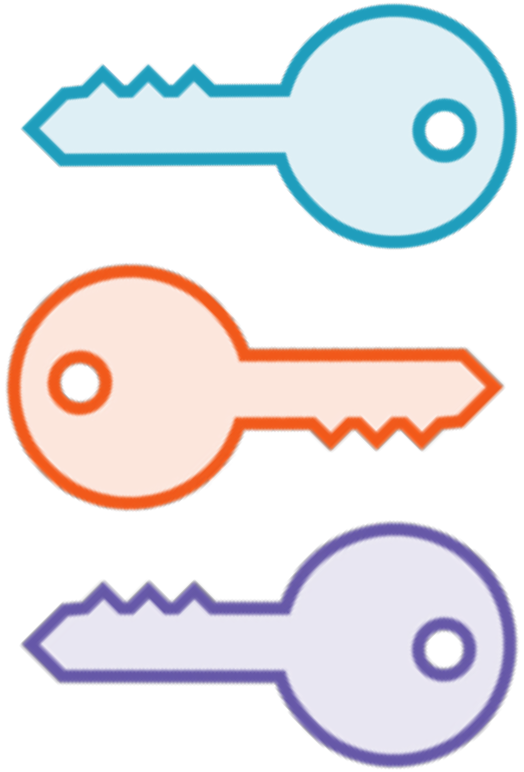
Causes

Independent variables

Effect

Dependent variable

Success as a Salesperson



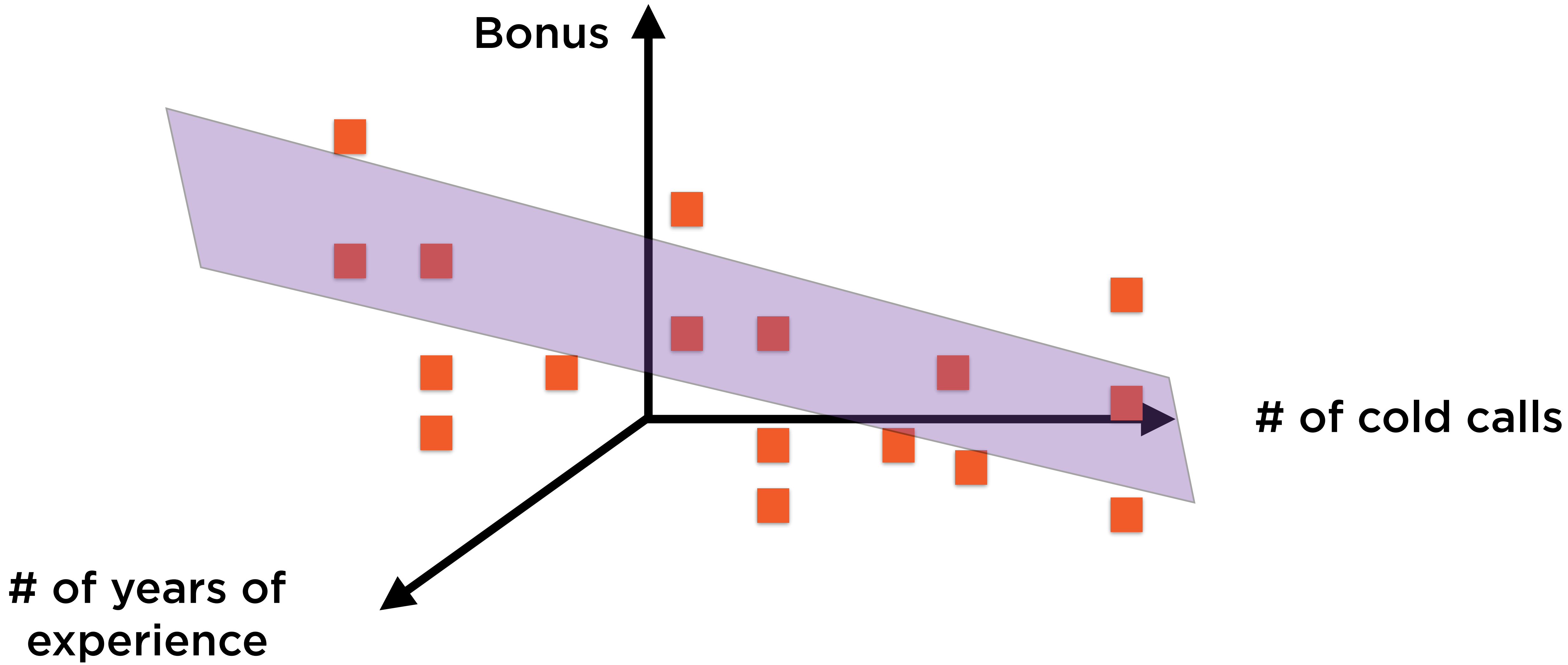
Causes

Number of cold calls, years of experience in sales jobs

Effect

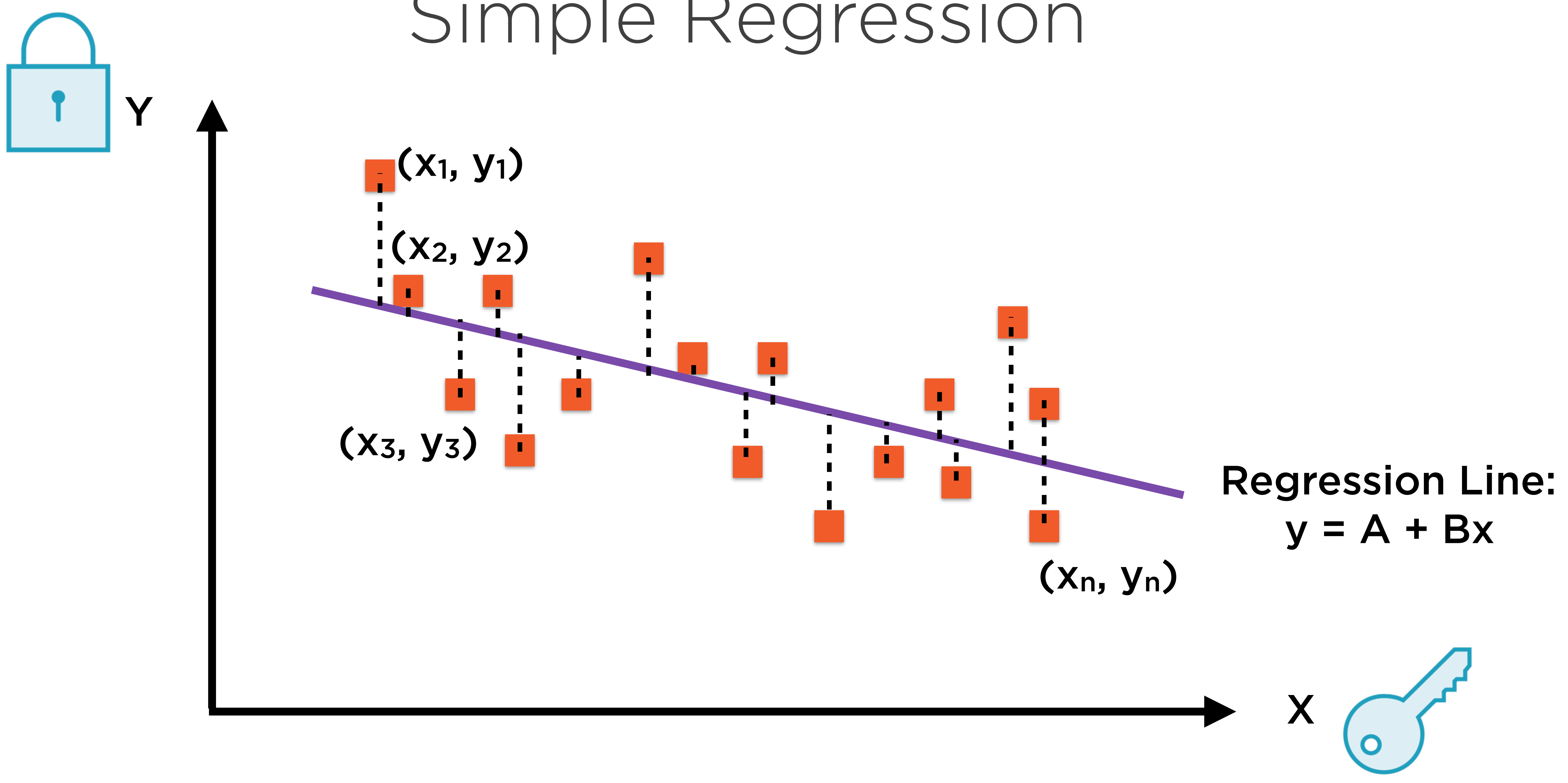
Bonus as member of sales team

Success as a Salesperson



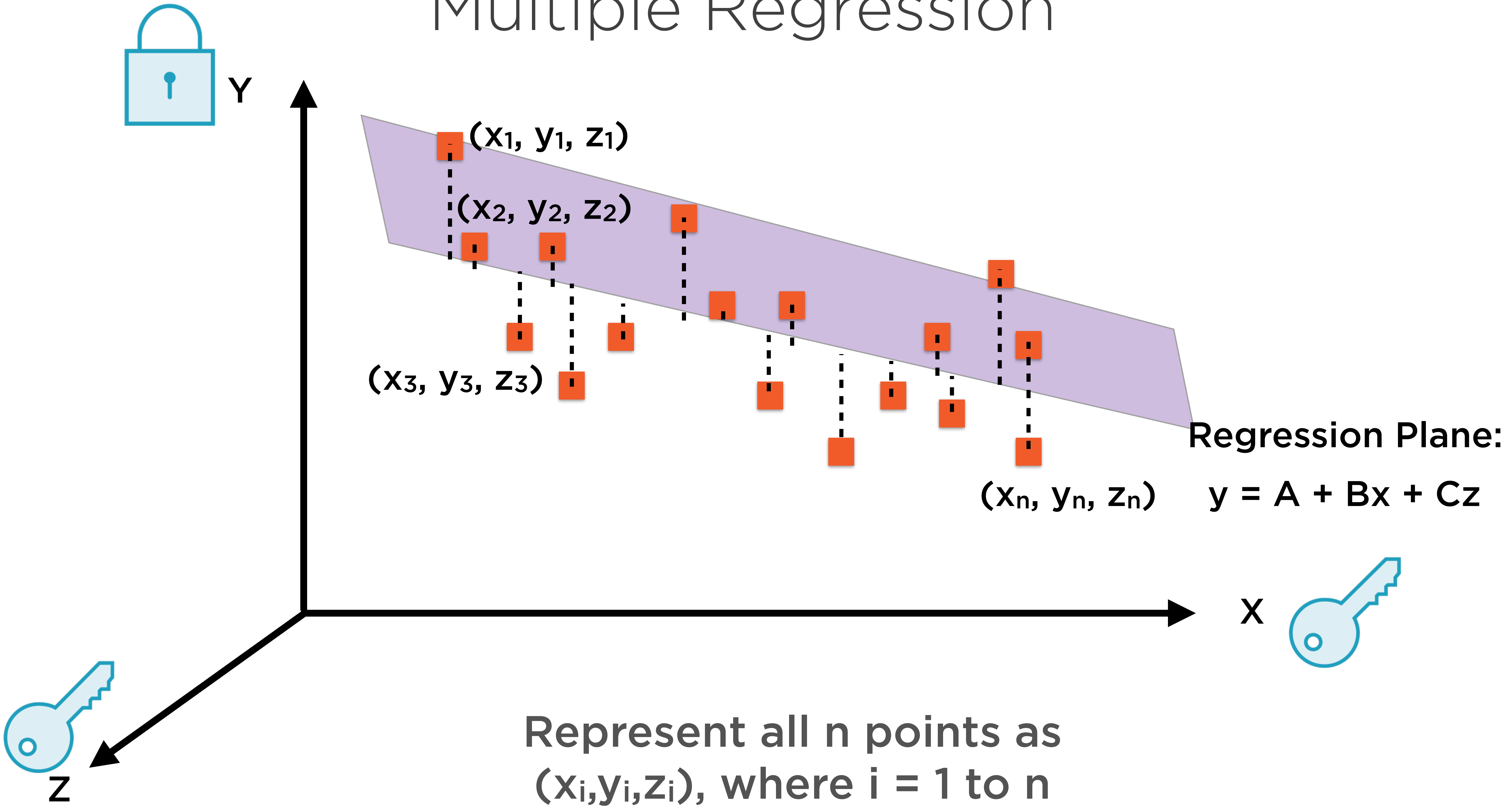
Many causes, one effect

Simple Regression



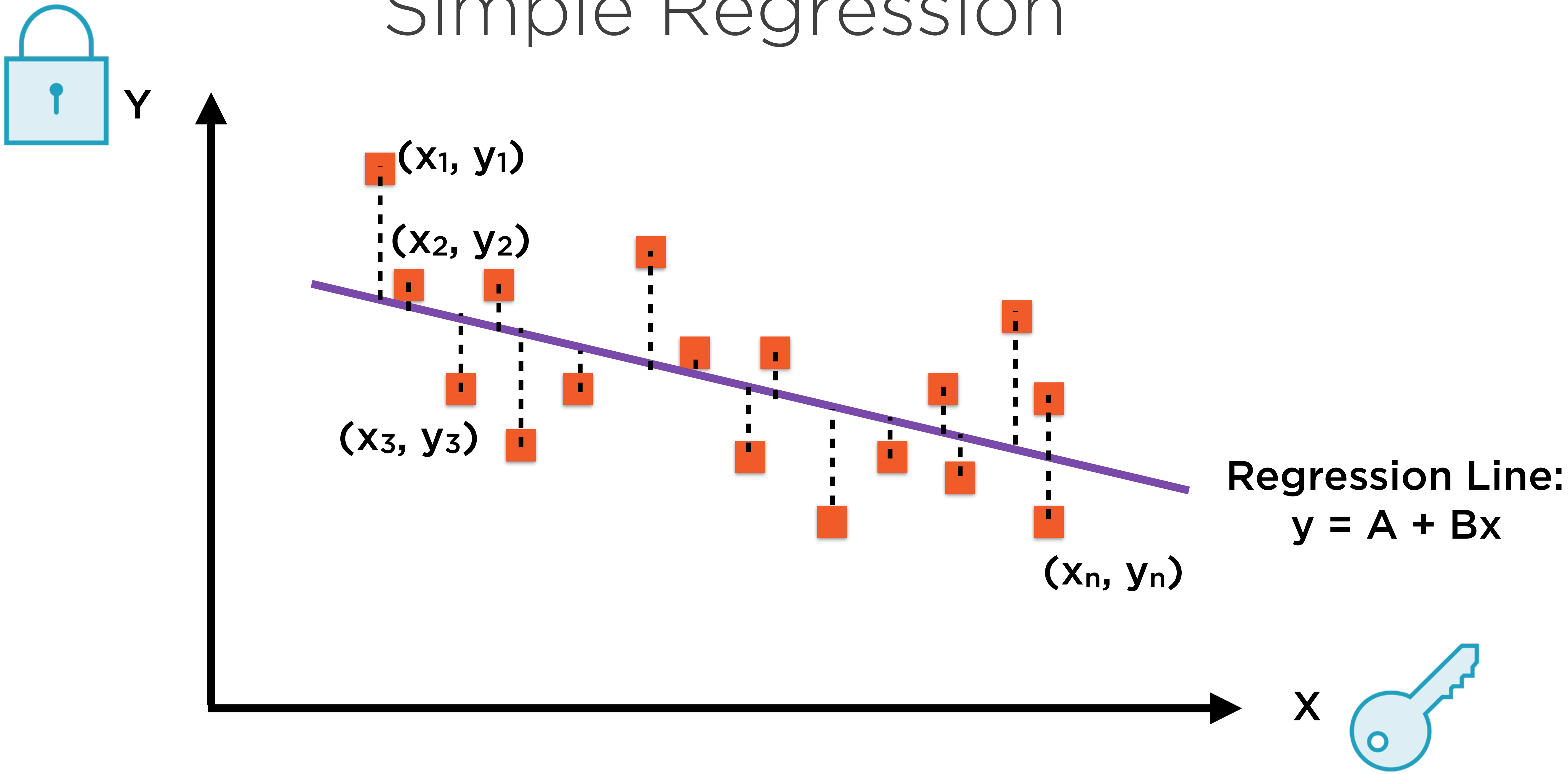
Represent all n points as (x_i, y_i), where i = 1 to n

Multiple Regression



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

Simple Regression



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

Simple Regression

Regression Equation:

$$y = A + Bx$$

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

Simple Regression

Regression Equation:

$$y = A + Bx$$

$$y_1 = A + Bx_1 + e_1$$

$$y_2 = A + Bx_2 + e_2$$

$$y_3 = A + Bx_3 + e_3$$

$$\ldots \qquad \ldots$$

$$y_n = A + Bx_n + e_n$$

Simple Regression

Regression Equation:

$$y = A + Bx$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} + B \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

Simple Regression

Regression Equation:

BONUS = A + B COLDCALLS

$$\begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \cdots \\ B_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} CC_1 \\ CC_2 \\ CC_3 \\ \cdots \\ CC_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_n \end{bmatrix}$$

B_i = Bonus of salesperson i

CC_i = Number of cold calls made by salesperson i

Multiple Regression

Regression Equation:

BONUS = A + B COLDCALLS + C EXPERIENCE

$$\begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \cdots \\ B_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} CC_1 \\ CC_2 \\ CC_3 \\ \cdots \\ CC_n \end{bmatrix} + C \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \cdots \\ E_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_n \end{bmatrix}$$

B_i = Bonus of salesperson i

CC_i = Number of cold calls made by salesperson i

E_i = Number of years of experience of salesperson i

Multiple Regression

Regression Equation:

$$y = A + Bx + Cz$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} + B \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} + C \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \cdots \\ z_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_n \end{bmatrix}$$

Multiple Regression

Regression Equation:

$$y = A + Bx + Cz$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ \dots & \dots & \dots \\ 1 & x_n & z_n \end{bmatrix} * \begin{bmatrix} A \\ B \\ C \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

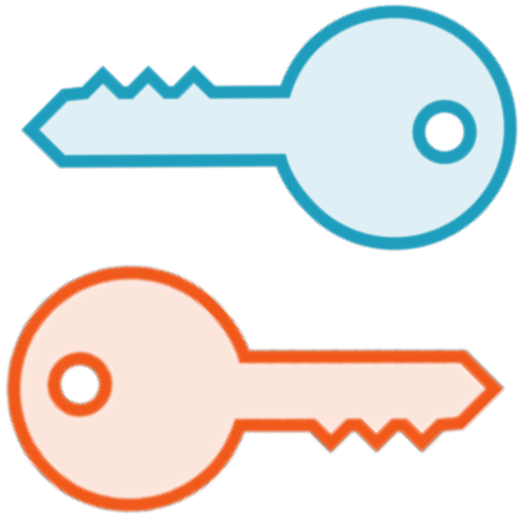
n Rows, 1 Column

n Rows, 3 Columns

3 Rows, 1 Column

n Rows, 1 Column

Multiple Regression



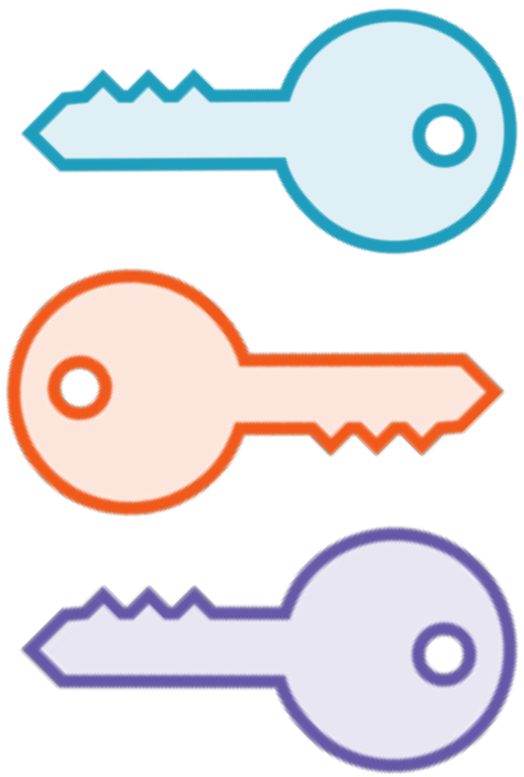
2 Causes

Cold calls, experience

1 Effect

Bonus in sales team

Multiple Regression



k Causes

Cold calls, experience, perceived honesty...

1 Effect

Bonus in sales team

Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \ldots + C_{k+1}x_k$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} = C_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \ldots \\ 1 \end{bmatrix} + C_2 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \ldots \\ x_{n1} \end{bmatrix} + \ldots C_{k+1} \begin{bmatrix} x_{1k} \\ x_{2k} \\ x_{3k} \\ \ldots \\ x_{nk} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{bmatrix}$$

Multiple Regression

Regression Equation:

$$y = C_1 + C_2 x_1 + \ldots + C_{k+1} x_k$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & & x_{1k} \\ 1 & x_{21} & & x_{2k} \\ 1 & x_{31} & \ldots & x_{3k} \\ \ldots & \ldots & & \ldots \\ 1 & x_{n1} & & x_{nk} \end{bmatrix} * \begin{bmatrix} C_1 \\ C_2 \\ \ldots \\ C_{k+1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{bmatrix}$$

n Rows, 1 Column	n Rows, k+1 Columns	k+1 Rows, 1 Column	n Rows, 1 Column

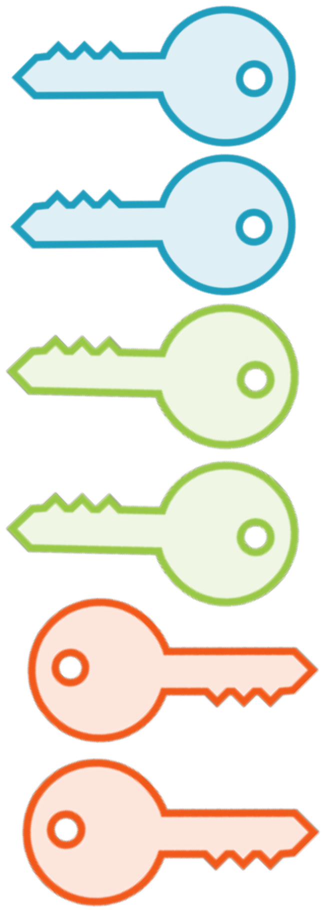
Multiple Regression

Regression Equation:

$$y = C_1 + C_2x_1 + \ldots + C_{k+1}x_k$$

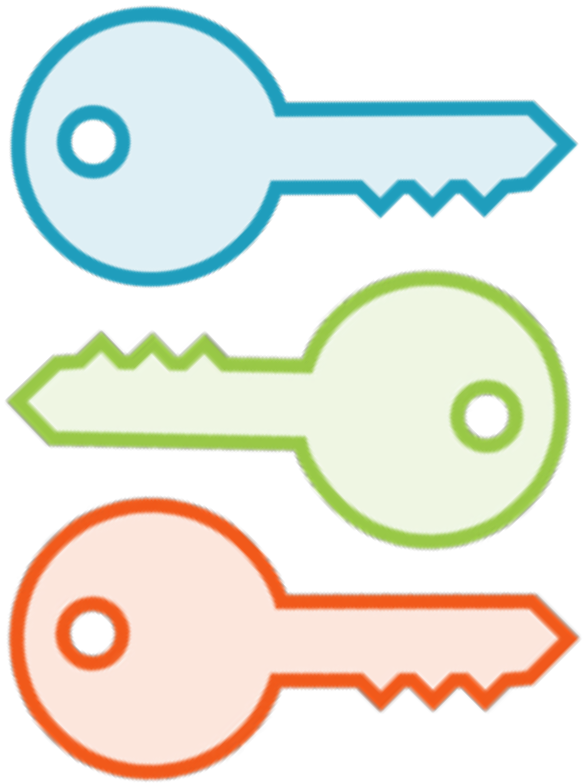
Linear regression involves finding k+1 coefficients, k for the explanatory variables, and 1 for the intercept

Similar, yet Different



Regression

Connect the dots



Factor Analysis

Cut through the clutter

Simplistic

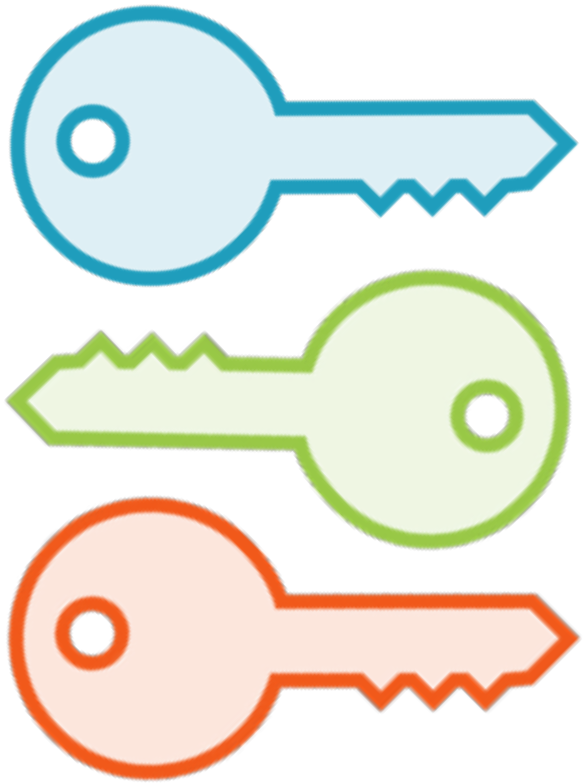
Causes

Independent variables

Effect

Dependent variable

Simple



Causes

Independent variables

Effect

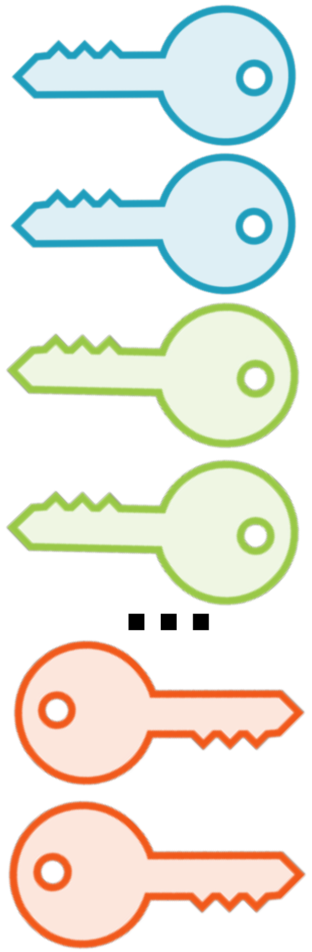
Dependent variable

Kitchen Sink Regression

Proposed Regression Equation:

BONUS = A + B COLDCALLS + C EXPERIENCE + D NUMFOLLOWERS + E HONESTY + F PUNCTUALITY + ...

Kitchen Sink Regression



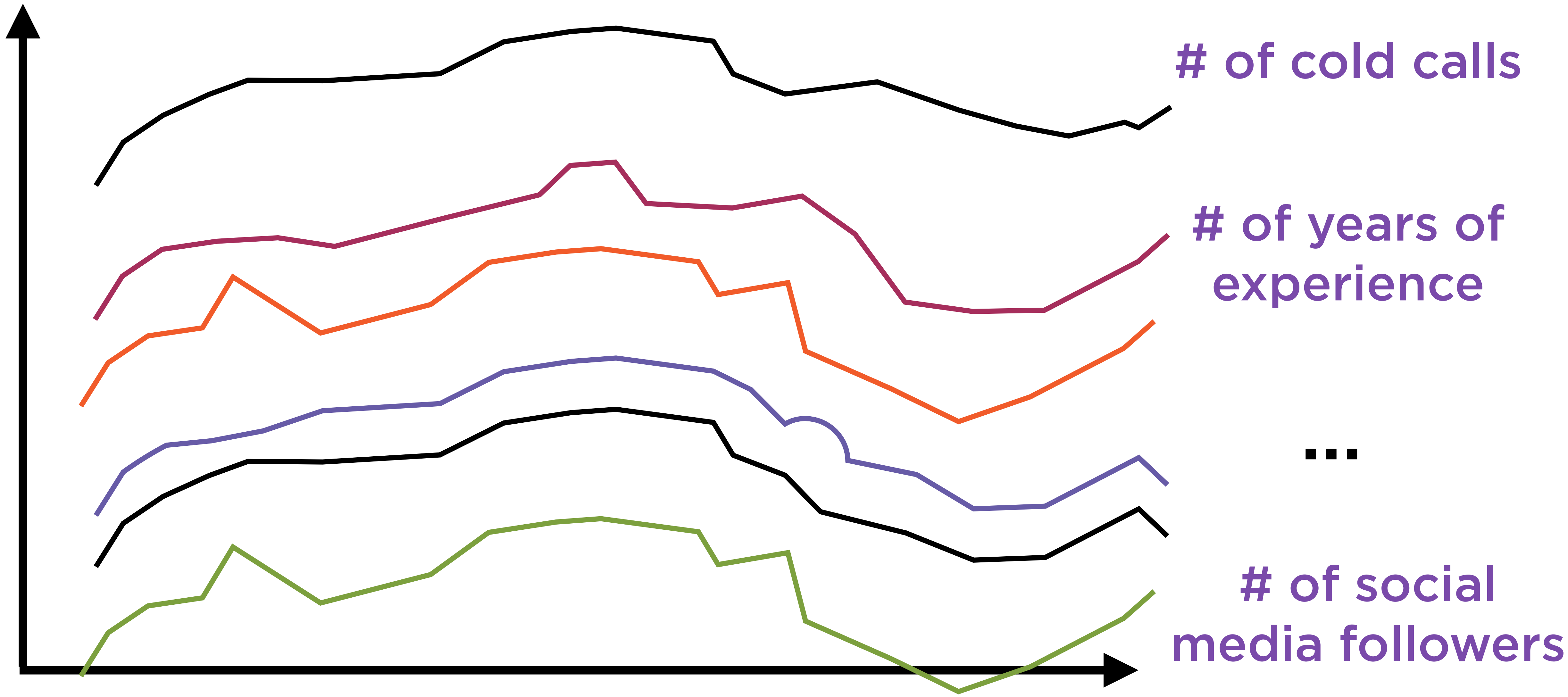
10 Causes

Cold calls, experience, social media followers, perceived honesty, billing punctuality...

1 Effect

Bonus in sales team

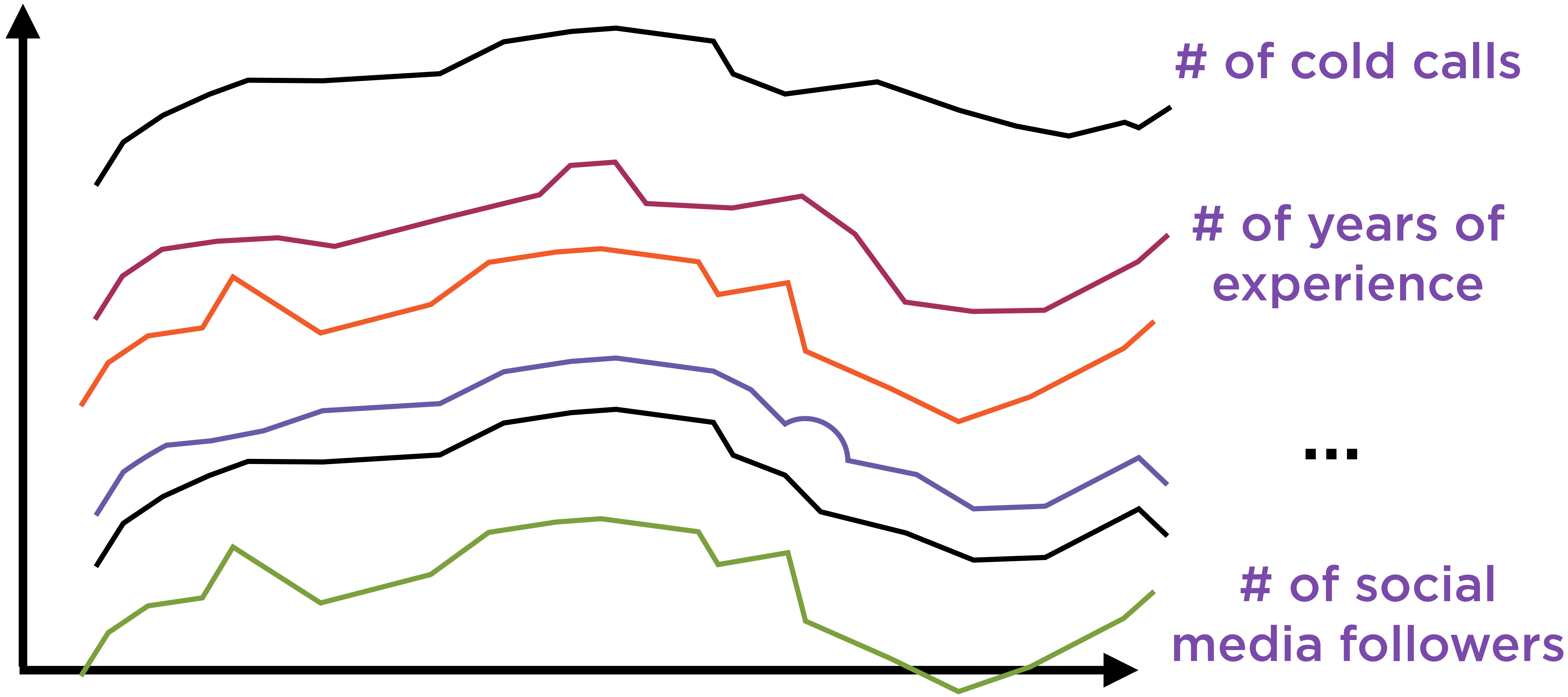
Bad News: Multicollinearity Detected



6 of 10 explanatory variables are highly correlated with each other

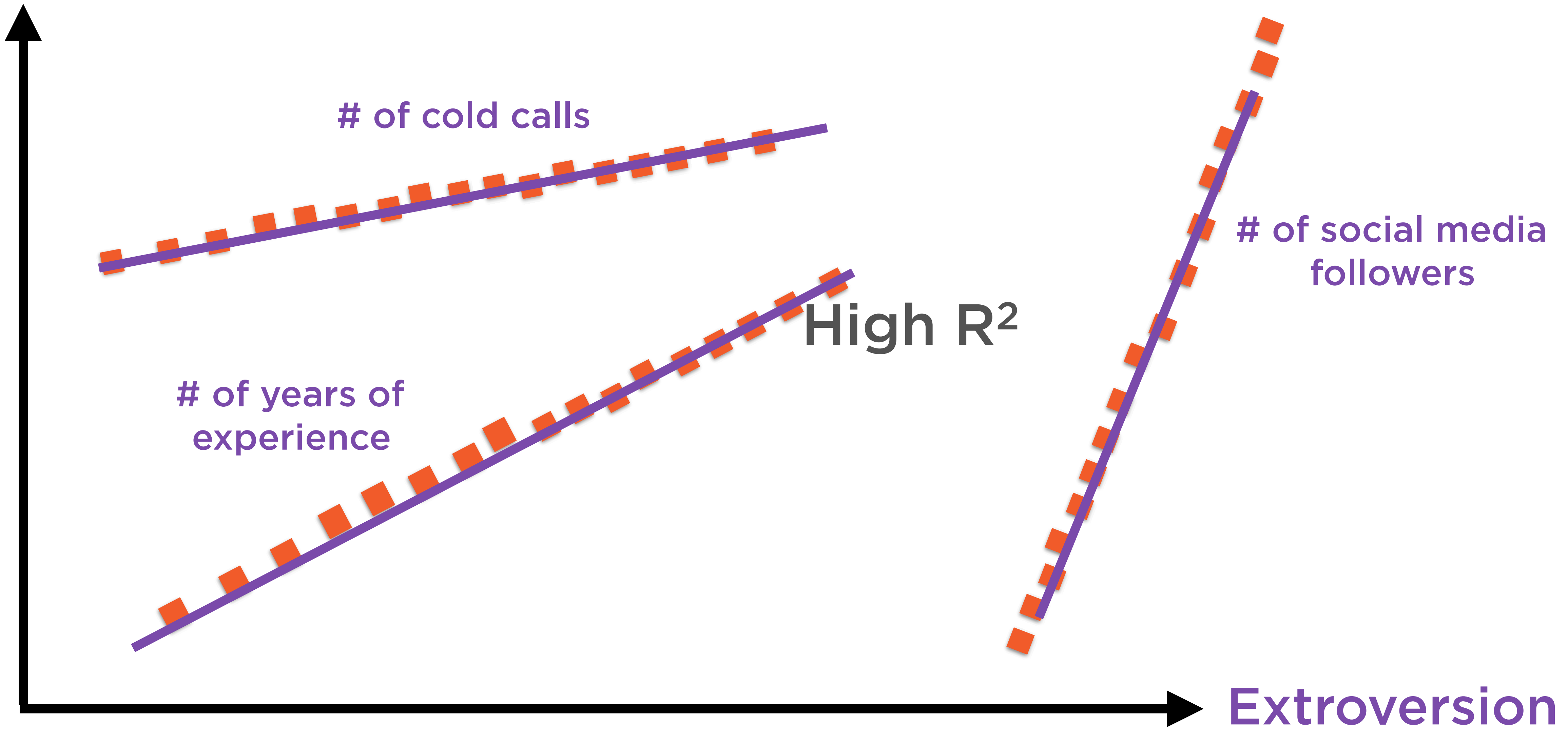
A big risk with regression is **multicollinearity**: X variables containing the same information

Underlying Cause: Extroversion



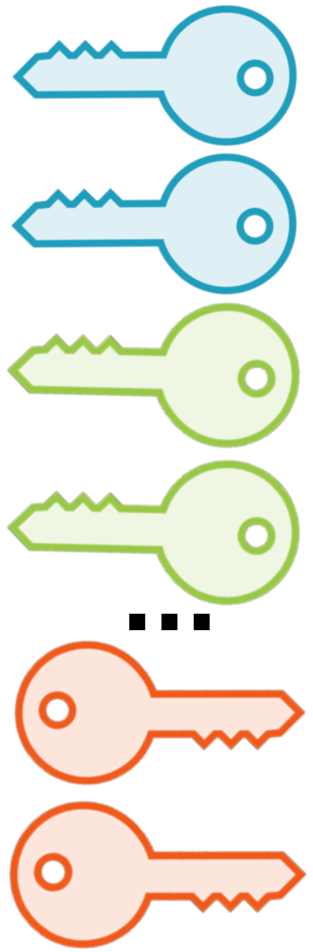
Each of these explanatory variables is caused by an underlying personality trait

Underlying Cause: Extroversion



Simply measure extroversion and use it instead of the correlated explanatory variables

Kitchen Sink Regression



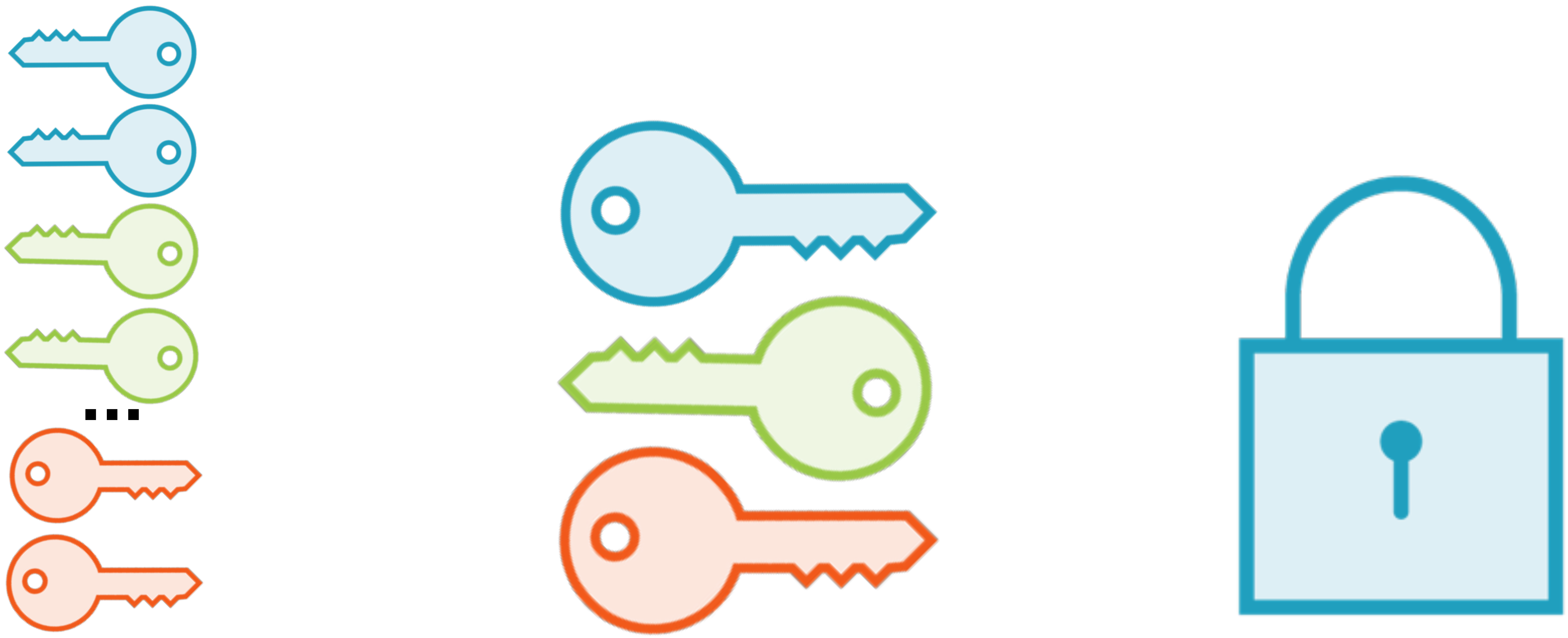
10 Causes

Cold calls, experience, social media followers, perceived honesty, billing punctuality...

1 Effect

Bonus in sales team

Factor Analysis

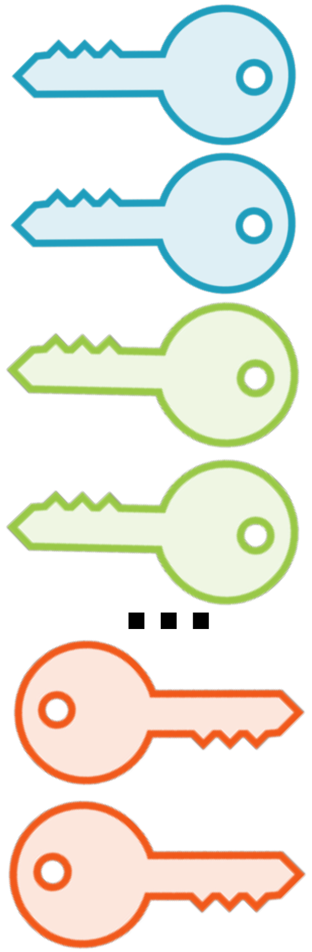


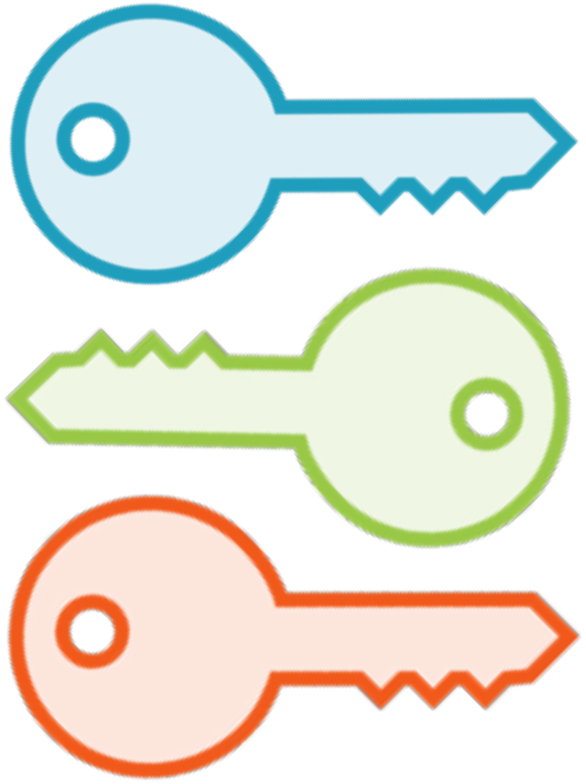
Many Observed Causes

Few Underlying Causes

One Effect

Success as a Salesperson





Many Observed Causes

Cold calls, experience, social media followers, perceived honesty, billing punctuality...

Few Underlying Causes

Personality traits

One Effect

Success as a salesperson

What and How: Factor Analysis and PCA

What and How

Cut through clutter

Extract underlying factors from a set of data

Principal components analysis (PCA)

Cookie-cutter technique that finds the ‘good’ factors from a set of data points

PCA is one solution to the factor-extraction problem - a cookie-cutter solution

What and How

Connect the dots

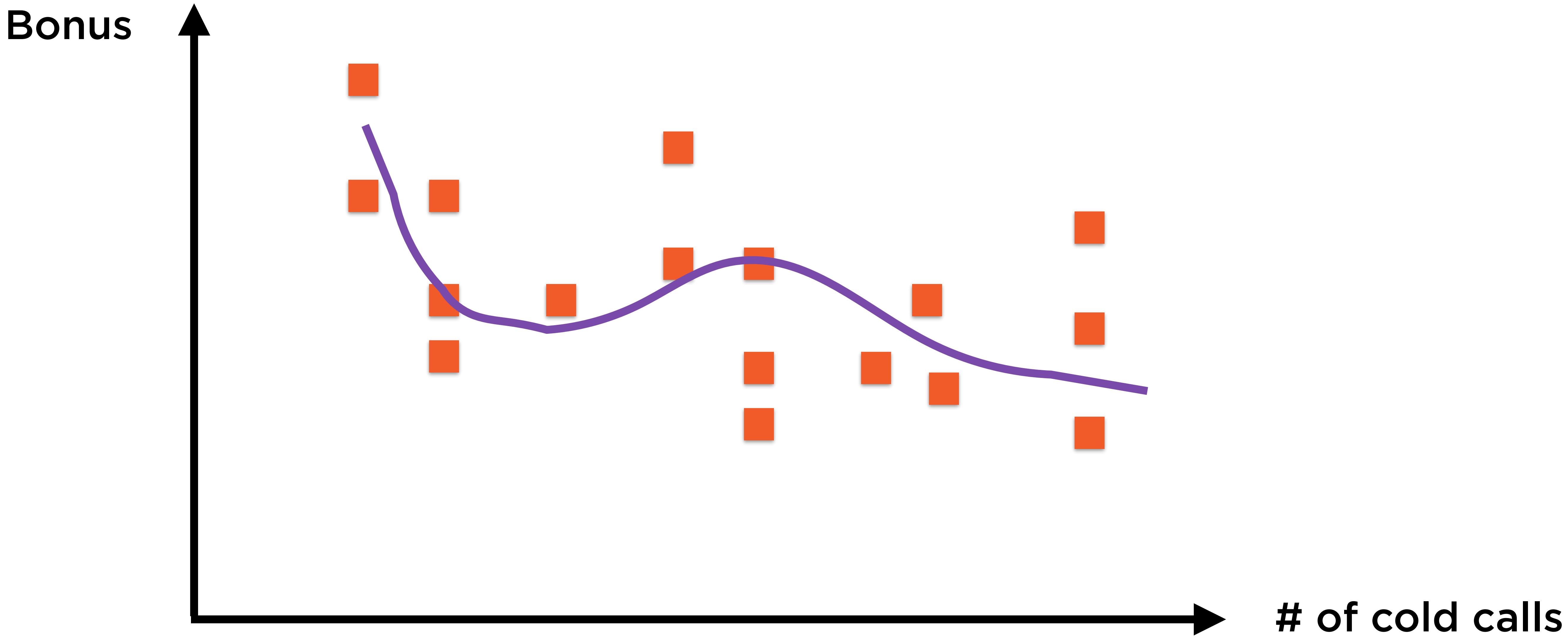
Fit a curve through a set of data

Regression

Cookie-cutter technique that finds the ‘best-fit’ line through a set of data points

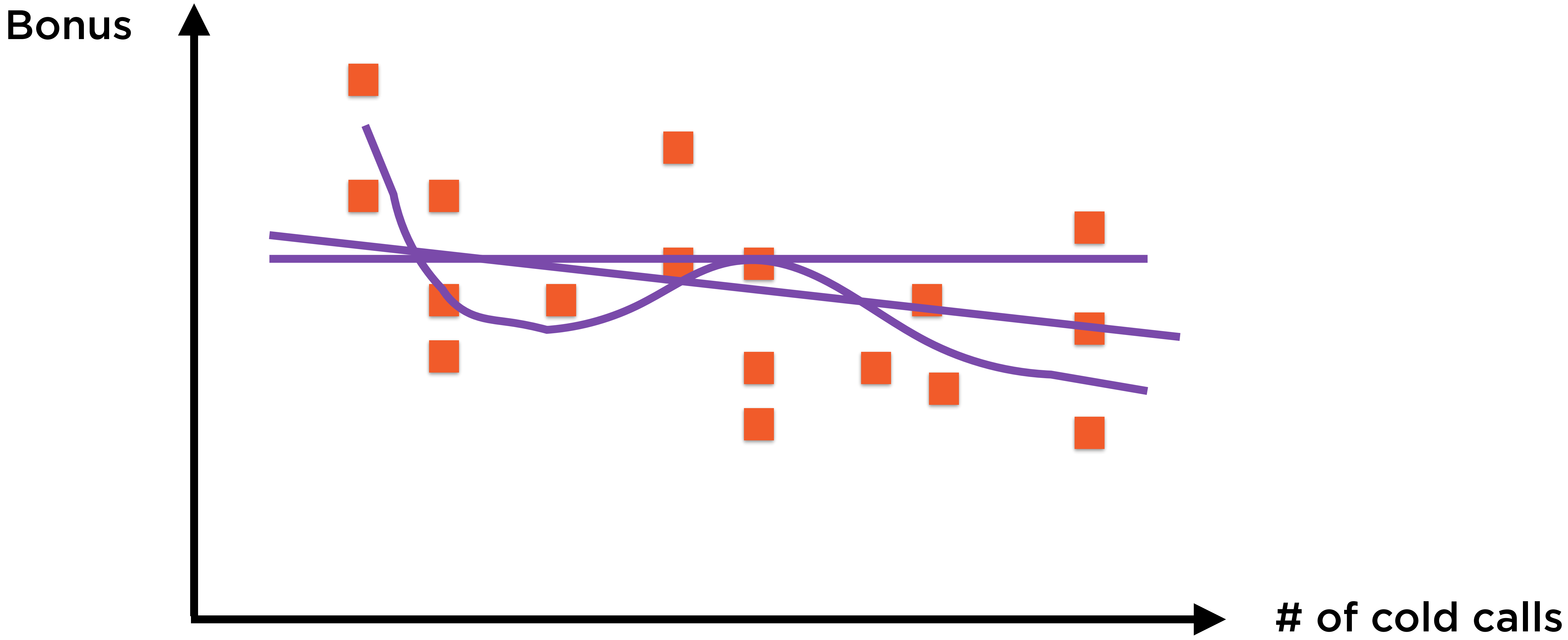
Regression is one solution to the data-fitting problem - a cookie-cutter solution

Connecting the Dots



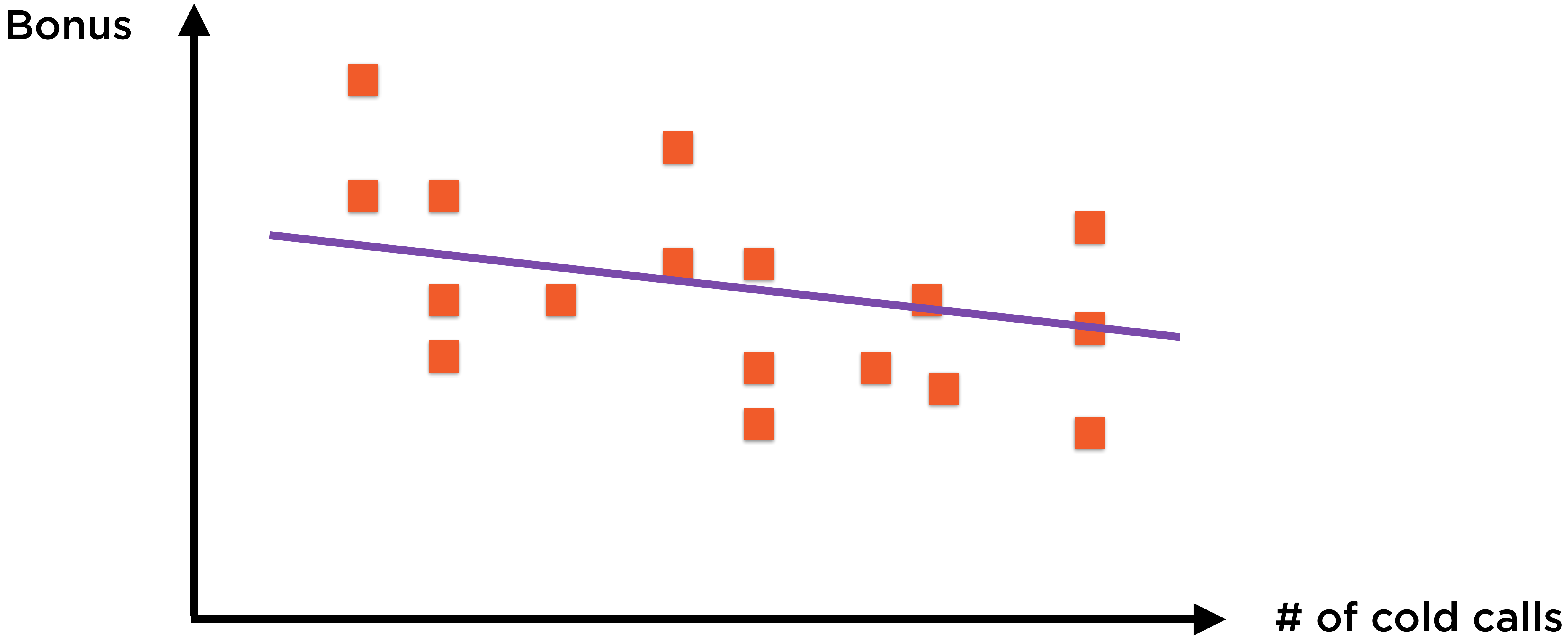
We can draw any number of curves to fit such data

Connecting the Dots



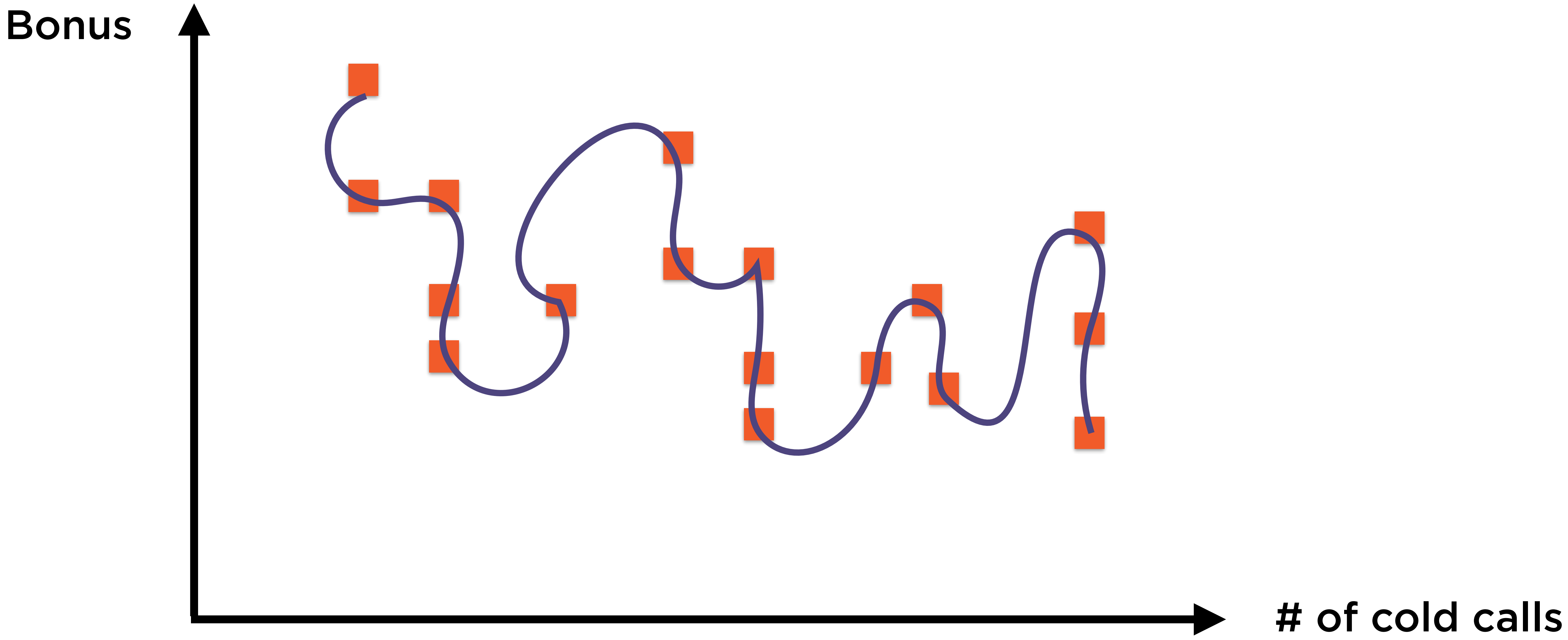
We can draw any number of curves to fit such data

Connecting the Dots



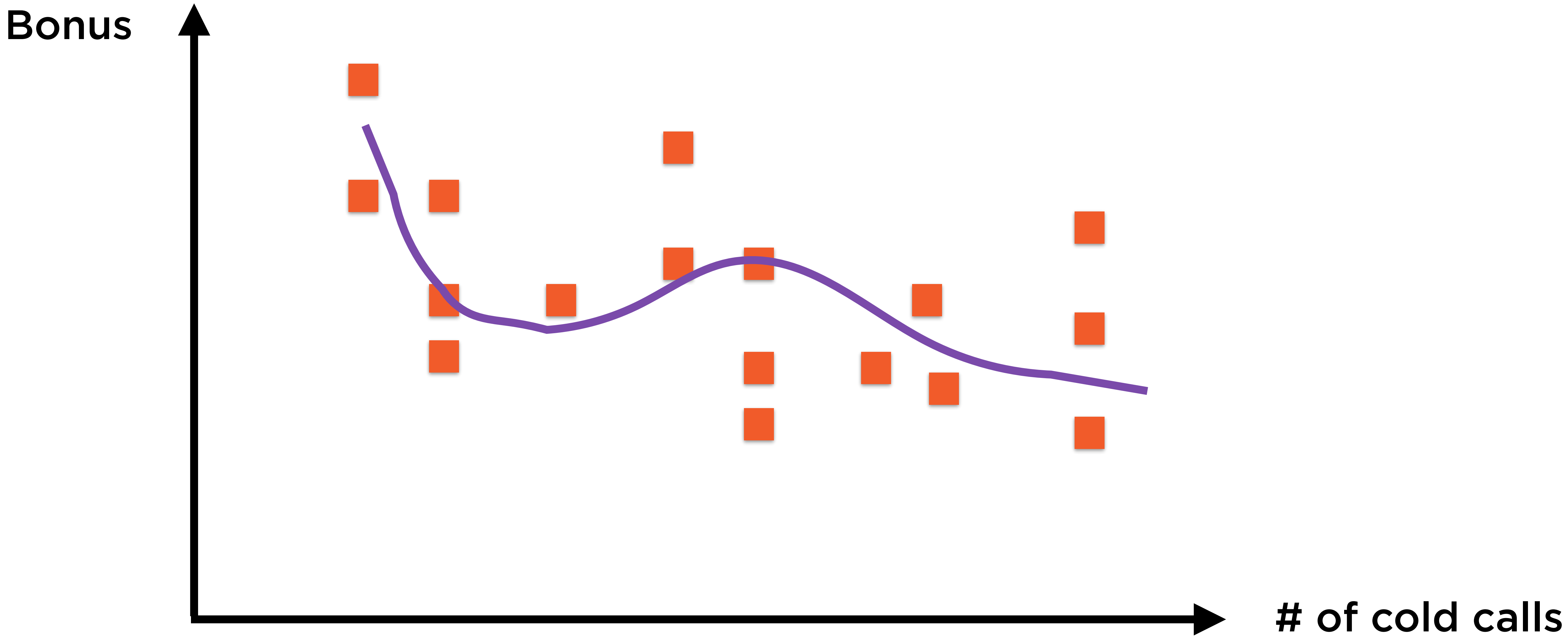
A straight line represents a linear relationship

Connecting the Dots



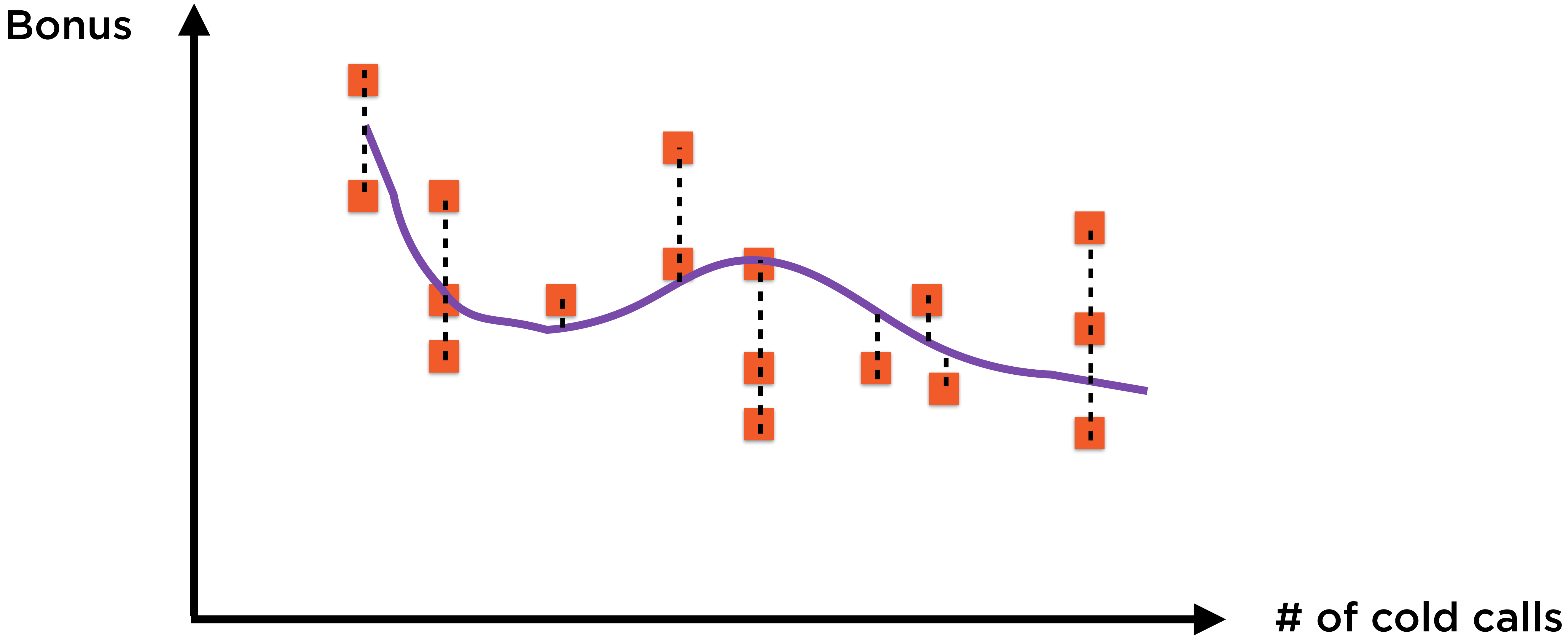
We could either make this curve pass through each point...

Connecting the Dots



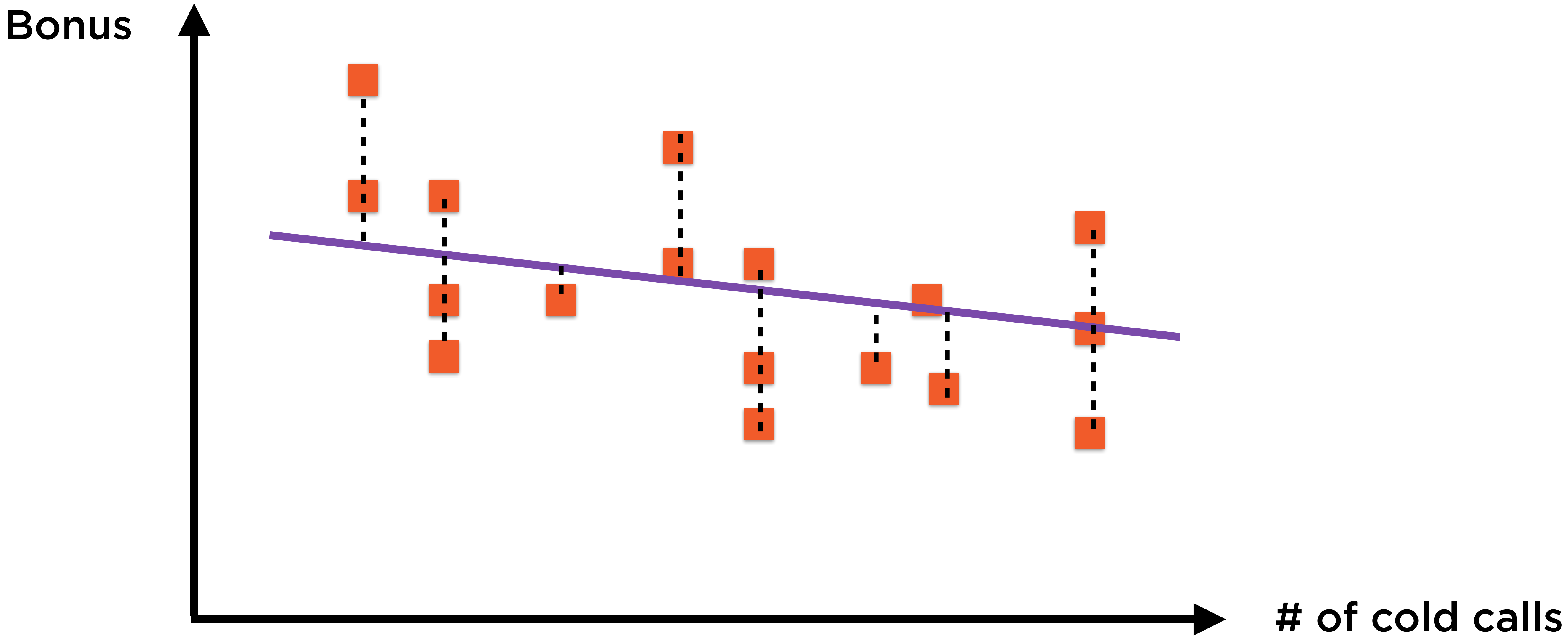
...Or in some sense “fit” the data in aggregate

Connecting the Dots



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

Connecting the Dots



Finding the "best" such straight line is called Linear Regression

What and How

Cut through clutter

Extract underlying factors from a set of data

Principal components analysis (PCA)

Cookie-cutter technique that finds the 'good' factors from a set of data points

PCA is one solution to the factor-extraction problem - a cookie-cutter solution

Two Approaches to Factor Extraction

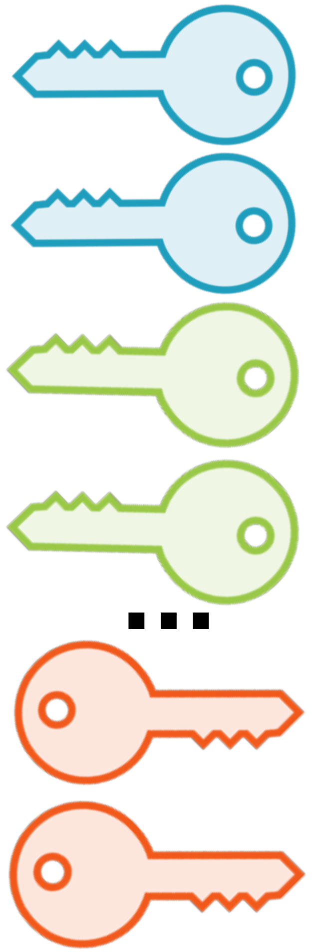
Rule-based

Human experts identify and extract factors

ML-based

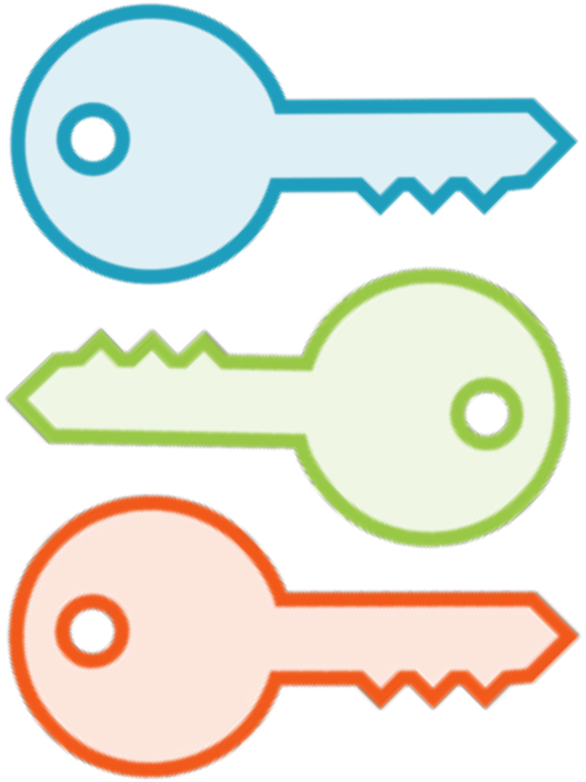
Algorithm identifies and extracts factors

Success as a Salesperson



Many Observed Causes

Cold calls, experience, social media followers, perceived honesty, billing punctuality...



Few Underlying Causes

Personality traits

One Effect

Success as a salesperson

Personality Profiles

Individual

Personality Assessment

Personality Profile

Personality Profiles

Gregariousness	Warmth	Assertiveness	Excitement-seeking	Modesty	Order	…
High	Medium	High	High	Low	High	…

1 row

100 columns

Individual

Personality Profile

Information Overload

Sales Team

Gregariousness	Warmth	Assertiveness	Excitement -seeking	Modesty	Order	...
High	Medium	High	High	Low	High	...

10,000 rows

100 columns

Personality Profile Database

The Big Five Personality Traits

Openness

Conscientiousness

Extraversion

Agreeableness

Neuroticism

The Big Five Personality Traits

100-dimensional data

1 column for each personality trait out there

Rule-based Factor Extraction Technique

5-dimensional data

5 columns, 1 for each big-five factor

The Big Five Personality Traits

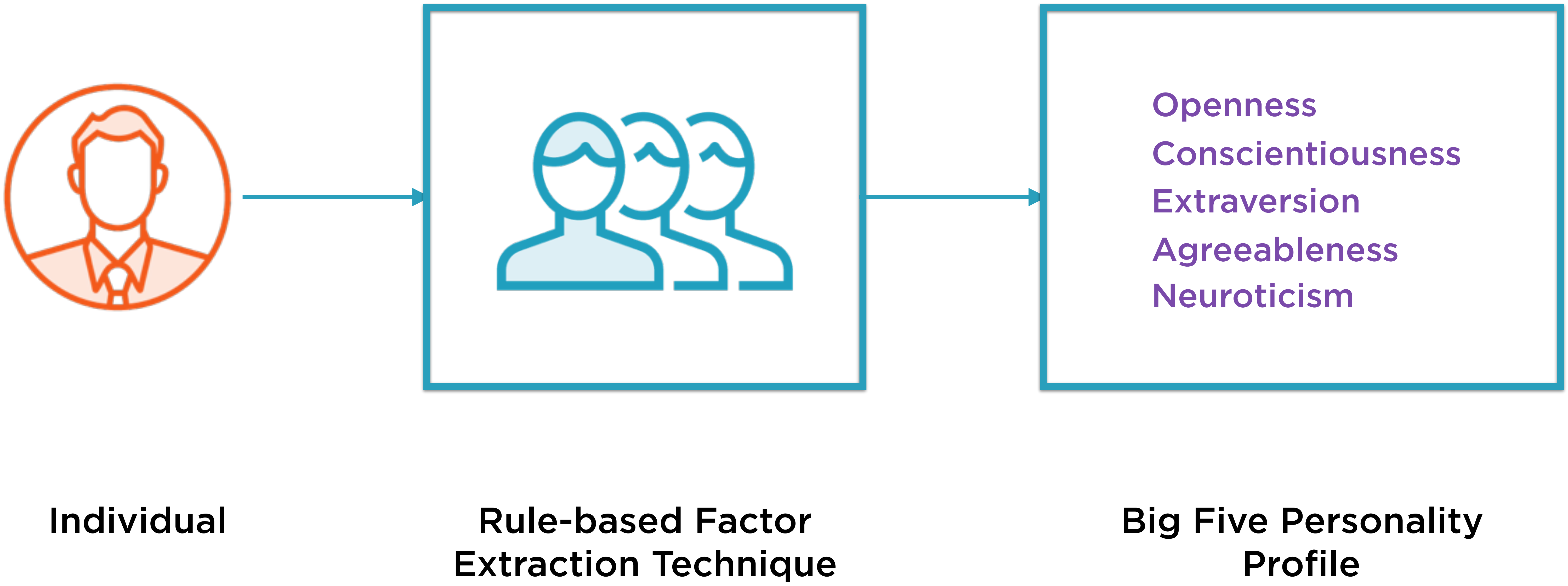
$$\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \cdots & x_{3k} \\ \cdots & & \cdots \\ x_{n1} & & x_{nk} \end{bmatrix} \rightarrow \boxed{\text{Rule-based Factor Extraction Technique}} \rightarrow \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \\ \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} & f_{n3} \end{bmatrix}$$

1 column for each personality trait out there

Rule-based Factor Extraction Technique

5 columns, 1 for each big-five factor

The Big Five Personality Traits



The Big Five Personality Traits

Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
High	Medium	High	High	Low

1 row

5 columns

Individual

Personality Profile

Two Approaches to Factor Extraction

Rule-based

Human experts identify and extract factors

ML-based

Algorithm identifies and extracts factors

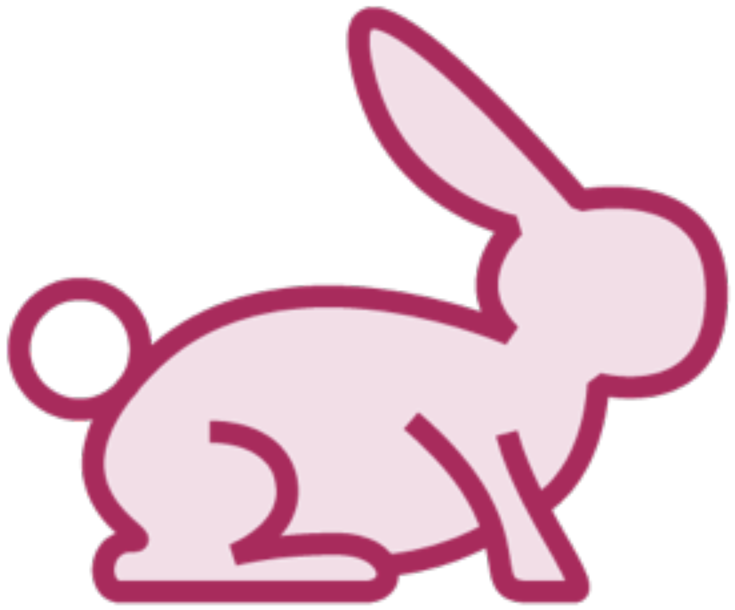
PCA and Factor Analysis

Principal Component Analysis is one procedure for factor analysis

It is mathematically guaranteed to result in independent factors

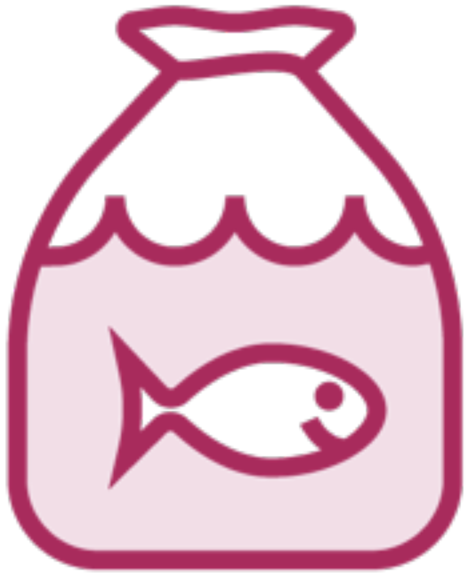
However, those factors may not actually correspond to intuition

Whales: Fish or Mammals?



Mammals

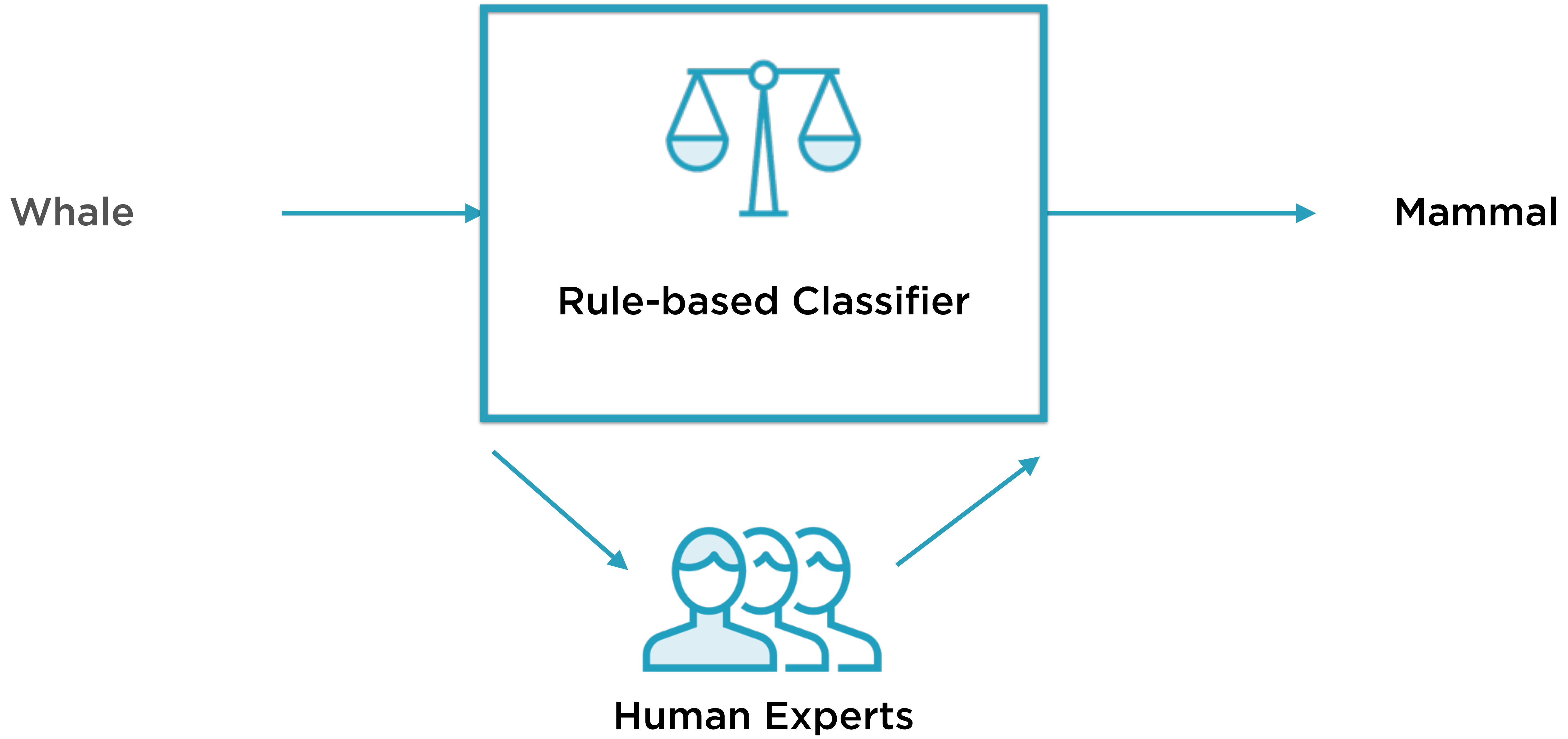
Members of the infraorder *Cetacea*



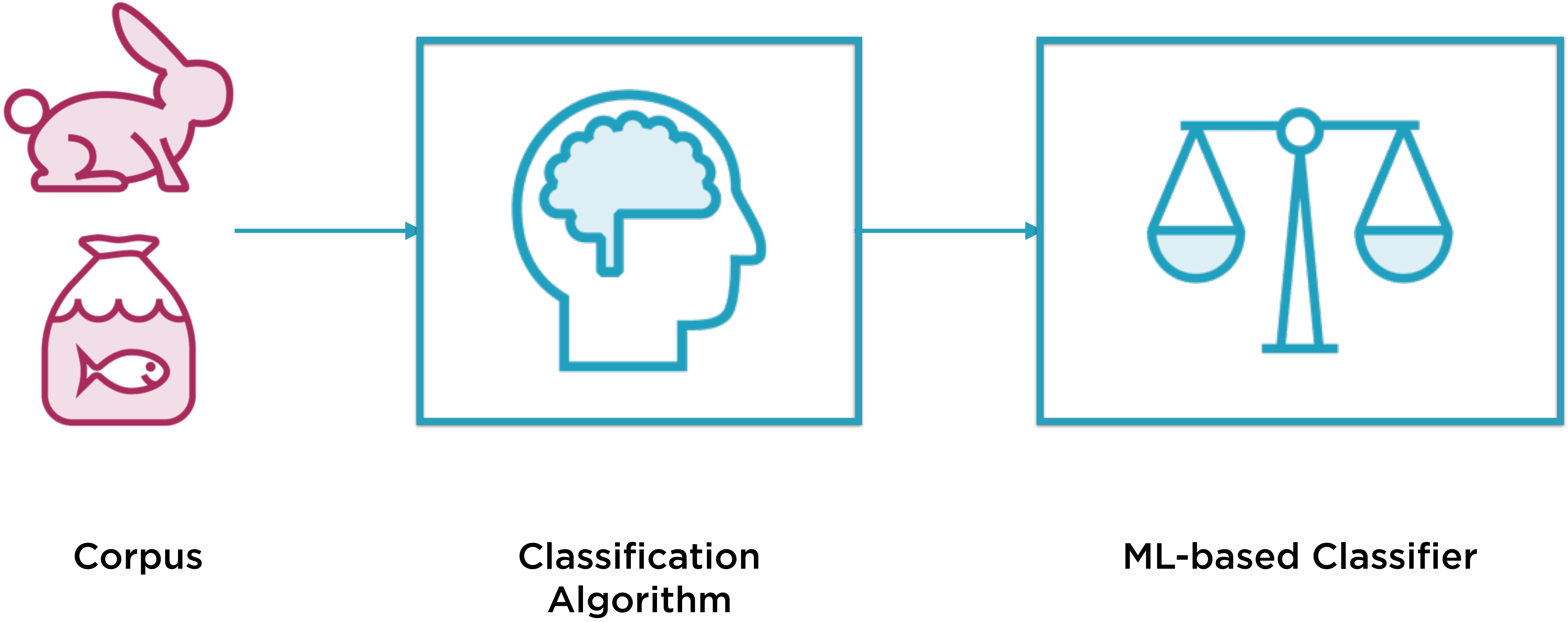
Fish

Look like fish, swim like fish, move like fish

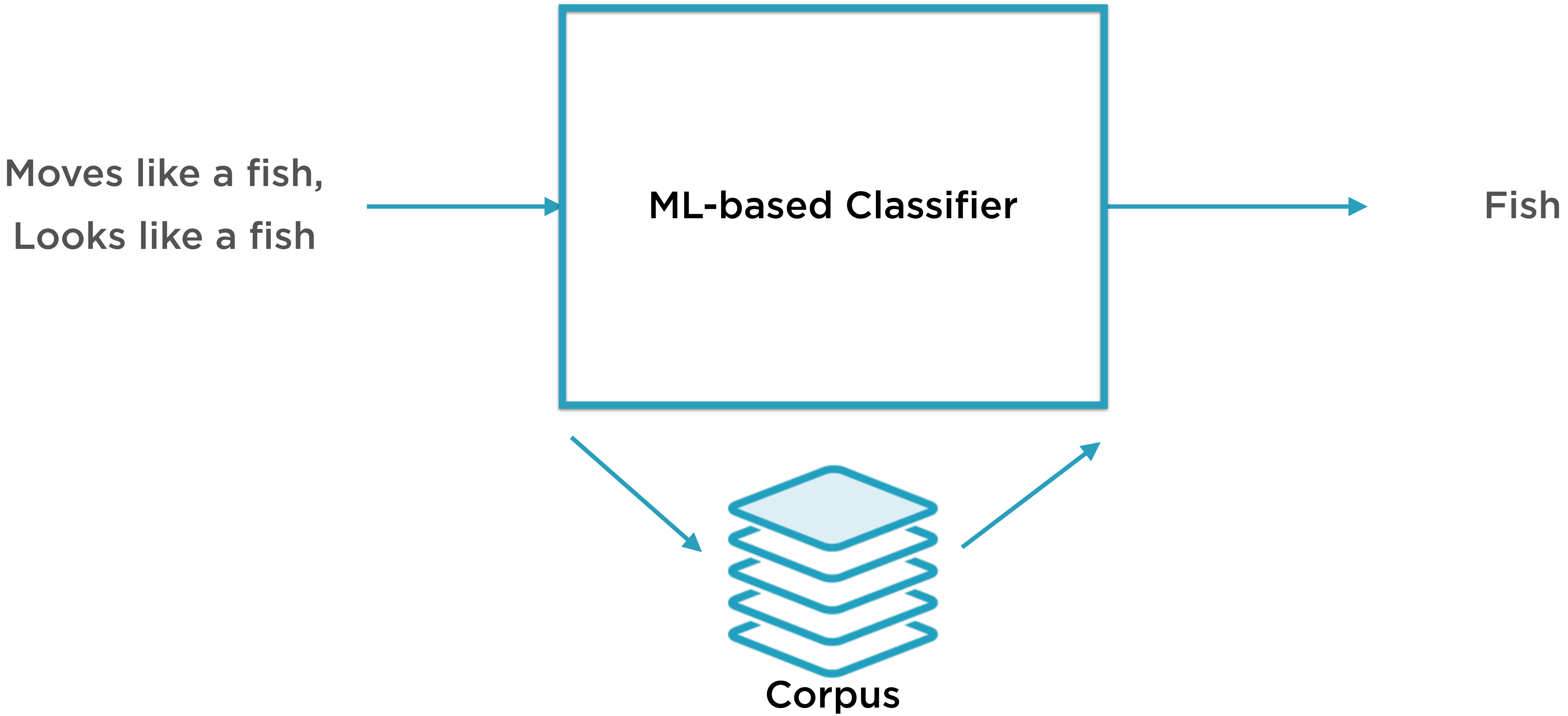
Rule-based Binary Classifier



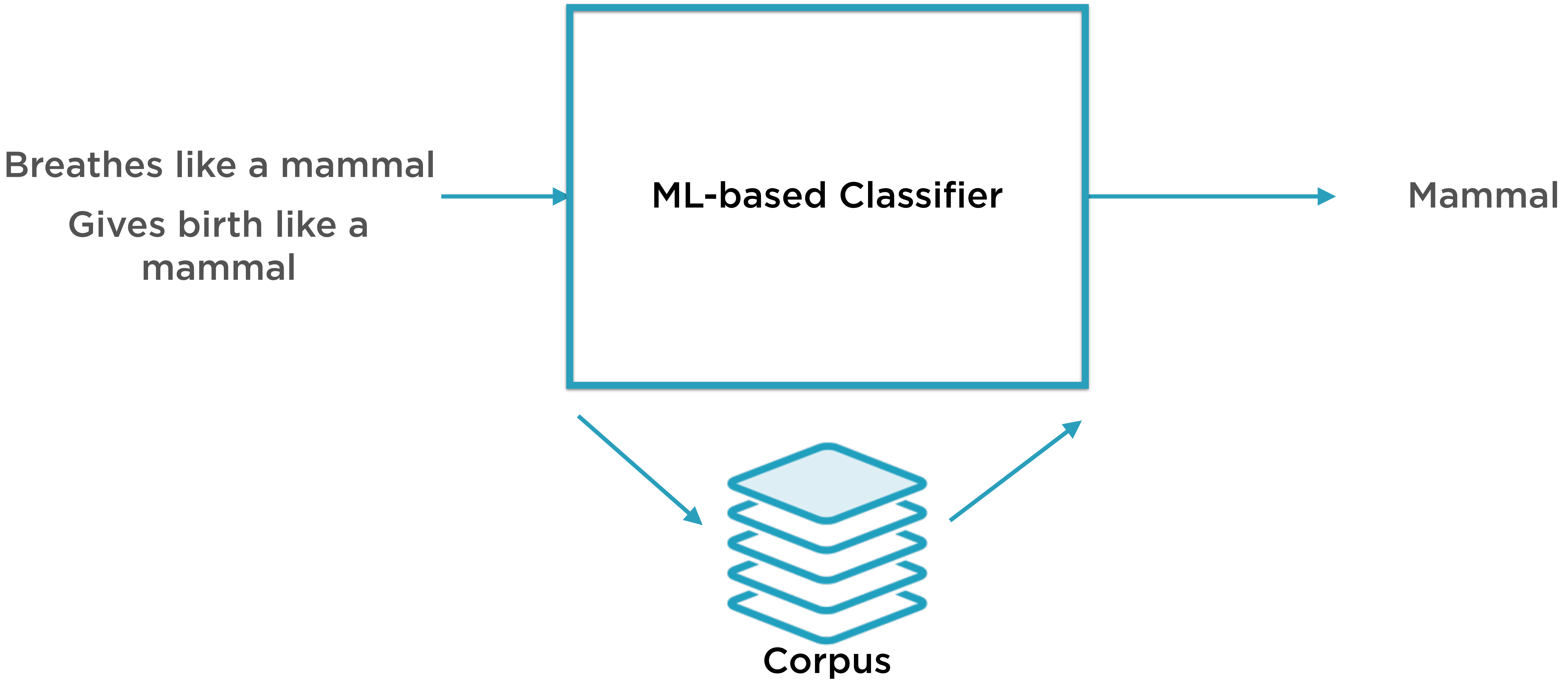
ML-based Binary Classifier



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

Two Approaches to Factor Extraction

Rule-based

Human experts identify and extract factors

ML-based

Algorithm identifies and extracts factors

What and How

Cut through clutter

Extract underlying factors from a set of data

Principal components analysis (PCA)

Cookie-cutter technique that finds the ‘good’ factors from a set of data points

PCA is one solution to the factor-extraction problem - a cookie-cutter solution

Applications of PCA

Dimensionality reduction

Cut through the clutter

Sparse data estimation

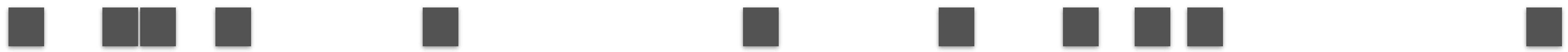
Estimate missing data

What-if risk analysis

Evaluate extreme scenarios

Mean and Variance

Data in One Dimension



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

Mean as Headline

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

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

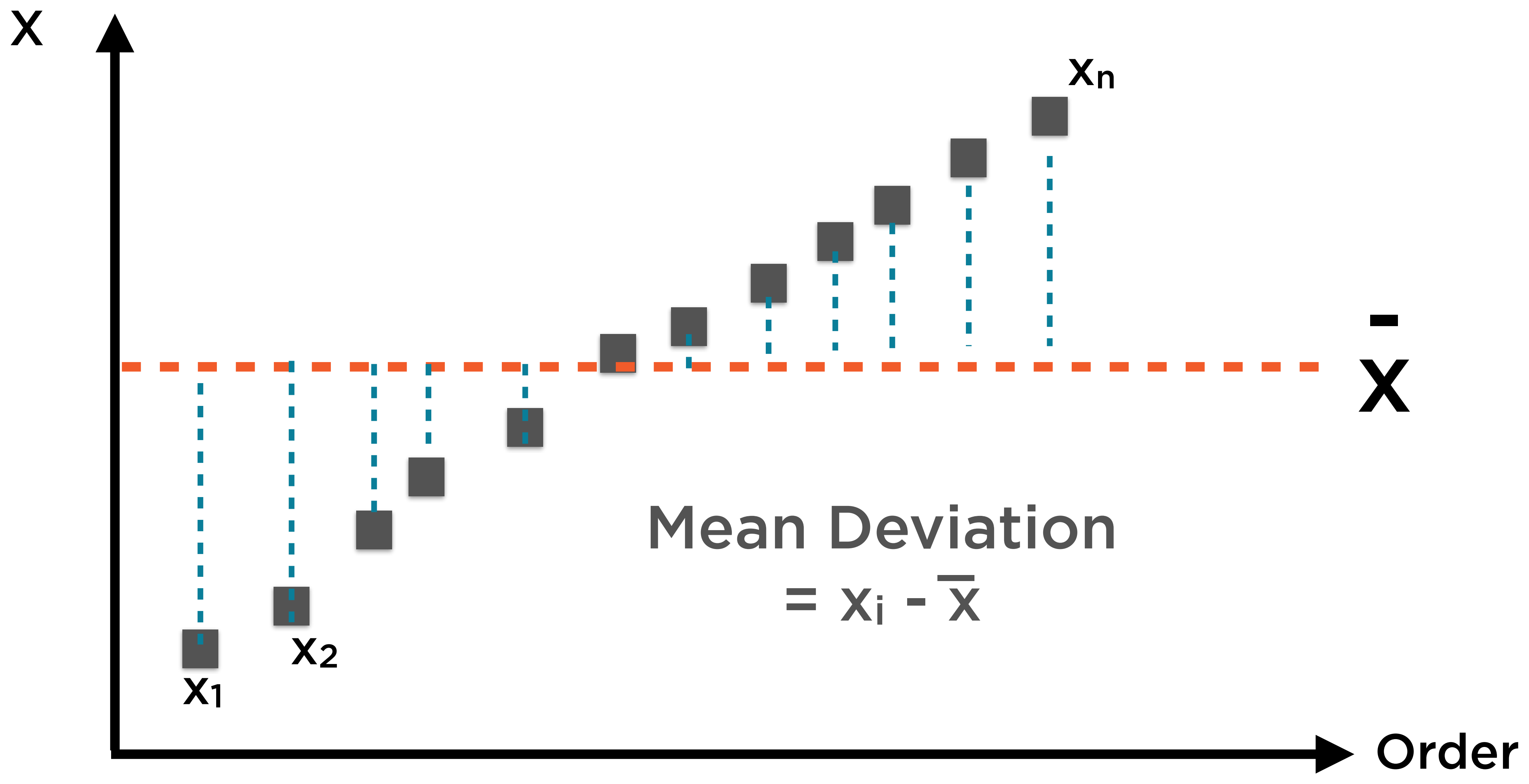
Variation Is Important Too

"Do the numbers jump around?"

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

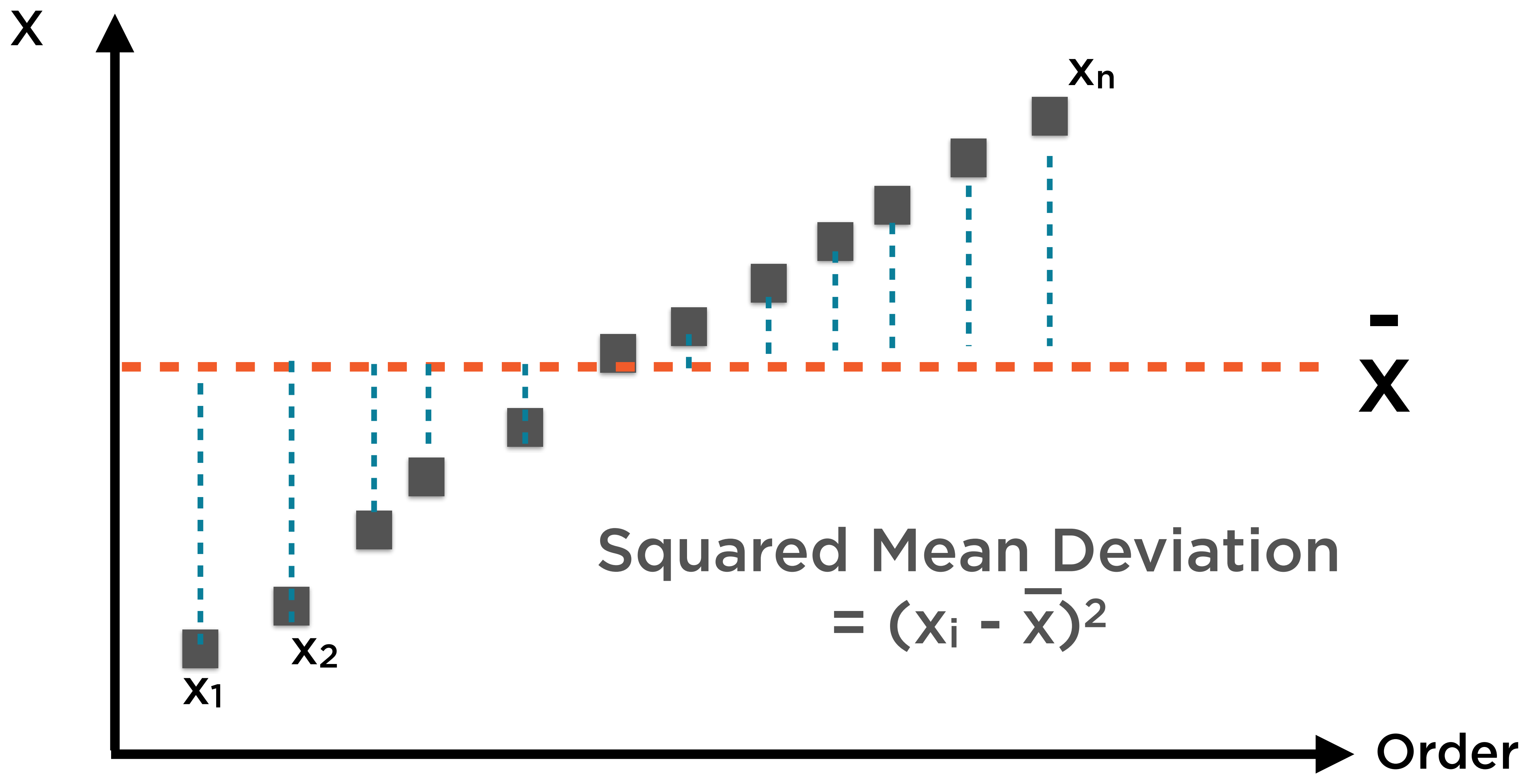
The range ignores the mean, and is swayed by outliers - that's where variance comes in

Variance as Asterisk



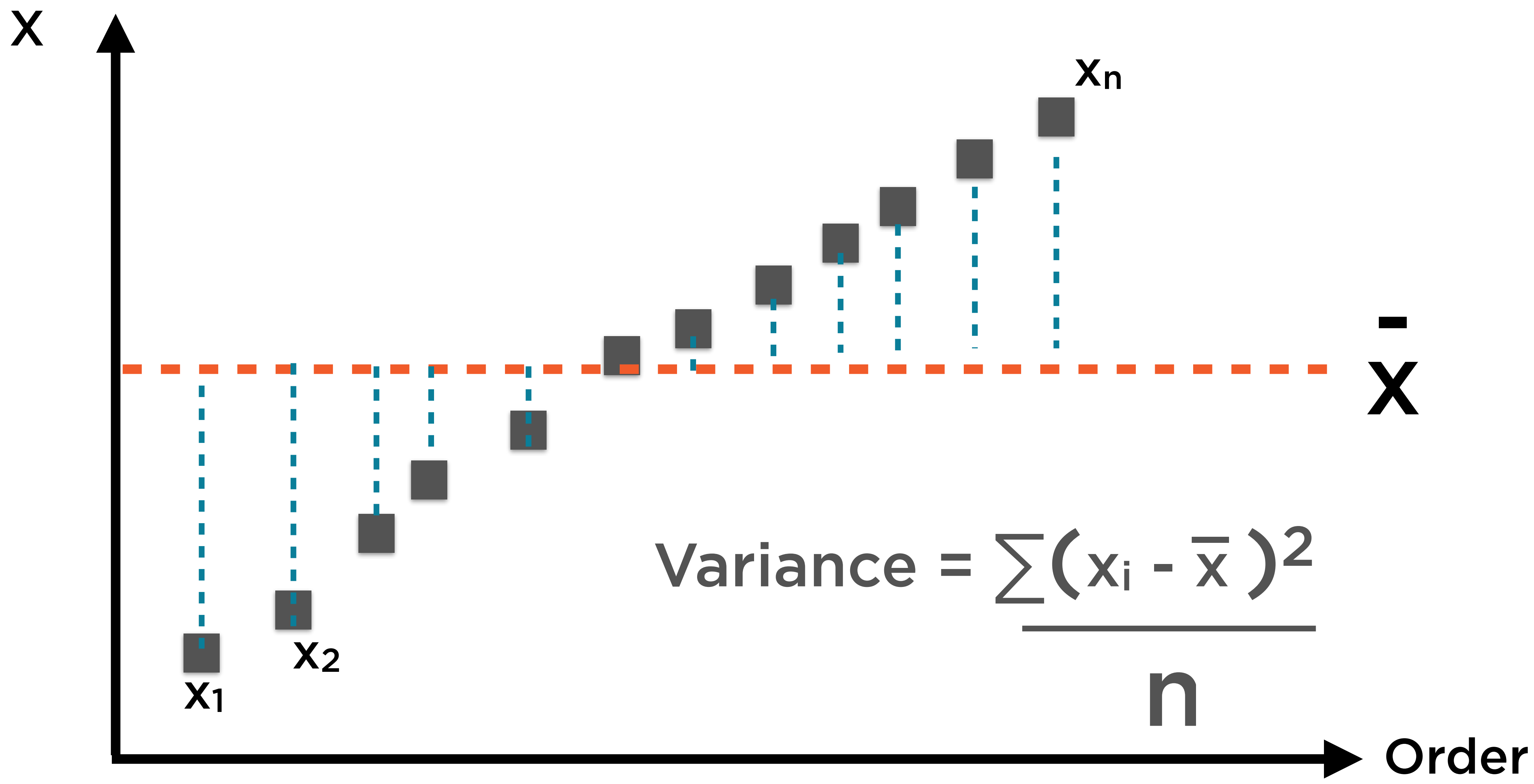
Variance is the second-most important number to summarise this set of data points

Variance as Asterisk



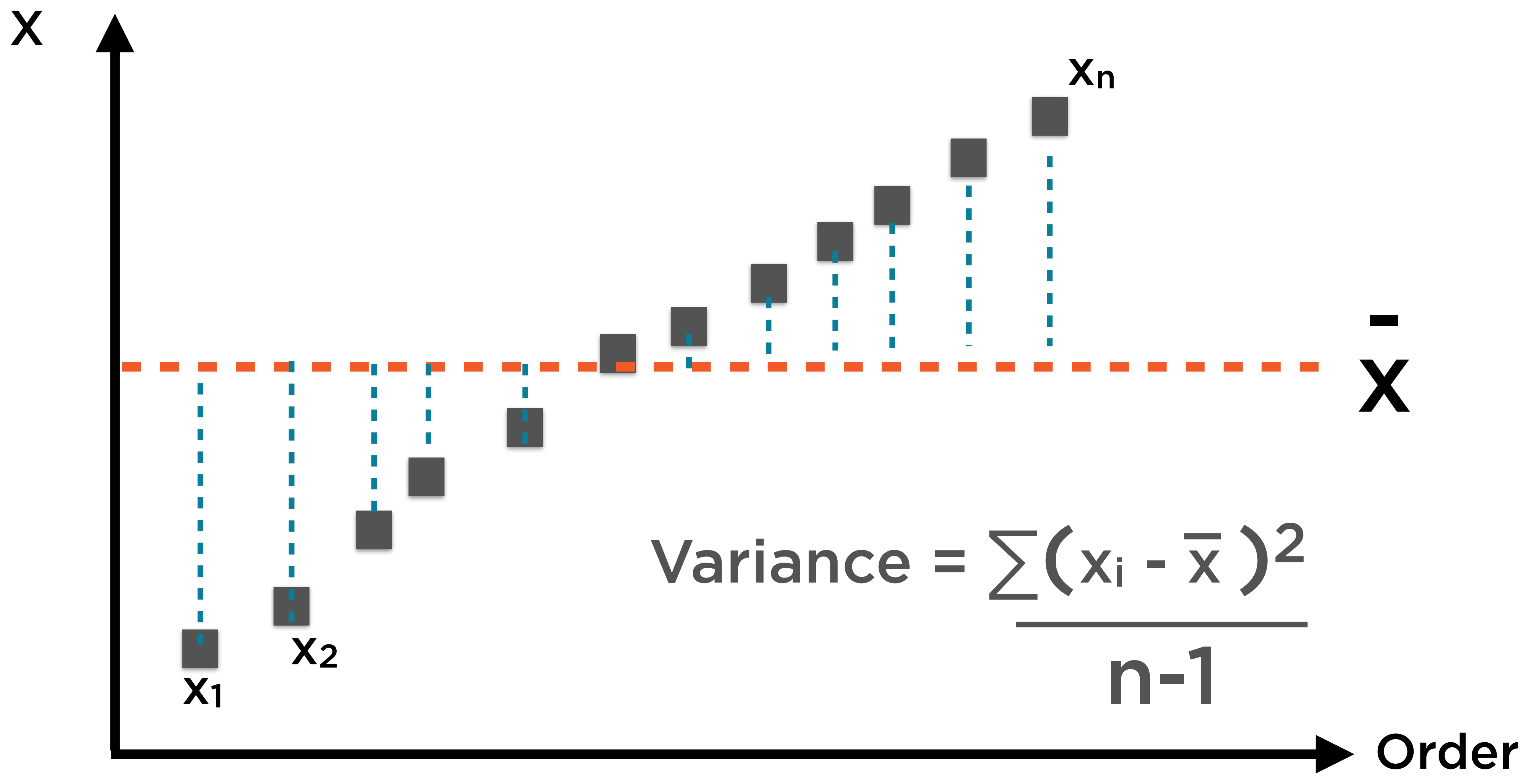
Variance is the second-most important number to summarise this set of data points

Variance as Asterisk



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

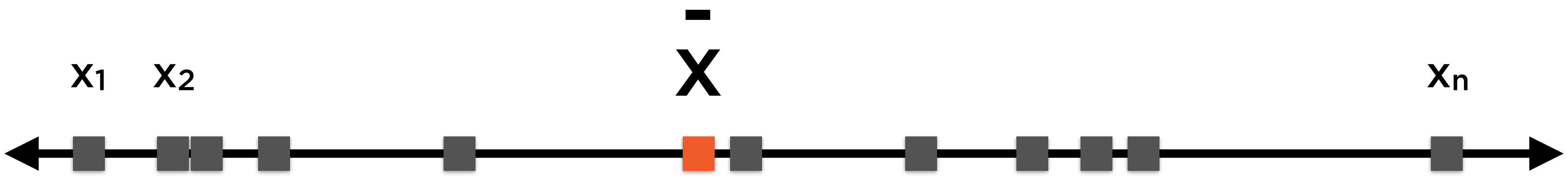
Variance as Asterisk



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

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

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

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

Tossing Two Coins

Coin X

Heads = $1

Tails = -$1

Coin Y

Heads = $1,000

Tails = -$1,000

Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

Tabulate the possible outcomes (assume each coin is a fair one)

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$$\bar{x} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$ $\bar{y} = 0$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0 \quad \bar{y} = 0$

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

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$ $\bar{y} = 0$

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
$1	1
$1	1
-$1	1
-$1	1

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

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$ $\bar{y} = 0$

$y_i - \bar{y}$	$(y_i - \bar{y})^2$
$1,000	1000000
-$1,000	1000000
$1,000	1000000
-$1,000	1000000

$$\text{Variance} = \frac{\sum(y_i - \bar{y})^2}{n} = 1{,}000{,}000$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$

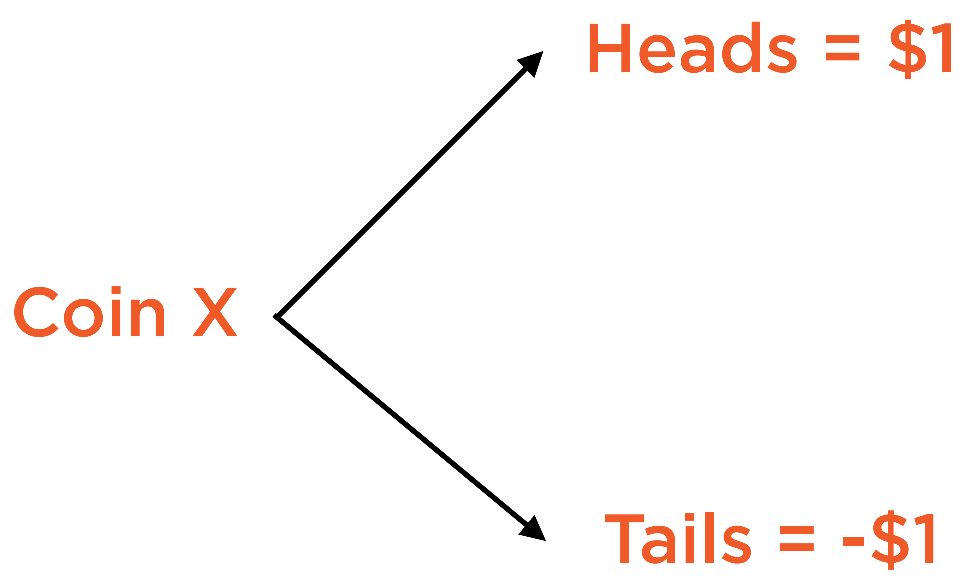
$\bar{y} = 0$

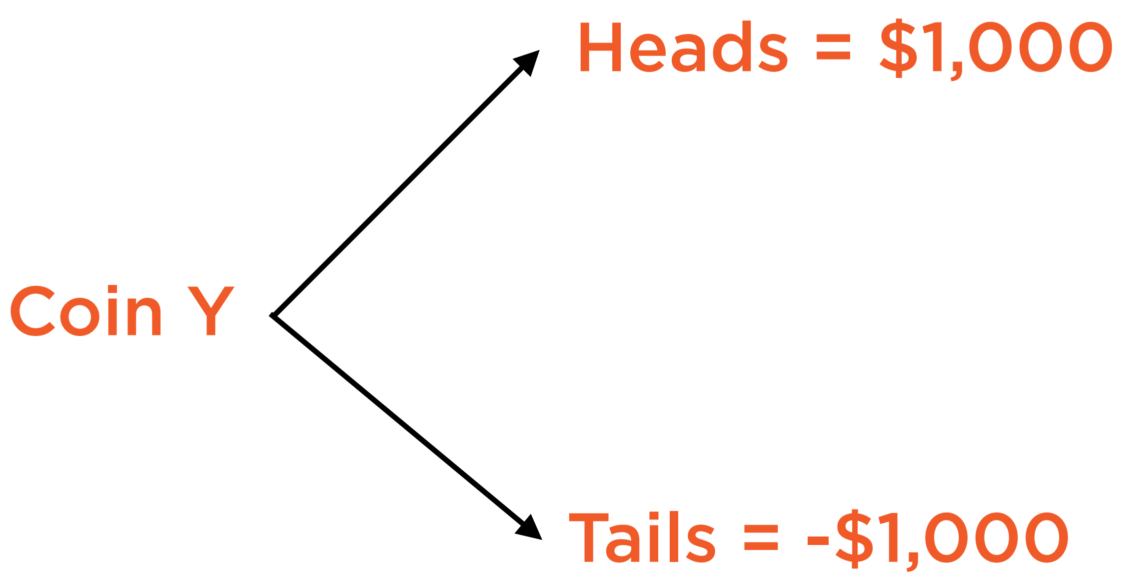
$Var(x) = 1$

$Var(y) = 1{,}000{,}000$

As stakes grow, variance gets big faster than the mean

Tossing Two Coins





Small Stakes

Loser pays $1, winner takes $1

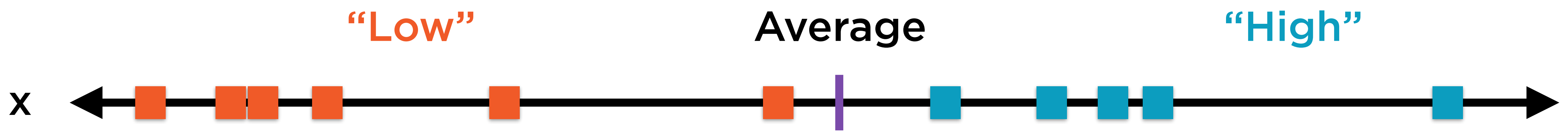
High Stakes

Loser pays $1000, winner takes $1000

As stakes grow 1000x, variance grows 1,000,000x

Covariance and Correlation

Data in One Dimension

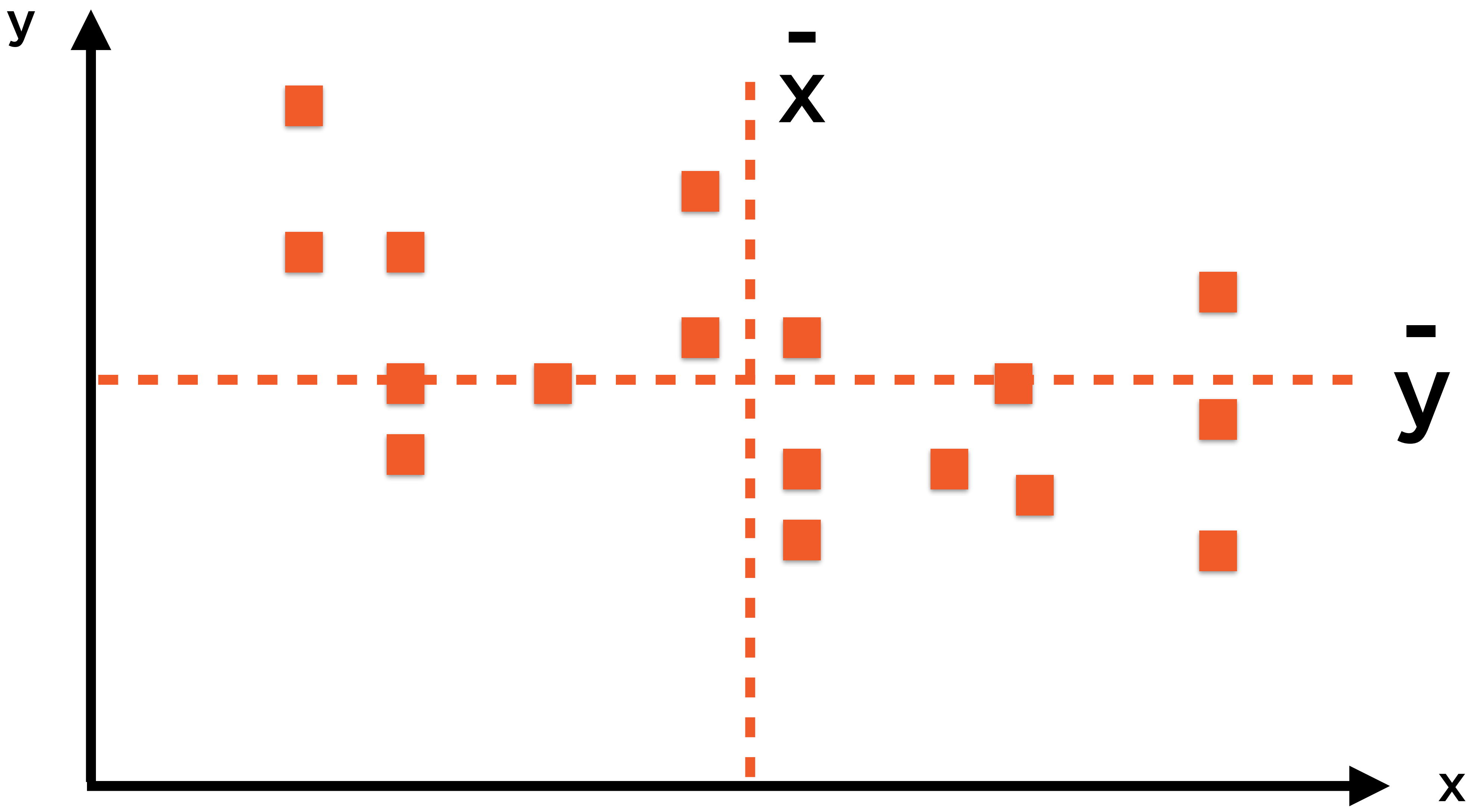


Unidimensional data is analysed using statistics such as mean, median, standard deviation

Data in Two Dimensions

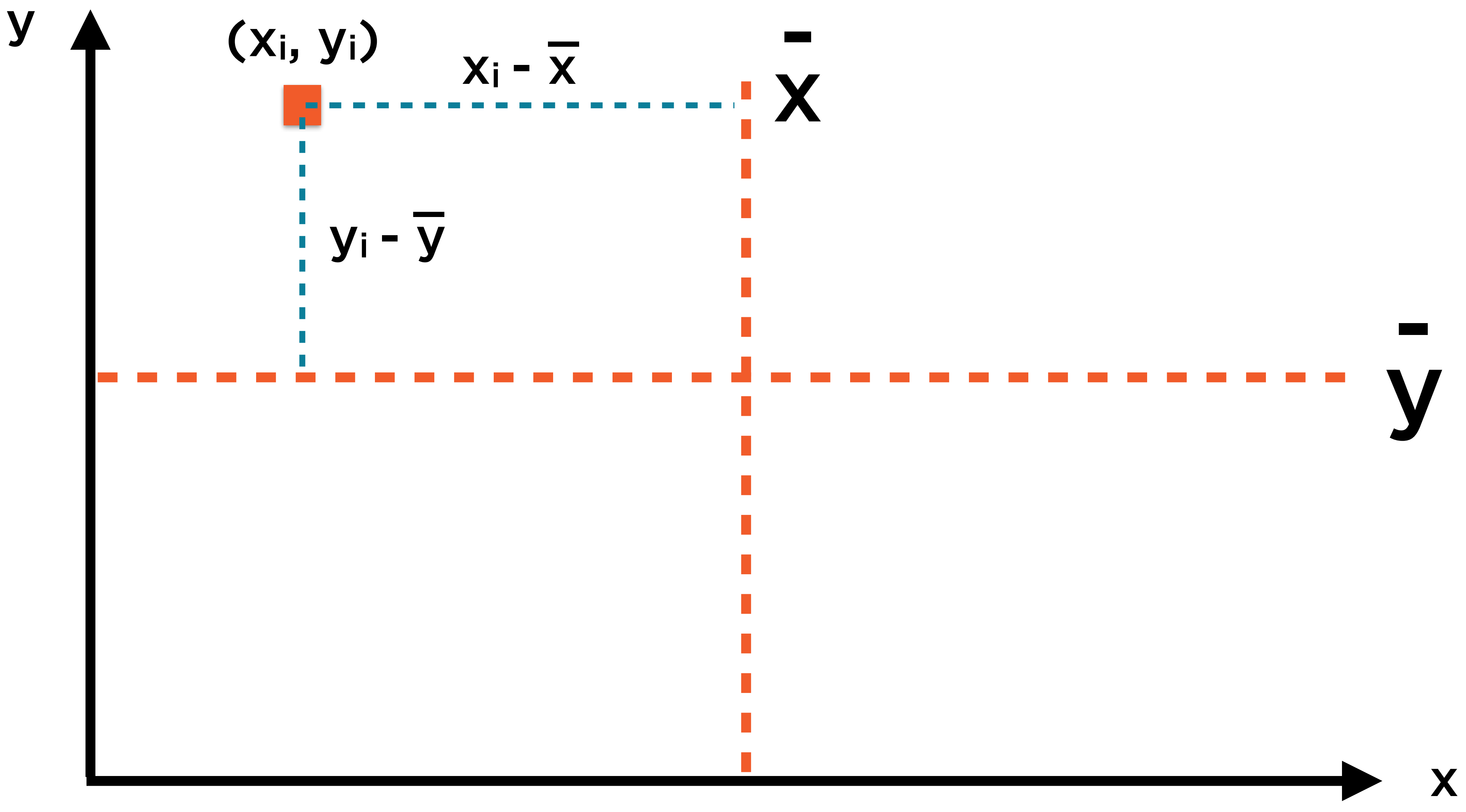
It's often more insightful to view data in relation to some other, related data

Covariance as Variance in Two Dimensions



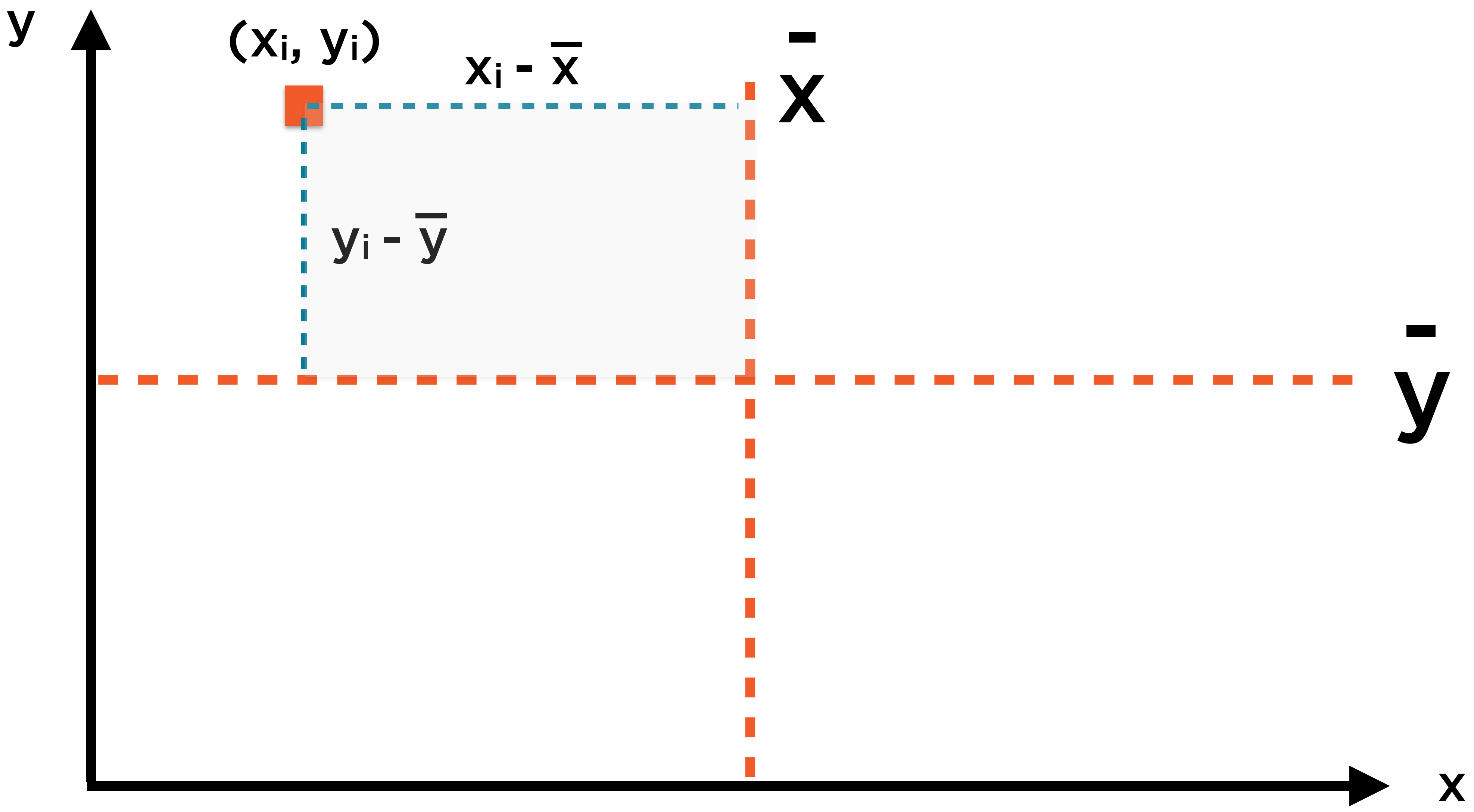
$$\text{Covariance}\ (x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariance as Variance in Two Dimensions



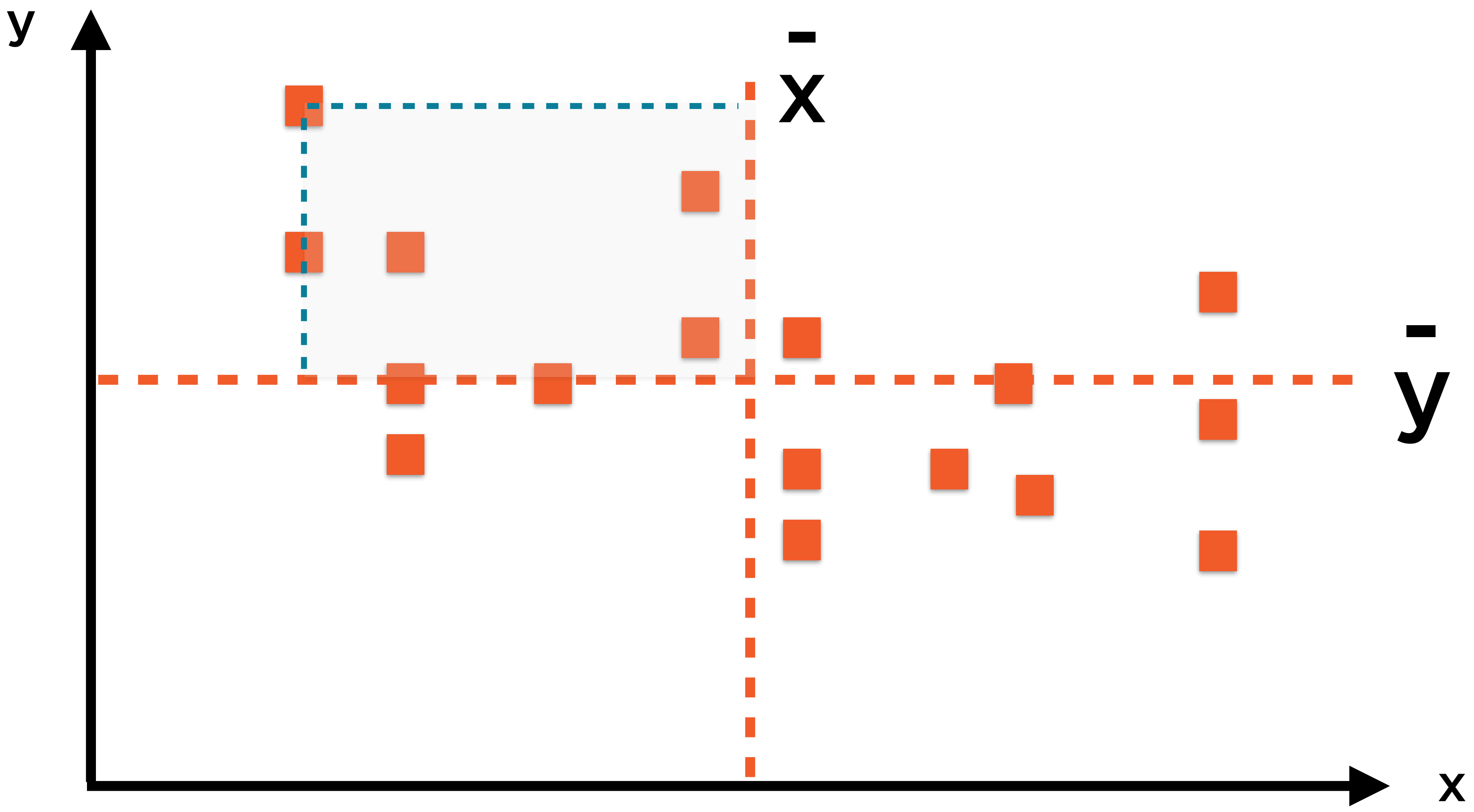
$$\text{Covariance}\ (x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariance as Variance in Two Dimensions



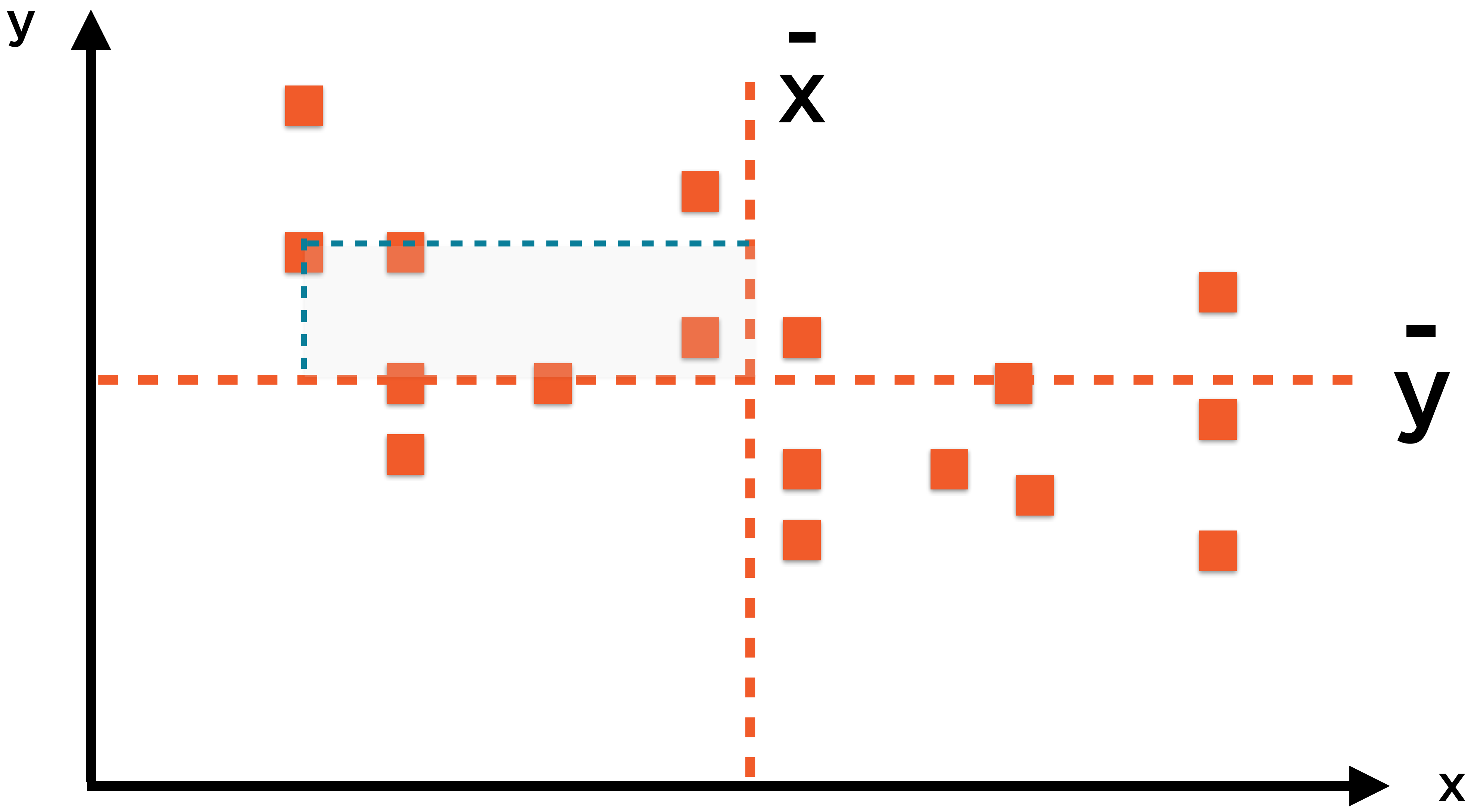
$$\text{Covariance }(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariance as Variance in Two Dimensions



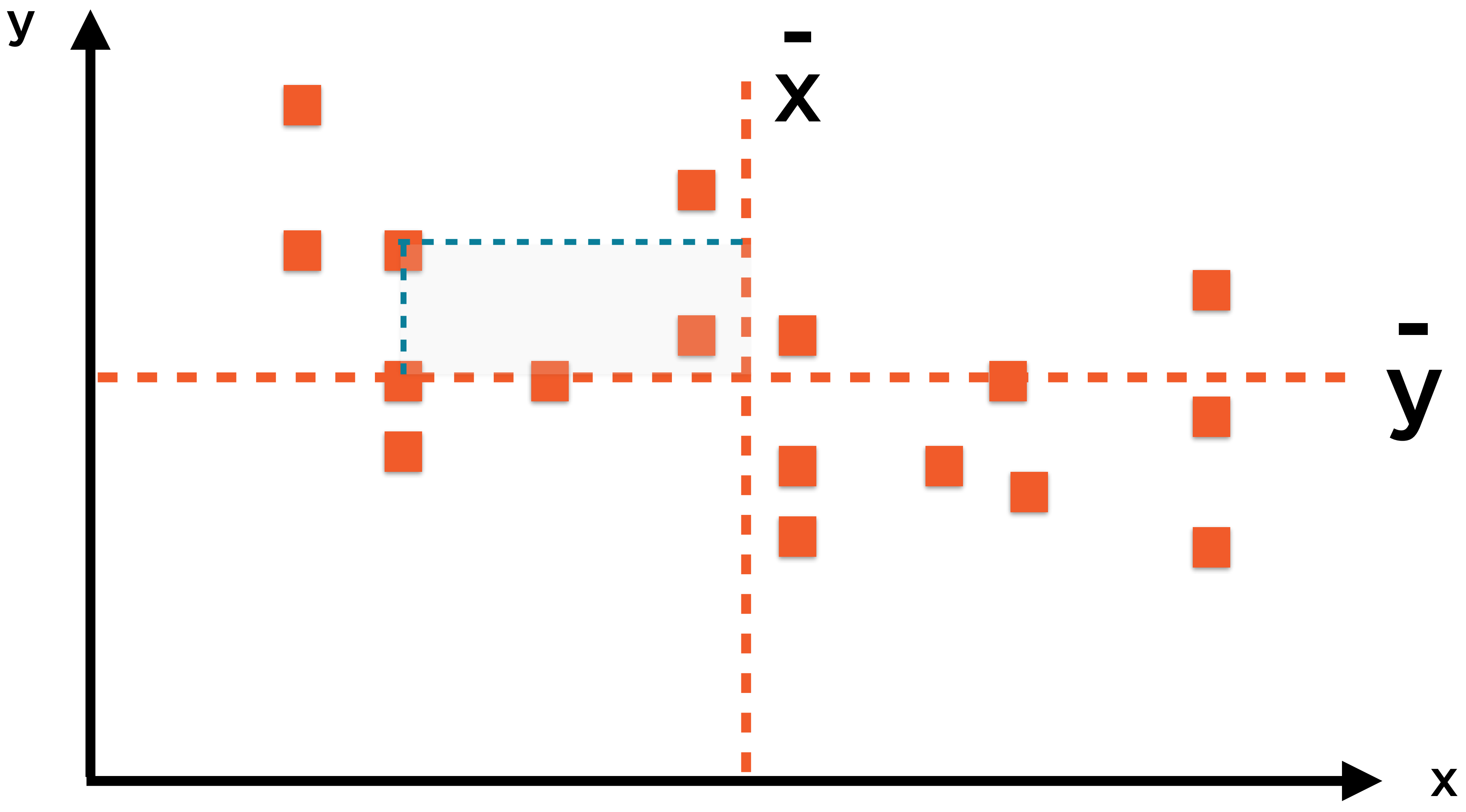
$$\text{Covariance}\ (x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariance as Variance in Two Dimensions



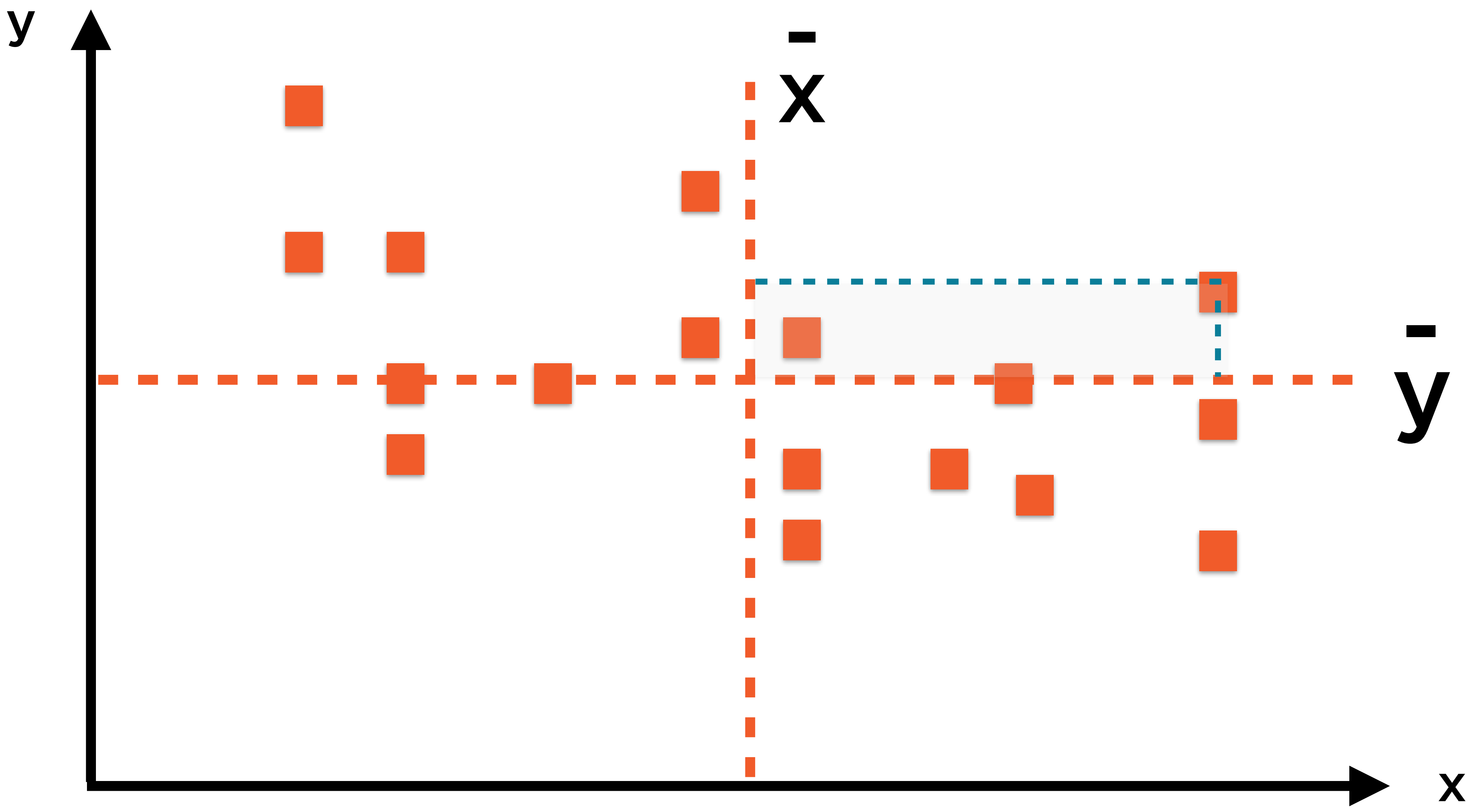
$$\text{Covariance}(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariance as Variance in Two Dimensions



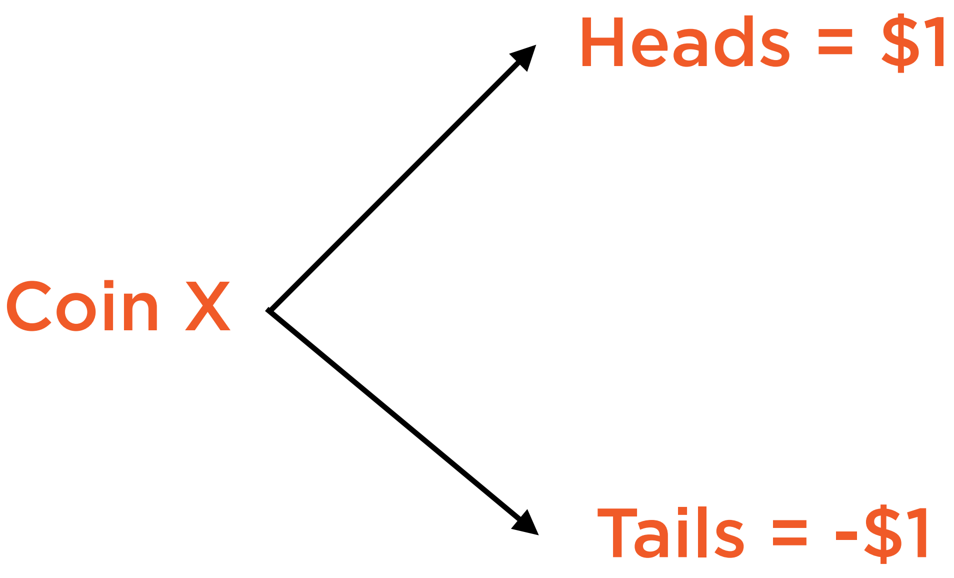
$$\text{Covariance}(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

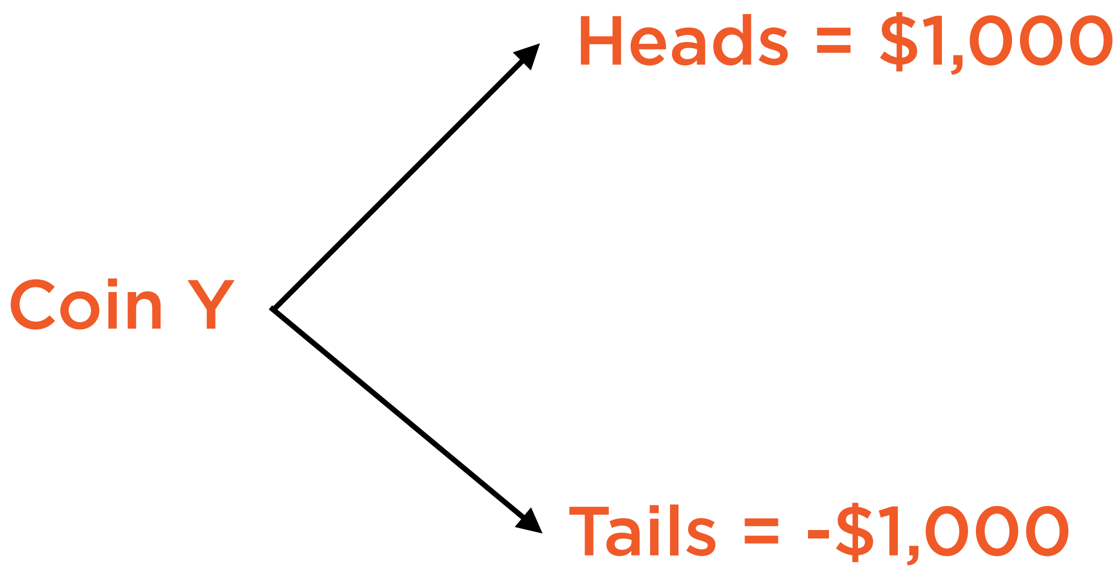
Covariance as Variance in Two Dimensions



$$\text{Covariance}\,(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Tossing Two Coins





Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$
Var(x) = 1

$\bar{y} = 0$
Var(y) = 1,000,000

$$\text{Covariance }(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
$1	$1,000	1,000
$1	-$1,000	-1,000
-$1	$1,000	-1,000
-$1	-$1,000	1,000

$\bar{x} = 0$
Var(x) = 1

$\bar{y} = 0$
Var(y) = 1,000,000

$$\text{Covariance }(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

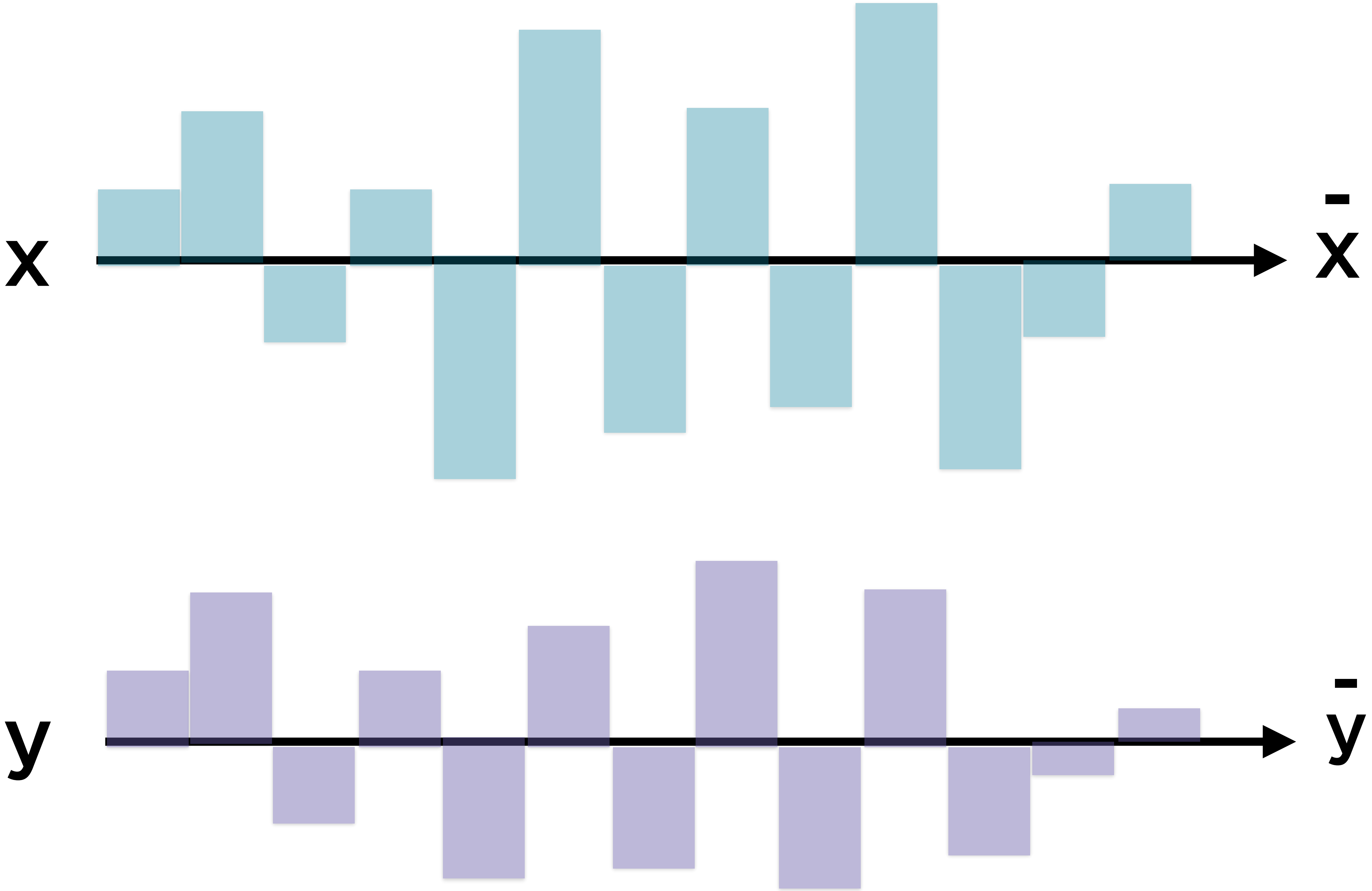
$\bar{x} = 0$ $\bar{y} = 0$

Var(x) = 1 Var(y) = 1,000,000

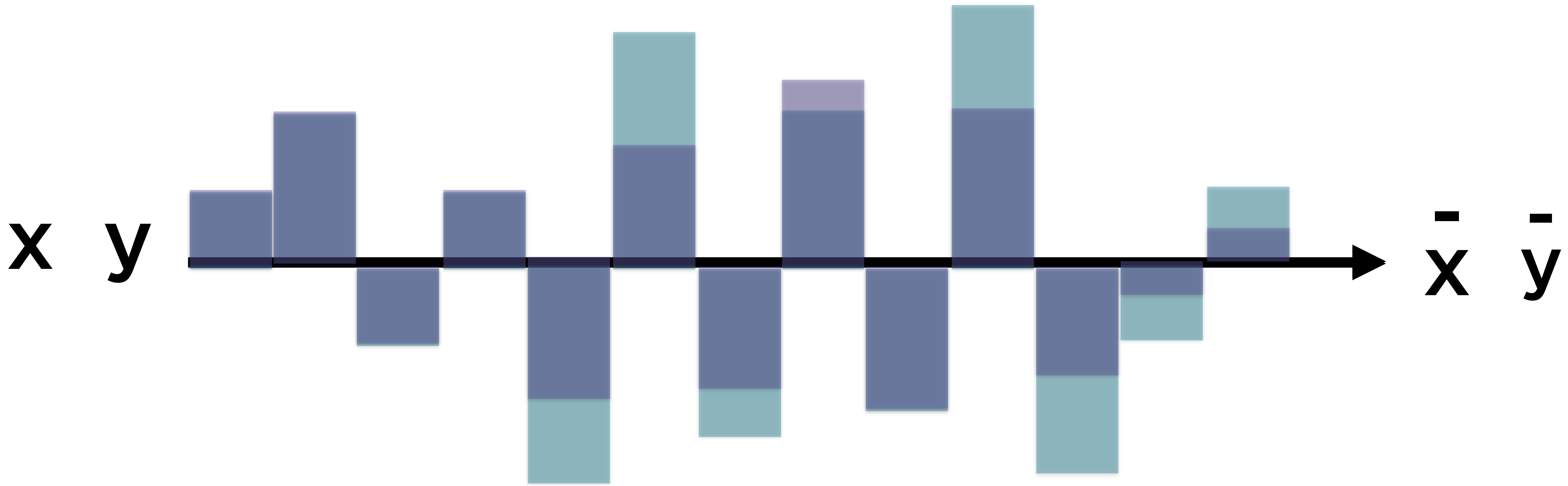
Covariance(x,y) = 0

Independent variables have zero covariance

Intuition: Positive Covariance

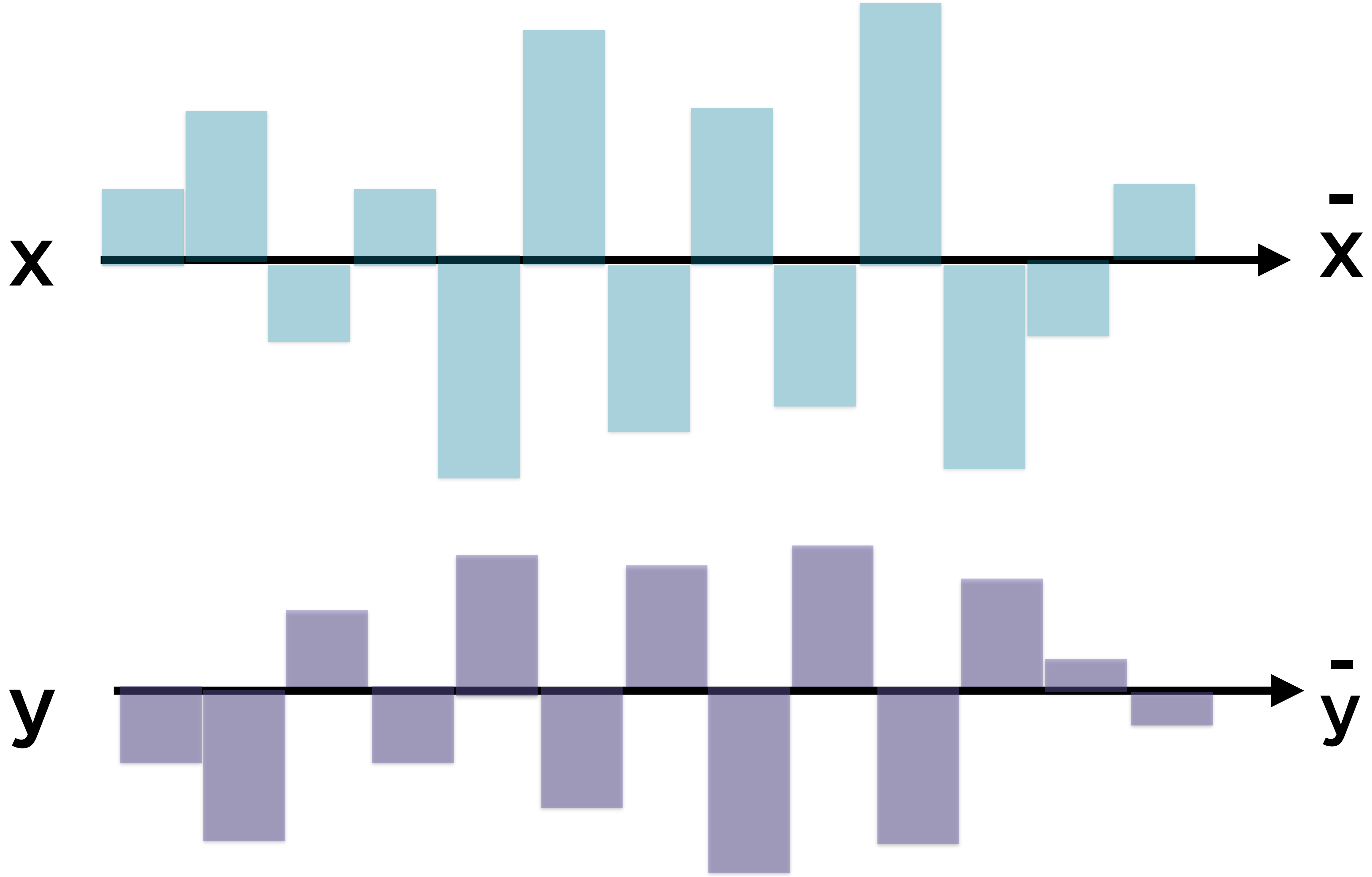


Intuition: Positive Covariance

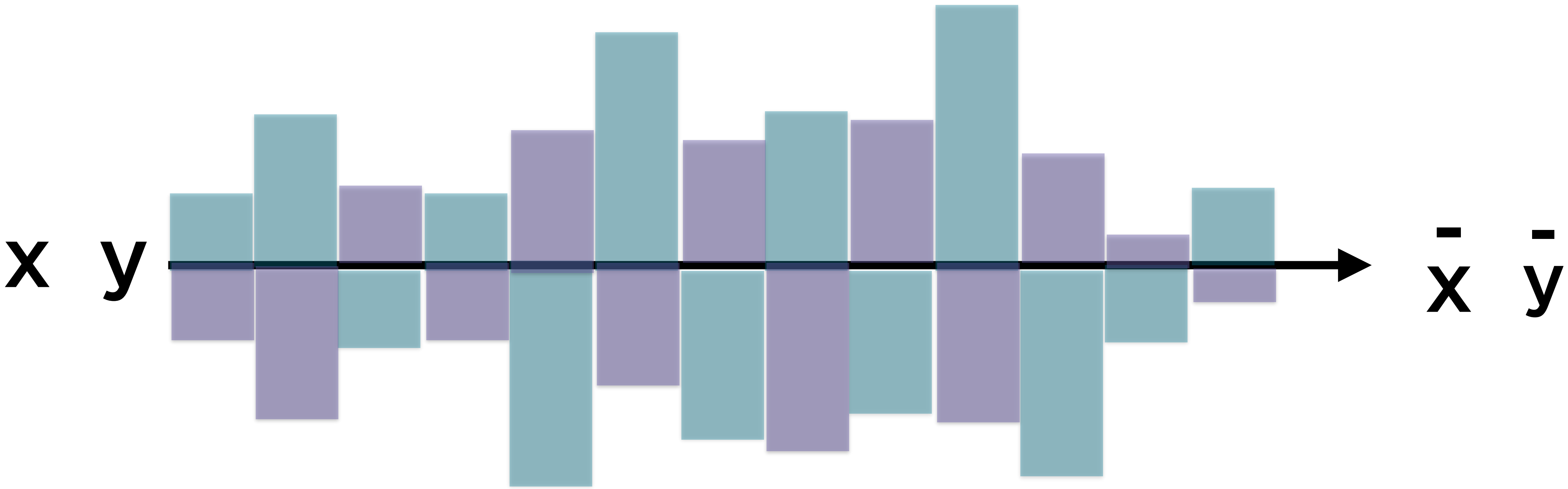


The deviations around the means of the two series are in-sync

Intuition: Negative Covariance

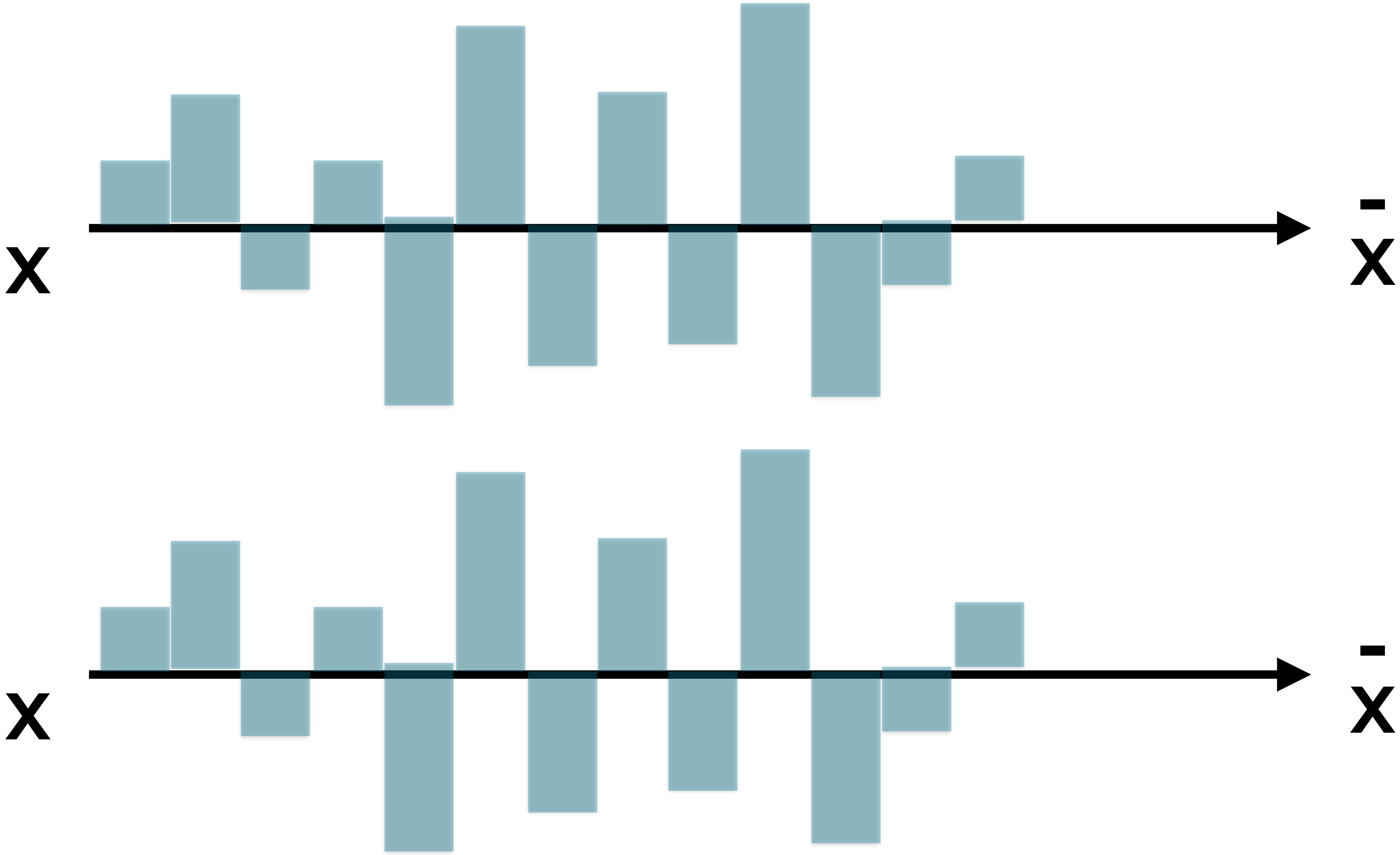


Intuition: Negative Covariance

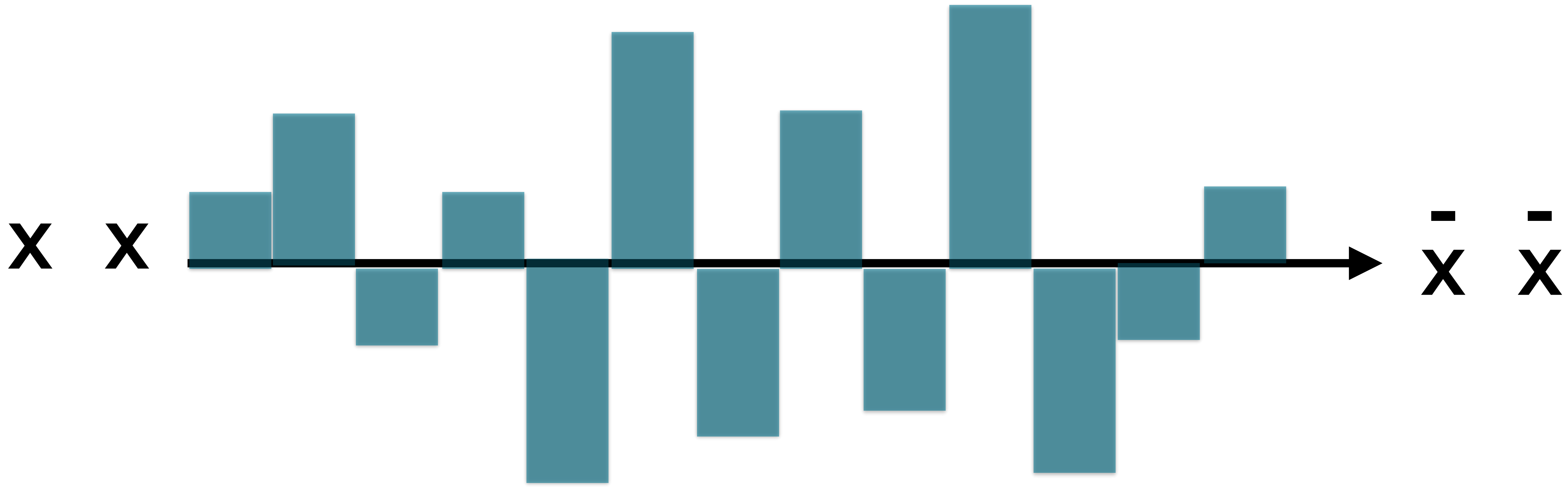


The deviations around the means of the two series are out-of-sync

Intuition: Covariance and Variance



Intuition: Positive Covariance



Variance is the covariance of a series with itself

Covariance and Variance

$$\text{Covariance}\,(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

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

$$\text{Variance}\,(y) = \sum \frac{(y_i - \bar{y})^2}{n} = \text{Covariance}\,(y,y)$$

Random variables are outcomes of uncertain events

- Coin tosses
- Dice rolls
- Sporting events
- Stock returns

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \dots \\ E_n \end{bmatrix} \qquad \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \dots \\ D_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

Returns (percentage changes) in the prices of two financial assets over time

- Exxon stock
- Dow Jones equity index

These returns are related to each other

Many Random Variables

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \dots \\ E_n \end{bmatrix} \quad \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \dots \\ D_n \end{bmatrix} \quad \begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ \dots \\ G_n \end{bmatrix} \quad \dots \quad \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \dots \\ A_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

G_i = % return of Google stock on day i

A_i = % return of Apple stock on day i

Summarising into a Matrix

$$\begin{bmatrix} E_1 & D_1 & G_1 & & A_1 \\ E_2 & D_2 & G_2 & & A_2 \\ E_3 & D_3 & G_3 & \cdots & A_3 \\ \cdots & \cdots & \cdots & & \cdots \\ E_n & D_n & G_n & & A_n \end{bmatrix}$$

n rows

k columns

Summarise the returns of k stocks, each over n days, into an nxk matrix

Summarising into a Matrix

$$\begin{array}{cl} \begin{bmatrix} x_{11} & x_{12} & x_{13} & & x_{1k} \\ x_{21} & x_{22} & x_{23} & & x_{2k} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & x_{n3} & & x_{nk} \end{bmatrix} & \text{n rows} \\ \text{k columns} & \end{array}$$

Summarise the returns of k stocks, each over n days, into an nxk matrix

Summarising into a Matrix

$$
\begin{bmatrix}
X_{11} & X_{12} & X_{13} & & X_{1k} \\
X_{21} & X_{22} & X_{23} & & X_{2k} \\
X_{31} & X_{32} & X_{33} & \dots & X_{3k} \\
\dots & \dots & \dots & & \dots \\
X_{n1} & X_{n2} & X_{n3} & & X_{nk}
\end{bmatrix}
\quad \text{n rows}
$$

X_1 (n rows, 1 column)

k columns

Summarising into a Matrix

$$\left[\begin{array}{ccccc} X_{11} & \boxed{X_{12}} & X_{13} & & X_{1k} \\ X_{21} & X_{22} & X_{23} & & X_{2k} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ X_{n1} & X_{n2} & X_{n3} & & X_{nk} \end{array}\right] \quad \text{n rows}$$

X_2 (n rows, 1 column)

k columns

Summarising into a Matrix

$$\begin{array}{c} \left[\begin{array}{ccccc} X_{11} & X_{12} & X_{13} & & X_{1k} \\ X_{21} & X_{22} & X_{23} & & X_{2k} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3k} \\ \cdots & \cdots & \cdots & & \cdots \\ X_{n1} & X_{n2} & X_{n3} & & X_{nk} \end{array}\right] \quad \text{n rows} \\ \text{k columns} \end{array}$$

X_k (n rows, 1 column)

Summarising into a Matrix

$$[\ X_1 \quad X_2 \quad X_3 \quad \dots \quad X_k\ ]$$

n rows

k columns

Each element X_i of this matrix is a vector with 1 column and n rows

A covariance matrix summarises the covariances of columns in a data matrix

Covariance Matrix

$$[\;X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\;]$$

$$\begin{array}{c}
\begin{bmatrix}
Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\
Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_k) \\
Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Cov(X_k, X_k)
\end{bmatrix} \text{ k rows} \\
\text{k columns}
\end{array}$$

Each element of the covariance matrix contains the covariance of a pair of vectors from the original data

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\underbrace{\begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_k) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_k) \\ \mathrm{Cov}(X_k, X_1) & \mathrm{Cov}(X_k, X_2) & \cdots & \mathrm{Cov}(X_k, X_k) \end{bmatrix}}_{\text{k columns}} \;\text{k rows}$$

The first row contains the covariance of the first column X_1 with each of the columns (including itself)

Covariance Matrix

$$[\;X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\;]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Cov(X_k, X_k) \end{bmatrix} \text{k rows}$$

k columns

The first row contains the covariance of the first column X_1 with each of the columns (including itself)

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The first row contains the covariance of the first column X_1 with each of the columns (including itself)

Covariance Matrix

$$[\ X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\ ]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The last row contains the covariance of the last column X_k with each of the columns (including itself)

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Cov(X_k, X_k) \end{bmatrix} \quad \text{k rows}$$

k columns

The last row contains the covariance of the last column X_k with each of the columns (including itself)

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The last row contains the covariance of the last column X_k with each of the columns (including itself)

Covariance Matrix

$$[\ X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\ ]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \dots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \dots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The matrix is symmetric - the value at row i and column j is the same as that at row j and column i

Covariance Matrix

$$[\ X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\ ]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The matrix is symmetric - the value at row i and column j is the same as that at row j and column i

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The values along the diagonal are the variances of the corresponding columns

Covariance and Variance

$$\text{Covariance}(x,y) = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

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

$$\text{Variance}(y) = \sum \frac{(y_i - \bar{y})^2}{n} = \text{Covariance}(y,y)$$

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Cov(X_k, X_k) \end{bmatrix}$$

k rows

k columns

The values along the diagonal are the variances of the corresponding columns

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Var(X_k) \end{bmatrix}$$

k rows

k columns

The values along the diagonal are the variances of the corresponding columns

Covariance Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Var(X_k) \end{bmatrix}$$

k rows

k columns

Covariance Matrix

$$[\ X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\ ]$$

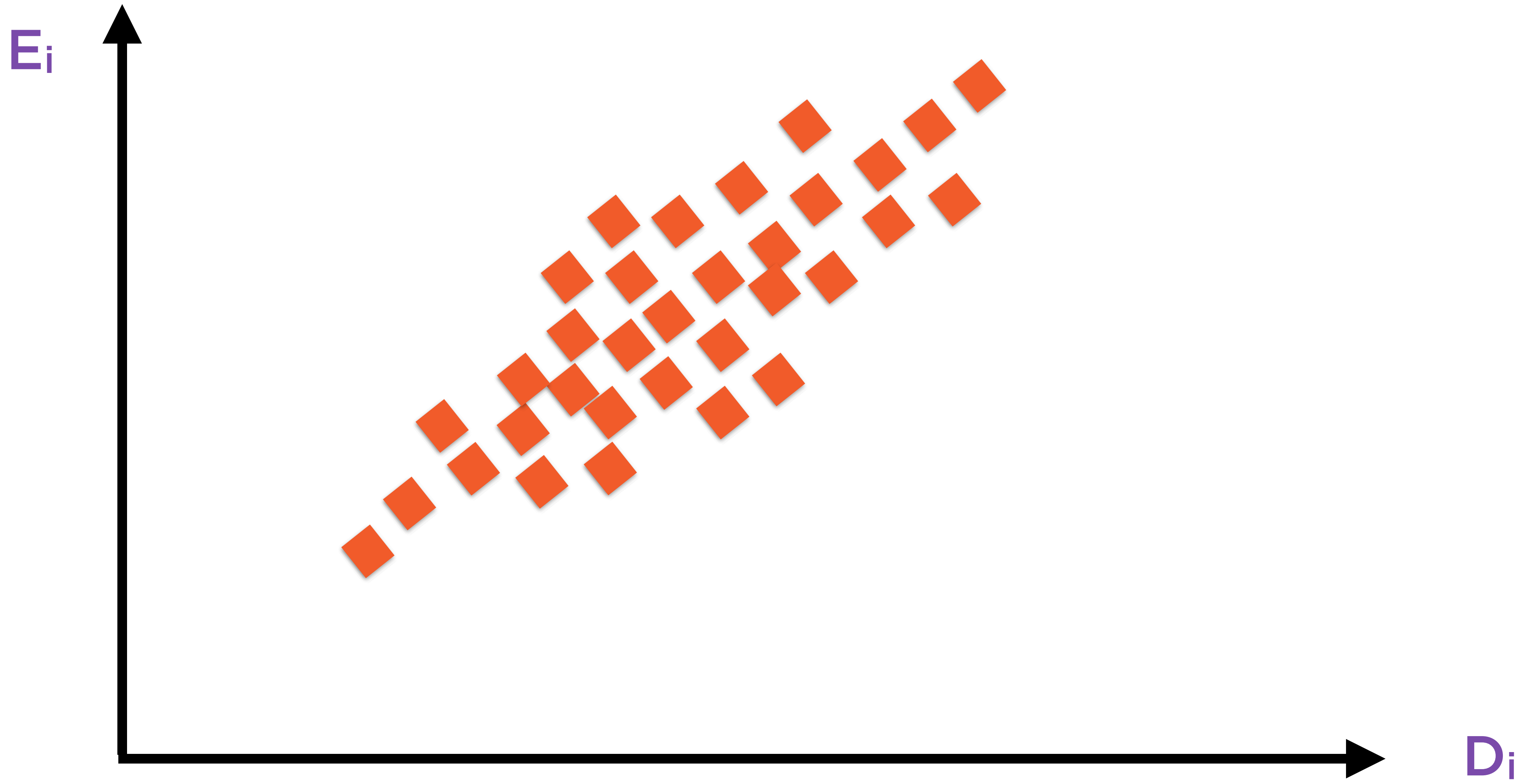
$$\begin{bmatrix} \sigma^2_{x_1} & \sigma^2_{x_1x_2} & \cdots & \sigma^2_{x_1x_k} \\ \sigma^2_{x_2x_1} & \sigma^2_{x_2} & \cdots & \sigma^2_{x_2x_k} \\ \sigma^2_{x_kx_1} & \sigma^2_{x_kx_2} & \cdots & \sigma^2_{x_k} \end{bmatrix}$$

k rows

k columns

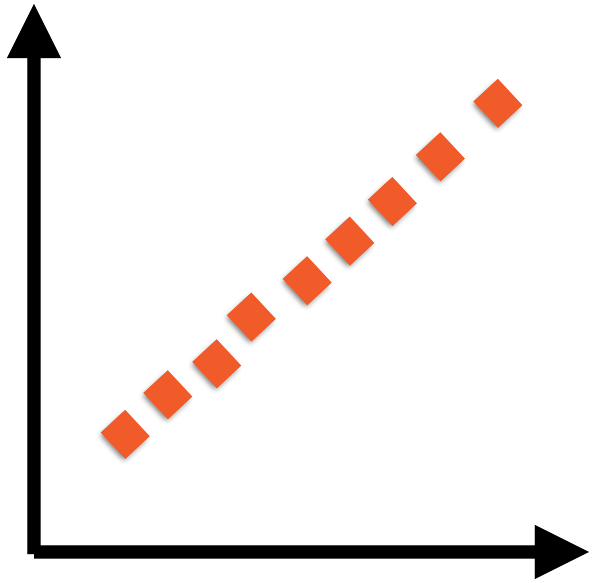
Each element of the covariance matrix contains the covariance of a pair of vectors from the original data

Correlated Random Variables



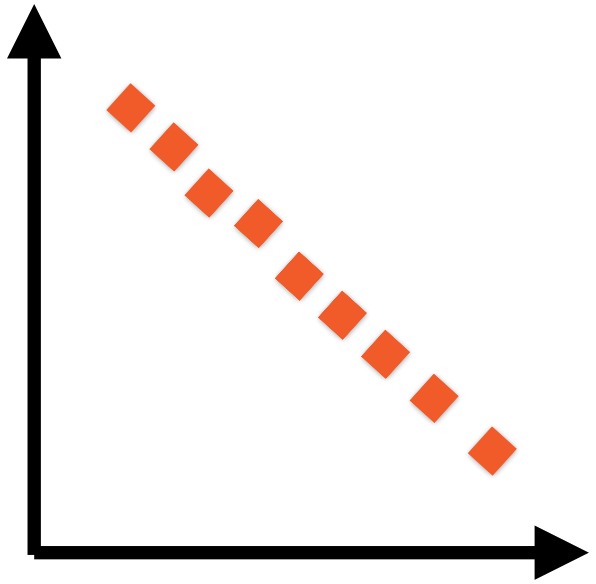
Returns on the Dow and on Exxon are related to each other

Correlation Captures Linear Relationships



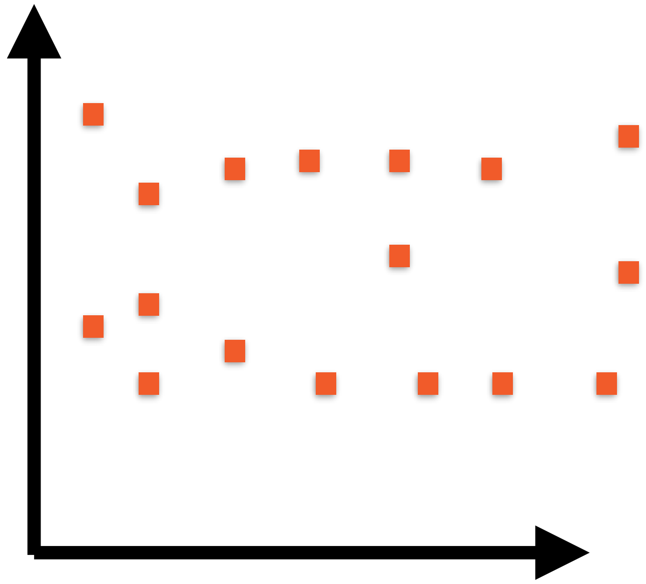
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y

Correlation and Covariance

$$\text{Correlation }(x,y) = \frac{\text{Covariance }(x,y)}{\sqrt{\text{Variance }(x)}\ \sqrt{\text{Variance }(y)}}$$

$$\rho_{xy} = \frac{\sigma^2_{xy}}{\sigma_x \sigma_y}$$

Correlation and Covariance

$$\rho_{xy} = \frac{\sigma^2_{xy}}{\sigma_x \sigma_y}$$

$$\rho_{xx} = \frac{\sigma^2_{x}}{\sigma_x \sigma_x}$$

$$= 1$$

Correlation of any series with itself is always +1

Covariance Matrix

$$[\ X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k\ ]$$

$$\begin{bmatrix} \sigma^2_{x_1} & \sigma^2_{x_1x_2} & \cdots & \sigma^2_{x_1x_k} \\ \sigma^2_{x_2x_1} & \sigma^2_{x_2} & \cdots & \sigma^2_{x_2x_k} \\ \sigma^2_{x_kx_1} & \sigma^2_{x_kx_2} & \cdots & \sigma^2_{x_k} \end{bmatrix}$$

k rows

k columns

Each element is the covariance of two random variables

Correlation Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{bmatrix} \rho_{x_1} & \rho_{x_1 x_2} & \cdots & \rho_{x_1 x_k} \\ \rho_{x_2 x_1} & \rho_{x_2} & \cdots & \rho_{x_2 x_k} \\ \rho_{x_k x_1} & \rho_{x_k x_2} & \cdots & \rho_{x_k} \end{bmatrix}$$

k rows

k columns

Each element is the correlation of two random variables

Correlation Matrix

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

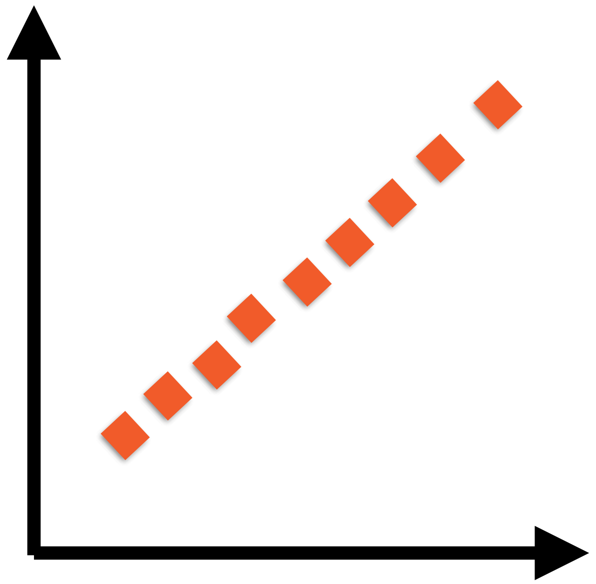
$$\begin{bmatrix} 1 & \rho_{x_1x_2} & \cdots & \rho_{x_1x_k} \\ \rho_{x_2x_1} & 1 & \cdots & \rho_{x_2x_k} \\ \rho_{x_kx_1} & \rho_{x_kx_2} & \cdots & 1 \end{bmatrix} \quad \text{k rows}$$

k columns

Diagonal elements are always 1

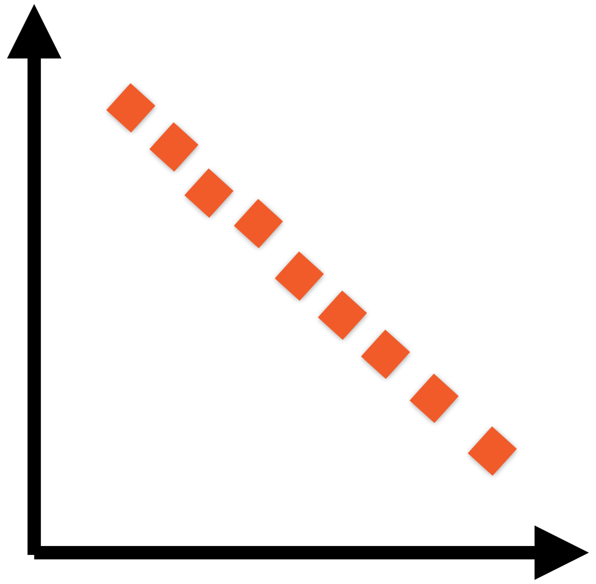
Independent variables have zero covariance and zero correlation

Correlation Captures Linear Relationships



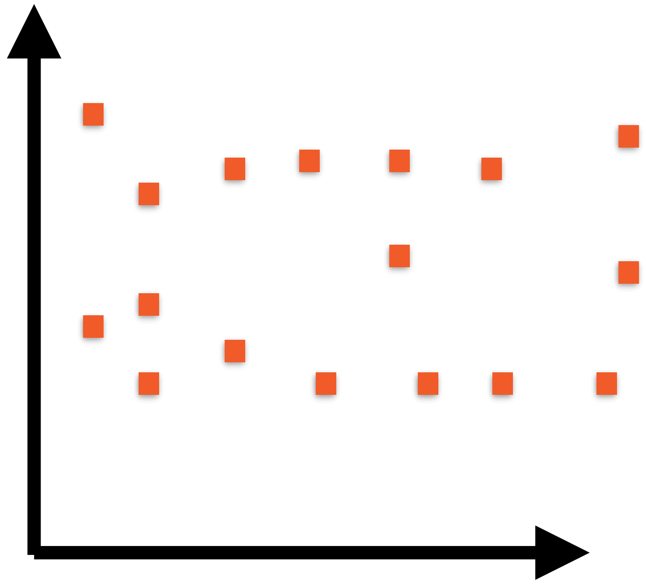
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y

Correlation Matrix of Independent Variables

$$\begin{array}{c} [\,X_1 \qquad X_2 \qquad X_3 \qquad \cdots \qquad X_k\,] \\ \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}$$

k rows

k columns

Correlation matrix of independent variables is the identity matrix

Covariance Matrix of Independent Variables

$$[\; X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_k \;]$$

$$\begin{array}{c}
\begin{bmatrix}
\sigma^2_{x_1} & 0 & \cdots & 0 \\
0 & \sigma^2_{x_2} & \cdots & 0 \\
0 & 0 & \cdots & \sigma^2_{x_k}
\end{bmatrix}
\end{array}
\quad \text{k rows}$$

k columns

Covariance matrix of independent variables is a **diagonal matrix**

Tossing Two Coins

Coin X
- Heads = $1
- Tails = -$1

Coin Y
- Heads = $1,000
- Tails = -$1,000

Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin X Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$ $\bar{y} = 0$

Var(x) = 1 Var(y) = 1,000,000

Covariance(x,y) = 0

Independent variables have zero covariance

Covariance Matrix of Two Coin Tosses

$$\begin{bmatrix} 1 & 0 \\ 0 & 1{,}000{,}000 \end{bmatrix}$$

2 rows

2 columns

Diagonal elements are variances, off-diagonal elements are covariances

Correlation Matrix of Two Coin Tosses

$$\begin{array}{cc} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \text{2 rows} \\ \text{2 columns} & \end{array}$$

Correlation matrix of independent variables is the identity matrix

Adding Random Variables

$$\begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \dots \\ E_n \end{bmatrix} \quad \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \dots \\ D_n \end{bmatrix} \quad \begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ \dots \\ G_n \end{bmatrix} \quad \dots \quad \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \dots \\ A_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

G_i = % return of Google stock on day i

A_i = % return of Apple stock on day i

Adding Random Variables

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ \cdots \\ P_n \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \cdots \\ E_n \end{bmatrix} + \begin{bmatrix} D_1 \\ D_2 \\ D_3 \\ \cdots \\ D_n \end{bmatrix} + \begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ \cdots \\ G_n \end{bmatrix} + \cdots + \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \cdots \\ A_n \end{bmatrix}$$

E_i = % return on Exxon stock on day i

D_i = % return of Dow Jones index on day i

G_i = % return of Google stock on day i

A_i = % return of Apple stock on day i

Adding Random Variables

$$P = E + D + G \ldots + A$$

P_i = % return of stock portfolio on day i

Portfolio P consists of 1 stock each of Exxon, the Dow, Google and Apple

Adding Random Variables

$$P = w_1E + w_2D + w_3G \ldots + w_kA$$

P_i = % return of stock portfolio on day i

Portfolio P consists of w_1 stocks of Exxon, w_2 of the Dow, w_3 of Google and w_k of Apple

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

Analysing the sum of random variables is an extremely common use-case

Adding Random Variables

k columns

$$[\ X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_k\ ]$$

n rows

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

n rows

1 column

Adding n variables, each of k-dimensional data, gives 1-dimensional data

Adding Random Variables

$$y = X_1 + X_2 + X_3 \dots + X_k$$

Mean(y) = ?

Variance(y) = ?

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{aligned} \text{Mean}(y) = \ & \text{Mean}(X_1) + \\ & \text{Mean}(X_2) + \\ & \text{Mean}(X_3) + \\ & \ldots \\ & \text{Mean}(X_k) \end{aligned}$$

k terms

Mean of sum = sum of means

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

Mean(y)

Simple - mean of sum is sum of means

Variance(y) = ?

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{aligned}
\text{Variance}(y) = \;& \text{Covariance}(X_1,X_1) + \\
& \text{Covariance}(X_1,X_2) + \\
& \ldots \\
& \text{Covariance}(X_1,X_k) + \\
& \ldots \\
& \text{Covariance}(X_k,X_1) + \\
& \text{Covariance}(X_k,X_2) + \\
& \ldots \\
& \text{Covariance}(X_k,X_k)
\end{aligned}$$

k^2 terms

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\text{Variance}(y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Covariance}(X_i, X_j)$$

k^2 terms

Variance of sum can be found from the covariance matrix

Covariance Matrix

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{array}{cl}
\begin{bmatrix}
Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\
Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\
Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k)
\end{bmatrix} & \text{k rows} \\
\text{k columns} &
\end{array}$$

Diagonal elements are the variances

Covariance Matrix

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k) \end{bmatrix}$$

k rows

k columns

Add all the diagonal elements...

Covariance Matrix

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{array}{c} \begin{array}{cc} \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k) \end{bmatrix} & \text{k rows} \end{array} \\ \text{k columns} \end{array}$$

...and half the sum of the off-diagonal entries

Adding Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

Mean(y)

Simple - mean of sum is sum of means

Variance(y)

Tricky - requires use of covariance matrix

Adding related variables is difficult,
adding independent variables is easy

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\text{Variance}(y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Covariance}(X_i, X_j)$$

k^2 terms

If the X variables are independent, we can easily find the variance of the sum

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{array}{cc} \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k) \end{bmatrix} & \text{k rows} \\ \text{k columns} & \end{array}$$

Diagonal elements are the variances

Adding Independent Random Variables

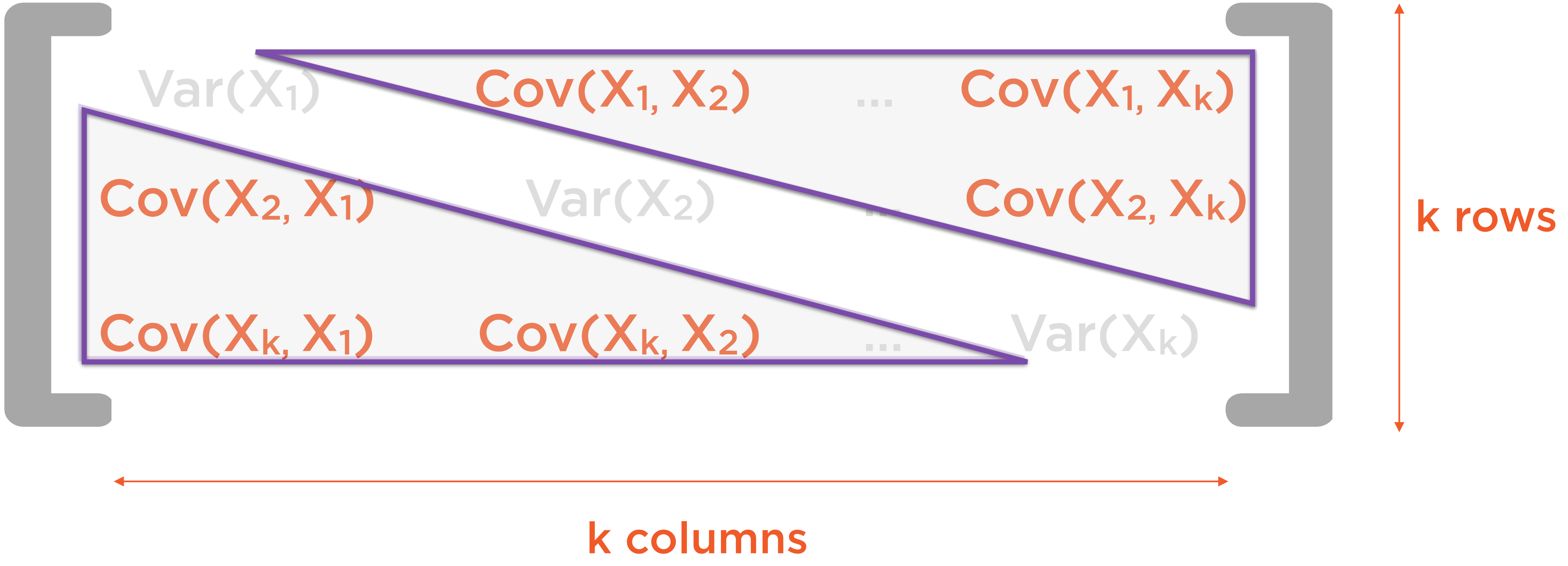
$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{array}{cc} \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_k) \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \ldots & Var(X_k) \end{bmatrix} & \text{k rows} \\ \text{k columns} & \end{array}$$

Add all the diagonal elements...

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$



...and half the sum of the off-diagonal entries

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{bmatrix} Var(X_1) & 0 & \cdots & 0 \\ 0 & Var(X_2) & \cdots & 0 \\ 0 & 0 & \cdots & Var(X_k) \end{bmatrix}$$

k rows

k columns

But all off-diagonal entries are zero!

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

$$\begin{array}{cc} \begin{bmatrix} Var(X_1) & 0 & \ldots & 0 \\ 0 & Var(X_2) & \ldots & 0 \\ 0 & 0 & \ldots & Var(X_k) \end{bmatrix} & \text{k rows} \\ \text{k columns} & \end{array}$$

Add all the diagonal elements...

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \dots + X_k$$

$$\text{Variance}(y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Covariance}(X_i, X_j)$$

k^2 terms

$$= \sum_{i=1}^{k} \text{Variance}(X_i)$$

k terms

For independent variables, variance of sum is sum of variances

Adding Independent Random Variables

$$y = X_1 + X_2 + X_3 \ldots + X_k$$

Mean(y)

Simple - mean of sum is sum of means

Variance(y)

Simple - variance of sum is sum of variances

Tossing Two Coins

Coin X → Heads = $1

Coin X → Tails = -$1

Coin Y → Heads = $1,000

Coin Y → Tails = -$1,000

Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$ $\bar{y} = 0$

Var(x) = 1 Var(y) = 1,000,000

Covariance(x,y) = 0

Independent variables have zero covariance

Combined Payoff

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

Combined Payoff
$1,001
-$999
$-999
-$1,001

$\bar{x} = 0$
Var(x) = 1

$\bar{y} = 0$
Var(y) = 1,000,000

$\bar{z} = 0$

Covariance(x,y) = 0

Combined Payoff

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

Combined Payoff
$1,001
-$999
$-999
-$1,001

$z_i - \bar{z}$	$(z_i - \bar{z})^2$
$1,001	1002001
-$999	998001
$-999	998001
-$1,001	1002001

$\bar{x} = 0$ $\bar{y} = 0$ $\bar{z} = 0$

Var(x) = 1 Var(y) = 1,000,000

Covariance(x,y) = 0

$$\text{Variance} = \frac{\sum(z_i - \bar{z})^2}{n} = 1{,}000{,}001$$

Combined Payoff

$$Z = X + Y$$

Mean(z)

Simple - mean of sum is sum of means

Variance(z)

Simple - variance of sum is sum of variances

Exploratory Factor Analysis: Experts trace back principal components to observable factors

Summary

Factor analysis is a way to find the underlying drivers of a large dataset

PCA is one of many techniques that can be used in factor analysis

PCA is powerful and versatile, so it is very popular indeed

Some linear algebra and statistics are helpful in using factor analysis and PCA