By the end of this section

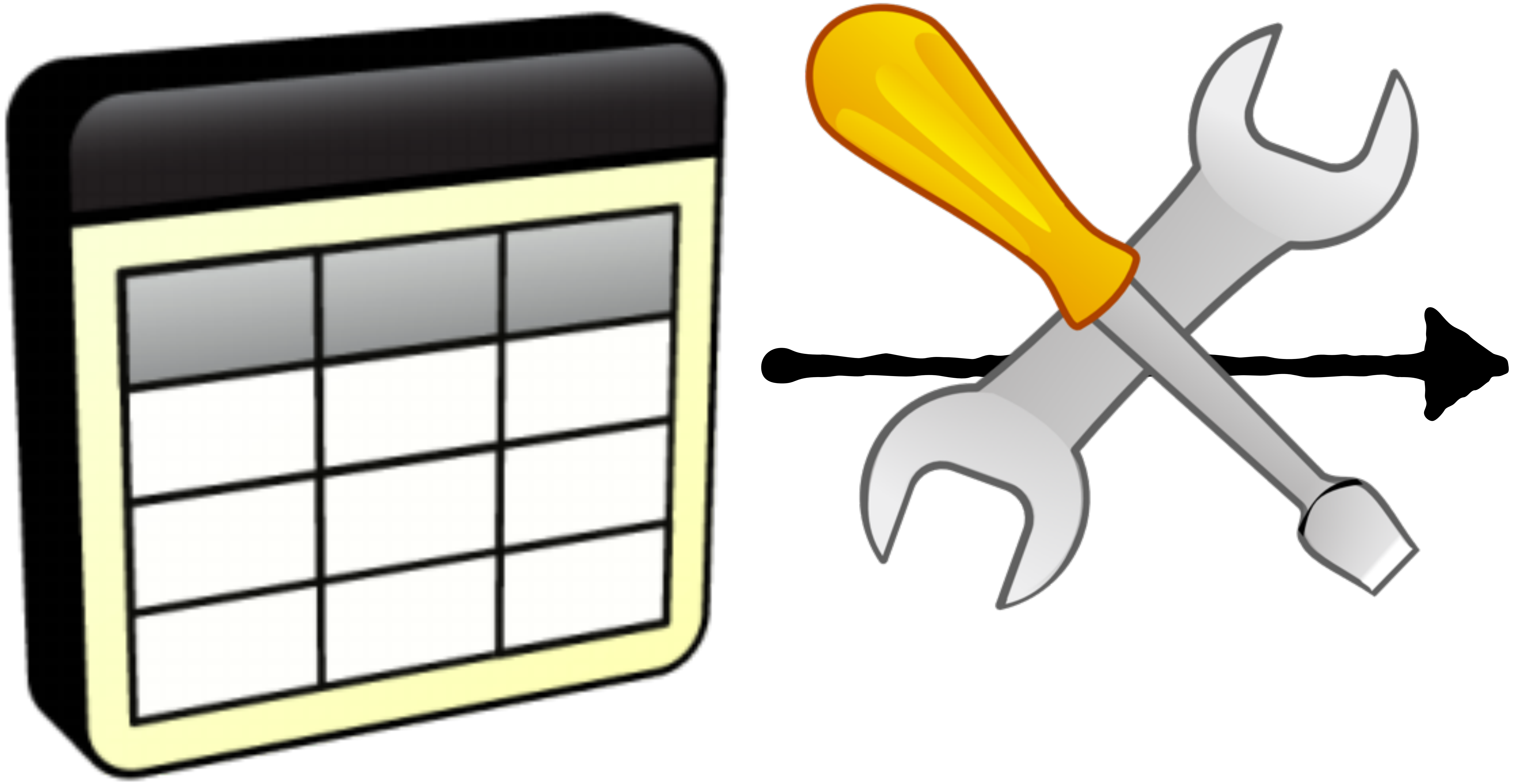
Understand the link between Factor analysis and Linear Regression

Understand when to use Factor analysis and PCA

Overview of statistics and linear algebra needed for PCA

Given a dataset

Connect the dots, recognize patterns, draw insights



Objective:

Build a fact-based, thoughtful point of view

Regression

Understand the relationships between variables

Factor Analysis

Understand the underlying drivers that influence the relationships

Regression

Connect the dots

Factor Analysis

Cut through the noise

Regression

Pageviews

Clicks

Add to Carts → Sales

Minutes browsed

Sessions

Cause

Independent/

Explanatory Variable

Effect

Dependent

Factor Analysis

Pageviews

Clicks

Add to Carts → Selection

Minutes browsed

Sessions

Marketing spend → Sales

Pricing

Many observed causes

Few underlying causes

Effect

Regression

All observed causes used to explain the effect

Simplistic

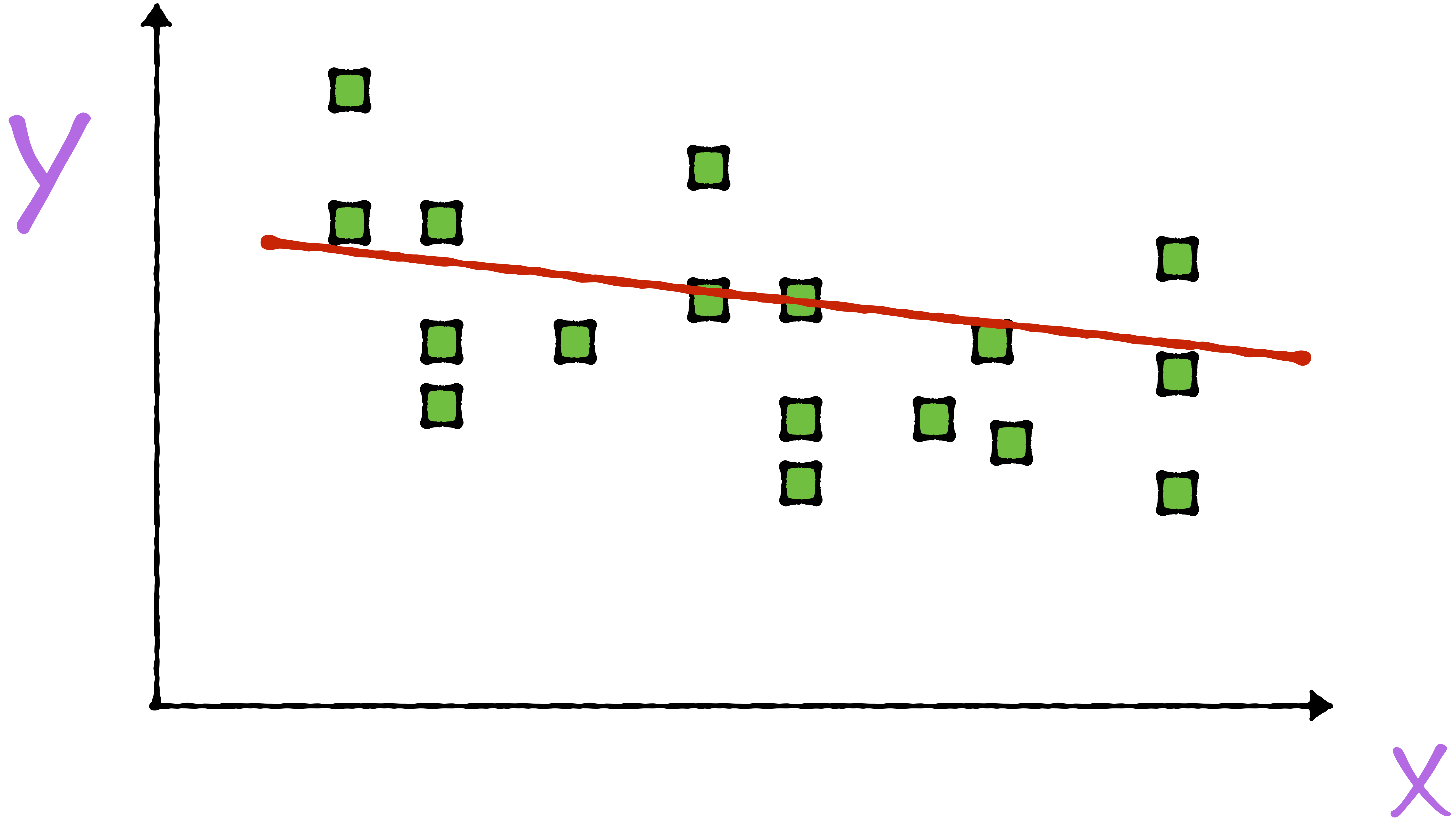
Factor Analysis

Few underlying drivers used to explain the effect

Simple

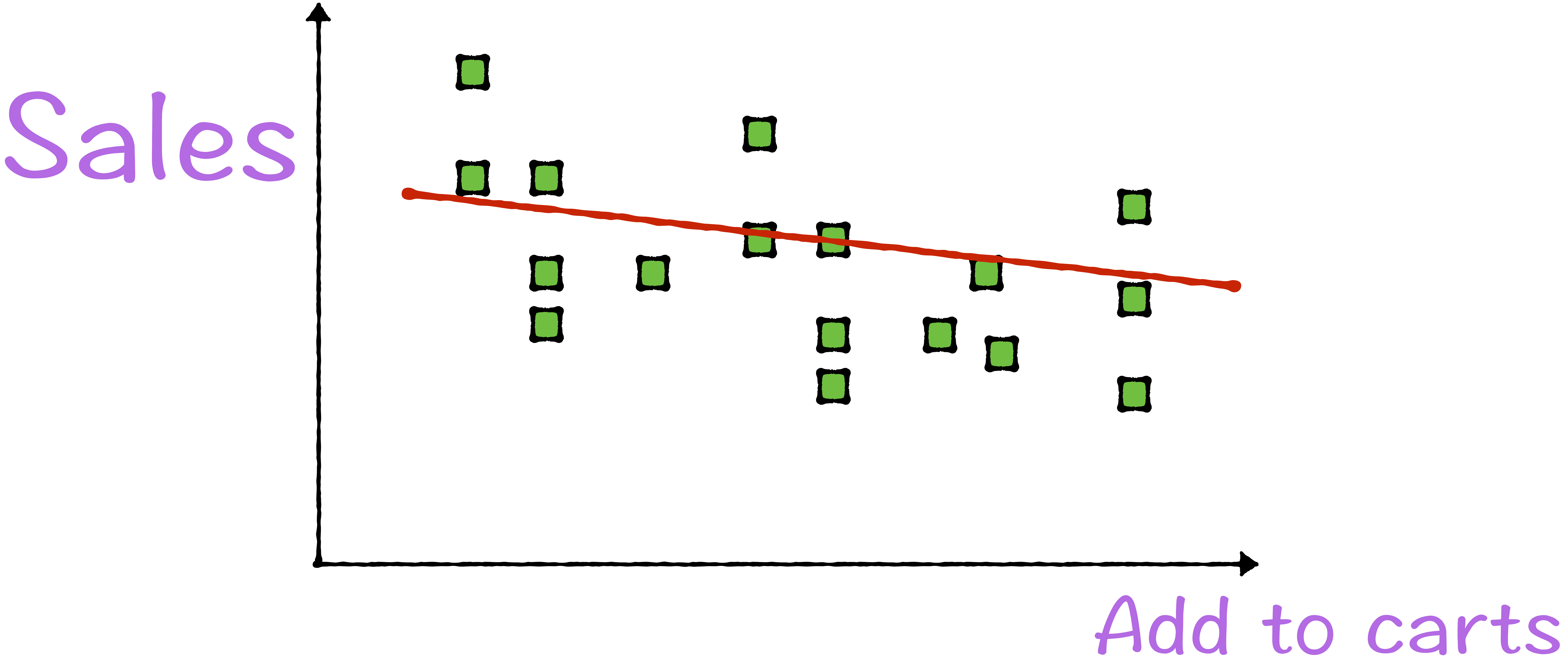
Regression

Find the best line through these data points



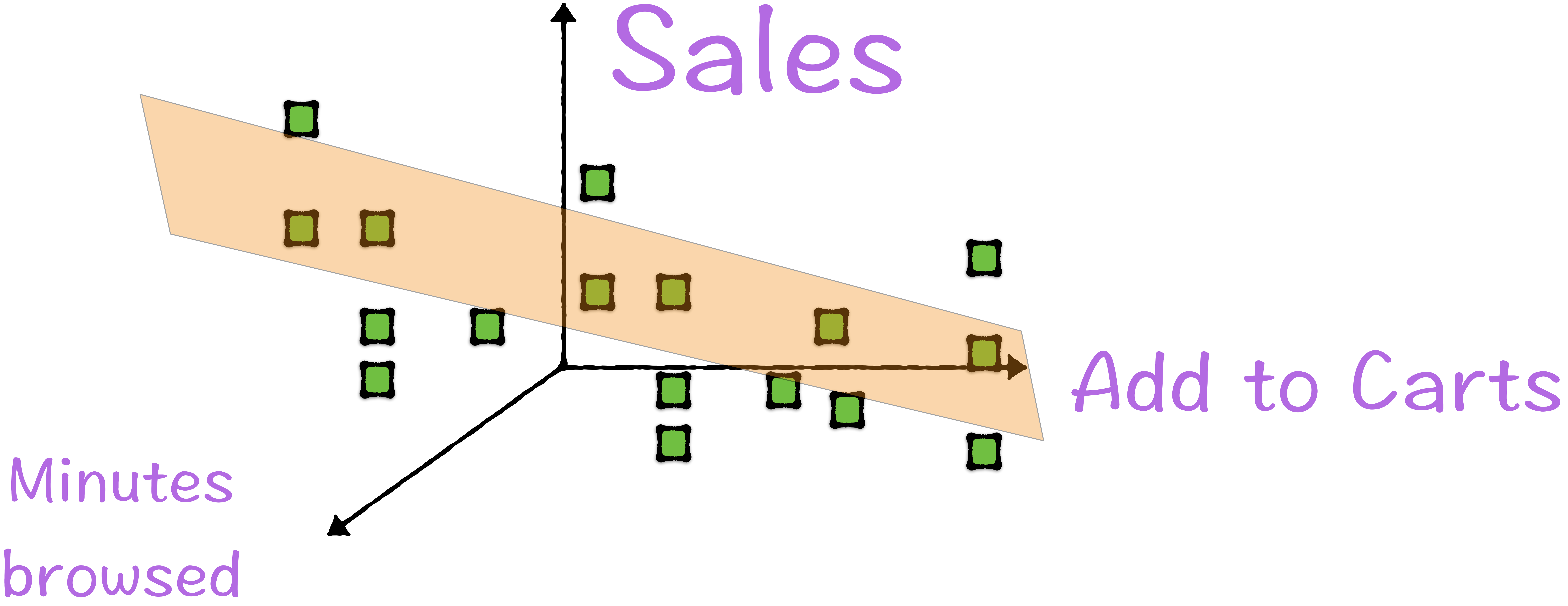
Regression

2 Dimensional data : One cause explains one effect



Regression

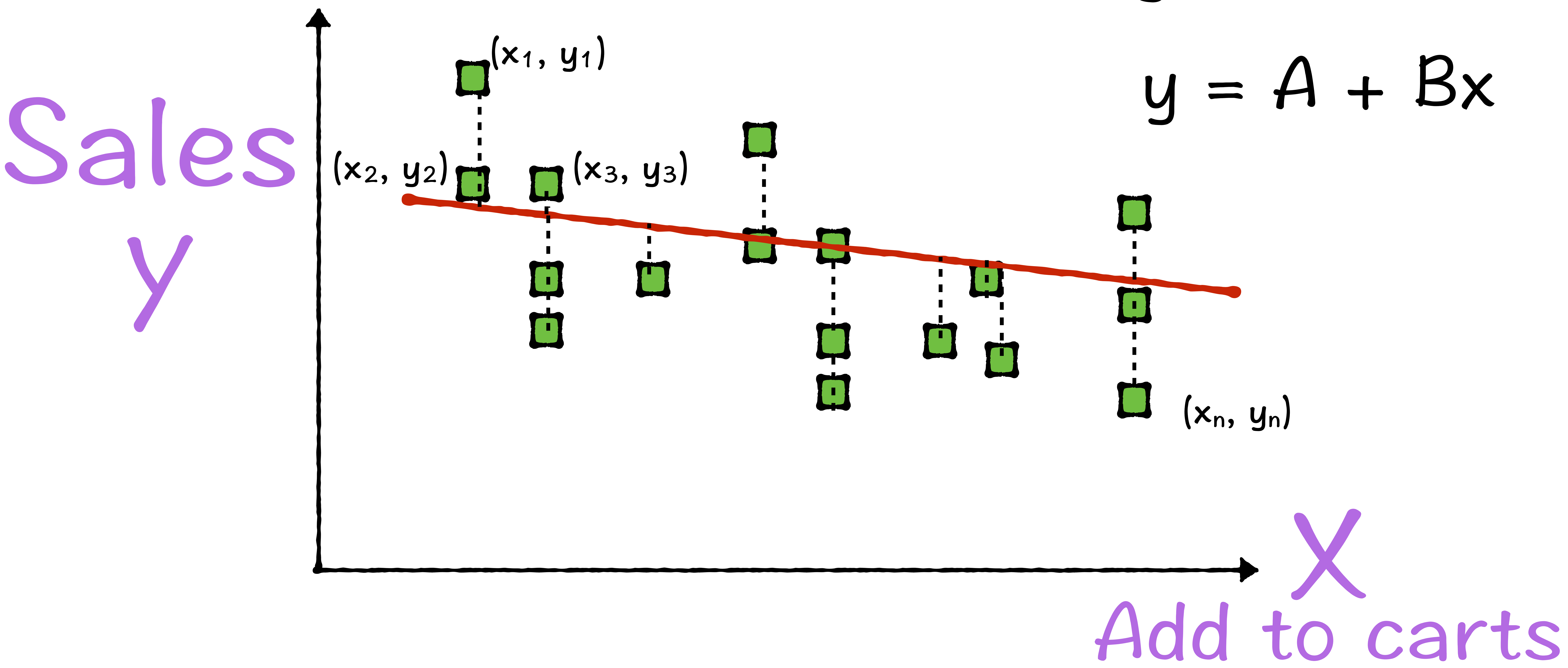
N-Dimensional data : multiple causes explain 1 effect



Simple Linear Regression

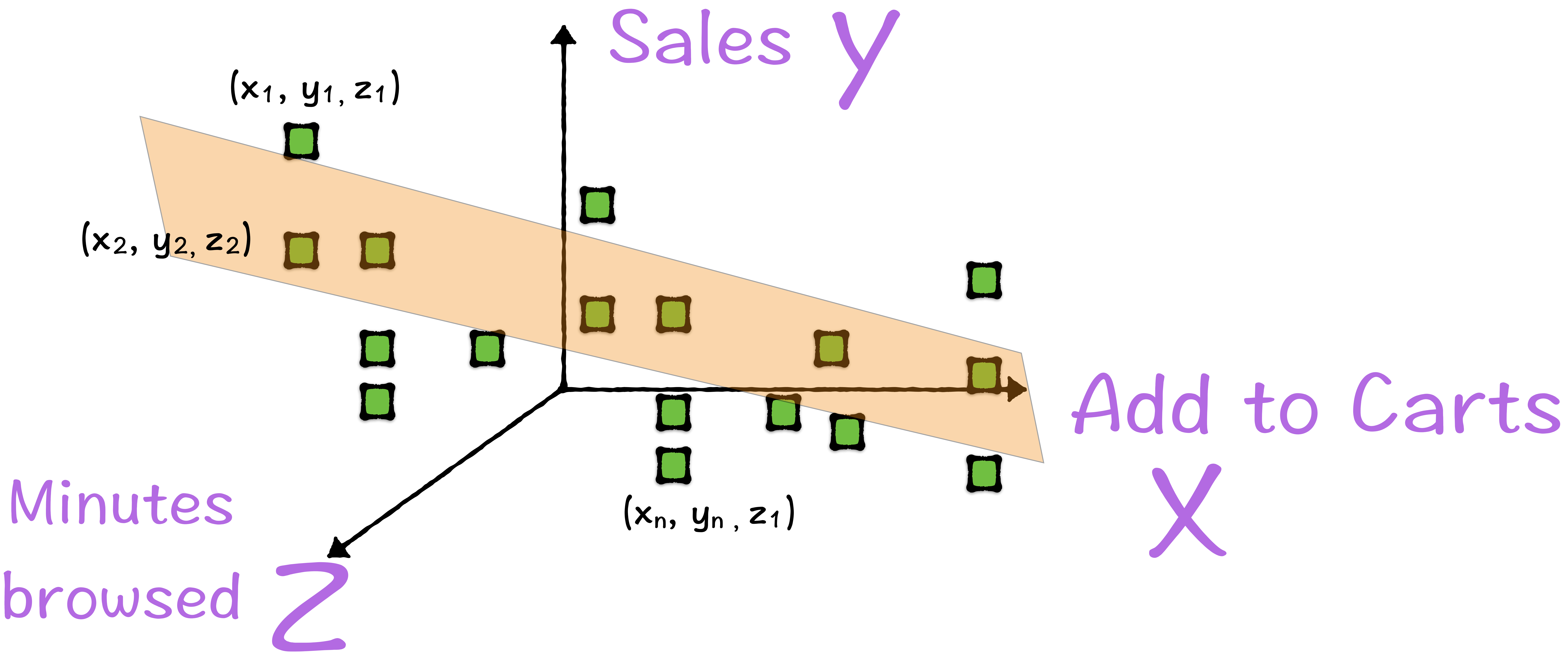
Regression Line:

$$y = A + Bx$$

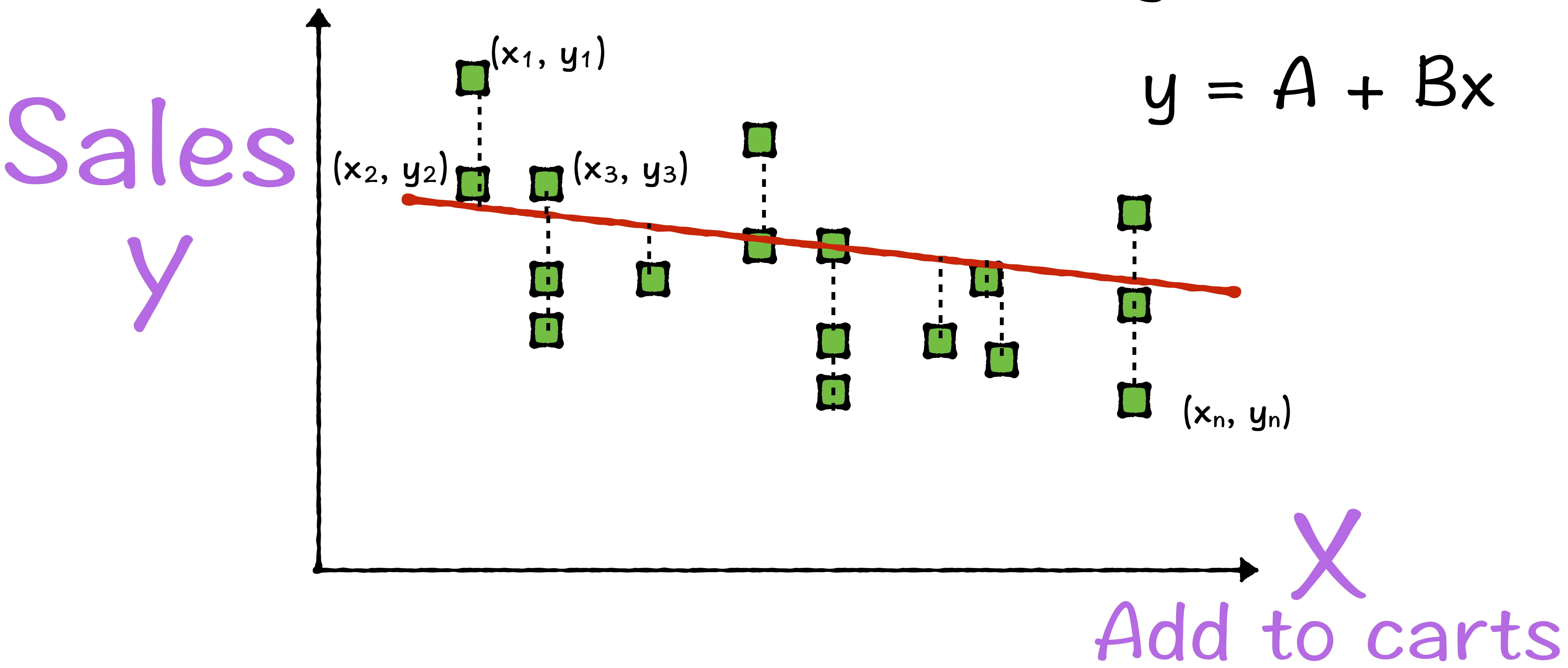


Multiple Linear Regression

Regression Plane: $y = A + Bx + cZ$



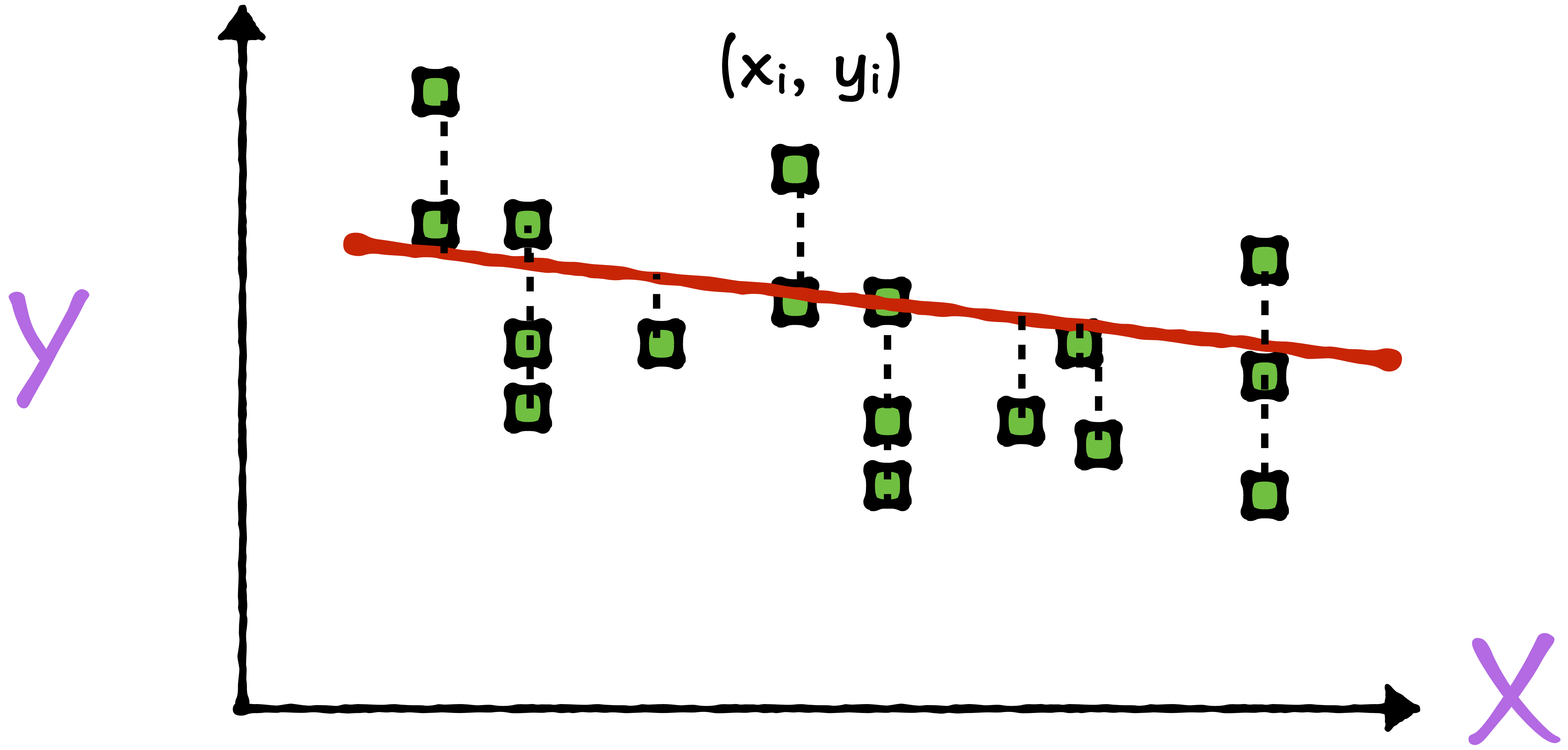
Simple Linear Regression



Regression Line:

$$y = A + Bx$$

Simple Linear Regression



Regression Line:

$$y = A + Bx$$

$$y_1 = A + Bx_1 + e_1$$

$$y_2 = A + Bx_2 + e_2$$

$$y_3 = A + Bx_3 + e_3$$

$$\dots \qquad \dots$$

$$y_n = A + Bx_n + e_n$$

Simple Linear Regression

Regression Line:

$$y = A + Bx$$

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

Simple Linear Regression

Regression Line:

$$y = A + Bx$$

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

Sales

Add to carts

Multiple Linear Regression

Regression Plane:

$$y = A + Bx + Cz$$

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

Sales

Add to carts

Minutes browsed

Multiple Linear Regression

Regression Plane:

$$y = A + Bx + Cz$$

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

n Rows, 1 Column

n Rows, 3 Columns

3 Rows, 1 Column

n Rows, 1 Column

Multiple Linear Regression

Regression Plane:

$$y = A + Bx + Cz$$

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

2 Causes

Multiple Linear Regression

Regression Plane:

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

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

K Causes

Multiple Linear Regression

Regression Plane:

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

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

n Rows, 1 Column

n Rows, k+1 Columns

k+1 Rows, 1 Column

n Rows, 1 Column

Multiple Linear Regression

Regression Plane:

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

$$\begin{bmatrix} C_1 \\ C_2 \\ \cdots \\ C_{k+1} \end{bmatrix}$$

Find k+1 coefficients,
k for the explanatory variables,
and 1 for the intercept

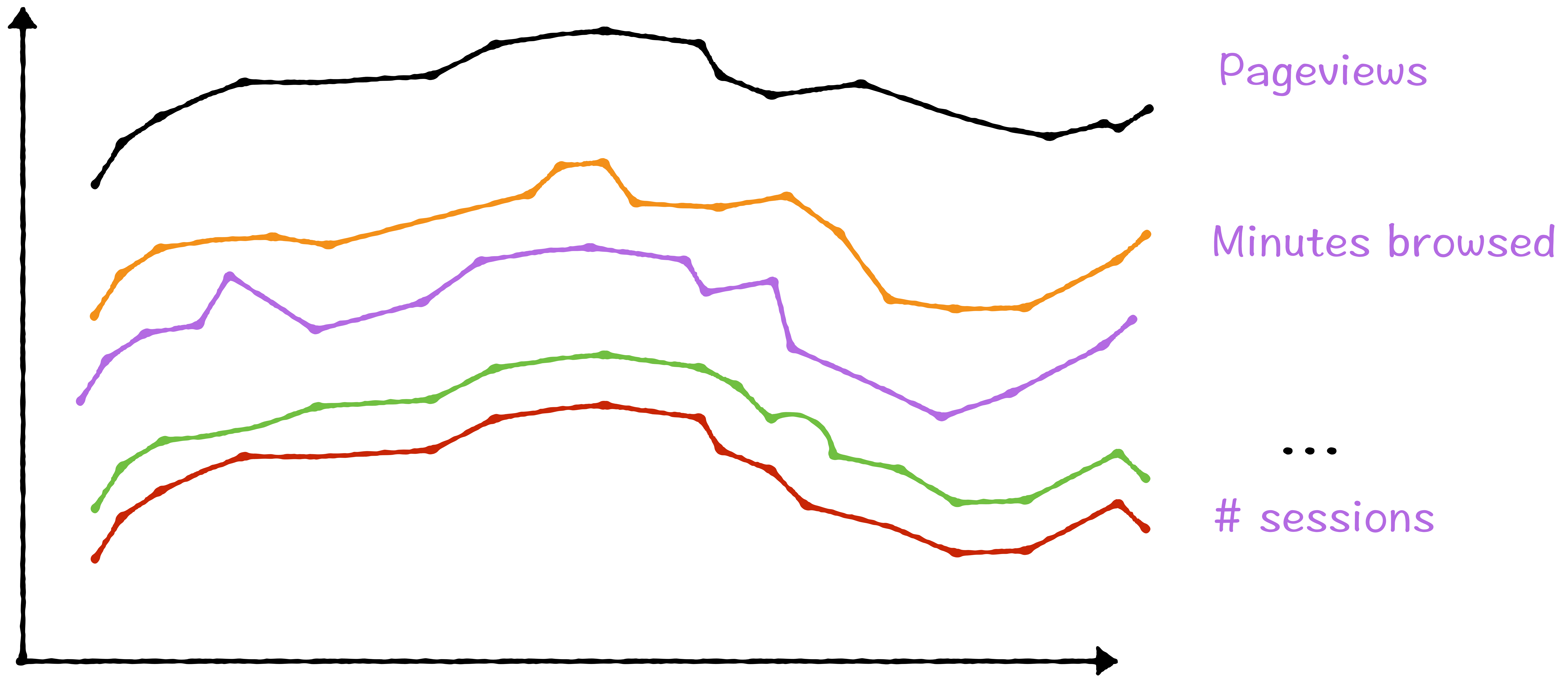
Kitchen sink Regression

Use all possible explanatory variables

Sales = A + B*Add to Cart + C*Minutes Browsed + D*Pageviews + E*# Sessions…

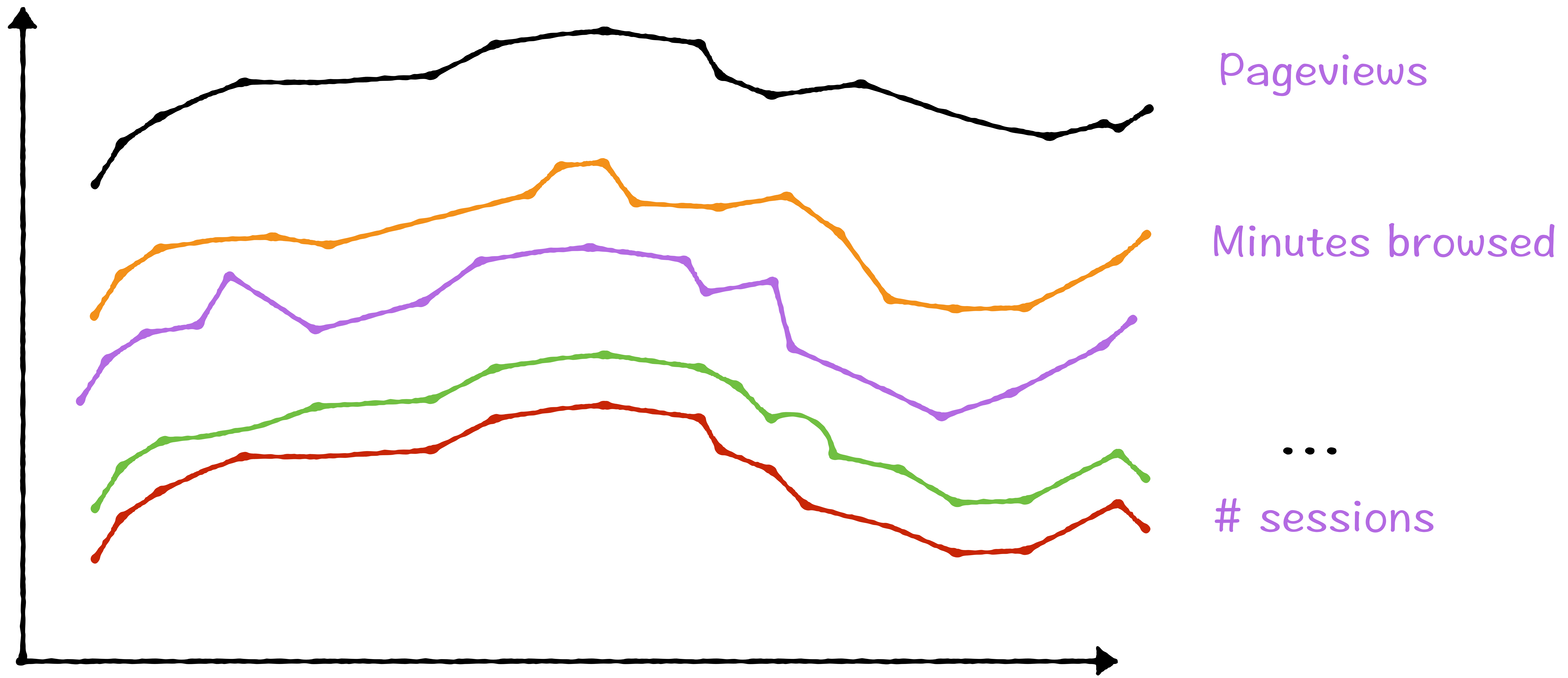
Problem -> Multicollinearity

Problem -> Multicollinearity



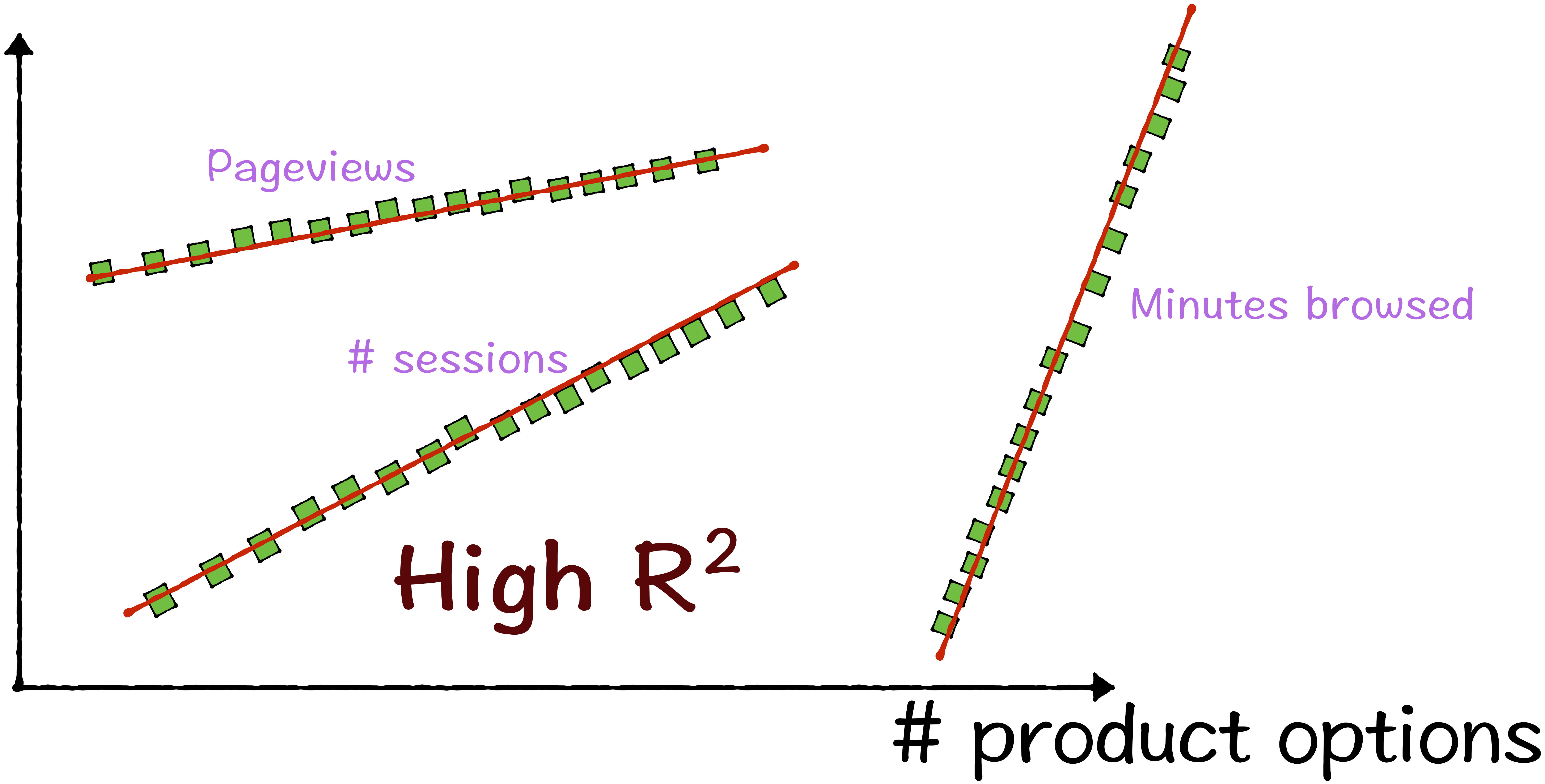
Many of the X variables contain the same information

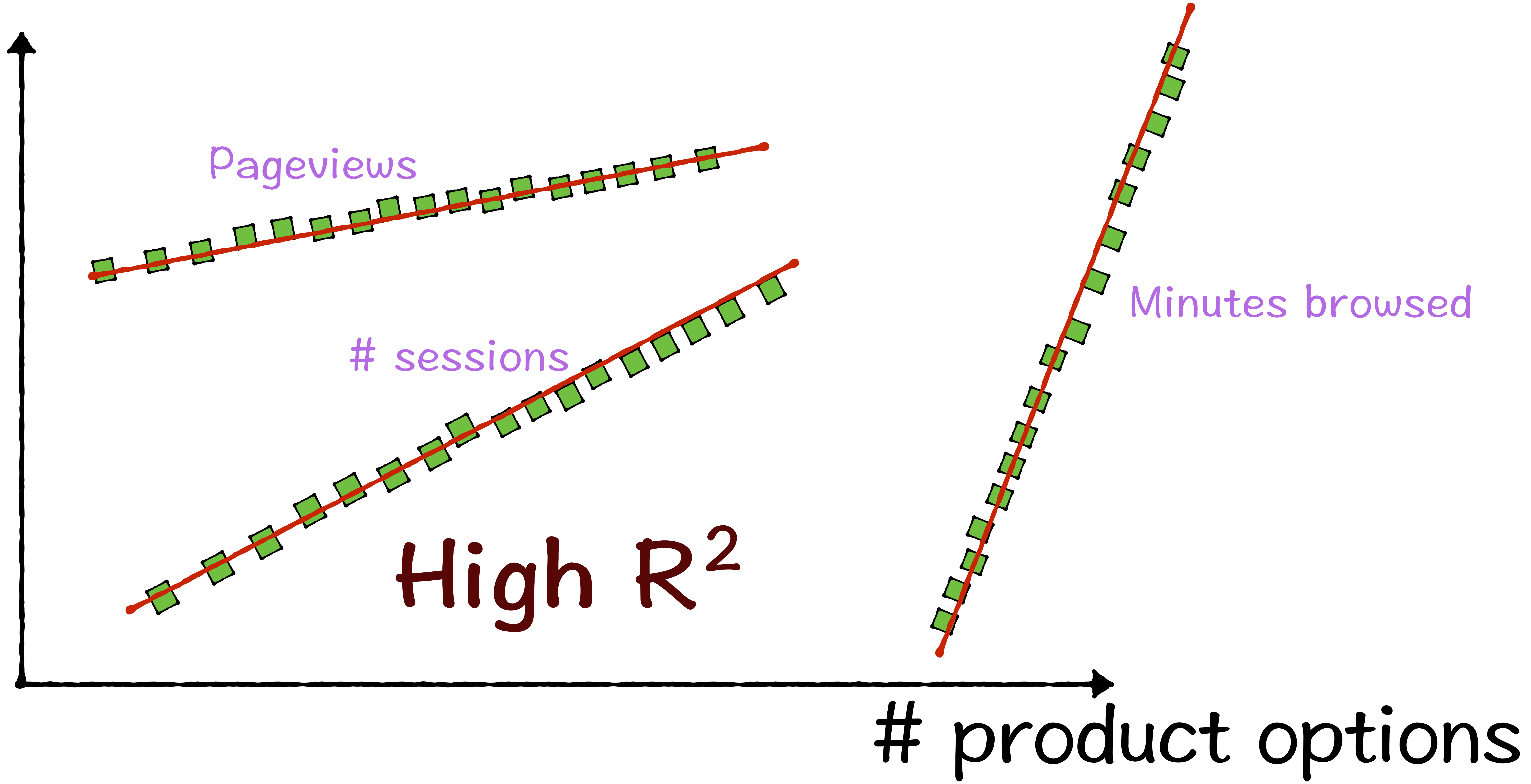
Problem -> Multicollinearity



There are underlying factors leading to this behavior

Underlying cause is selection (# product options)





Drop these 3 variables and use selection

Factor Analysis

Pageviews Clicks Add to Carts Minutes browsed # Sessions	→	Selection Marketing spend Pricing	→	Sales
Many observed causes		Few underlying causes		Effect

Principal Components Analysis

PCA

The problem to be solved

Fitting a curve through a set of data points

How it's solved

Linear Regression

The problem to be solved

Extract factors that explain the data

How it's solved

Principal components analysis (PCA)

Factor Extraction

Rule based

Use human experts to identify the factors

ML based

Extract the factors using an algorithm

Rule based

What factors influence success as a sales person ?

Rule based

Sales person → Personality assessment →

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

Personality Profile

Rule based

Each

Sales

person

Personality Profile

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

100 variables

Rule based

All employees →

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

100 variables X 10000 rows

Rule based

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

Openness

Conscientiousness

Extraversion

Agreeableness

Neuroticism

Map to 5 major underlying traits

Factor Extraction

Rule based

Use human experts to identify the factors

ML based

Extract the factors using an algorithm

ML based

Extract the factors using an algorithm

Gregarious ness	Warmth	Asserti veness	Exciteme nt-	Modes tu	Order	...
High	Medium	High	High	Low	High	...

PCA

F1	F2	F3	F4

Factors may or may not map to intuition

PCA

Identify latent factors

Dimensionality reduction

PCA

Identify latent factors

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

F1	F2	F3	F4

PCA

Identify latent factors

F1	F2	F3	F4

Human experts examine these factors

Openness

Conscientiousness

Extraversion

Agreeableness

Neuroticism

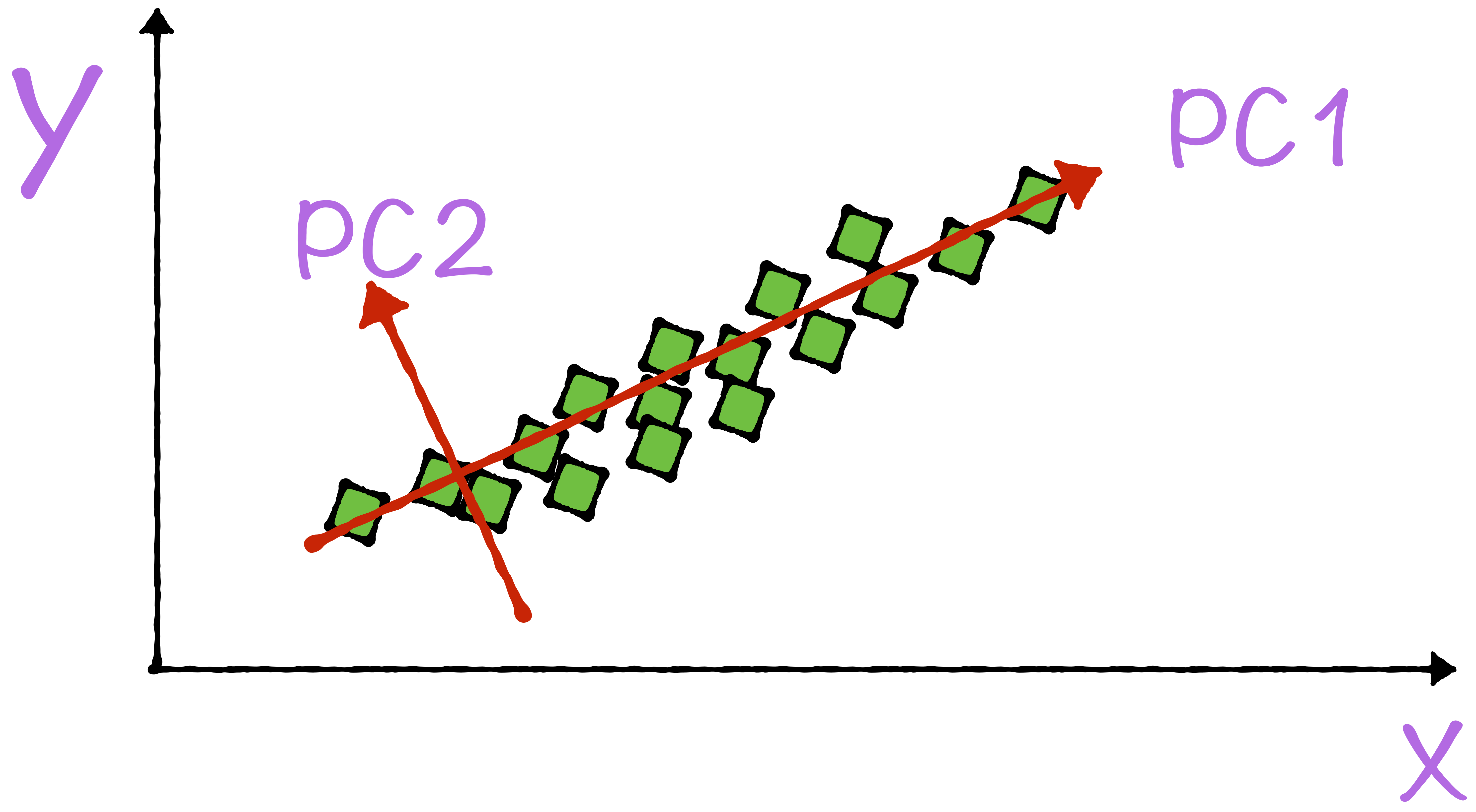
PCA

Identify latent factors

Dimensionality reduction

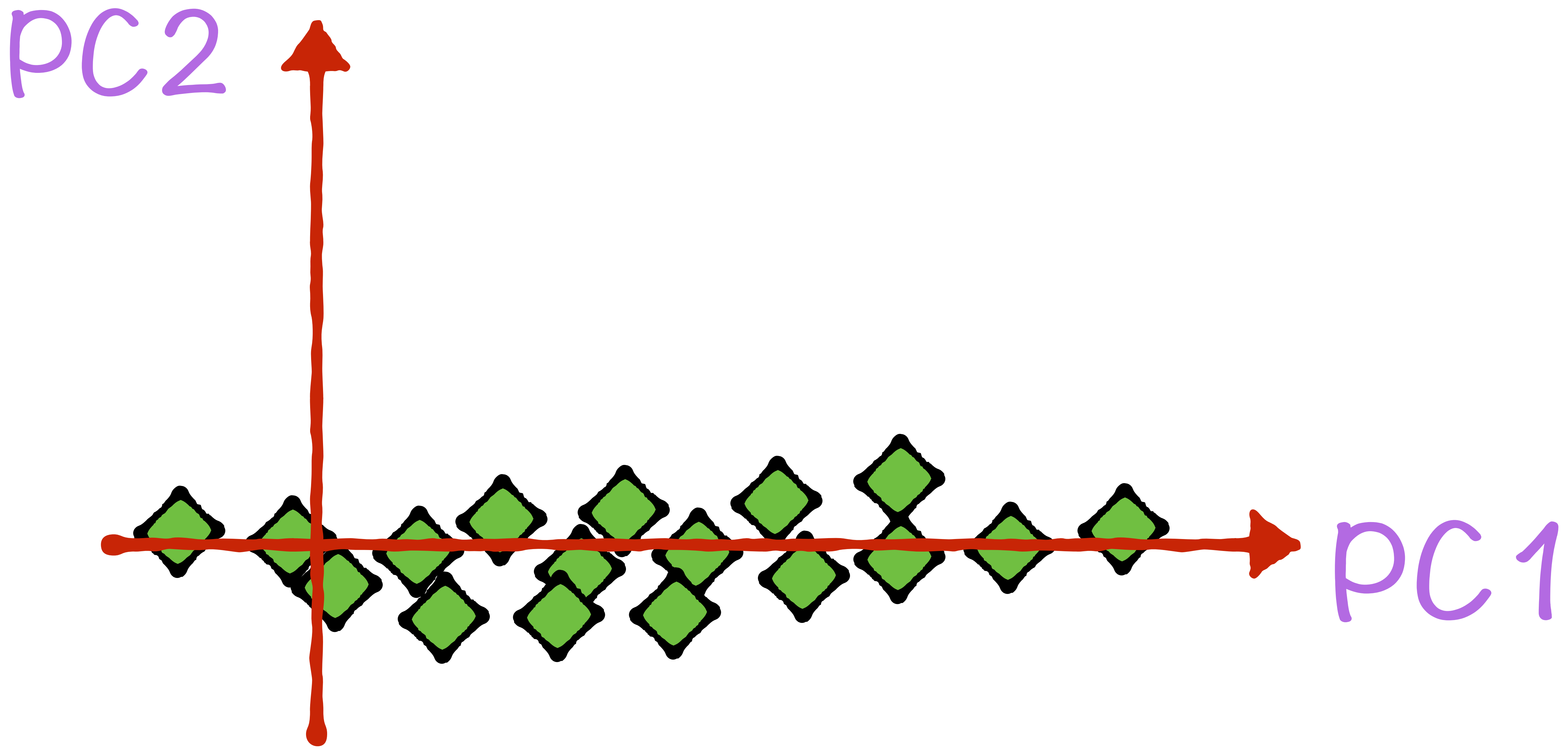
PCA

Dimensionality reduction



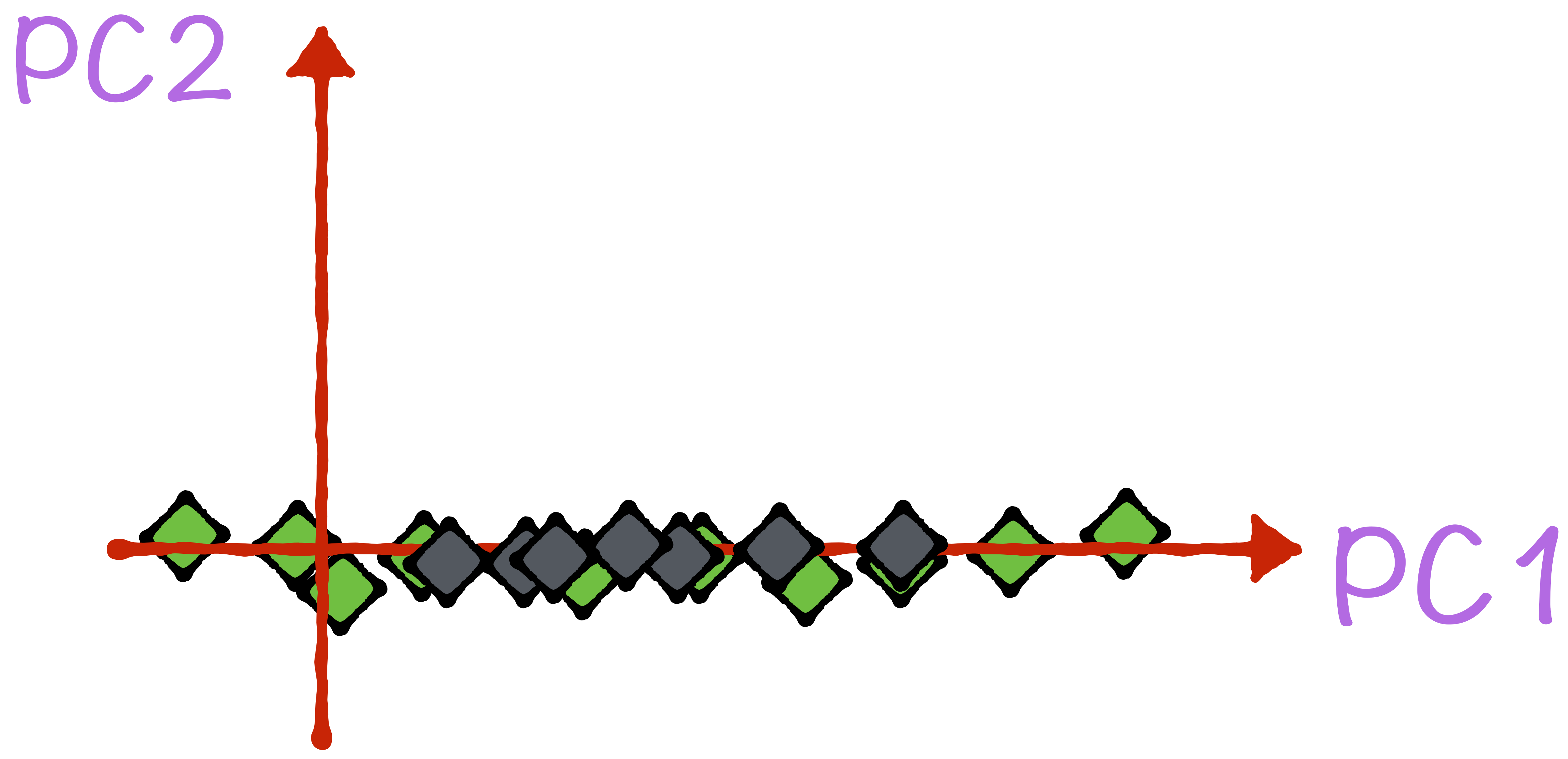
PCA

Dimensionality reduction



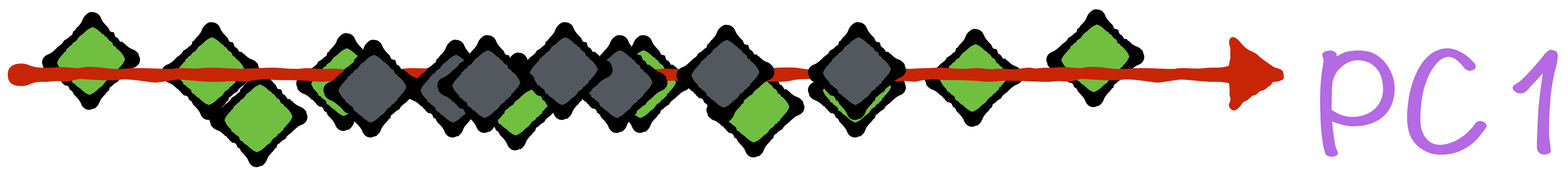
PCA

Dimensionality reduction

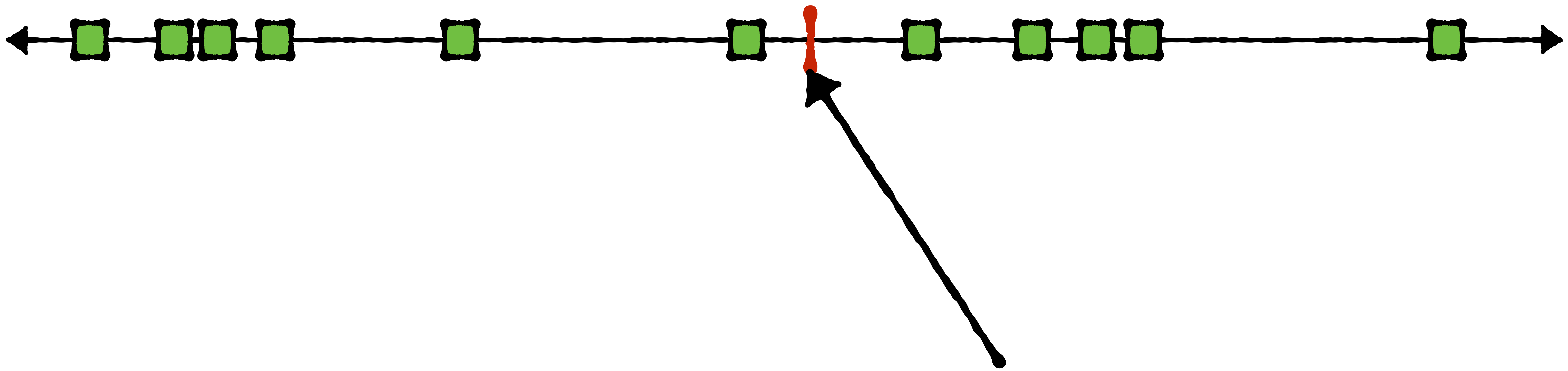


PCA

Dimensionality reduction



Data in one dimension

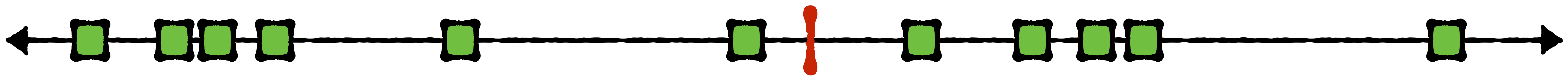


A central measure

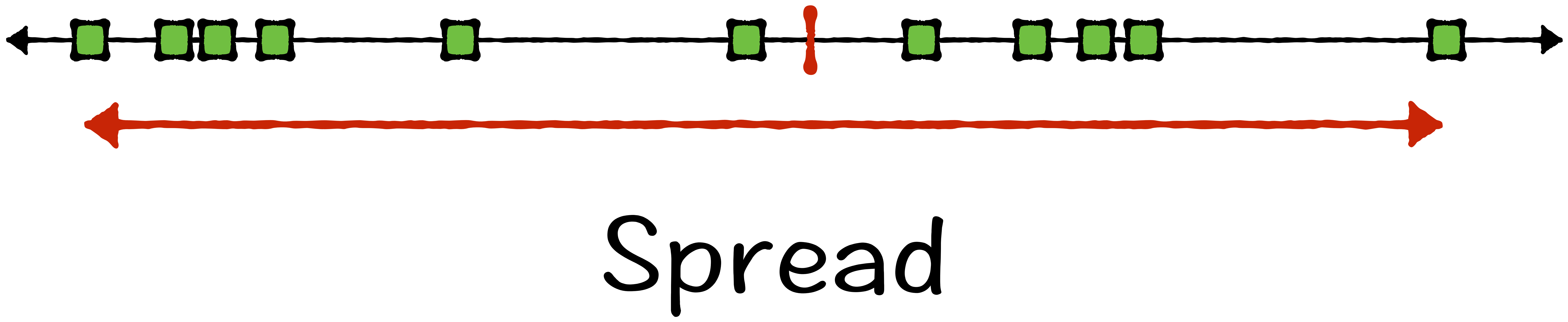
Mean/median

Mean

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

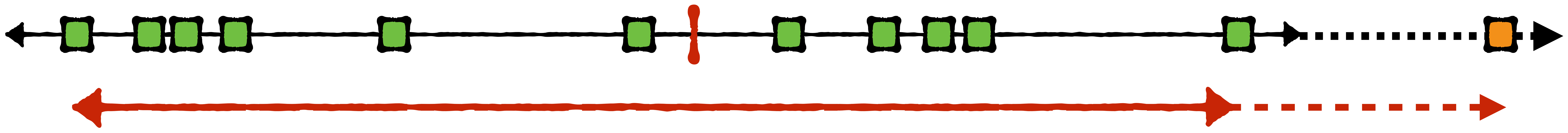


One number that best represents all the points


Spread

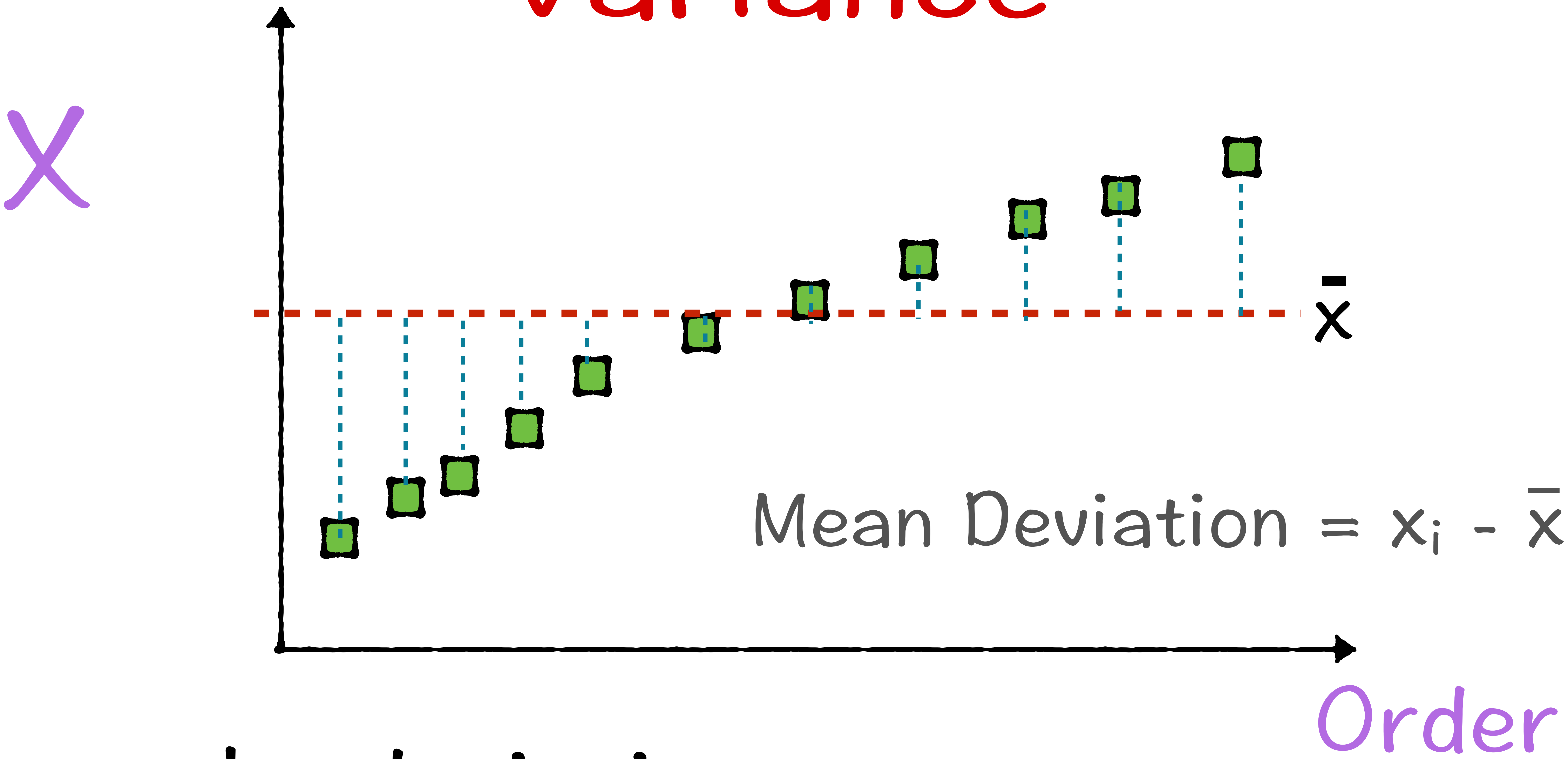
Range

$$x_{max} - x_{min}$$



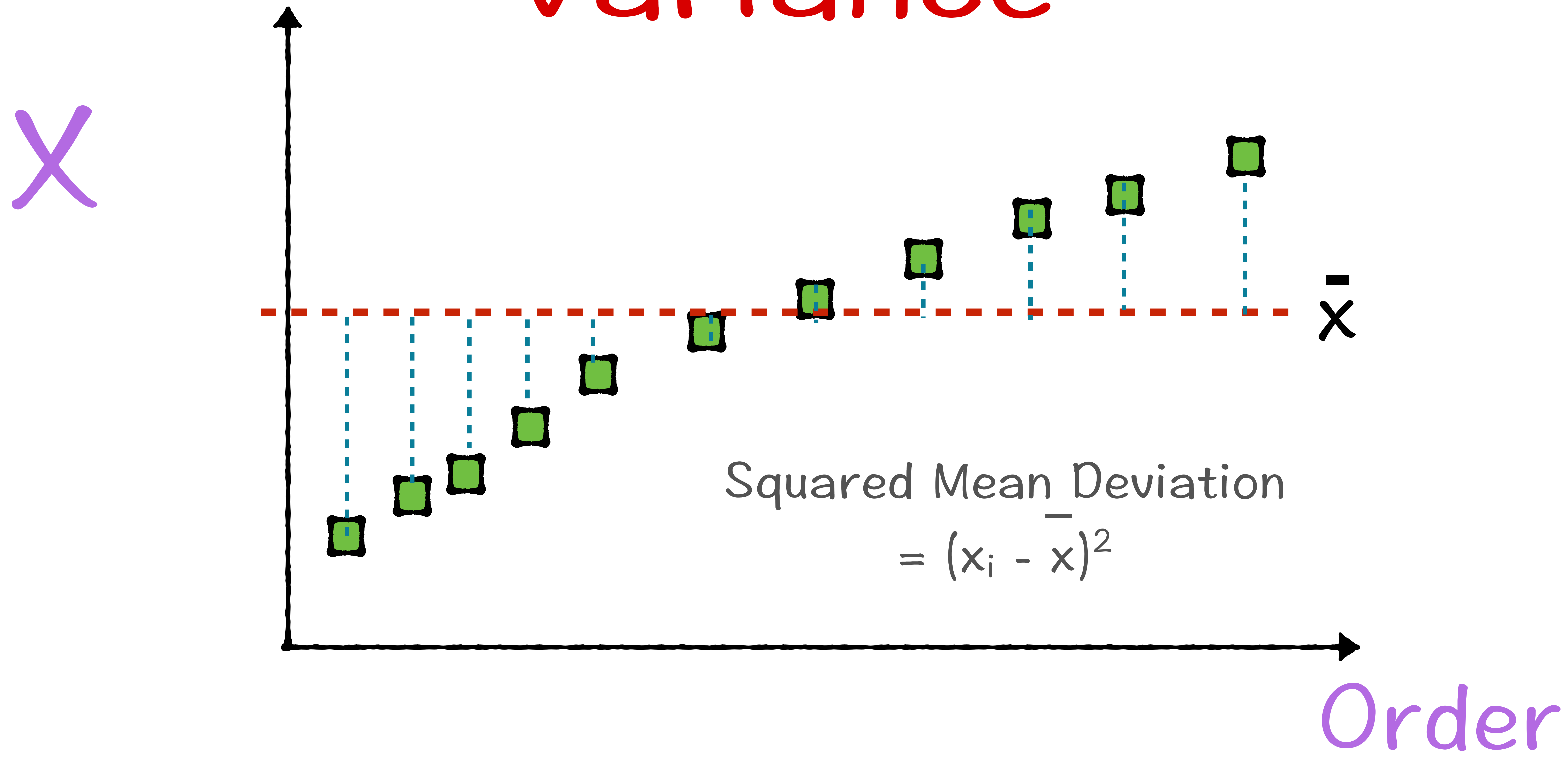
Ignores the mean,
affected by outliers

Variance

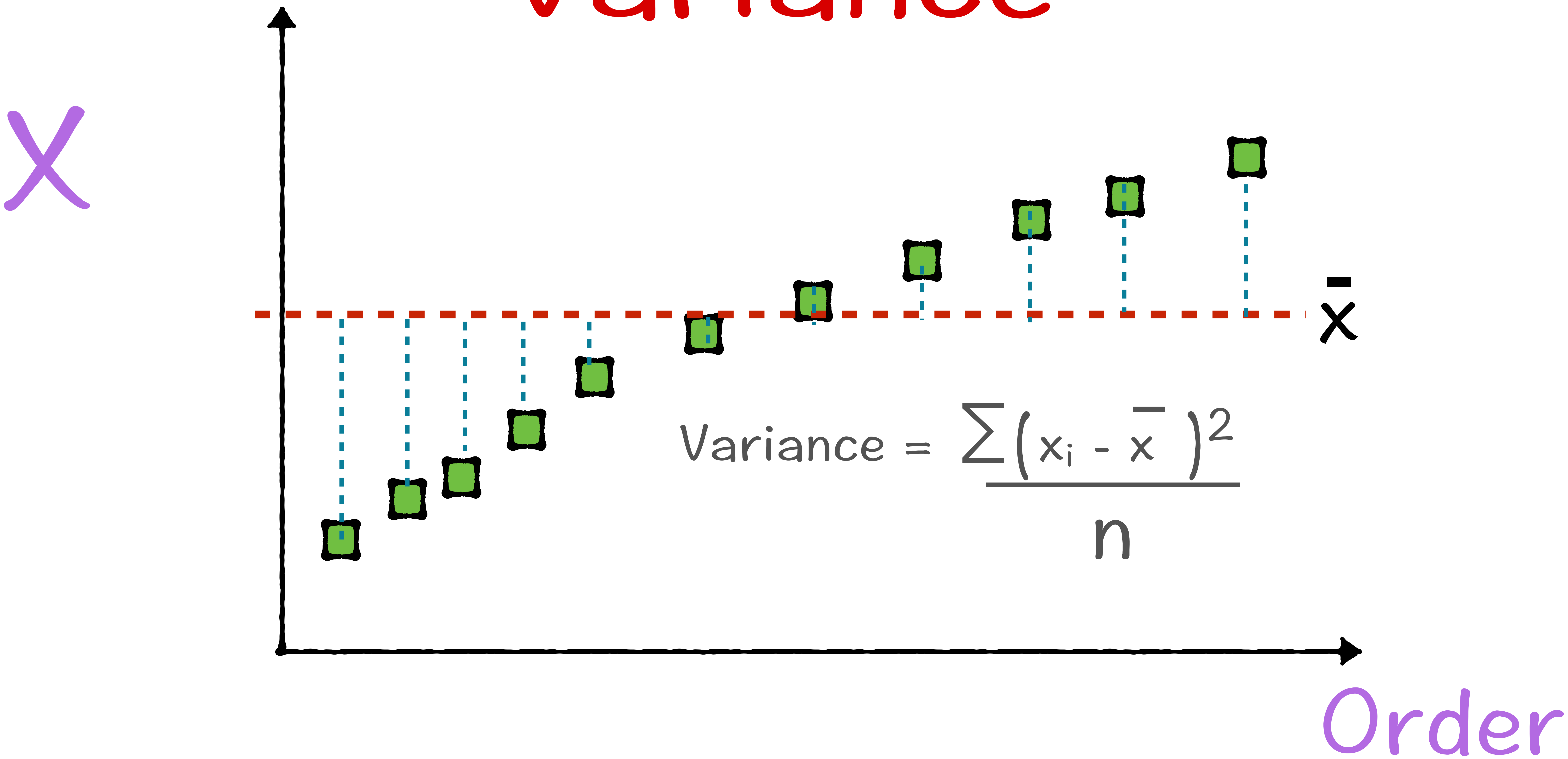


Measure the deviations from the mean

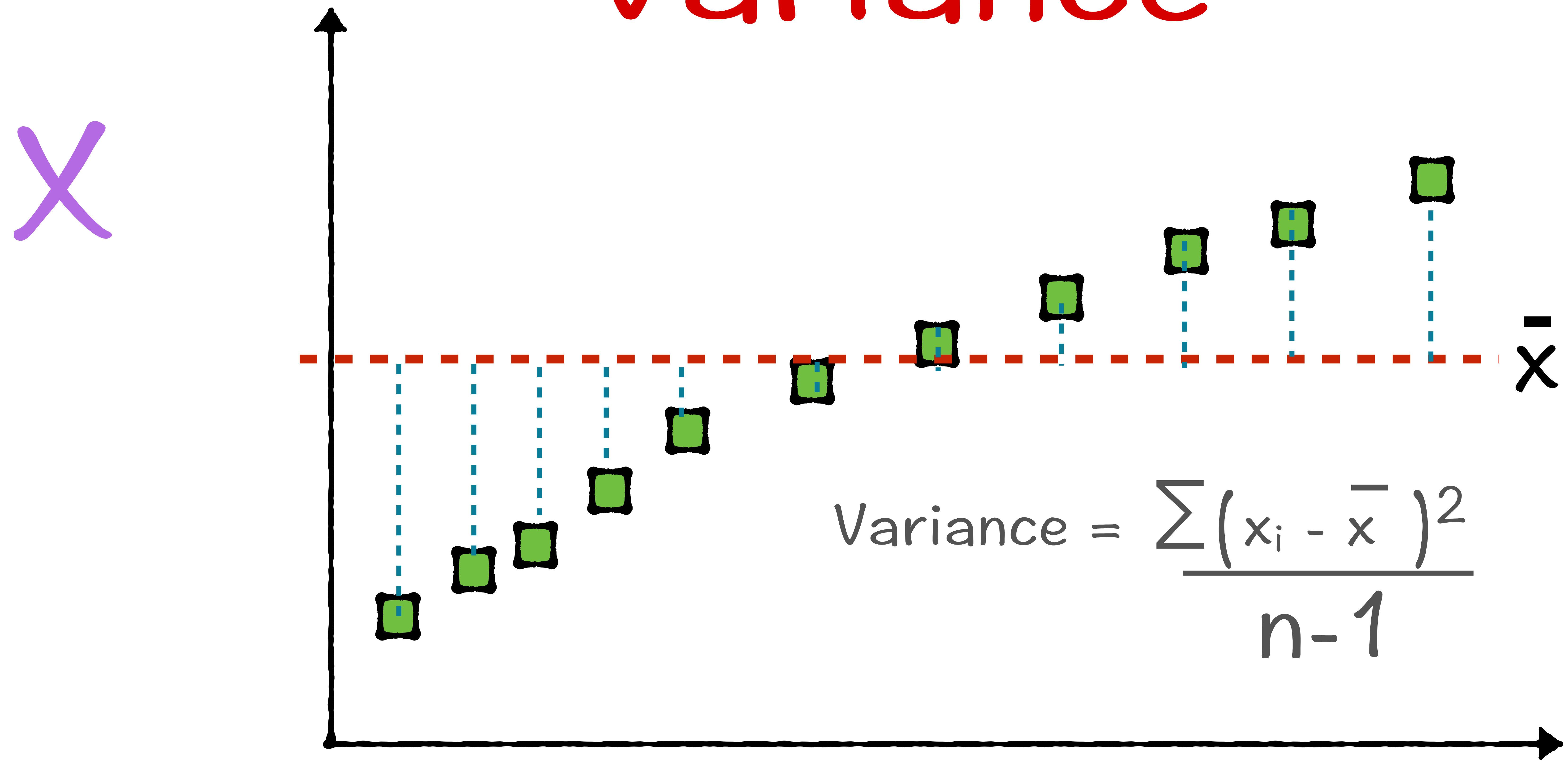
Variance



Variance

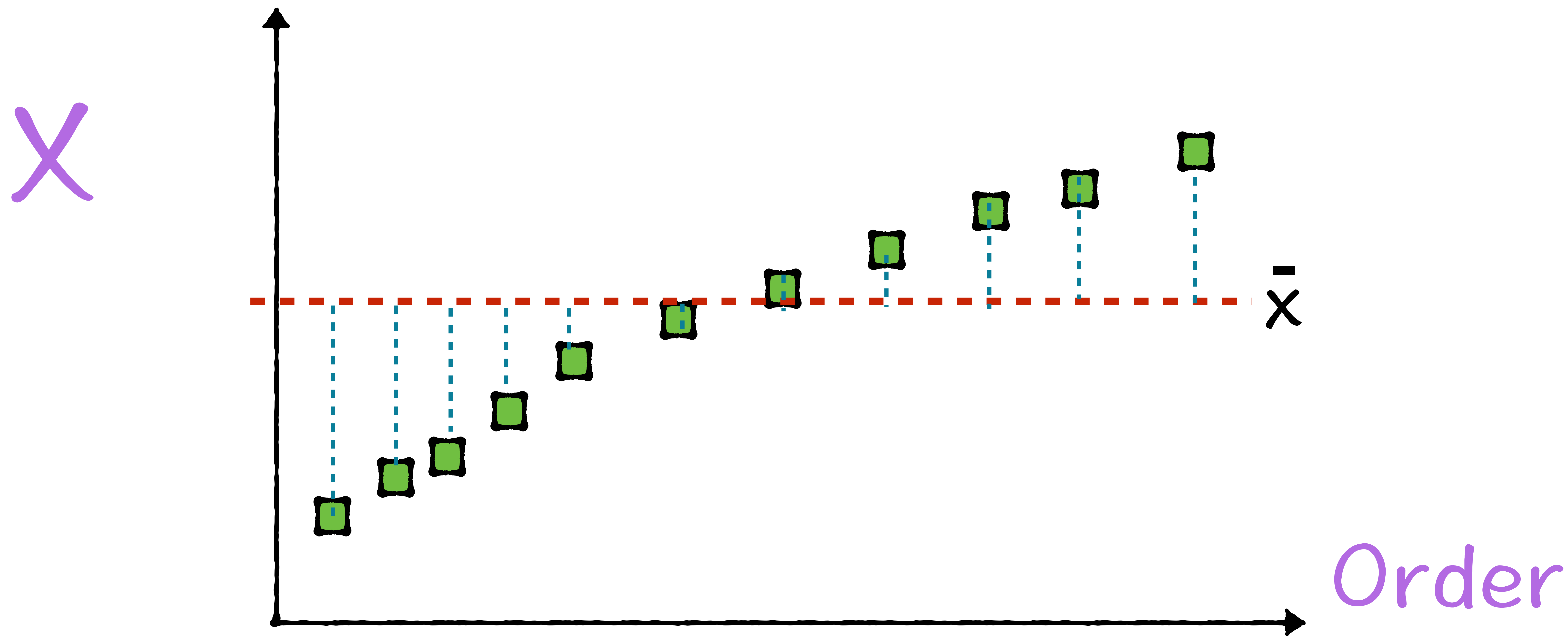


Variance



Bessel's correction

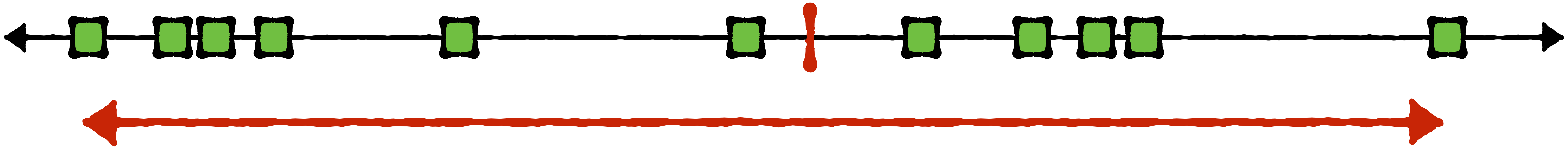
Variance and Standard Deviation



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

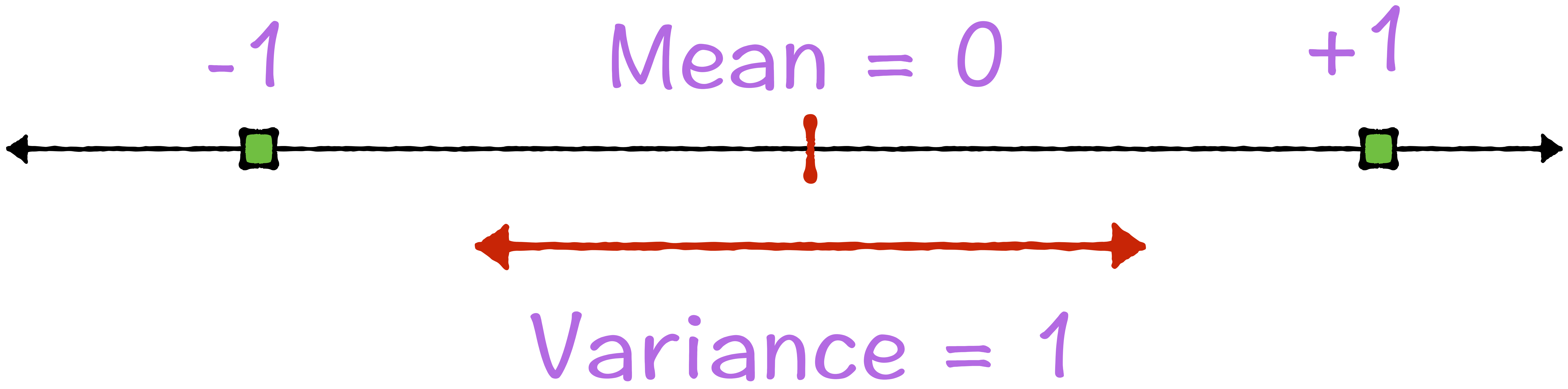
$$\text{Standard deviation} = \sqrt{\frac{\sum(\bar{x_i} - x)^2}{n-1}}$$

Mean vs Variance

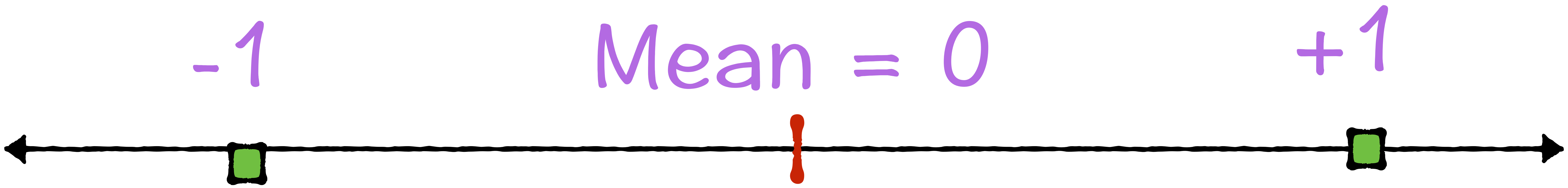


Variance measures risk

Small stakes game



Small stakes game

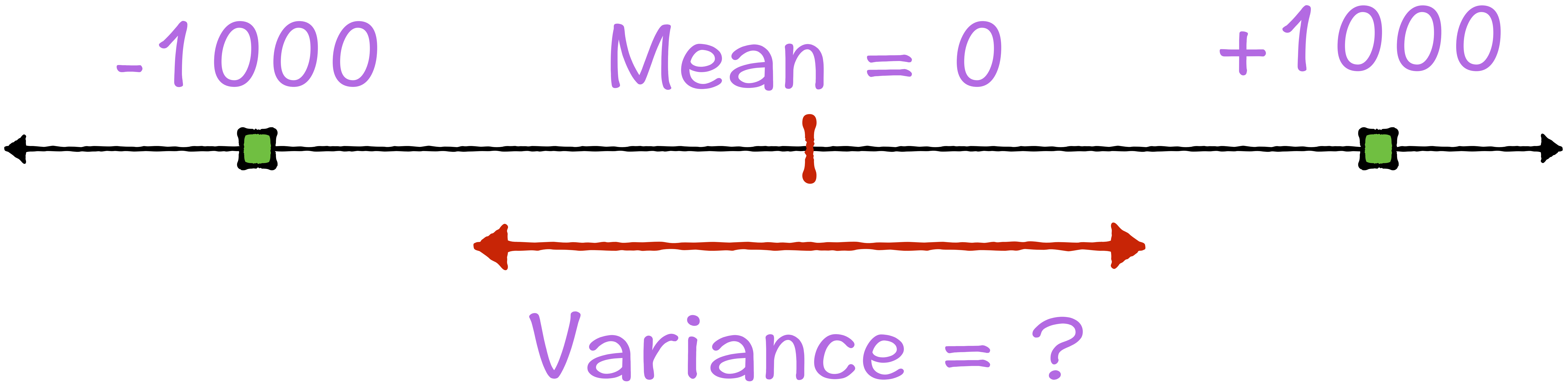


$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n}$$

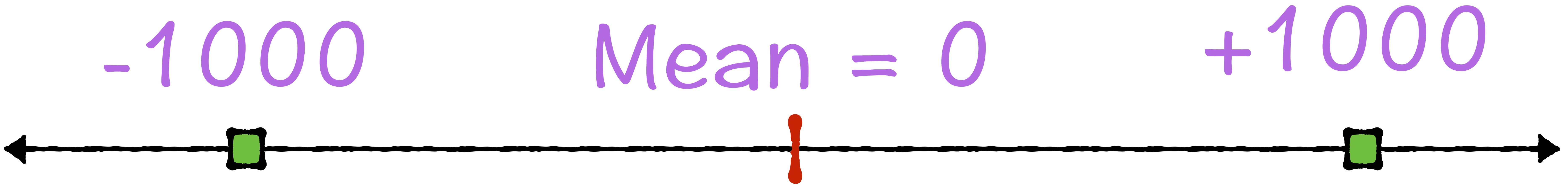
$$= \frac{(1-0)^2 + (-1-0)^2}{2}$$

$$= 1$$

Small stakes game



Small stakes game

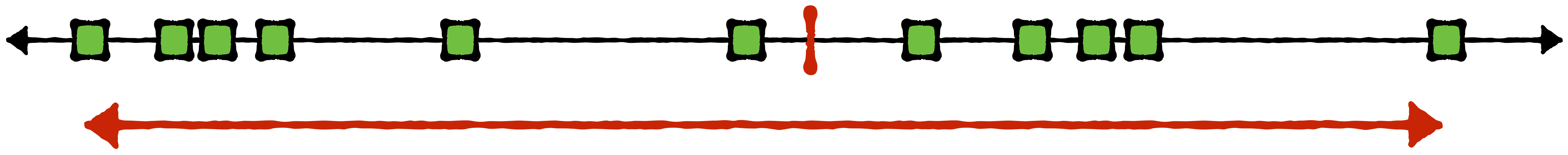


$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$= \frac{(1000-0)^2 + (-1000-0)^2}{2}$$

$$= 1000000$$

Mean vs Variance



Variance measures risk

Variance grows faster than the mean