# Distance & Similarity

Boston University CS 506 - Lance Galletti

| Refund | Marital Status | Income | Age |
| --- | --- | --- | --- |

| Refund | Marital Status | Income | Age |
|---|---|---|---|
| 1 | Single | 125k | 25 |

| Refund | Marital Status | Income | Age |
|---|---|---|---|
| 1 | Single | 125k | 25 |
| 0 | Married | 100k | 27 |

| Refund | Marital Status | Income | Age |
| --- | --- | --- | --- |
| 1 | Single | 125k | 25 |
| 0 | Married | 100k | 27 |
| 0 | Single | 70k | 22 |

| Refund | Marital Status | Income | Age |
|---|---|---|---|
| 1 | Single | 125k | 25 |
| 0 | Married | 100k | 27 |
| 0 | Single | 70k | 22 |
| 1 | Married | 120k | 30 |
| 0 | Divorced | 90k | 28 |
| 0 | Married | 60k | 37 |
| 1 | Divorced | 220k | 24 |
| 0 | Single | 85k | 23 |
| 0 | Married | 75k | 23 |
| 0 | Single | 90k | 26 |

# Data

$$\textbf{n} \text{ data points} \left\{ \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m} \text{ features}} \right.$$

# Data

$$\textbf{n}\text{ data points}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m}\text{ features}}\right.$$

**Data point i**

# Data

Attribute / feature j

**n** data points

$$\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}$$

Data point i

**m** features

# Data

**Attribute / feature j**

**Feature j of data point i**

$$\begin{array}{c}\textbf{n}\text{ data}\\\text{points}\end{array}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & \boxed{x_{ij}} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m}\text{ features}}\right.$$

**Data point i**

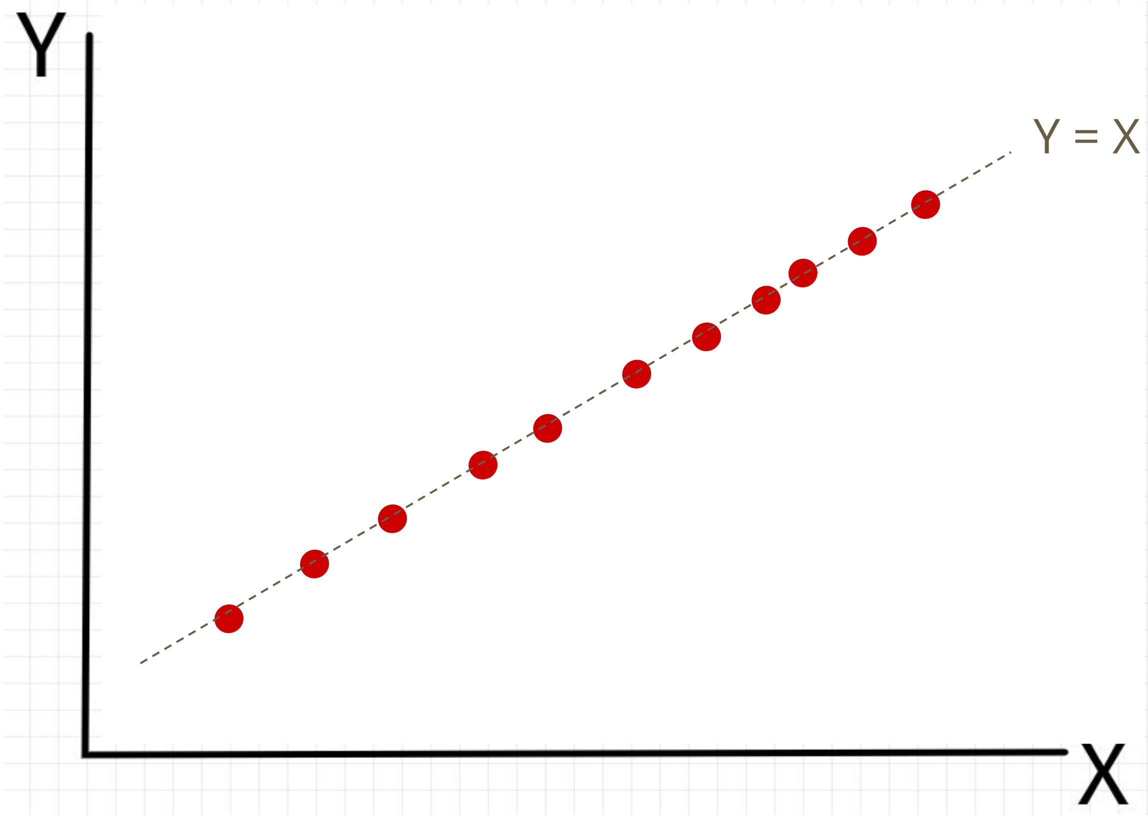
# Data

**How similar are feature j and feature m ?**

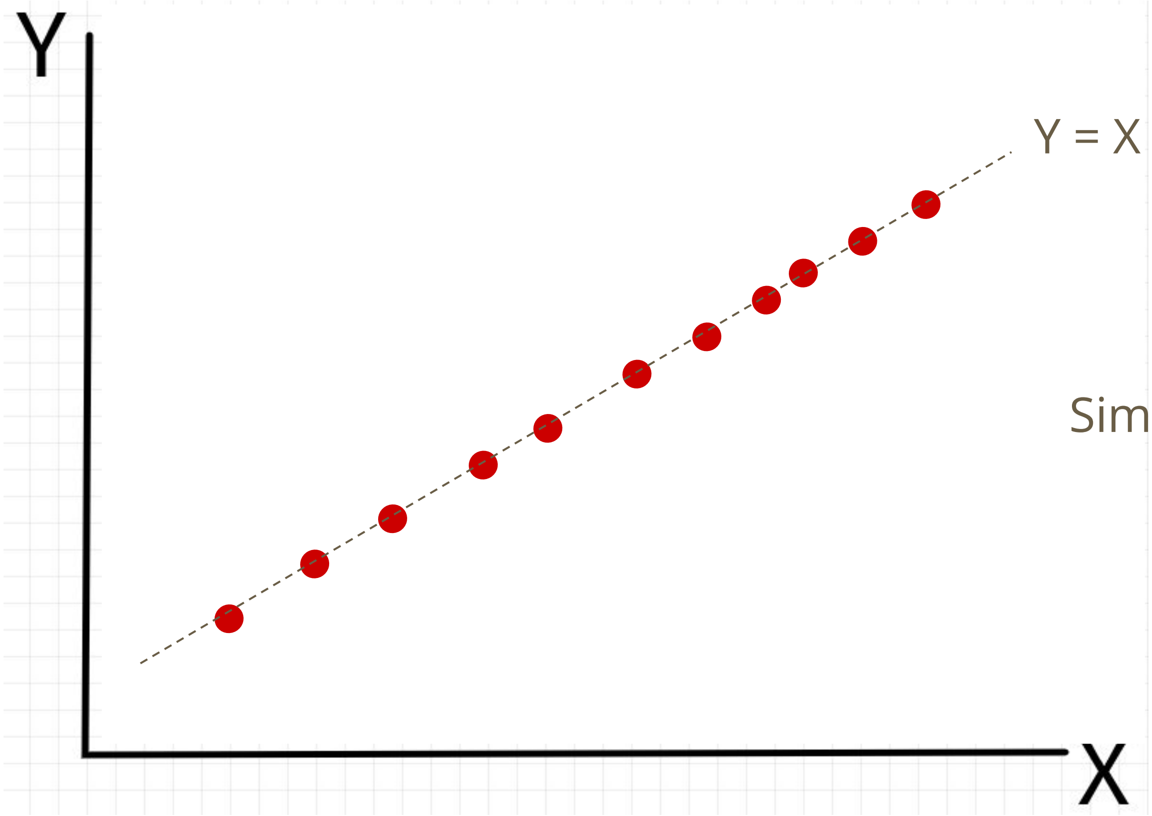
$$\textbf{n}\text{ data points}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m}\text{ features}}\right.$$

# What does it mean for two features to be similar?

# What does it mean for two features to be similar?

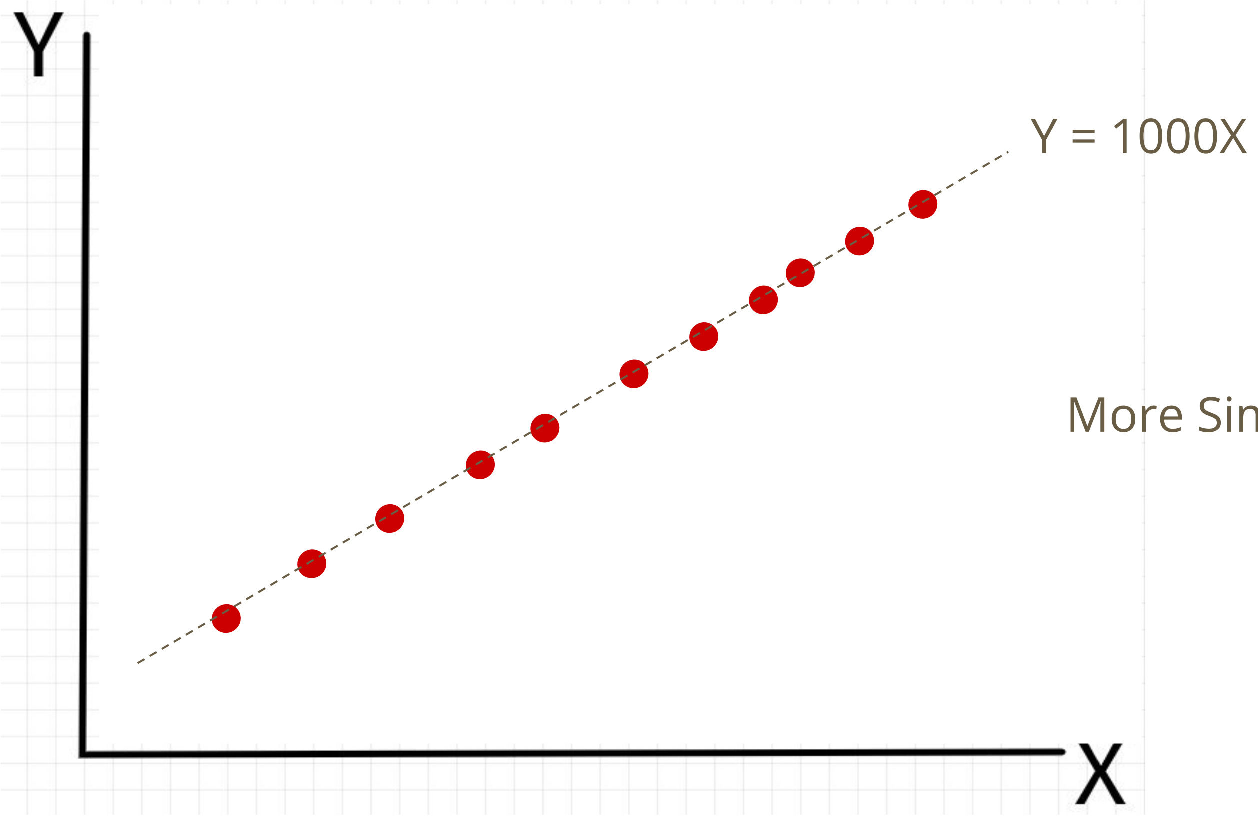


# What does it mean for two features to be similar?



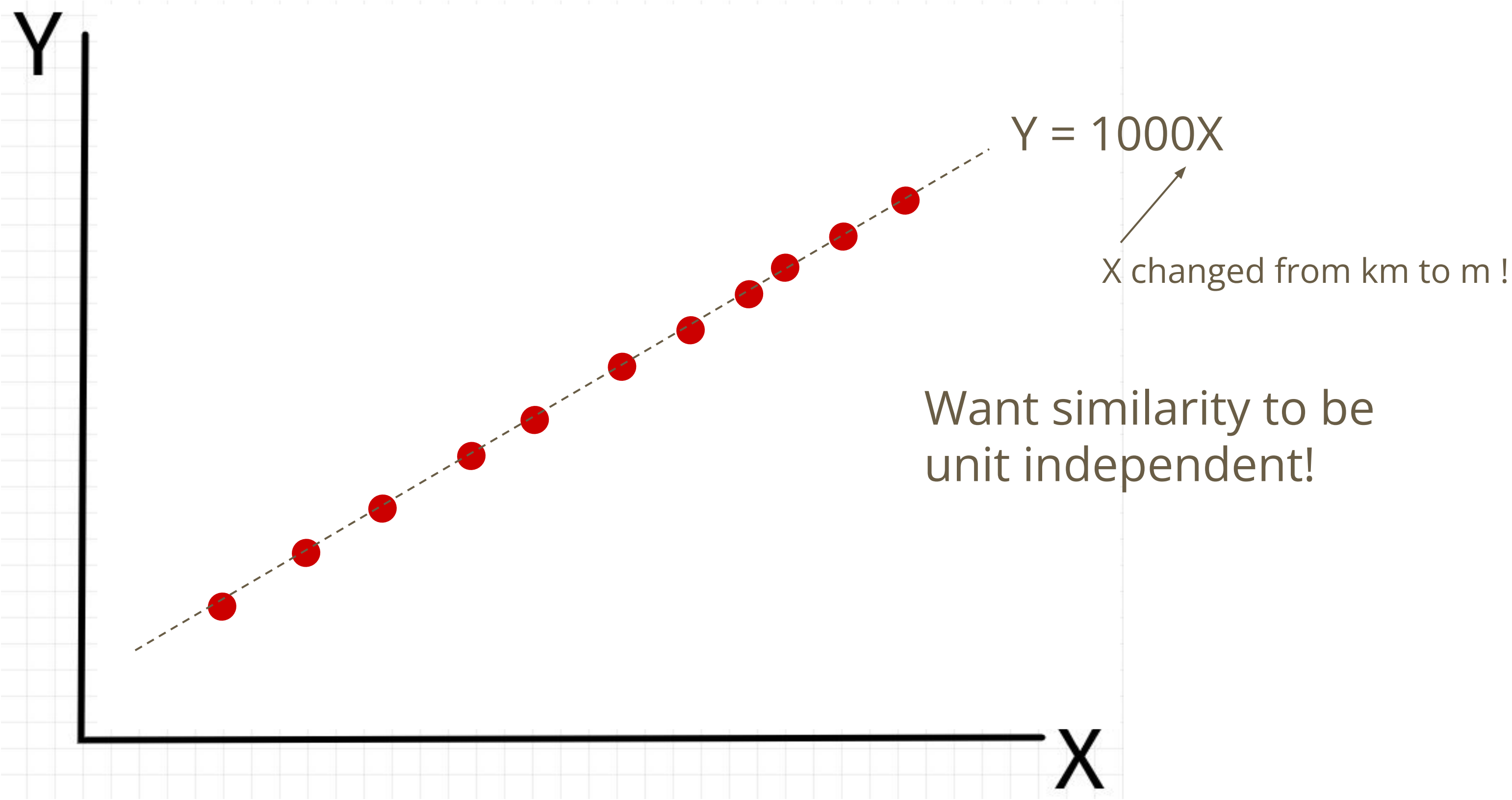

Similar

# What does it mean for two features to be similar?



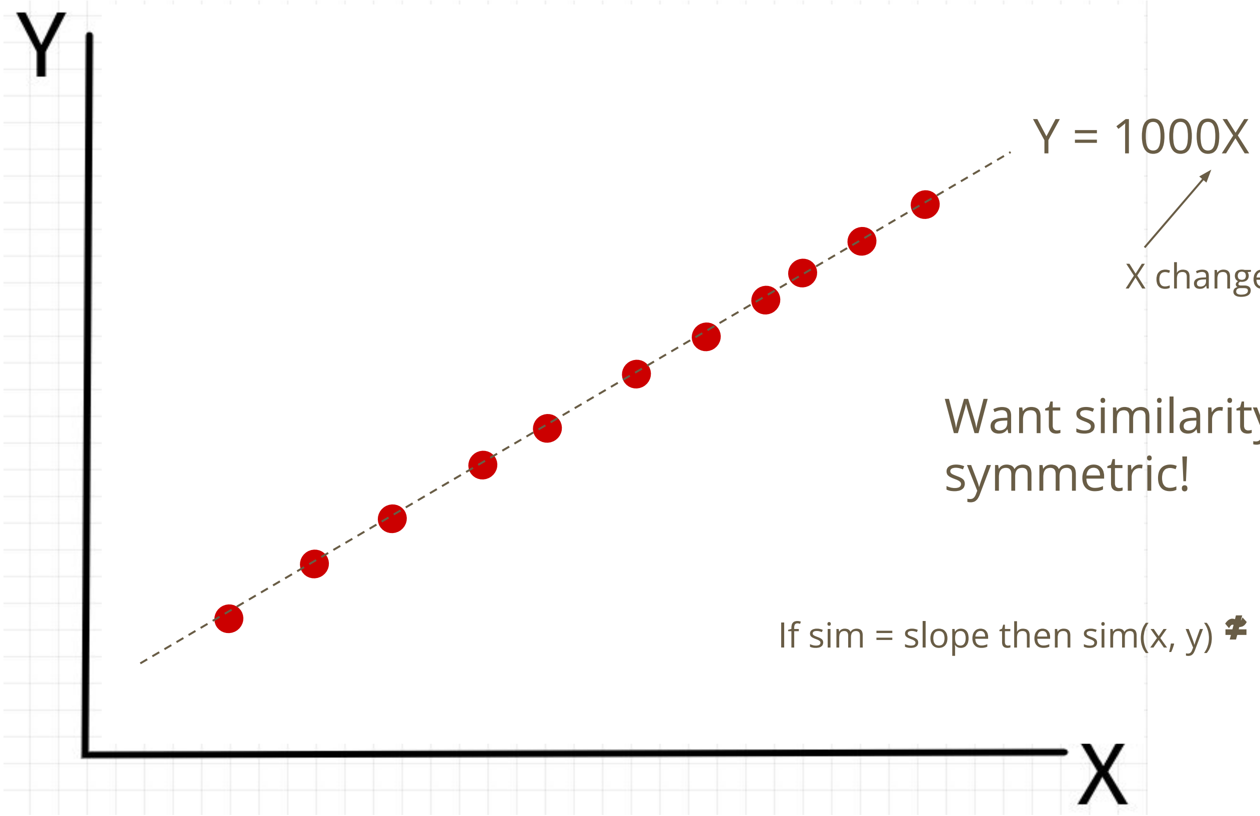

More Similar ?

# What does it mean for two features to be similar?




Want similarity to be unit independent!

# What does it mean for two features to be similar?

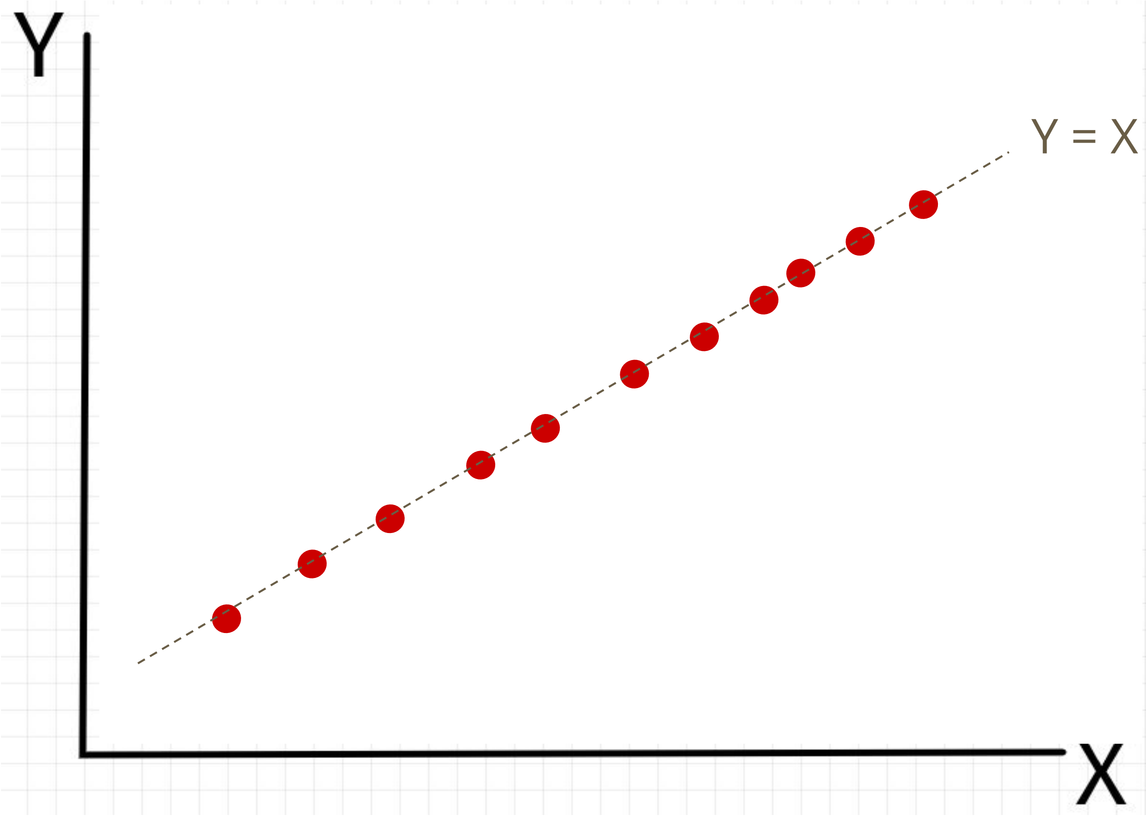



X changed from km to m !

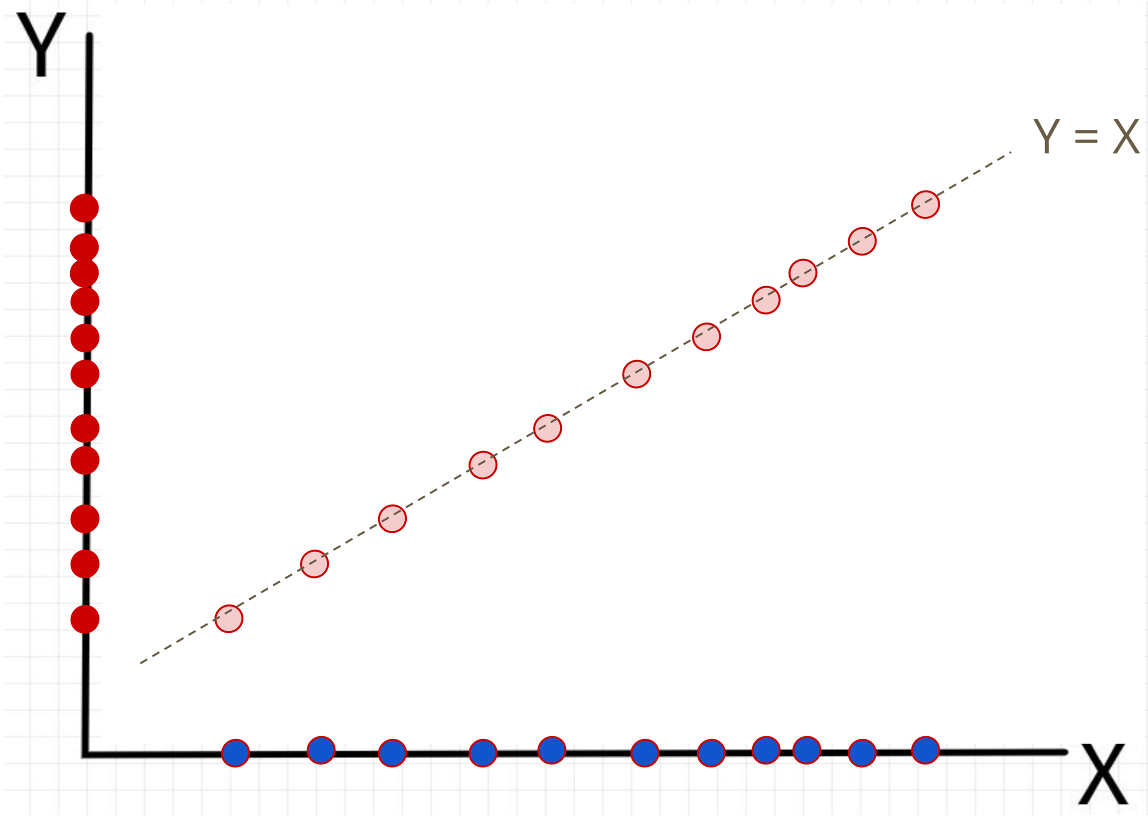
Want similarity to be symmetric!

If sim = slope then sim(x, y) $\neq$ sim(y,x)

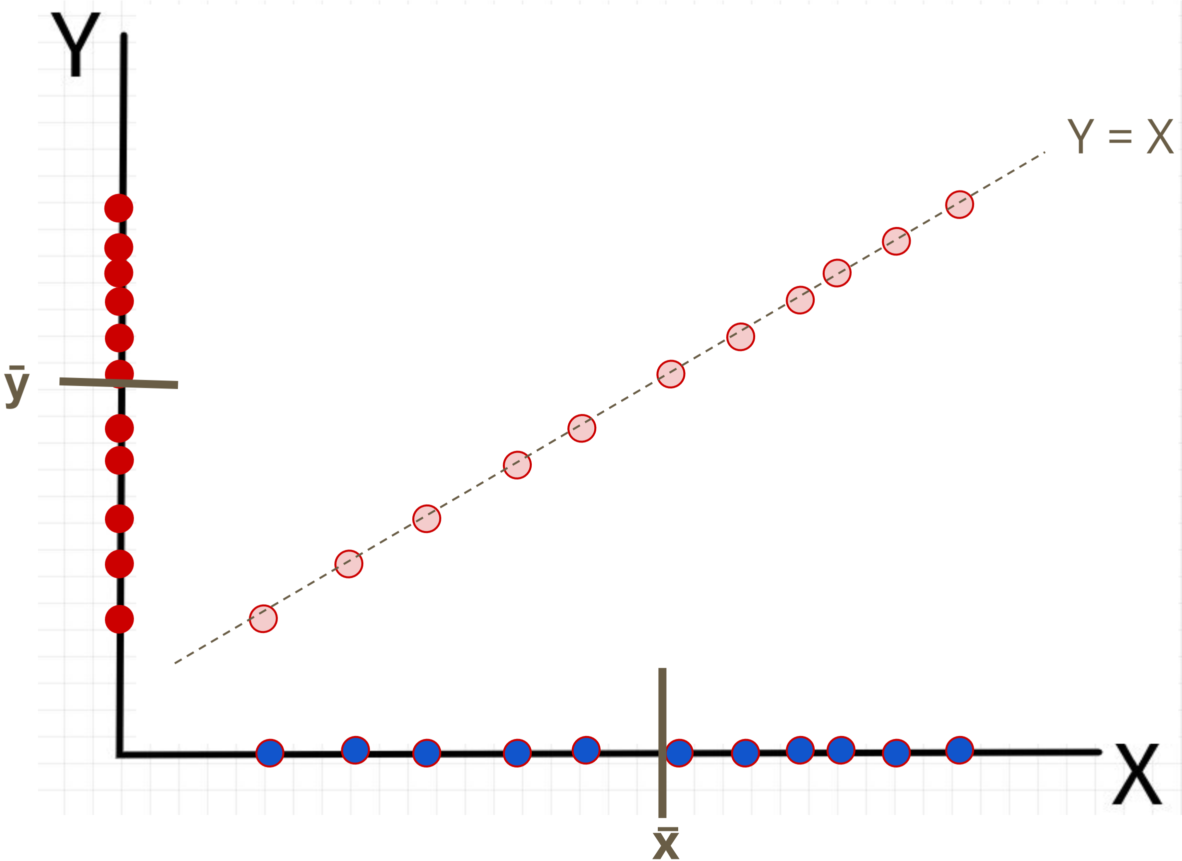
# Recall: Variance



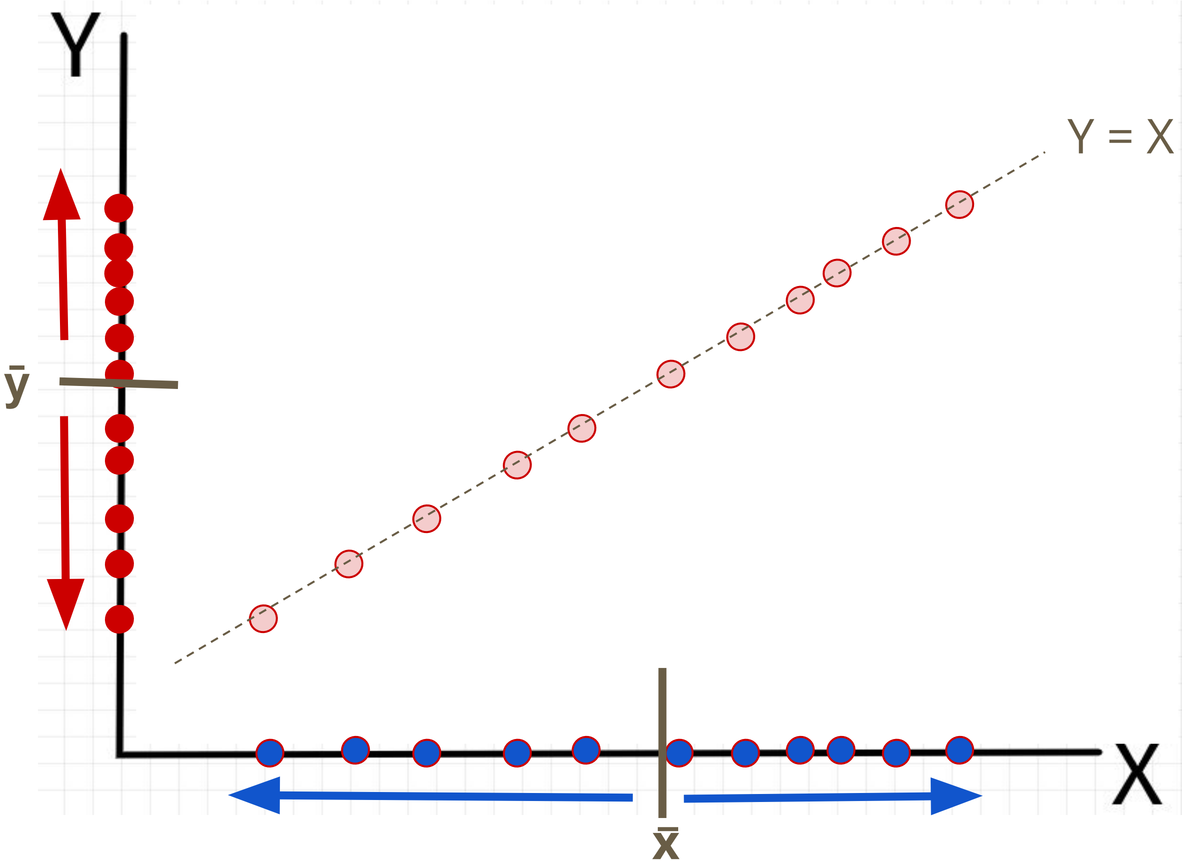
# Recall: Variance



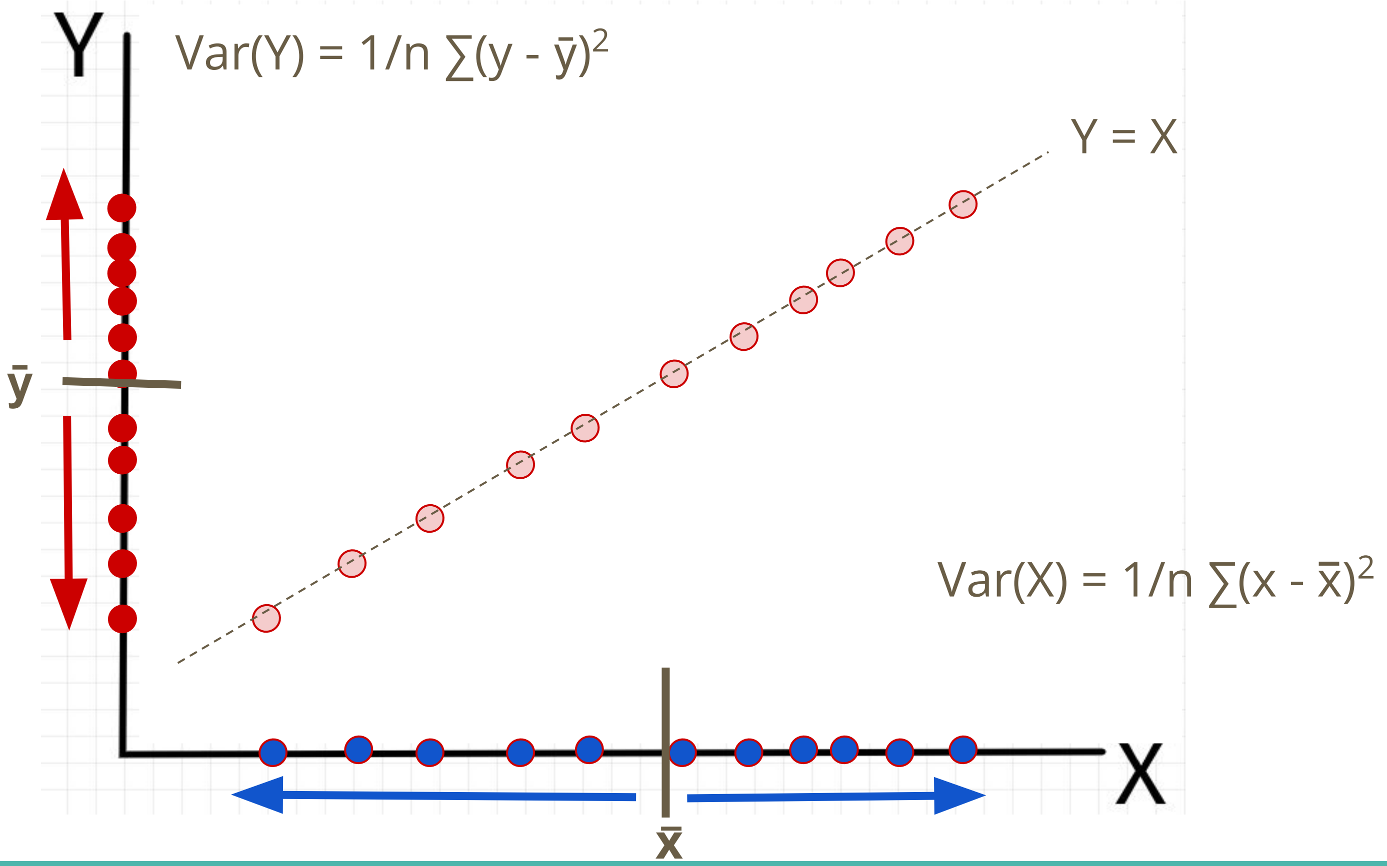
# Recall: Variance



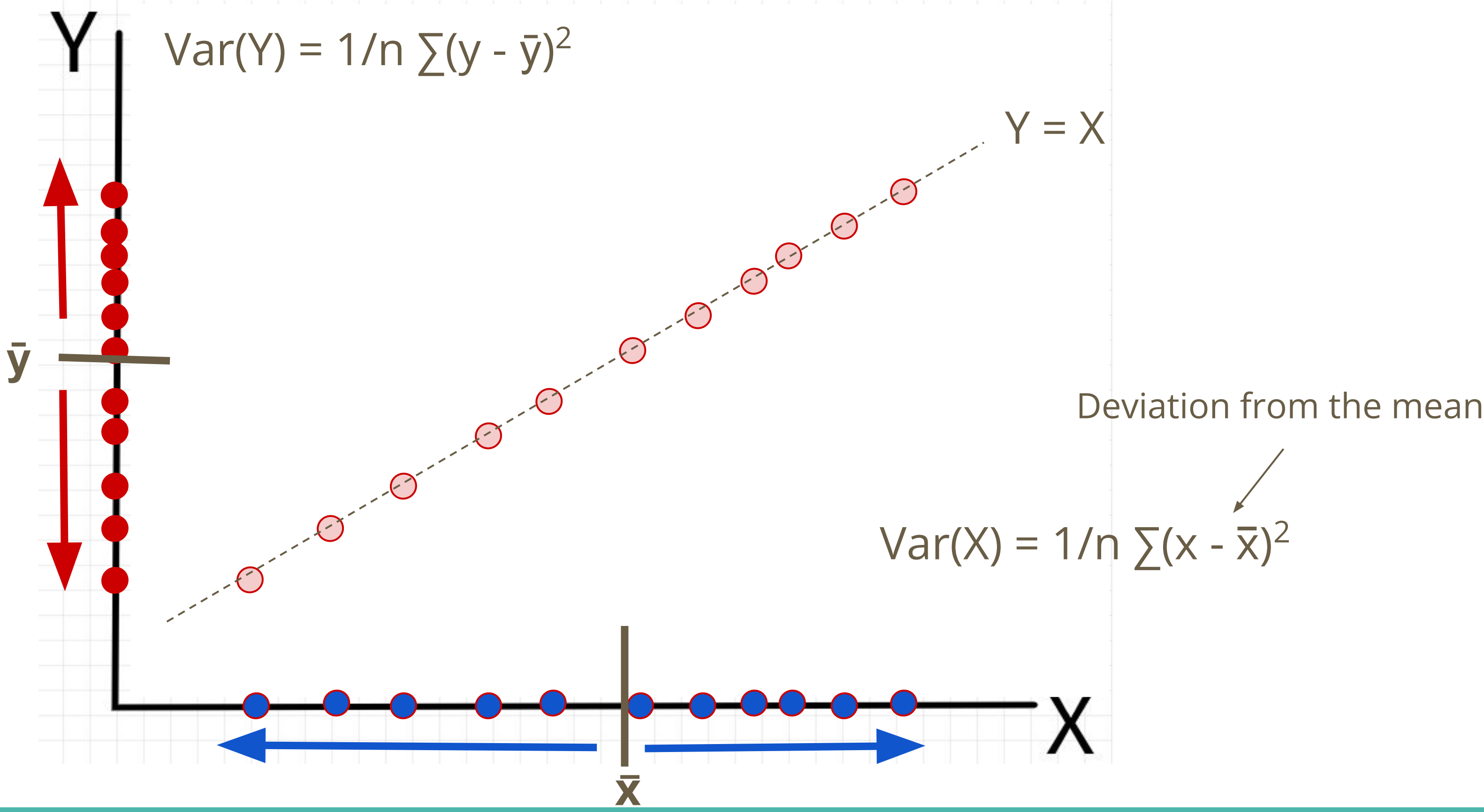
# Recall: Variance



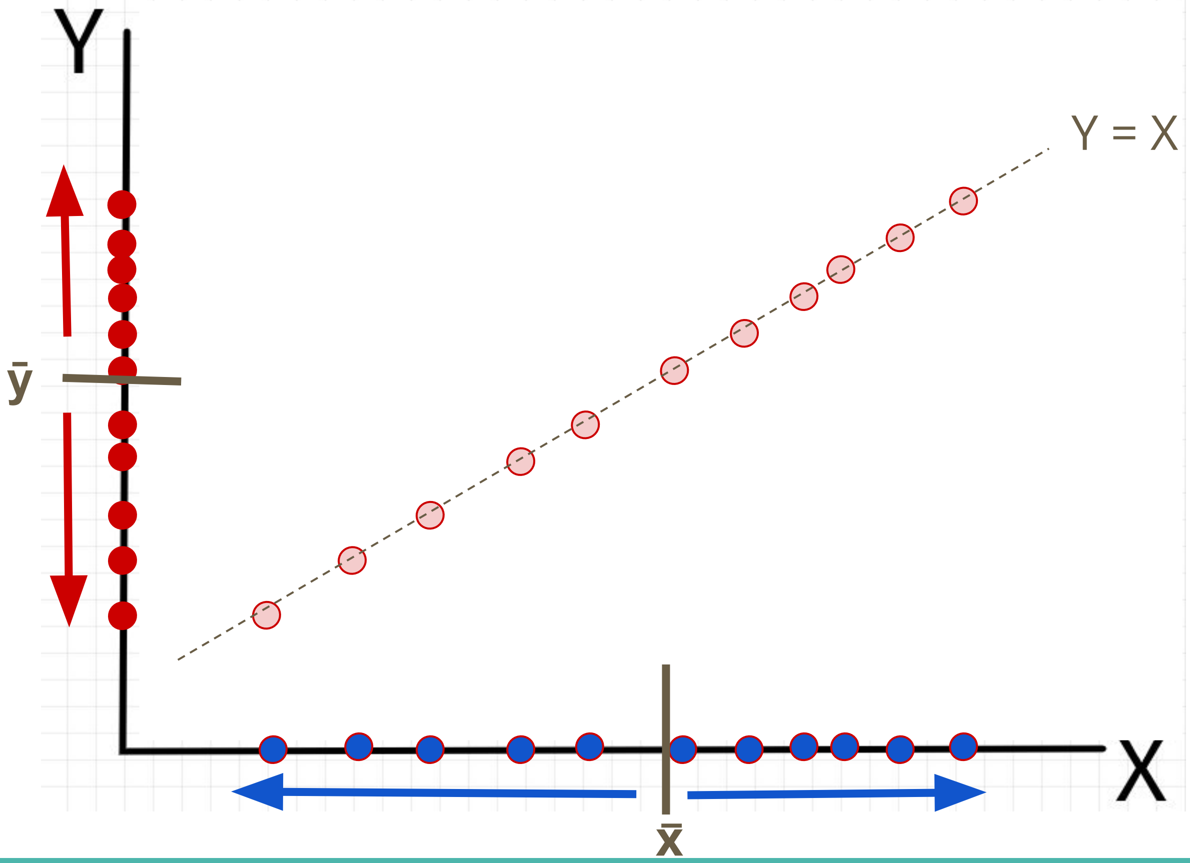
# Recall: Variance



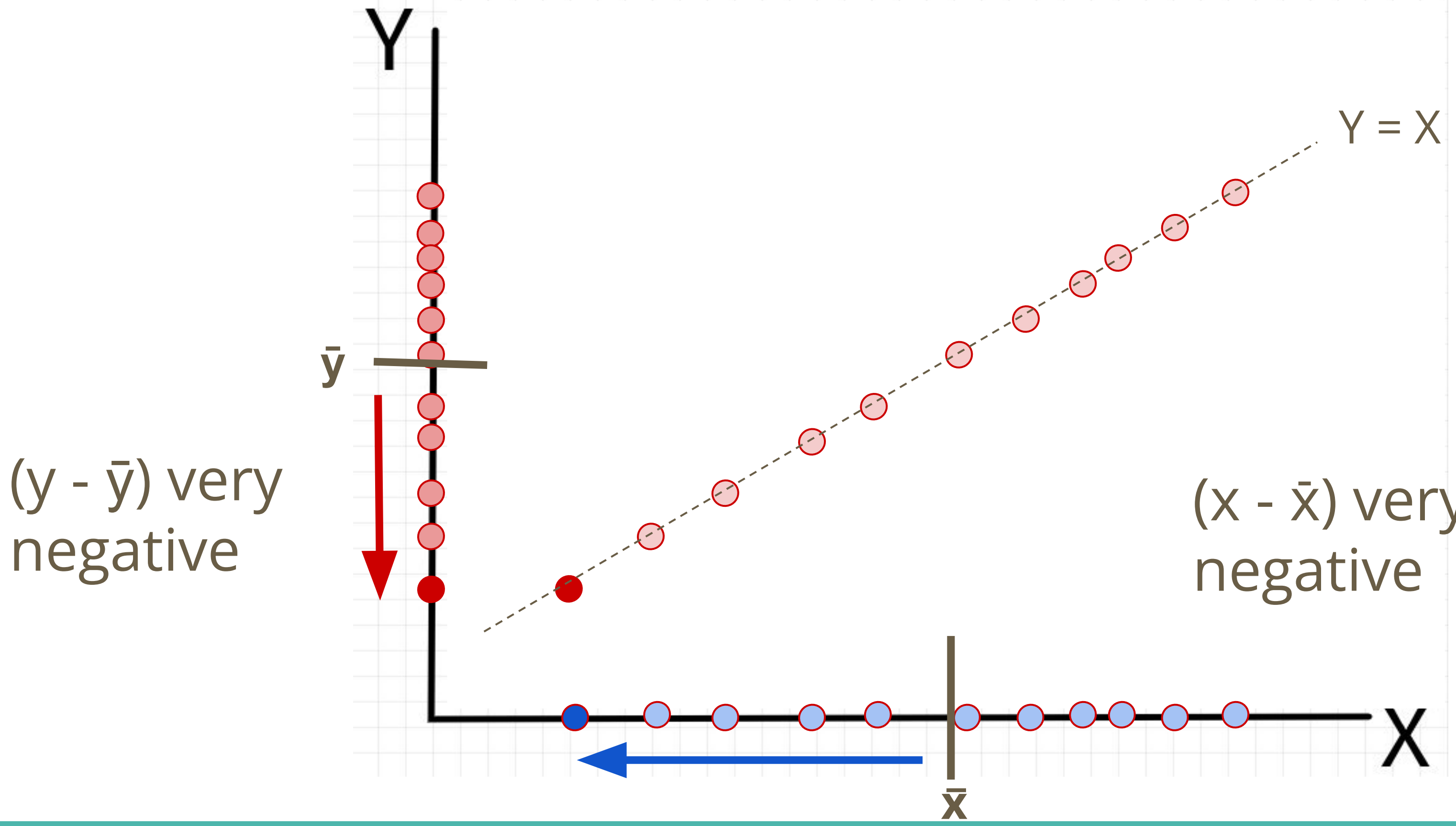
# Recall: Variance



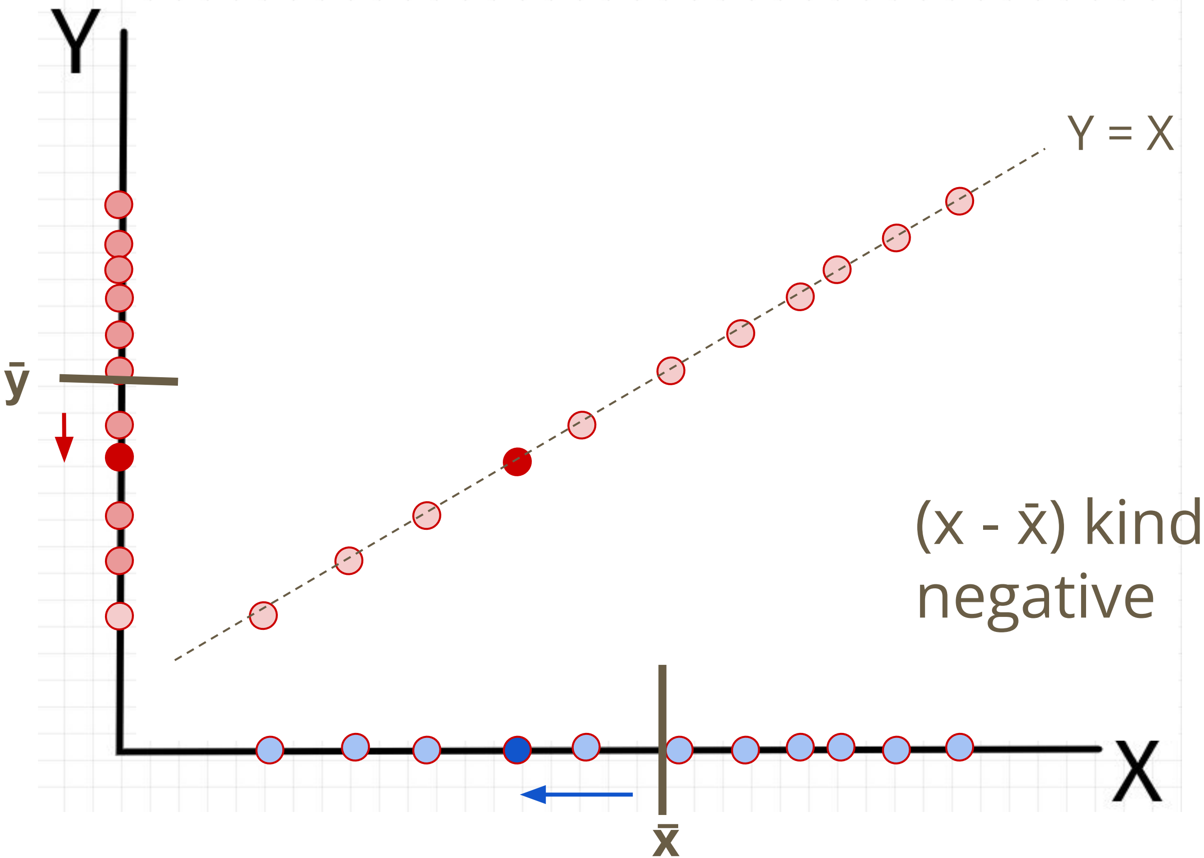
# Do deviations in X align with deviations in Y?



# Do deviations in X align with deviations in Y?



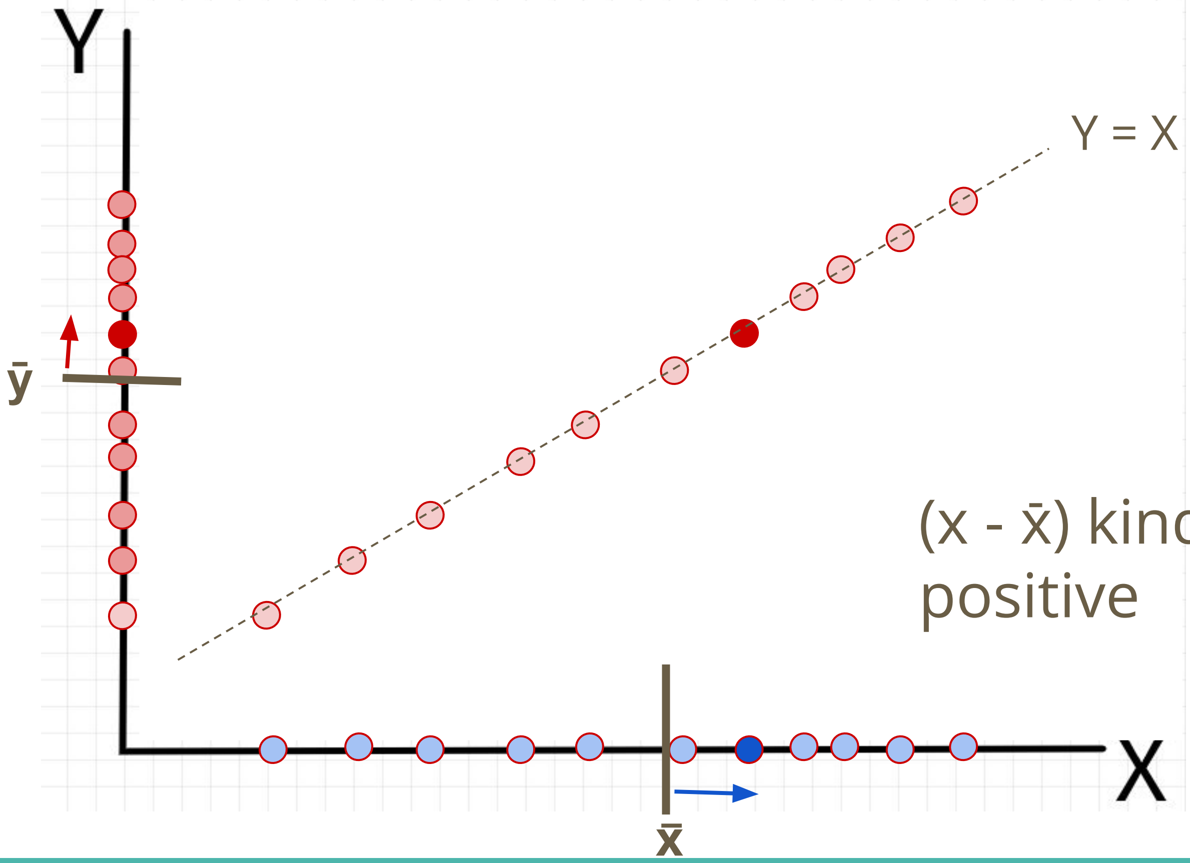
# Do deviations in X align with deviations in Y?




(y - ȳ) kinda negative

(x - x̄) kinda negative

# Do deviations in X align with deviations in Y?



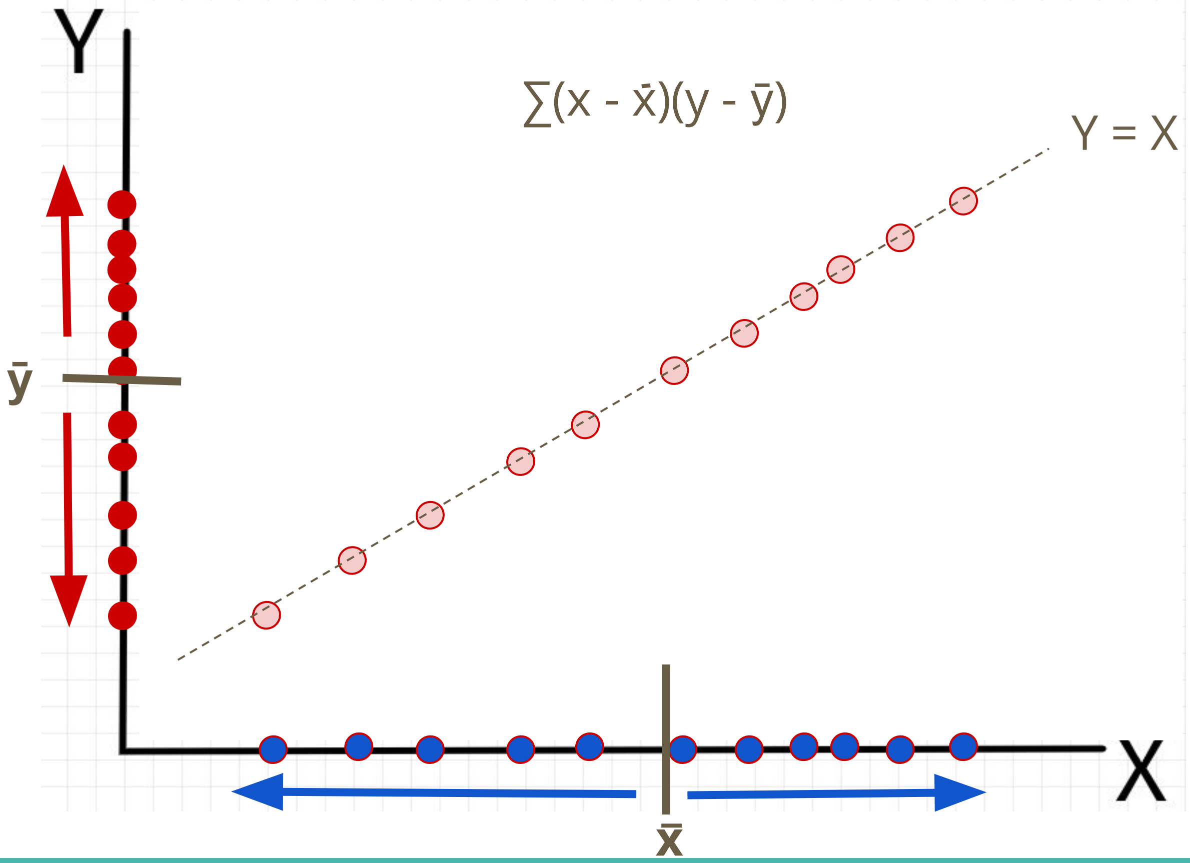

(y - ȳ) kinda positive

(x - x̄) kinda positive

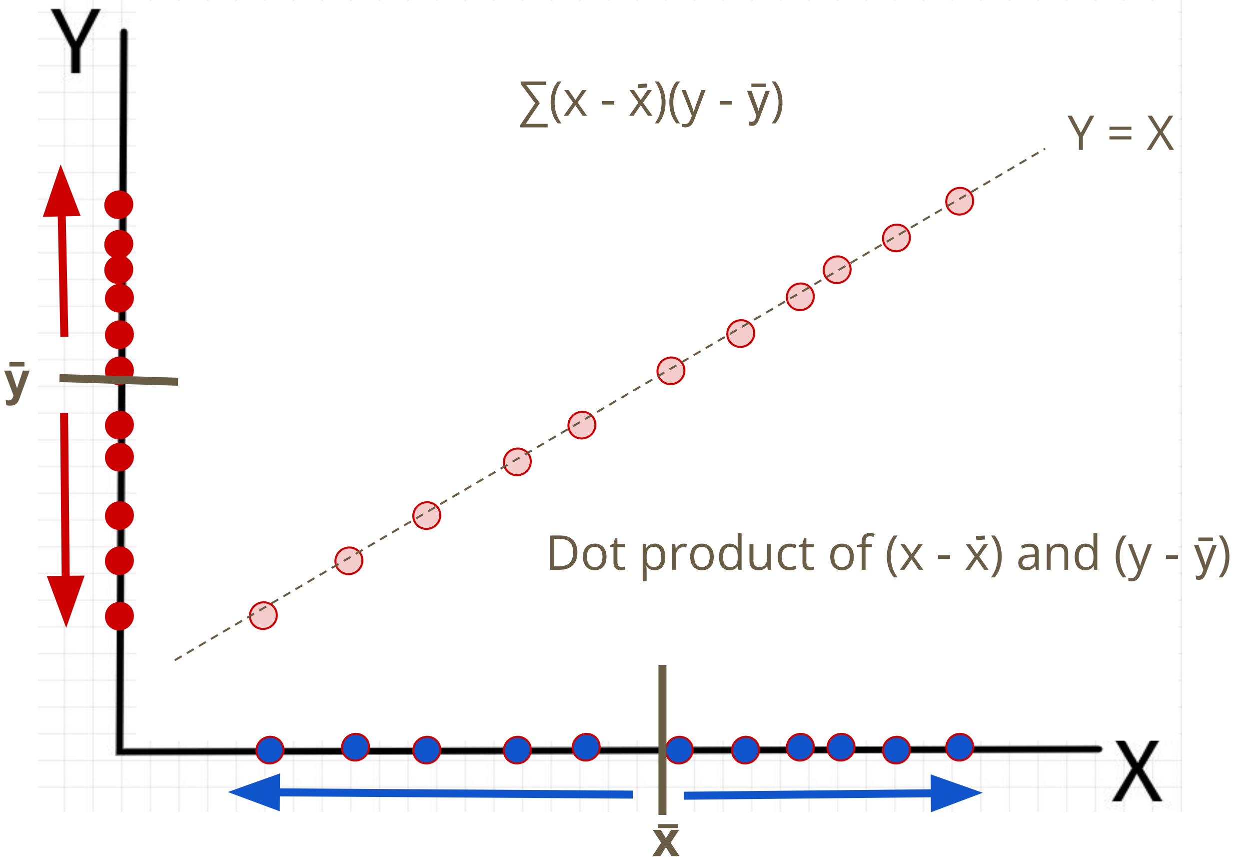
# Do deviations in X align with deviations in Y?

| $(y - \bar{y})$ | $(x - \bar{x})$ | $(y - \bar{y}) * (x - \bar{x})$ | **Interpretation** |
|---|---|---|---|
| + | + | + | move together |
| – | – | + | move together |
| + | – | – | move opposite |
| – | + | – | move opposite |

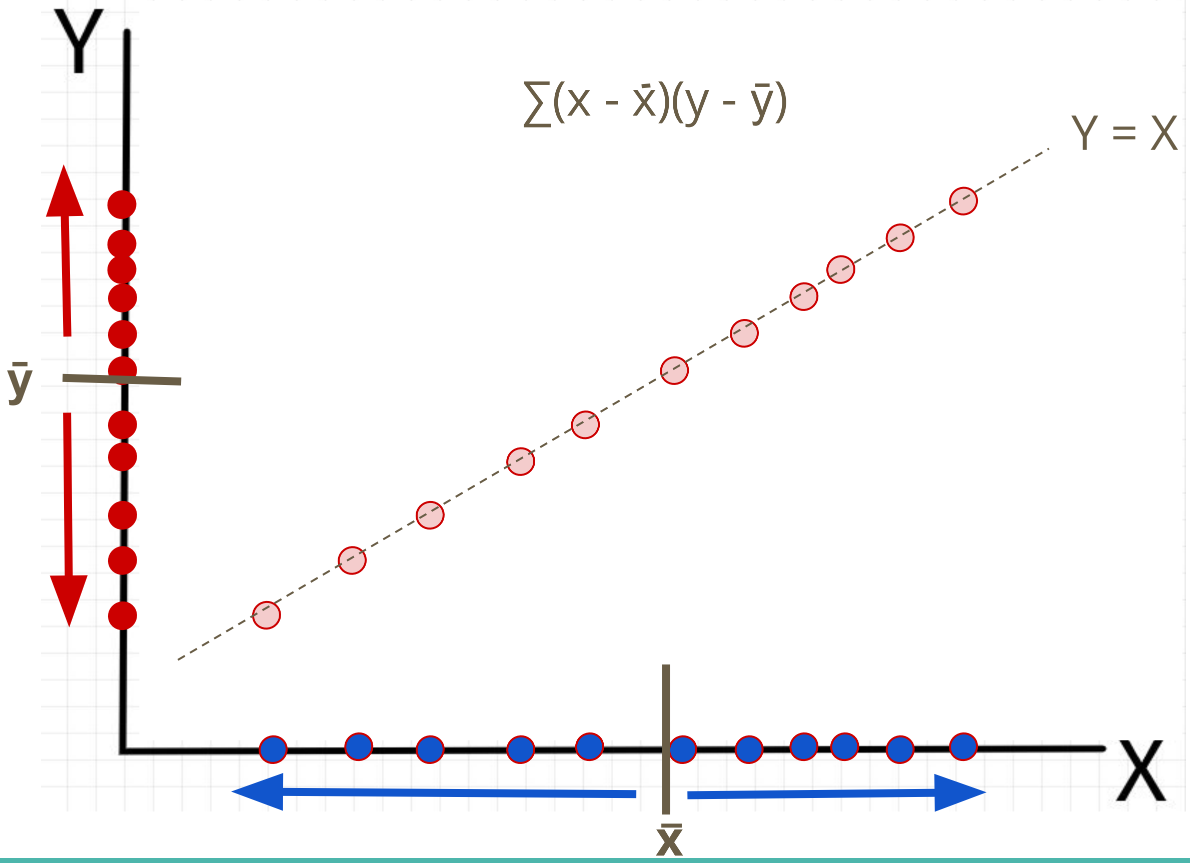
# Does this sum of products remind you of anything?



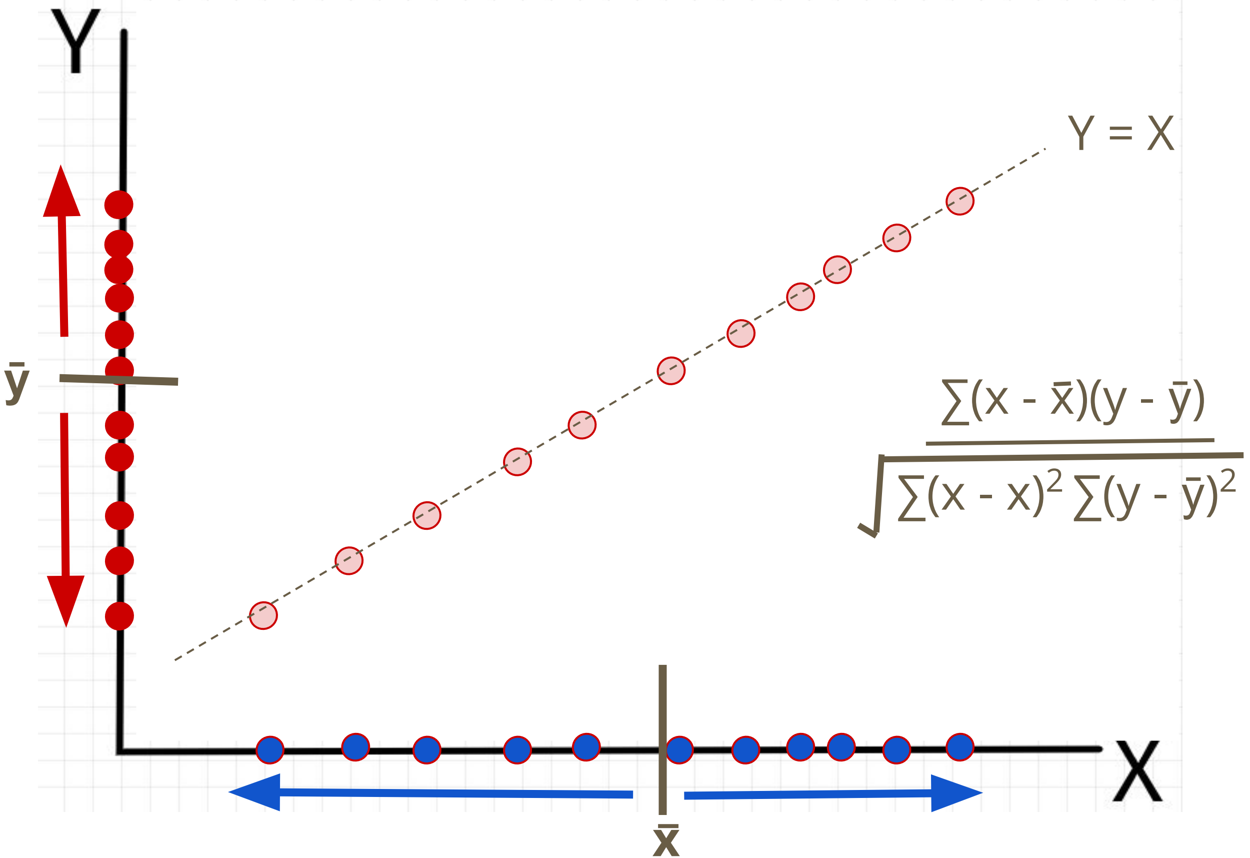
# Does this sum of products remind you of anything?



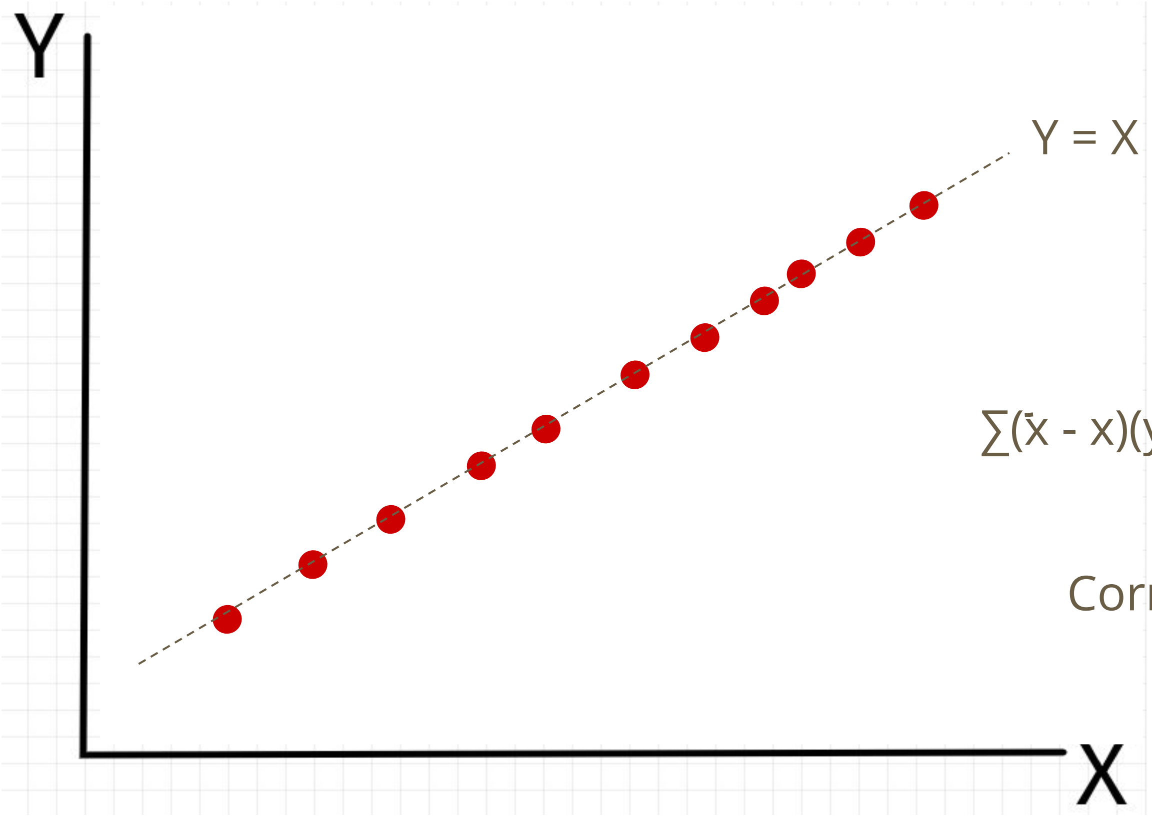
# But that product still depends on units!



# Make it relative to individual variances



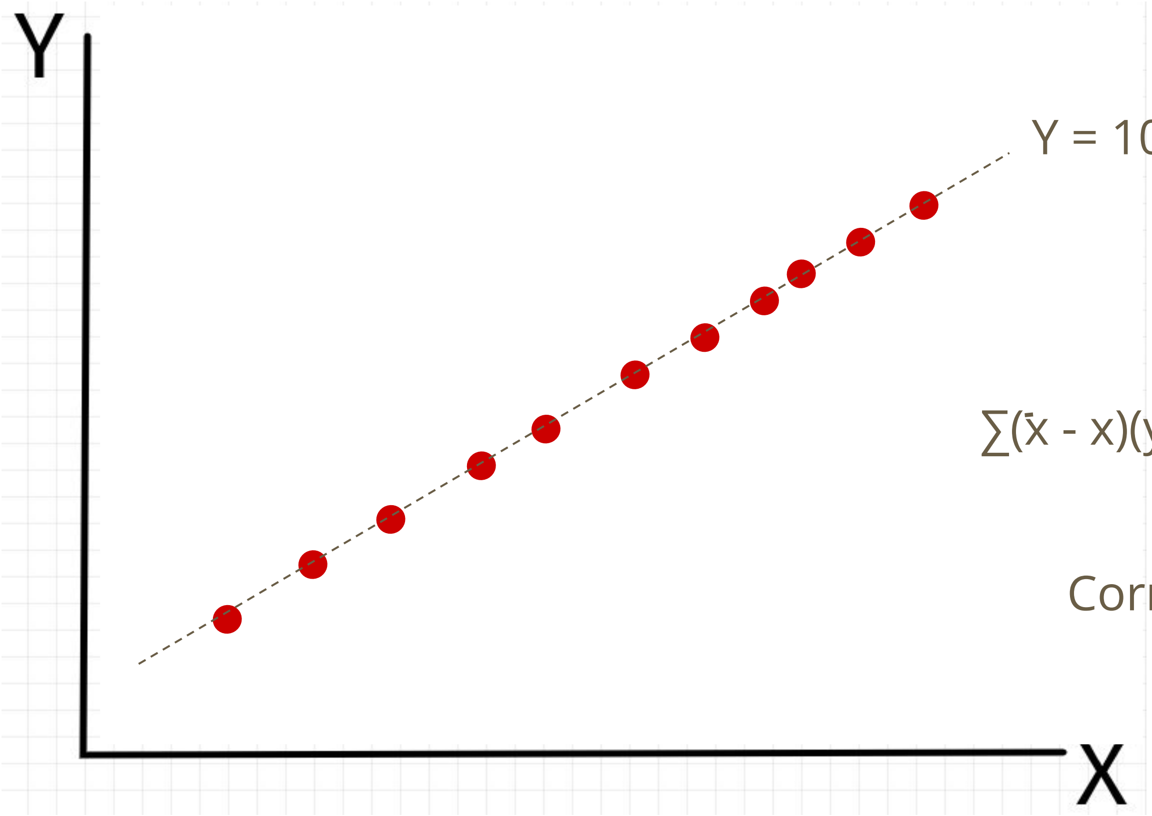
# Correlation




$$\sum(\bar{x} - x)(y - \bar{y}) \quad = \sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}$$

Correlation = 1

# Correlation

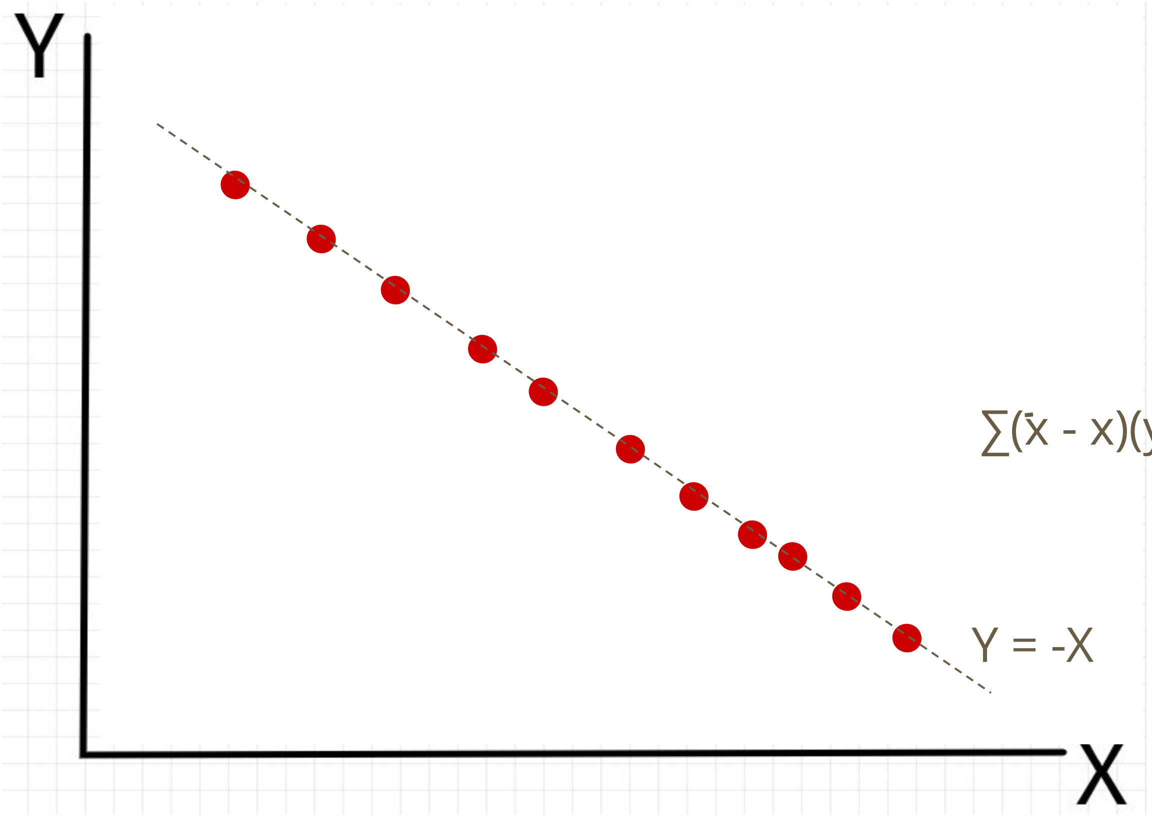



Y = 1000X

$$\sum(\bar{x} - x)(y - \bar{y}) \;=\; \sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}$$

Correlation = 1

# Correlation




Y = -X

$$\sum(\bar{x} - x)(y - \bar{y}) = -\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}$$

Correlation = -1

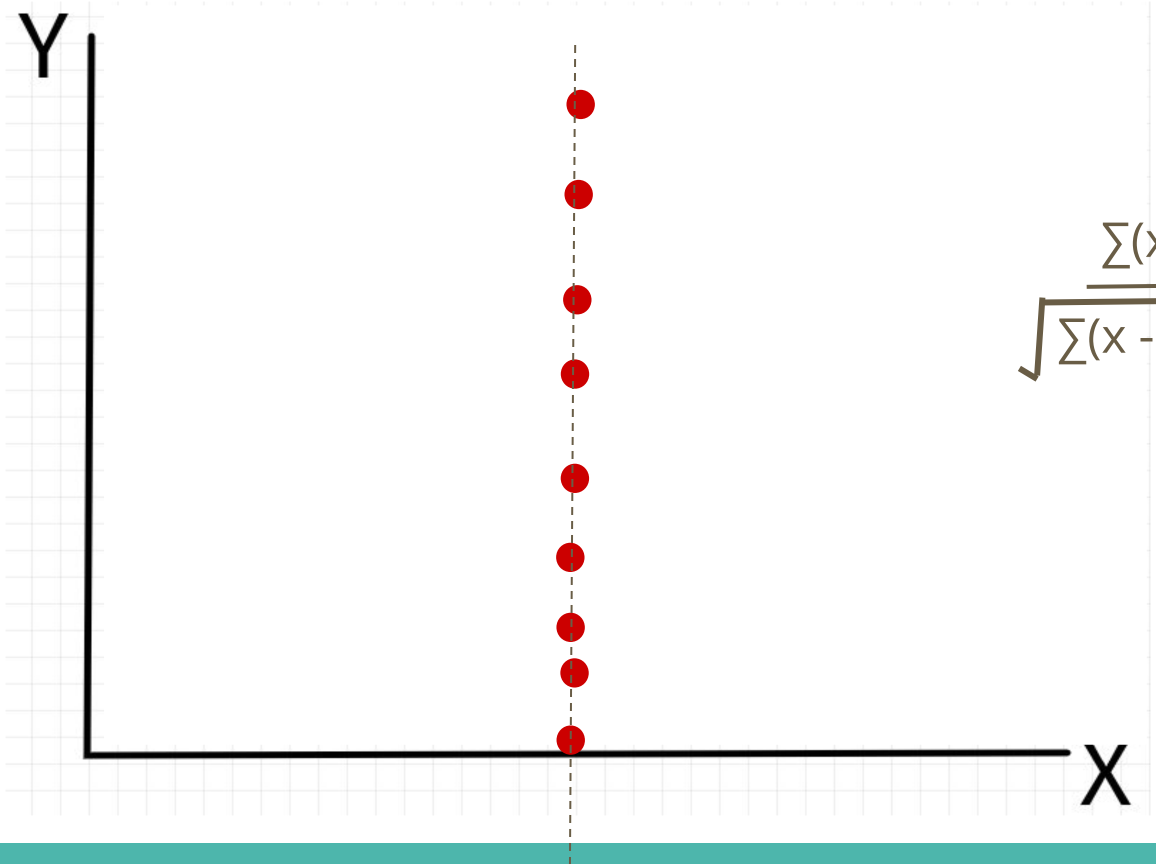
# Correlation = 0 ?

# Correlation = 0 ?


Always zero

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$

# Correlation = 0 ?




Always zero

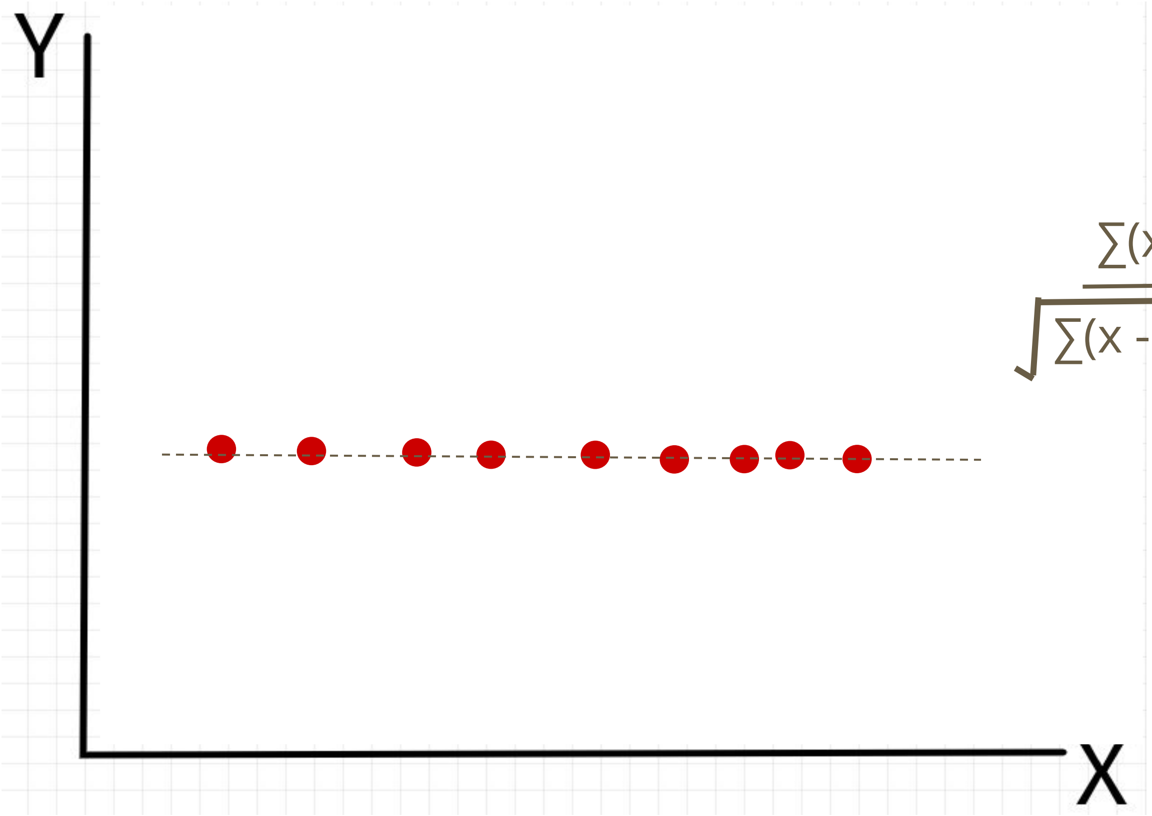
$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$

# Correlation = 0 ?


Always zero

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$

# Correlation = 0 ?




Always zero

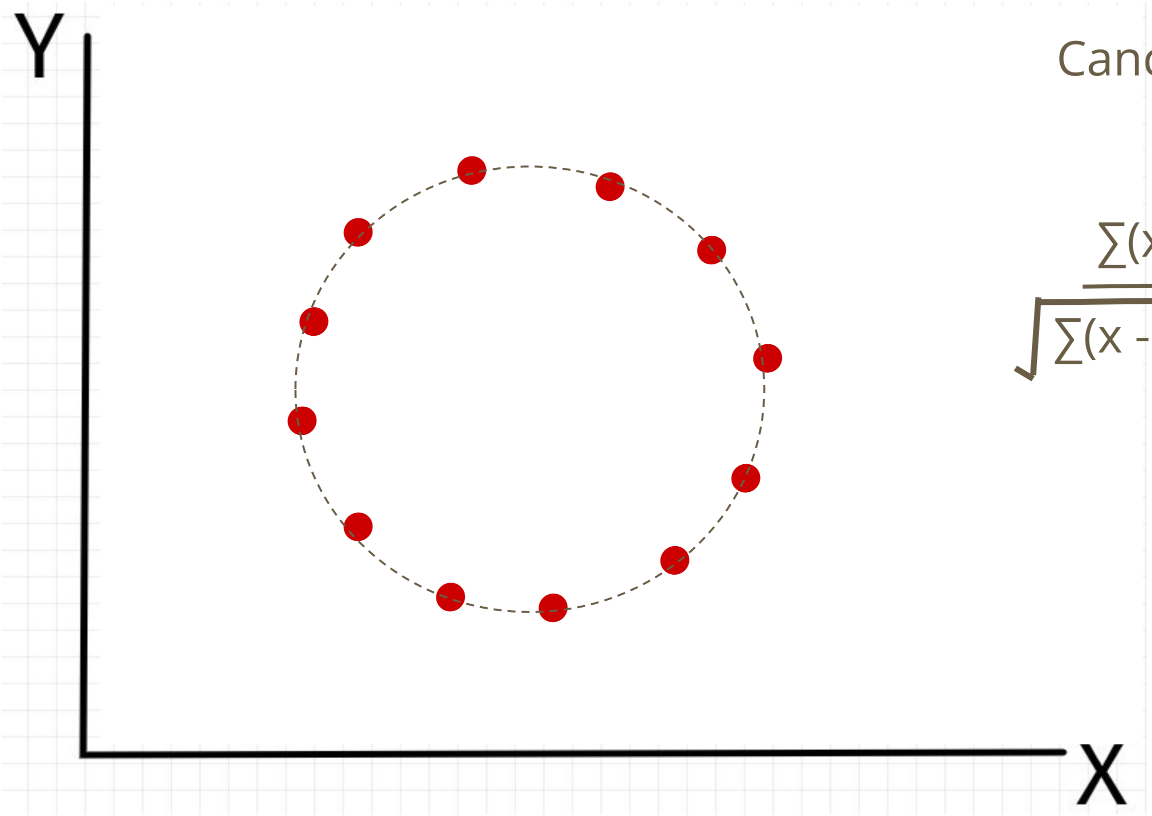
$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$

# Correlation = 0 ?


Cancel each other in the sum

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$

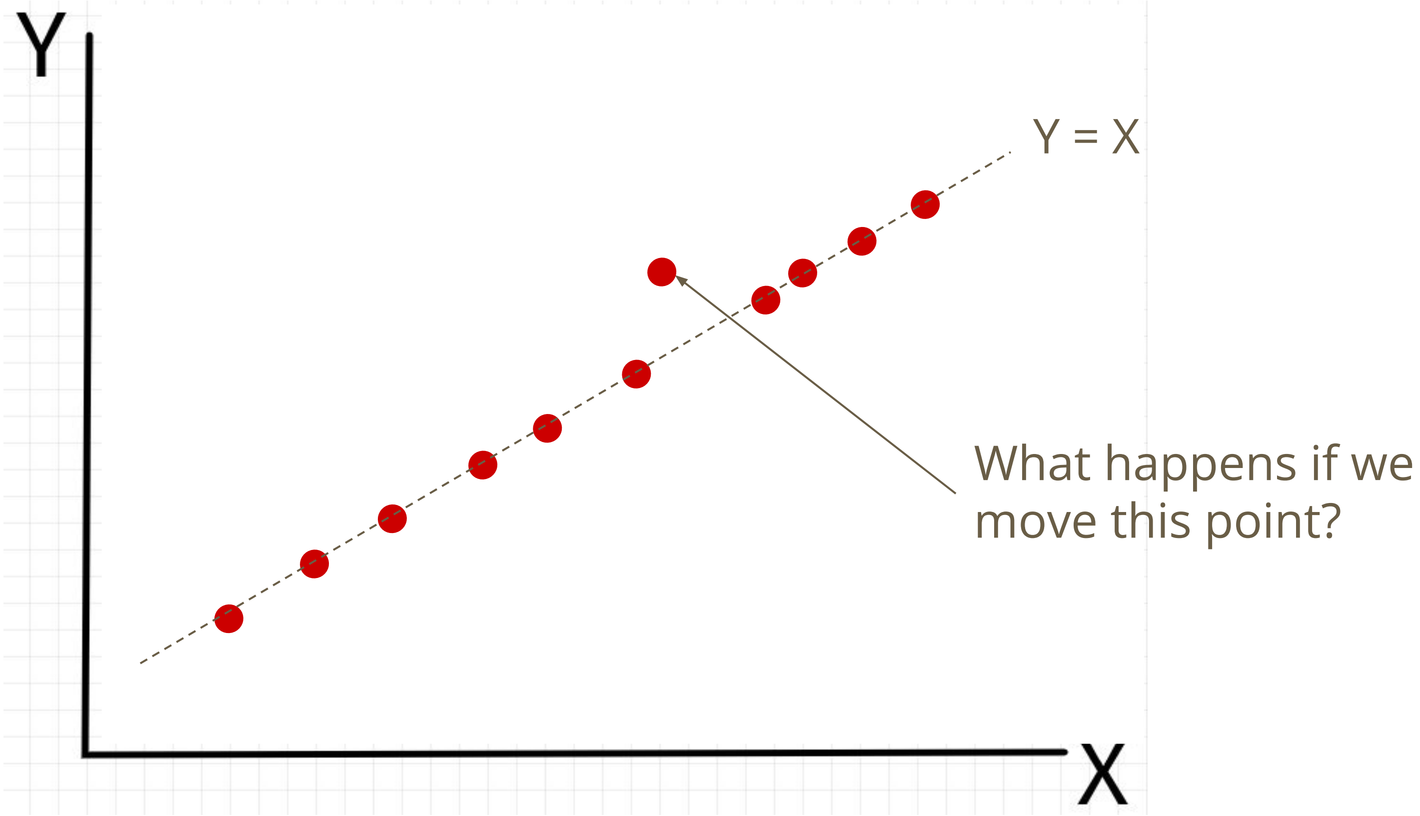
# Correlation = 0 ?



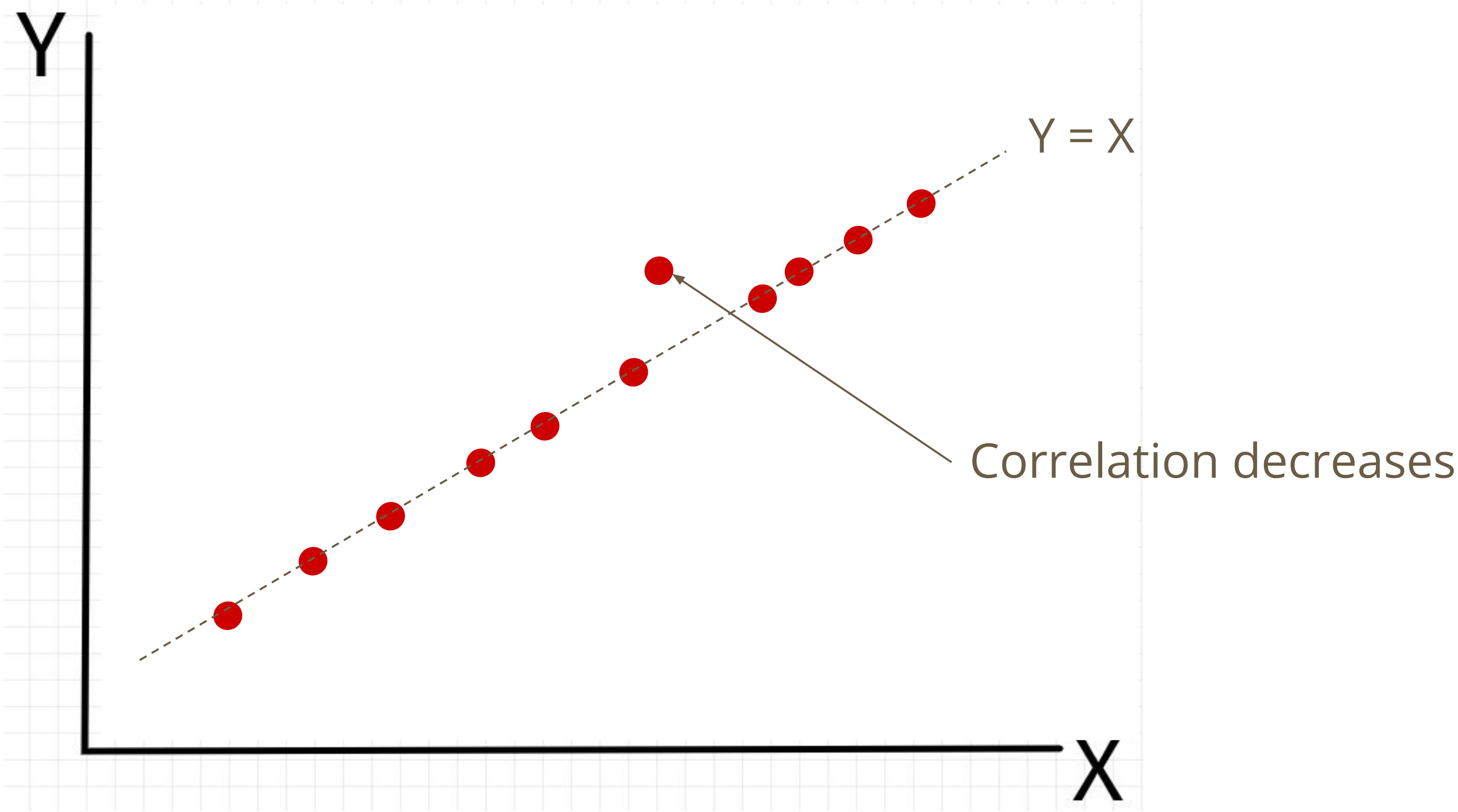

Cancel each other in the sum

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$

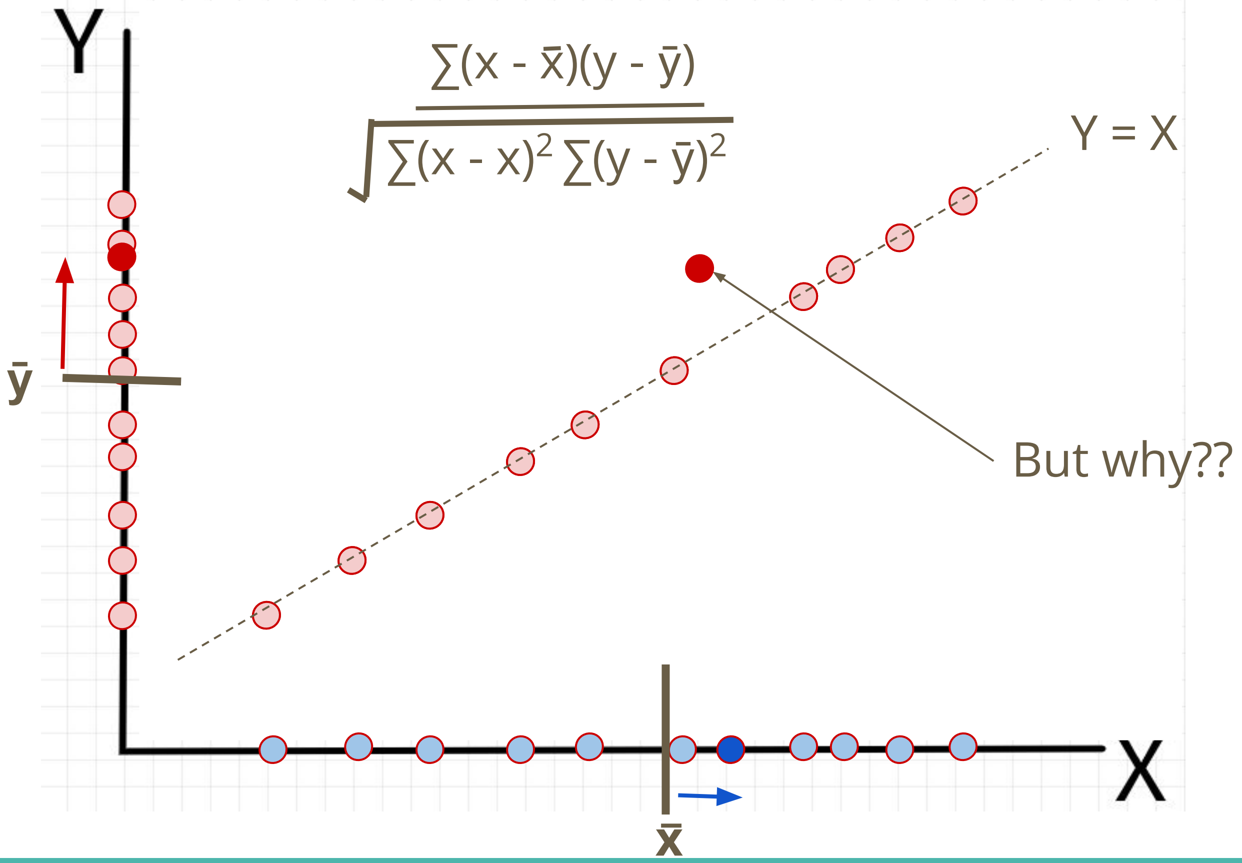
# What does it mean for two features to be similar?



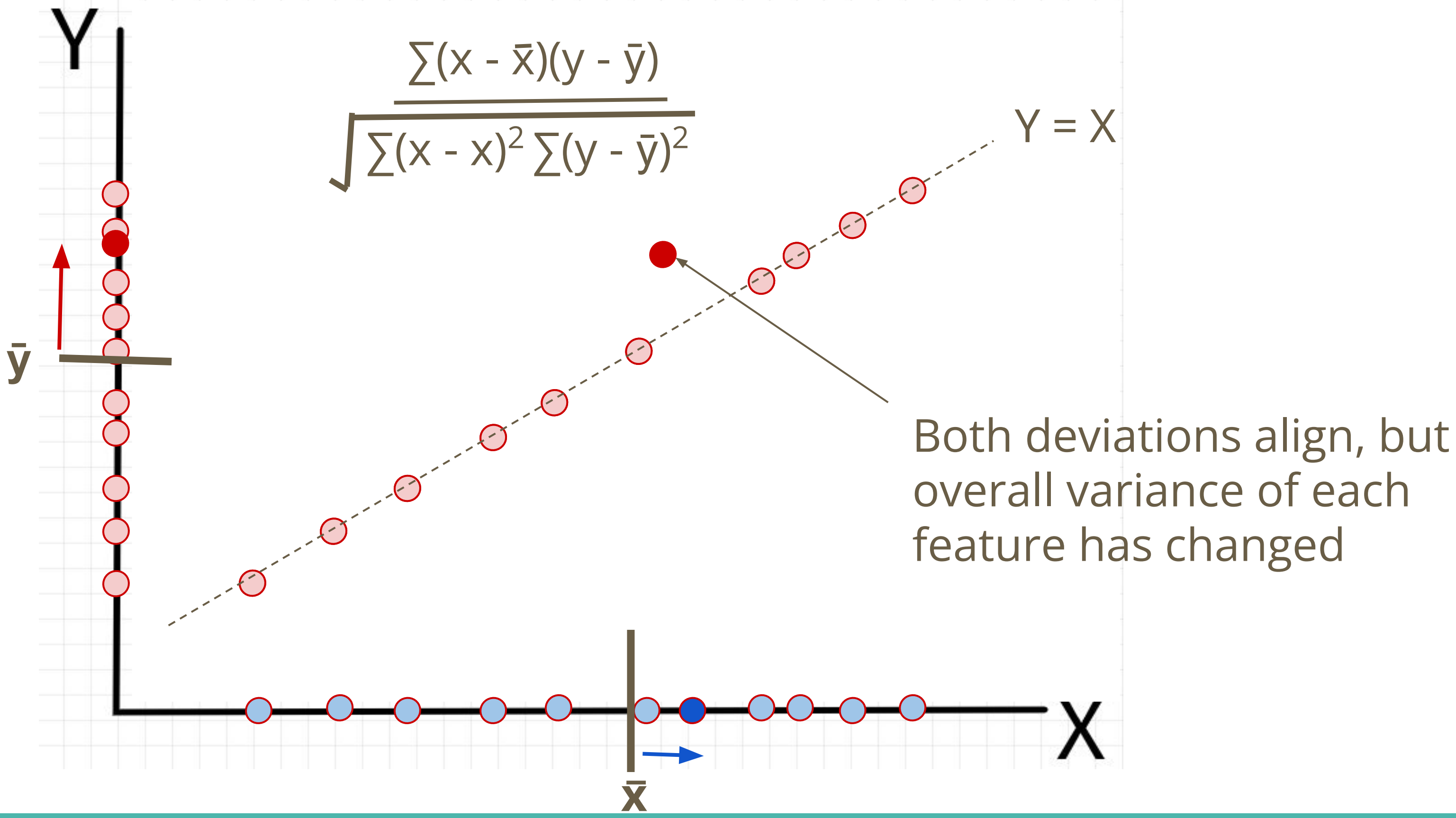
# What does it mean for two features to be similar?



# What does it mean for two features to be similar?

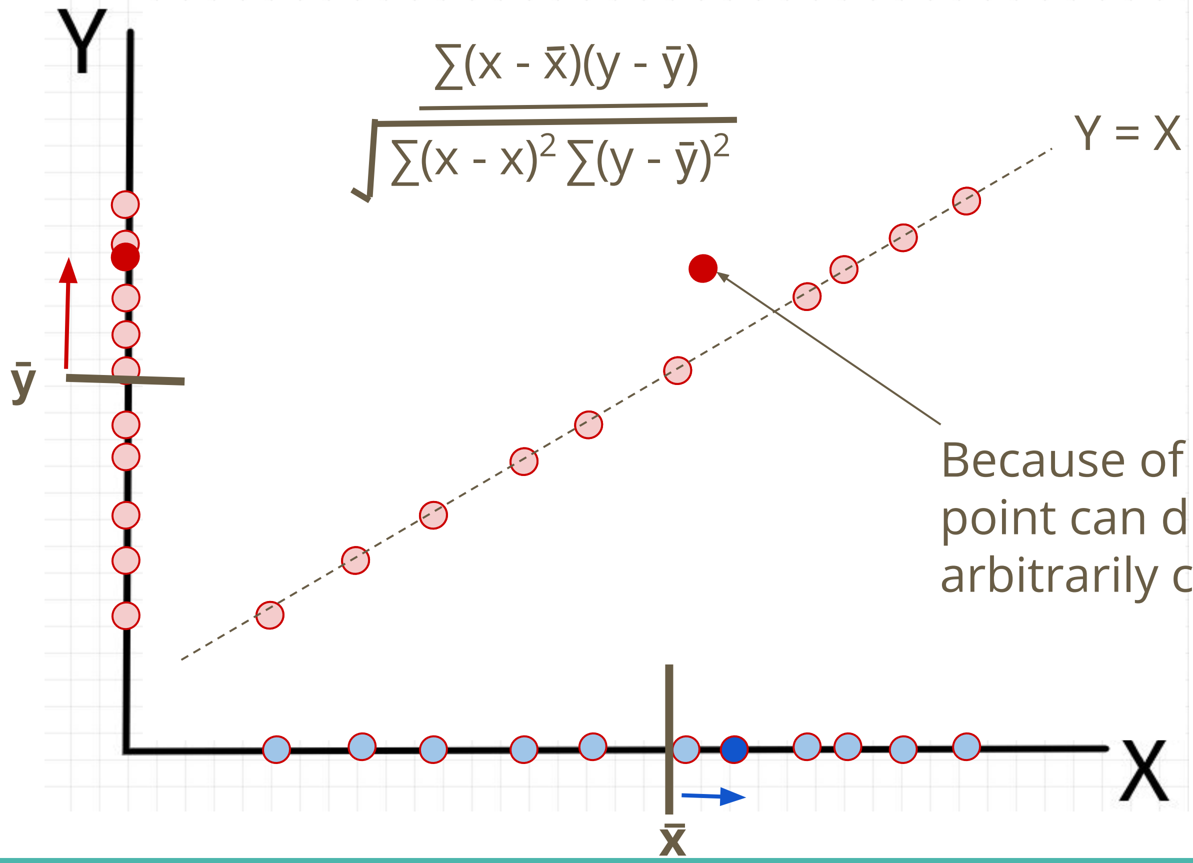


# What does it mean for two features to be similar?



# What does it mean for two features to be similar?

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - x)^2 \sum(y - \bar{y})^2}}$$



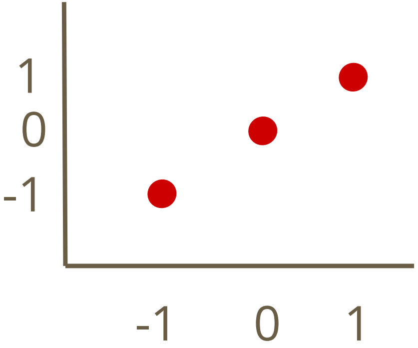

Because of that, a single point can drive correlation arbitrarily close to 1, 0, or -1

# Demo

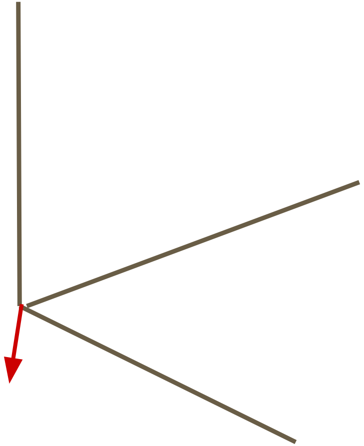
# Correlation = cosine(x, y) (if x,y centered)

Assume x & y are centered, then

$$\text{Correlation} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$



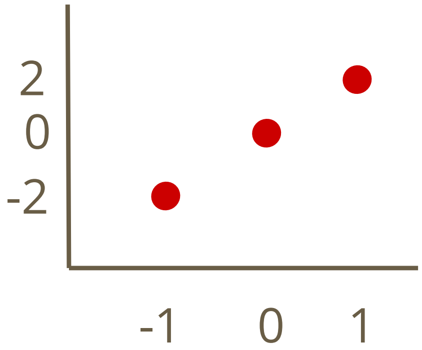

X = [-1 0 1]
Y = [-1 0 1]



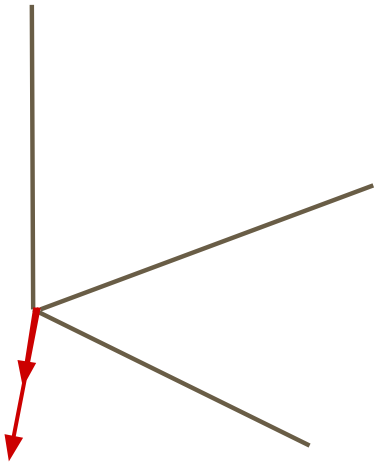
# Correlation = cosine(x, y) (if x,y centered)

Assume x & y are centered, then

$$\text{Correlation} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$



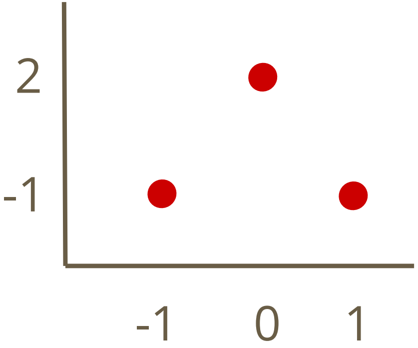

X = [-1 0 1]
Y = [-2 0 2]



# Correlation = cosine(x, y) (if x,y centered)

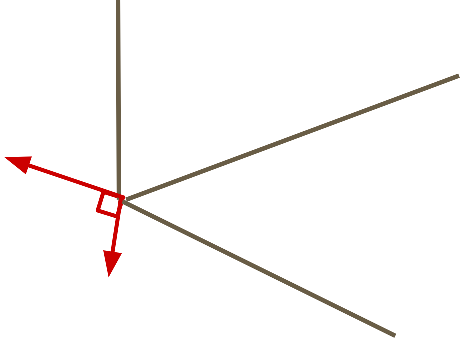
Assume x & y are centered, then

$$\text{Correlation} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

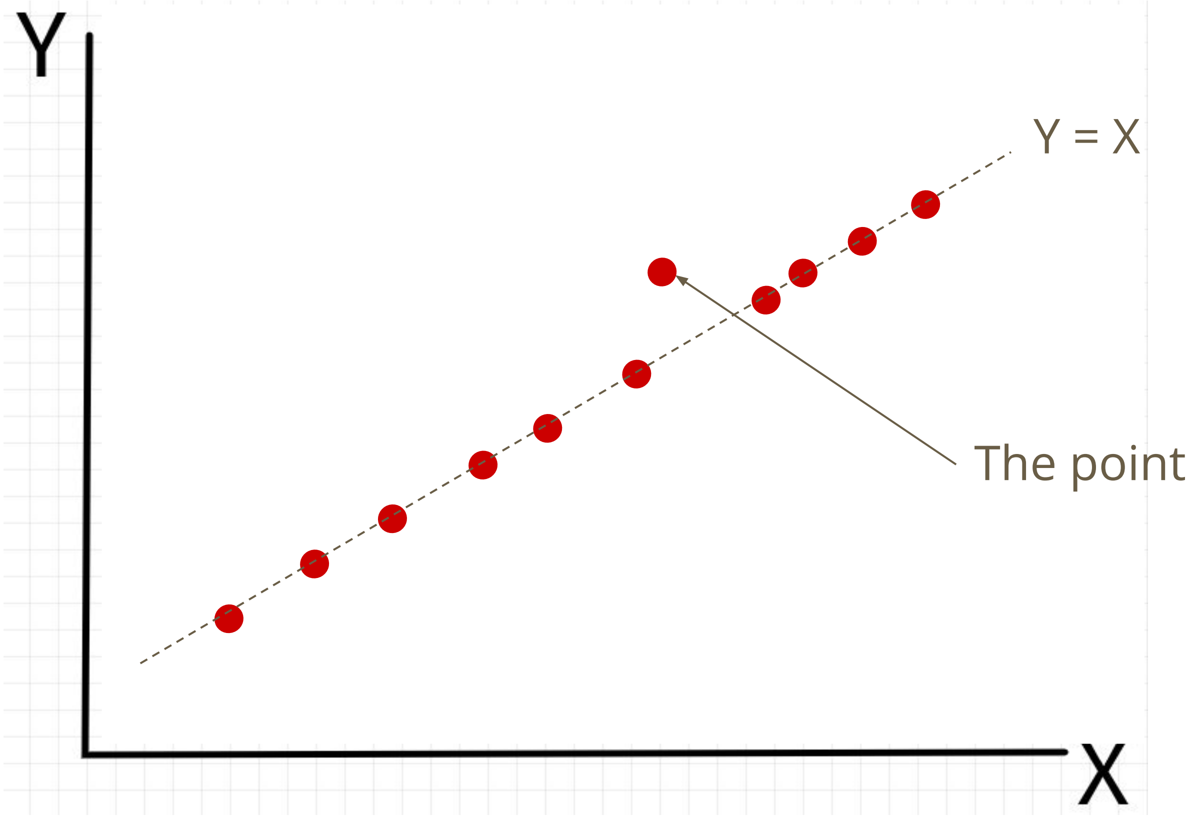



X = [-1 0 1]
Y = [-1 2 -1]

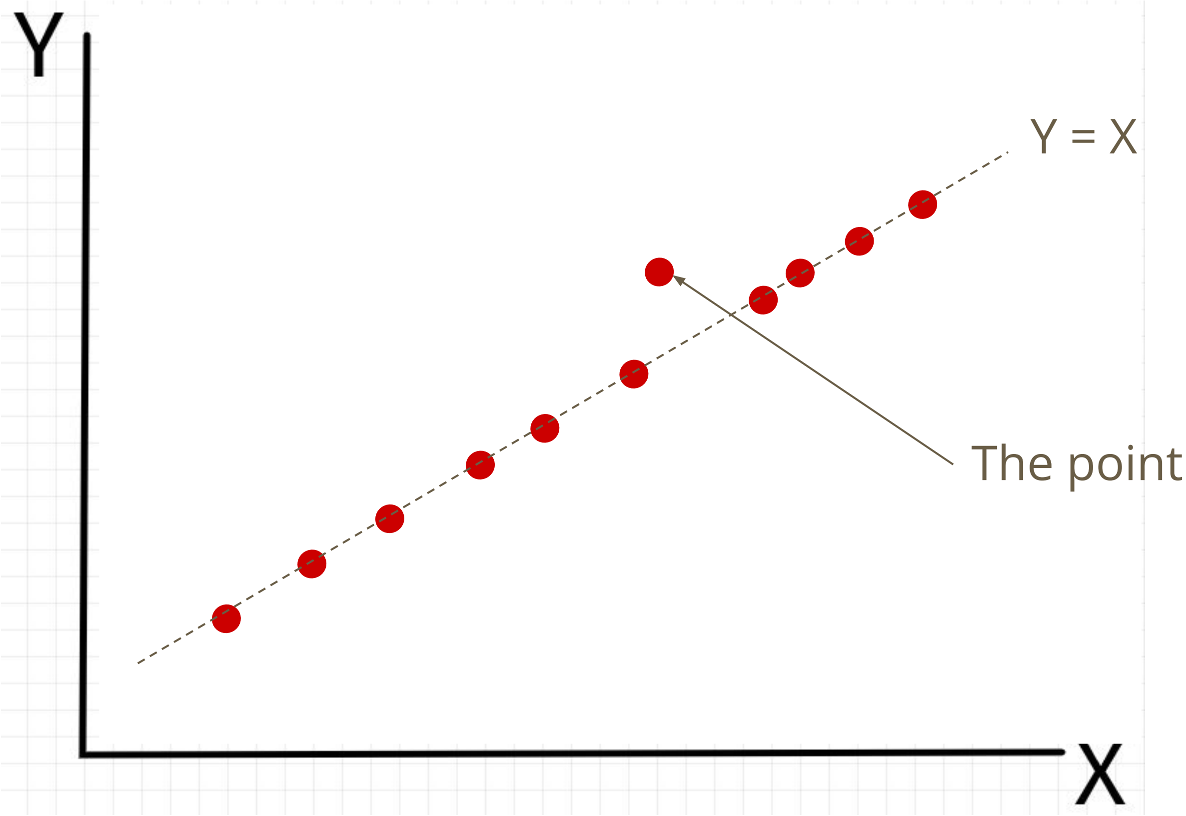
$\mathbf{x} \cdot \mathbf{y} = 1 + 0 + (-1) = 0$



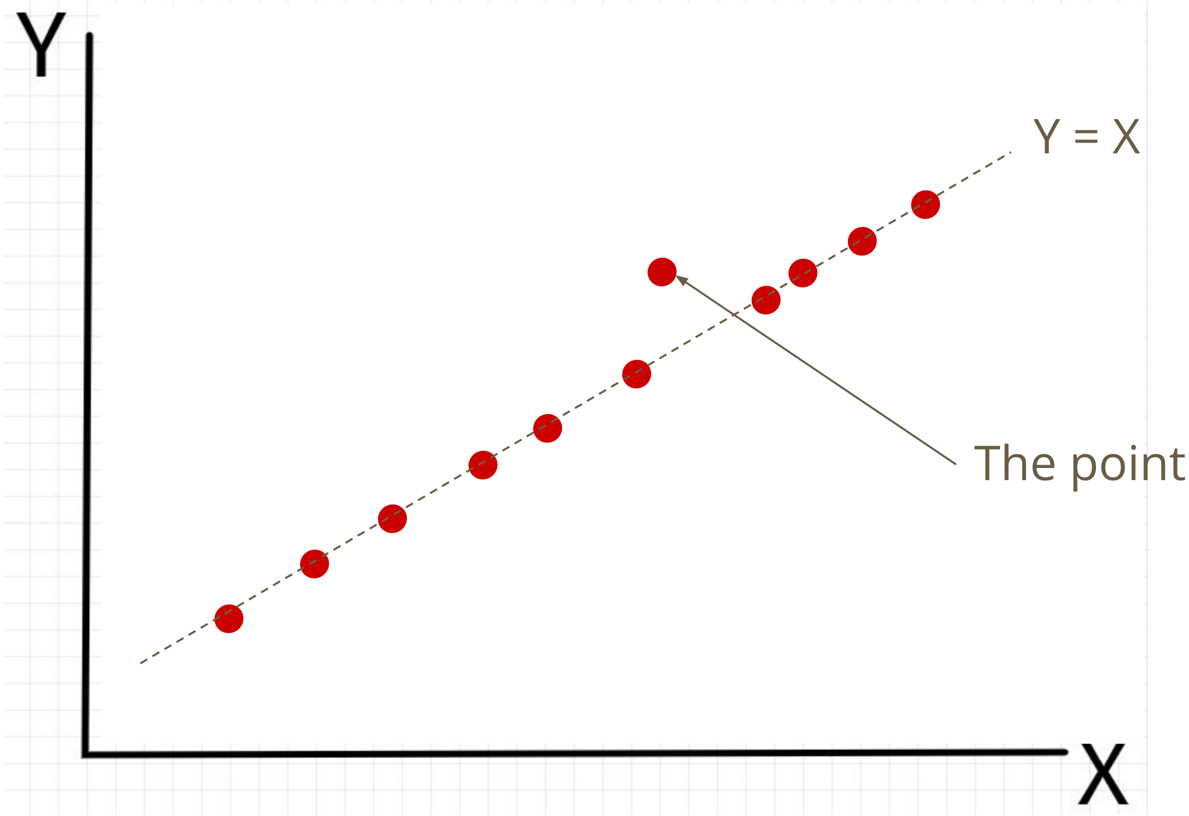
# Can I move this point so far that the correlation is 0?



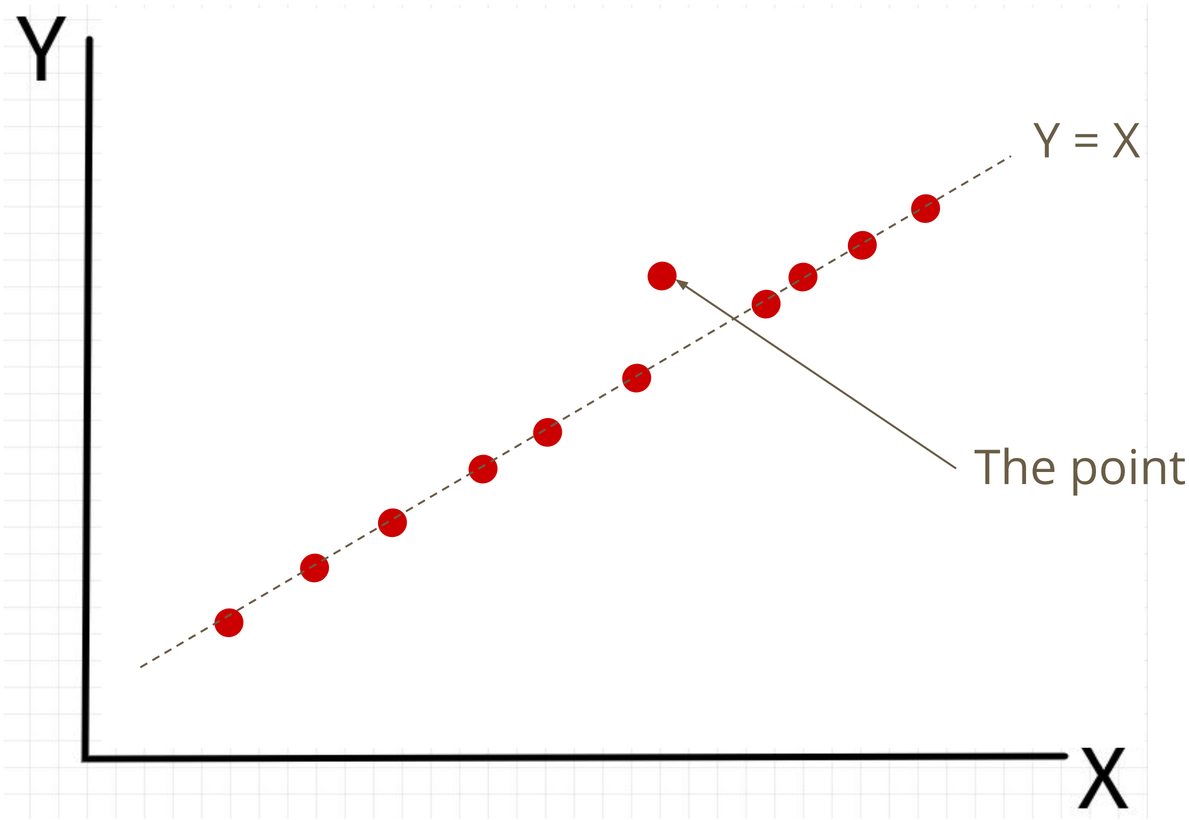
# Can I move this point so far that the correlation is reversed?



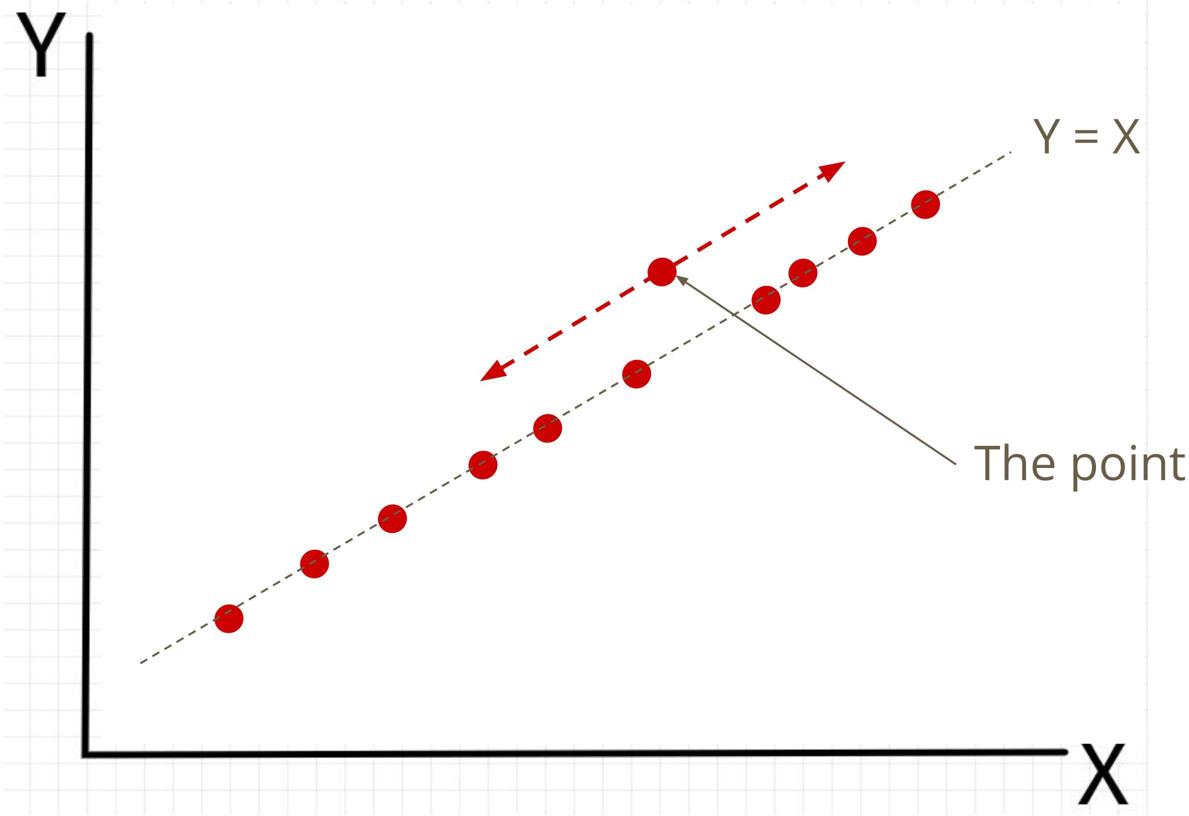
# Can I move this point so far that the correlation is -1?



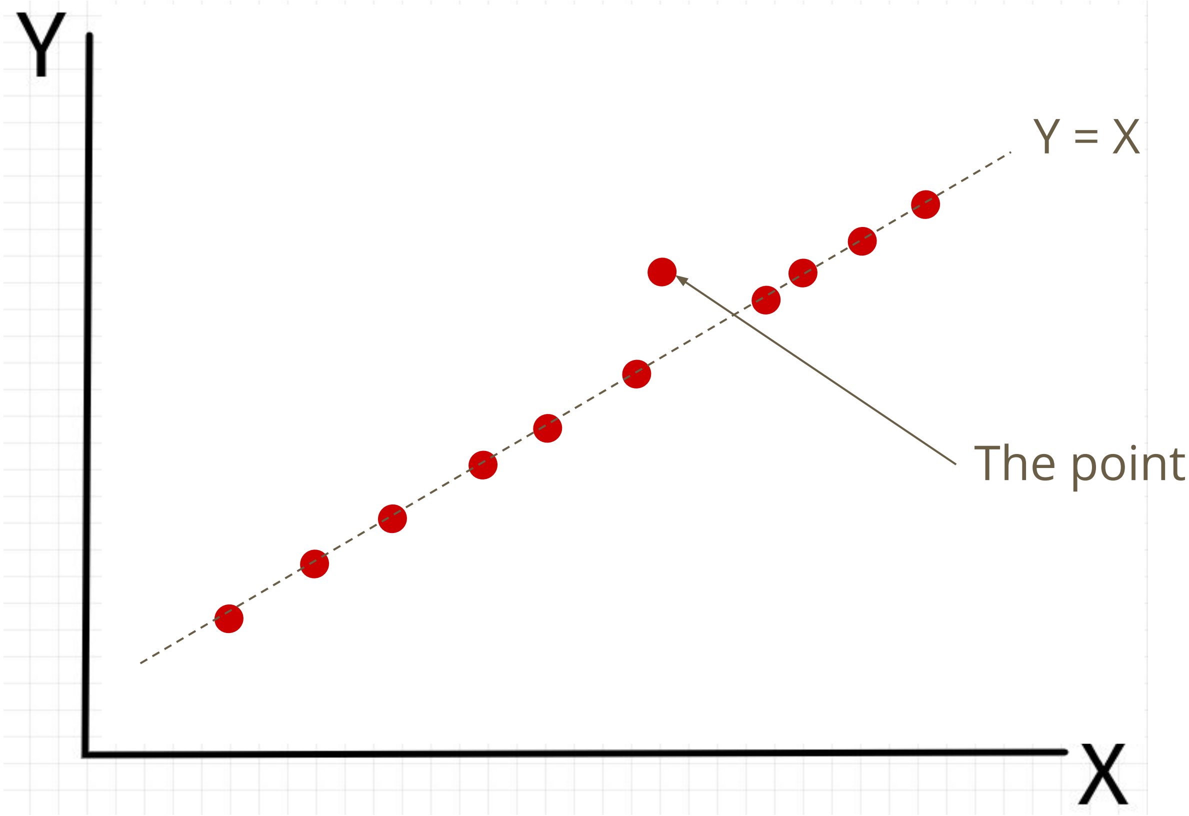
# Can I move this point without changing the correlation?



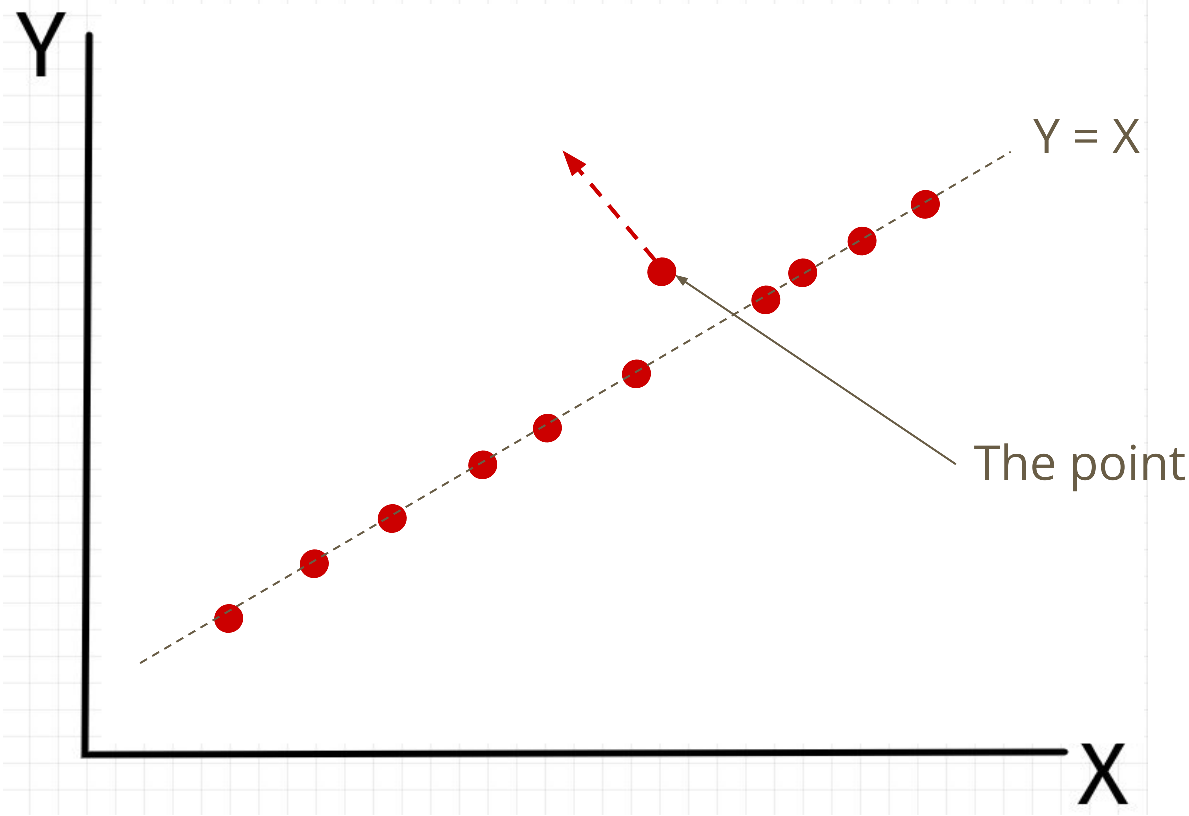
# Can I move this point without changing the correlation?



# Which direction would change the correlation the most?



# Which direction would change the correlation the most?



# Data

**How similar are data point i and feature n ?**

$$\textbf{n}\ \text{data points}\left\{\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix}}_{\textbf{m}\ \text{features}}\right.$$

# Feature Space

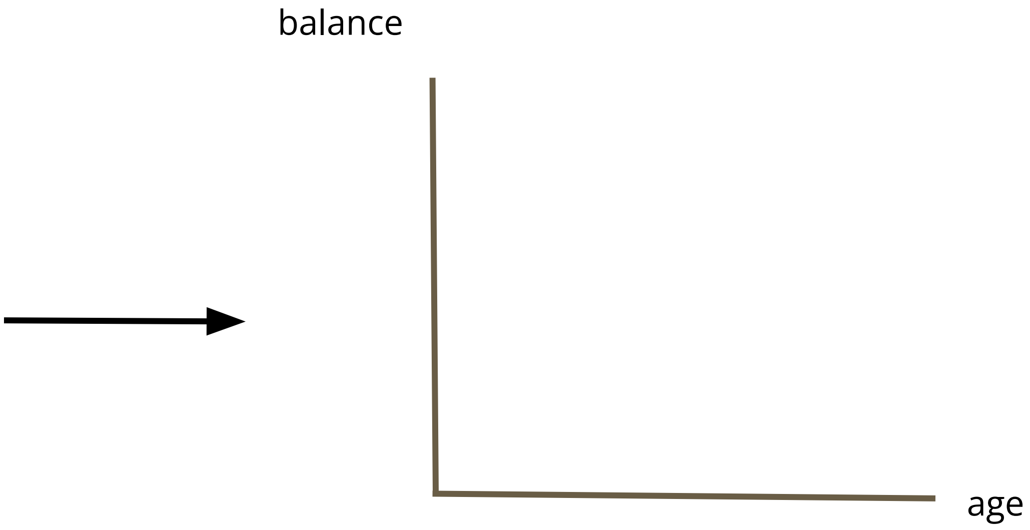
From our data we can generate a **feature space** of all possible values for the set of features in our data.

| name | age | balance |
|---|---|---|
| Jane | 25 | 150 |
| John | 30 | 100 |

# Feature Space

From our data we can generate a **feature space** of all possible values for the set of features in our data.

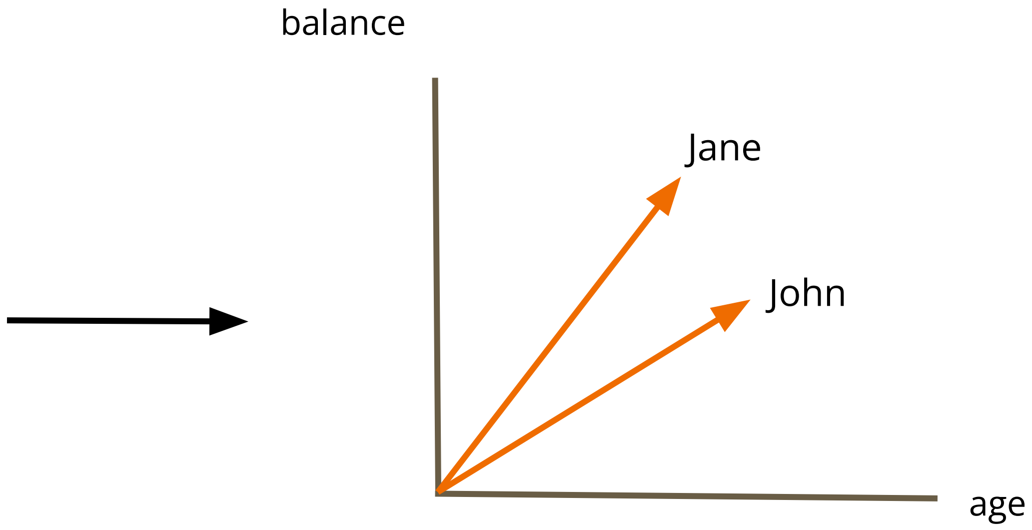
| name | age | balance |
|---|---|---|
| Jane | 25 | 150 |
| John | 30 | 100 |



# Feature Space

From our data we can generate a **feature space** of all possible values for the set of features in our data.

| name | age | balance |
|---|---|---|
| Jane | 25 | 150 |
| John | 30 | 100 |



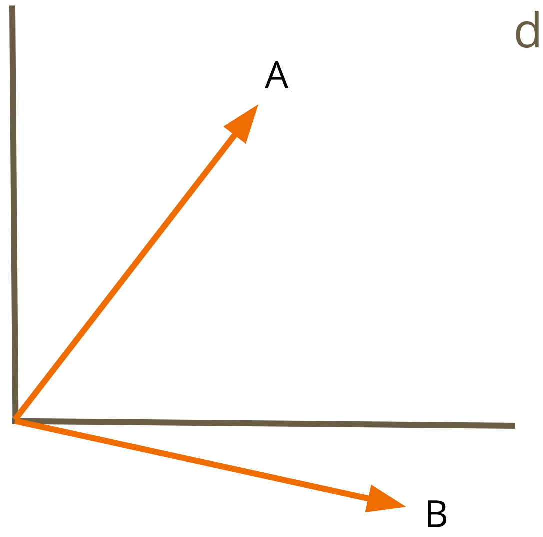

Our feature space is the Euclidean plane

# Dissimilarity

In order to uncover interesting structure from our data, we need a way to **compare** data points.

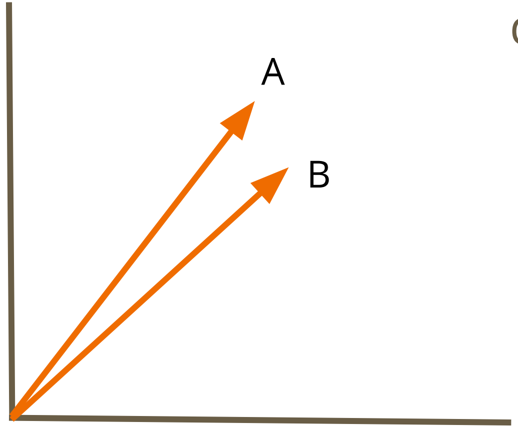
A **dissimilarity function** is a function that takes two objects (data points) and returns a **large value** if these objects are **dissimilar**.

# Dissimilarity




dissim(A, B) is large

# Dissimilarity




dissim(A, B) is small

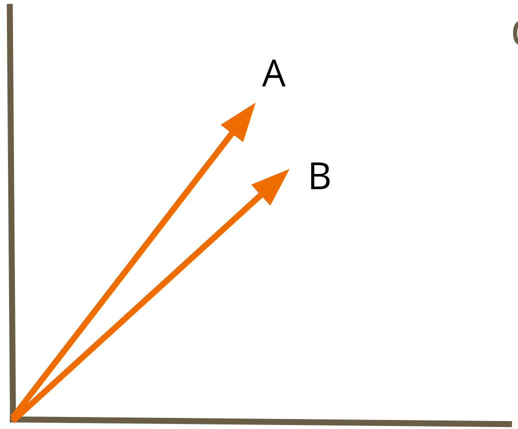
# Distance

A special type of dissimilarity function is a **distance** function

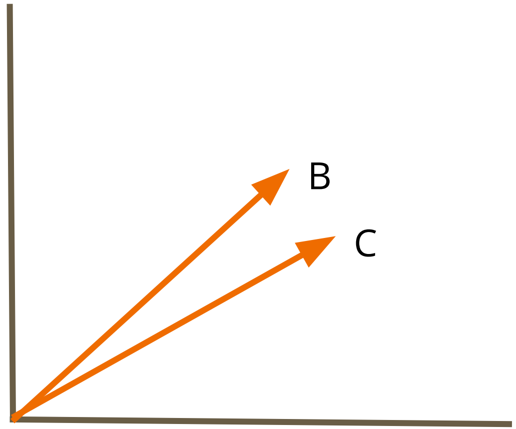
**d** is a distance function if and only if:

- $d(i, j) = 0$ if and only if $i = j$
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$

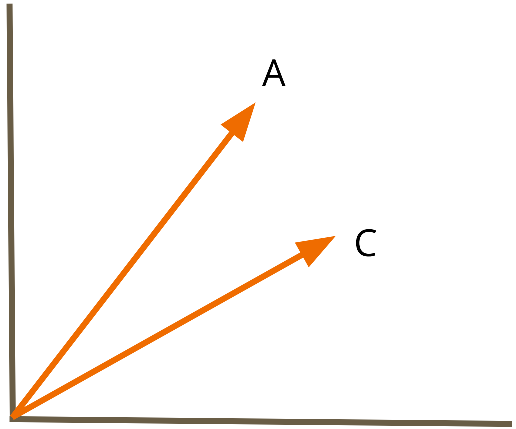
We don't **need** a distance function to compare data points, but why would we prefer using a distance function?



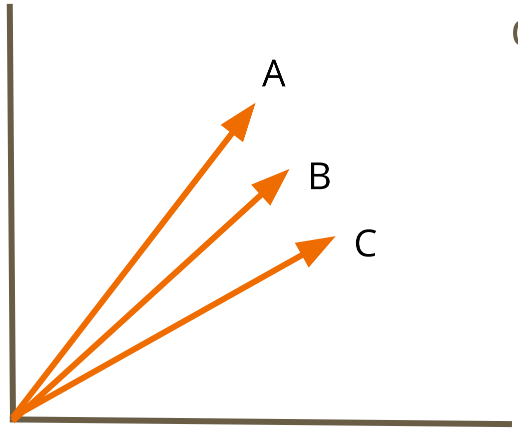

dissim(A, B) is small




dissim(B, C) is small




**dissim(A, C) not necessarily small**




d(A, B) is small

d(B, C) is small

**Triangle inequality guarantees d(A, C) small**

# Minkowski Distance

For **x**, **y** points in **d**-dimensional real space

I.e. $\mathbf{x} = [x_1, \dots, x_d]$ and $\mathbf{y} = [y_1, \dots, y_d]$

**p ≥ 1**

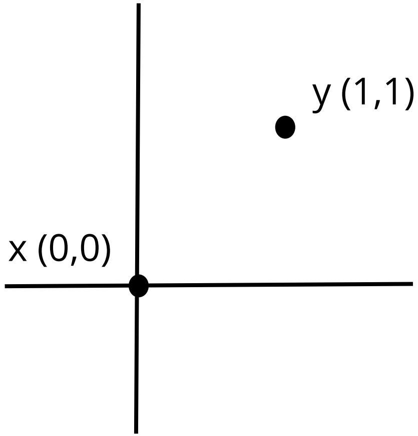
$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

When **p** = 2 -> Euclidean Distance

When **p** = 1 -> Manhattan Distance

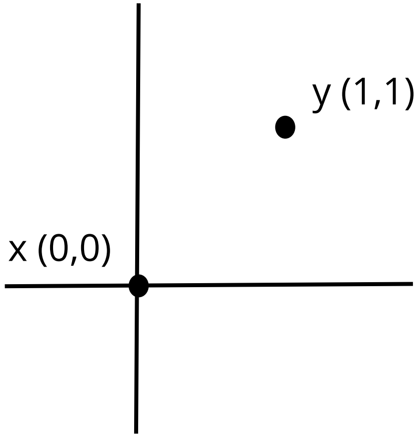
# Example

**d** = 2



# Example

**d** = 2

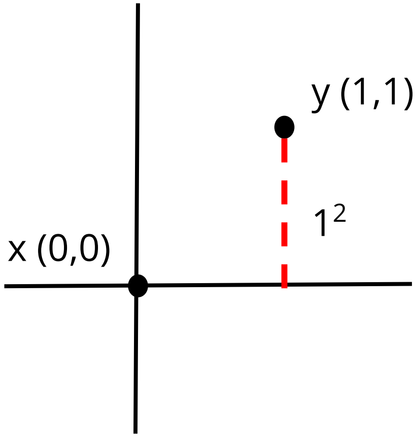



**p** = 2

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2

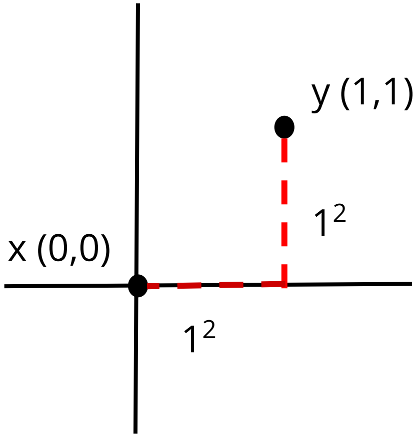



**p** = 2

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2

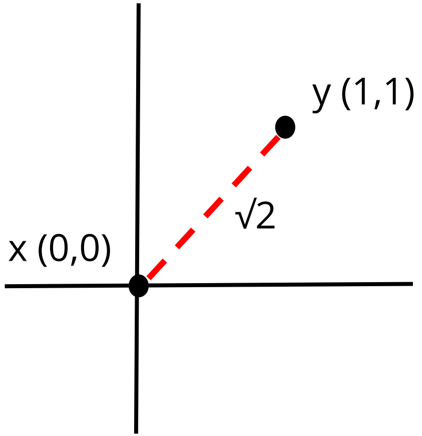



**p** = 2

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2

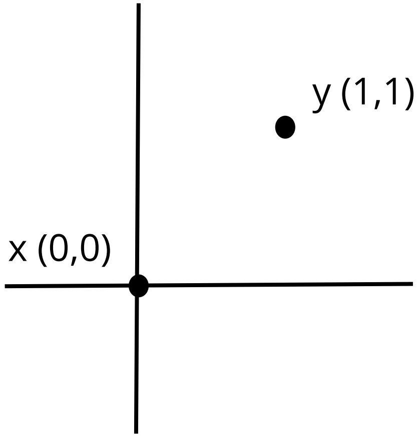



**p** = 2

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2

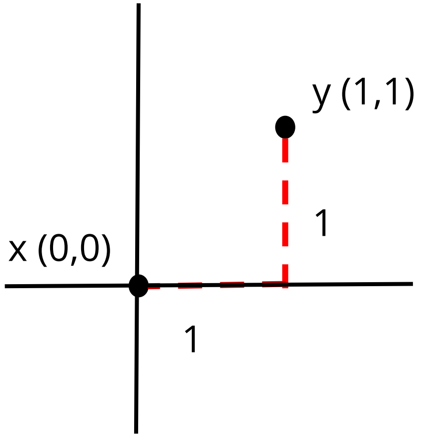



**p** = 1

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2

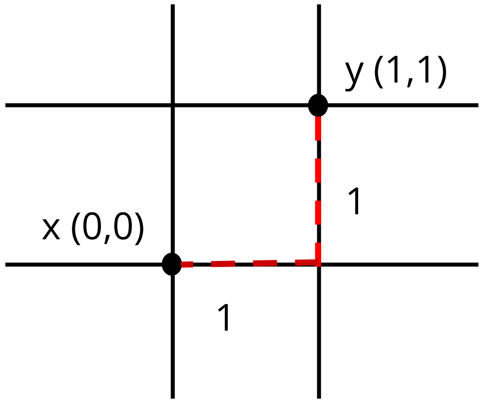



**p** = 1

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2



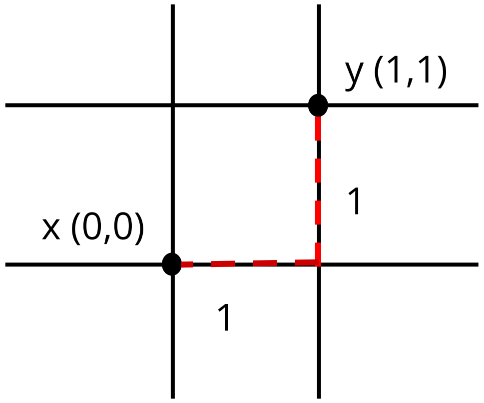

**p** = 1

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 2

But wait - there are infinitely many paths...

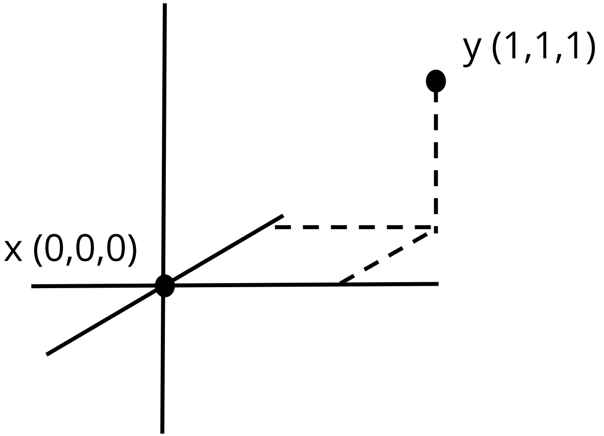



**p** = 1

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 3

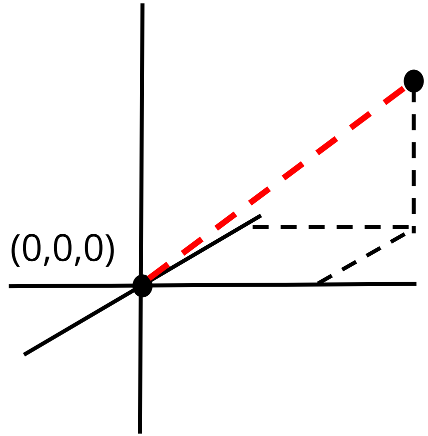



**p** = 2

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 3

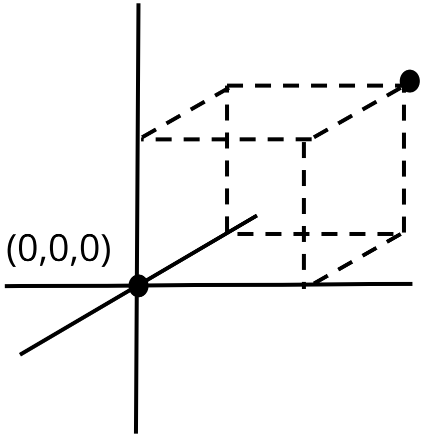



**p** = 2

$$L_p(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Example

**d** = 3

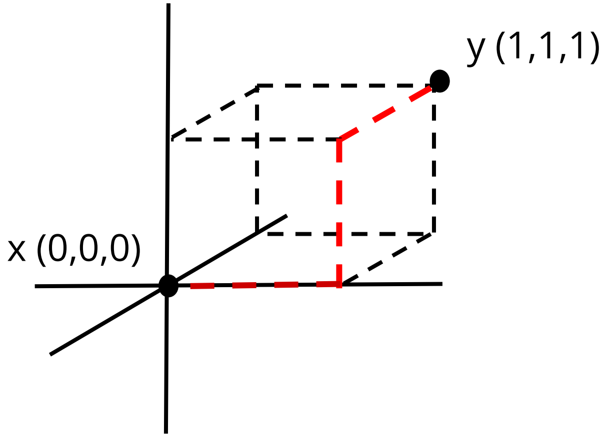



**p** = 1

$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

# Example

**d** = 3




**p** = 1

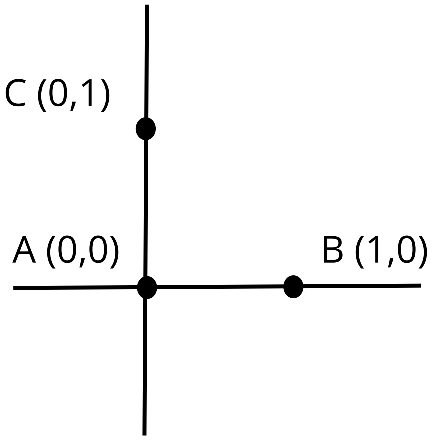
$$L_p(x,y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

# Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$ ?

# Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$ ?

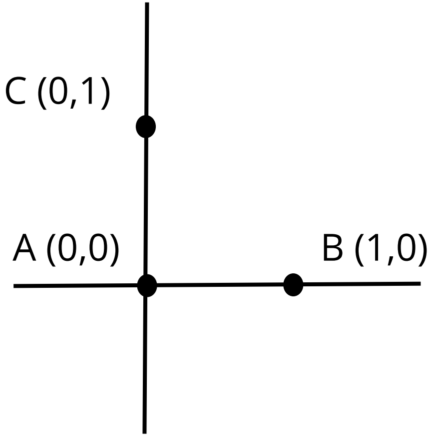


# Minkowski Distance

Is $L_p$ a distance function when $0 < p < 1$ ?

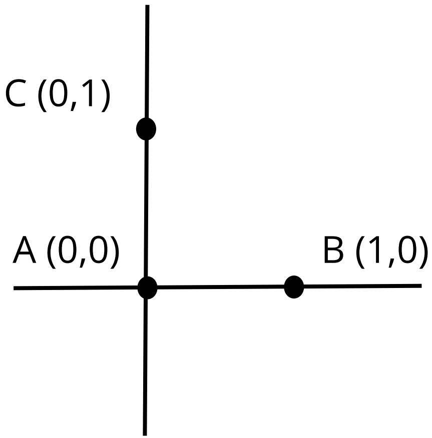
**D(B,A) = D(A, C) = 1**

**D(B, C) = $2^{1/p}$**



# Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$ ?



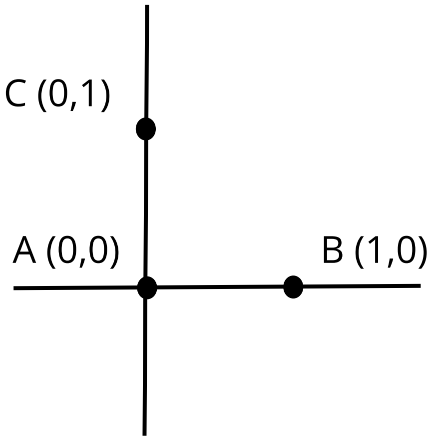

$\mathbf{D(B,A) + D(A, C) = 2}$

$\mathbf{D(B, C) = 2^{1/p}}$

But… if $\mathbf{p < 1}$ then $\mathbf{1/p > 1}$

# Minkowski Distance

Is $\mathbf{L_p}$ a distance function when $\mathbf{0 < p < 1}$ ?




**D(B,A) + D(A, C) = 2**

**D(B, C) =** $\mathbf{2^{1/p}}$

So **D(B, C) > D(B, A) + D(A, C)** which violates the triangle inequality

# L1 vs L2

For **x** in **d**-dimensional real space (i.e. $\mathbf{x} = [\mathbf{x}_1 , \ldots , \mathbf{x}_d]$), let's say you have a penalty **λ** you can distribute to each coordinate to increase the distance of **x** from the origin (or the magnitude of the vector).

# L1 vs L2

For **x** in **d**-dimensional real space (i.e. $\mathbf{x} = [x_1, \ldots, x_d]$), let's say you have a penalty **λ** you can distribute to each coordinate to increase the distance of **x** from the origin (or the magnitude of the vector).

Example: **x = [0, 0, 0, 0, 0]** and **λ = 10**

[10, 0, 0, 0, 0]  -> L2 distance is 10
[2, 2, 2, 2, 2] -> L2 distance is $\sqrt{20} \sim 4.5$

# L1 vs L2

For **x** in **d**-dimensional real space (i.e. $\mathbf{x} = [x_1, \ldots, x_d]$), let's say you have a penalty **λ** you can distribute to each coordinate to increase the distance of **x** from the origin (or the magnitude of the vector).

Example: **x = [0, 0, 0, 0, 0]** and **λ = 10**

[10, 0, 0, 0, 0] -> L1 distance is 10
[2, 2, 2, 2, 2] -> L1 distance is 10

# L1 vs L2 demo

# Jaccard Similarity

How similar are the following documents?

| | $w_1$ | $w_2$ | ... | $w_d$ |
|---|---|---|---|---|
| x | 1 | 0 | ... | 1 |
| y | 1 | 1 | ... | 0 |

# Jaccard Similarity

One way is to use the Manhattan distance which will return the size of the set difference

| | $w_1$ | $w_2$ | ... | $w_d$ |
|---|---|---|---|---|
| x | 1 | 0 | ... | 1 |
| y | 1 | 1 | ... | 0 |

$$L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

# Jaccard Similarity

One way is to use the Manhattan distance which will return the size of the set difference

| | $w_1$ | $w_2$ | ... | $w_d$ |
|---|---|---|---|---|
| x | 1 | 0 | ... | 1 |
| y | 1 | 1 | ... | 0 |

$$L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

Will only be 1 when $x_i \neq y_i$

# Jaccard Similarity

But how can we distinguish between these two cases?

| | $w_1$ | $w_2$ | … | $w_{d-1}$ | $w_d$ |
|---|---|---|---|---|---|
| x | 1 | 1 | 1 | 0 | 1 |
| y | 1 | 1 | 1 | 1 | 0 |

Only differ on the last two words

| | $w_1$ | $w_2$ |
|---|---|---|
| x | 0 | 1 |
| y | 1 | 0 |

Completely different

# Jaccard Similarity

But how can we distinguish between these two cases?

| | $w_1$ | $w_2$ | ... | $w_{d-1}$ | $w_d$ |
|---|---|---|---|---|---|
| x | 1 | 1 | 1 | 0 | 1 |
| y | 1 | 1 | 1 | 1 | 0 |

Only differ on the last two words

| | $w_1$ | $w_2$ |
|---|---|---|
| x | 0 | 1 |
| y | 1 | 0 |

Completely different

Both have Manhattan distance of 2

# Jaccard Similarity

We need to account for the size of the intersection!

Given two documents x and y:

$$JSim(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

# Jaccard Similarity

We need to account for the size of the intersection!

Given two documents x and y:

$$JSim(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

$$JDist(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

Here, x is the set of words (not the binary vector representation)

# Jaccard Similarity

$$JDist(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

| | $w_1$ | $w_2$ | ... | $w_{d-1}$ | $w_d$ |
|---|---|---|---|---|---|
| x | 1 | 1 | 1 | 0 | 1 |
| y | 1 | 1 | 1 | 1 | 0 |

Only differ on the last two words

| | $w_1$ | $w_2$ |
|---|---|---|
| x | 0 | 1 |
| y | 1 | 0 |

Completely different

What is the jaccard distance in each?

# Jaccard Similarity

$$JDist(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

Here, x is the set of words (not the binary vector representation)

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of:

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\boldsymbol{\theta})}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of:

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of: 0

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(\theta)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of: 0

two opposite vectors have a similarity of:

# Cosine Similarity

A **similarity** function is a function that takes two objects (data points) and returns a **large value** if these objects are **similar**.

$$\mathbf{s(x, y) = cos(}\boldsymbol{\theta}\mathbf{)}$$

where $\boldsymbol{\theta}$ is the angle between **x** and **y**

two proportional vectors have a cosine similarity of: 1

two orthogonal vectors have a similarity of: 0

two opposite vectors have a similarity of: - 1

# Cosine Similarity

To get a corresponding **dissimilarity** function, we can usually try

**d(x, y) = 1 / s(x, y)**

or

**d(x, y) = k - s(x, y)** for some **k**

Here, we can use

**d(x, y) = 1 - s(x, y)**

# Cosine Similarity

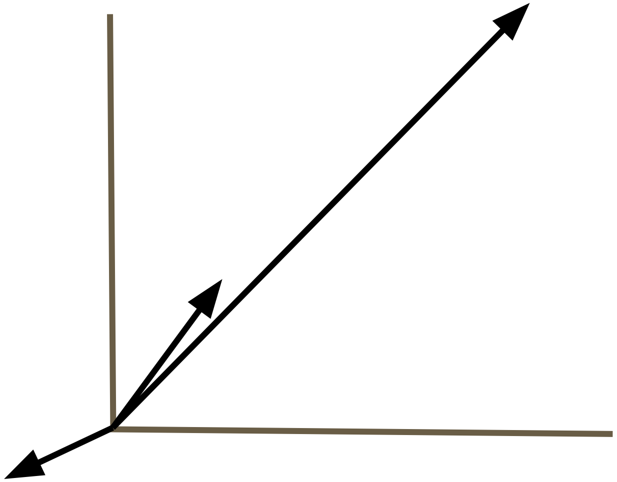
When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**

# Cosine Similarity

When should you use **cosine (dis)similarity** over **euclidean distance**?

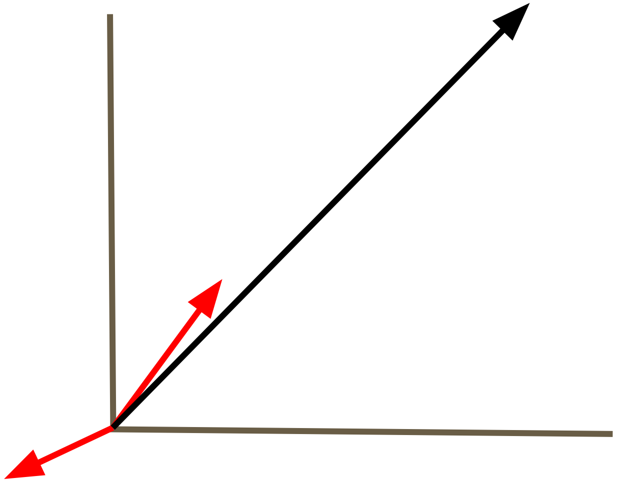
When **direction** matters more than **magnitude**



# Cosine Similarity

When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**

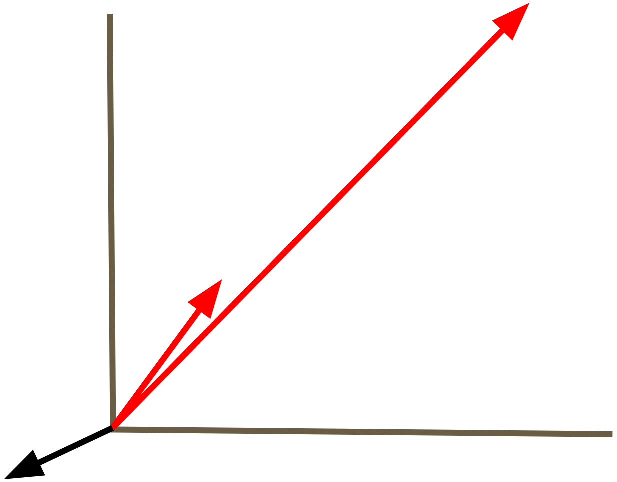


Close under
Euclidean distance

# Cosine Similarity

When should you use **cosine (dis)similarity** over **euclidean distance**?

When **direction** matters more than **magnitude**



Close under Cosine Similarity

# A quick Note on Norms

$$d(A, B) = \|A - B\|$$

Size = Distance from the origin $$d(0, X) = \|X\|$$

- Minkowski Distance <=> Lp Norm
- Not all distances can create a Norm