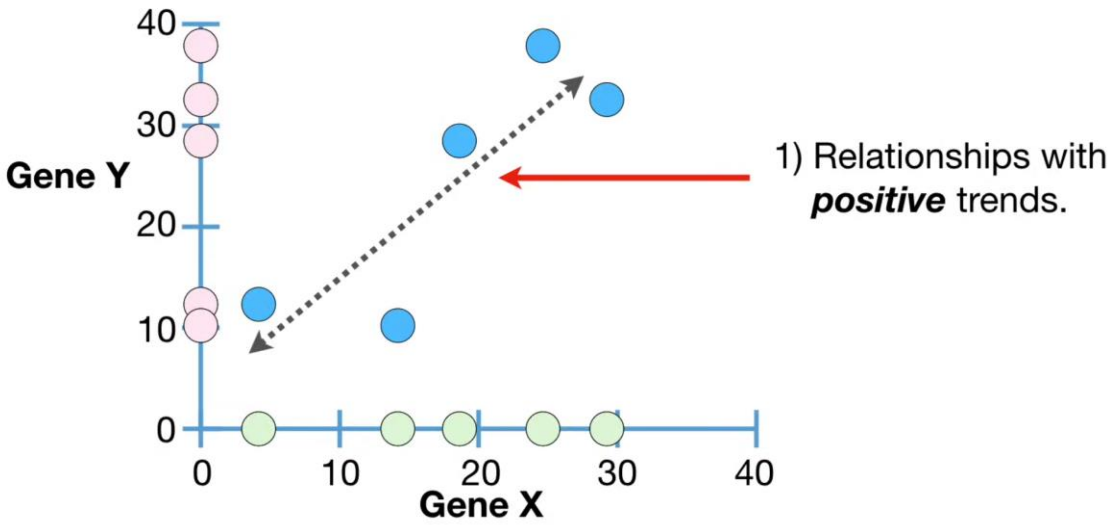
# Covariance

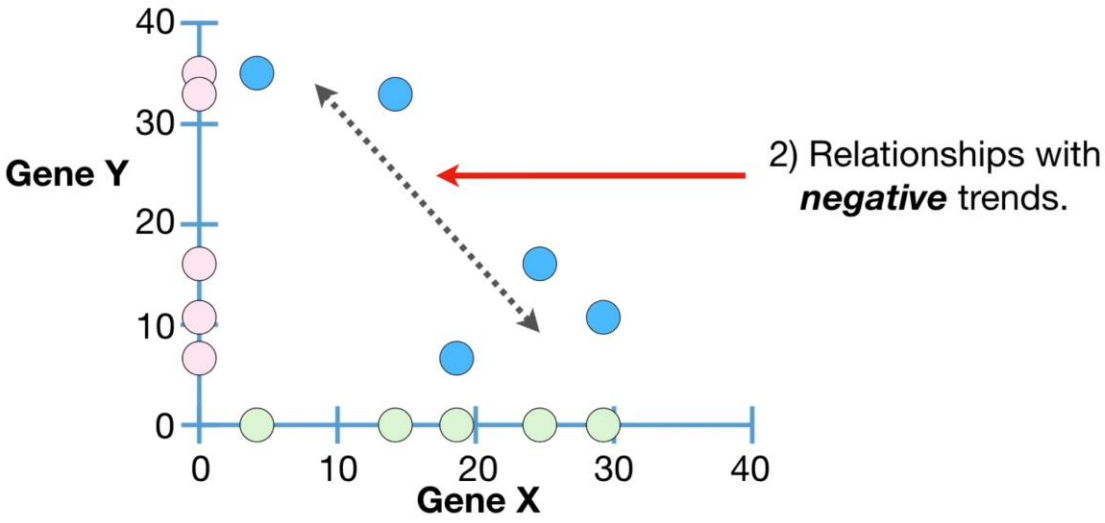
## Purpose of covariance

Covariance is used to measuring the relationship with pairs of data. It can capture the correlation that individual data can not tell us.

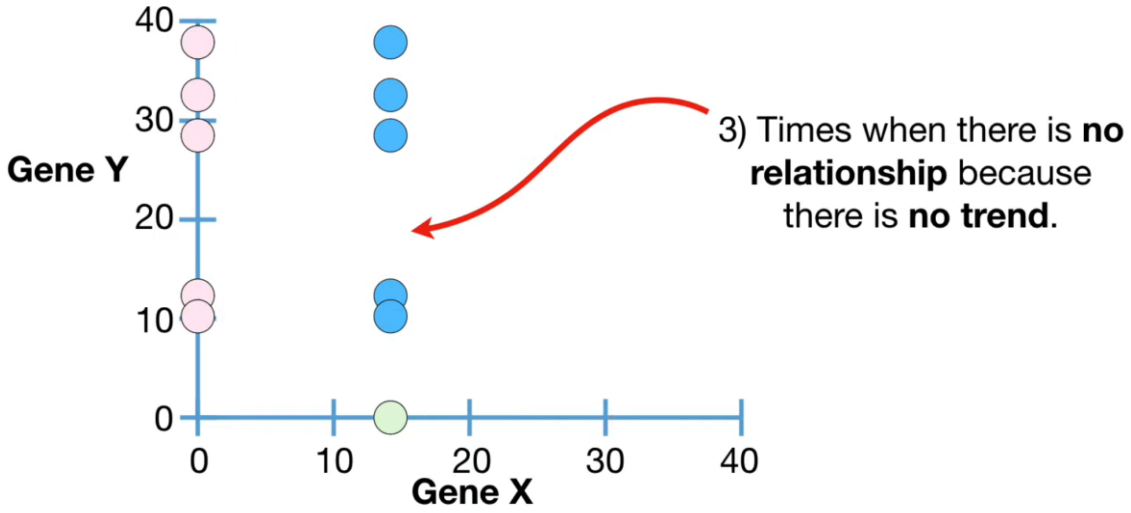
- Relationship with positive trends




- Relationship with negative trends



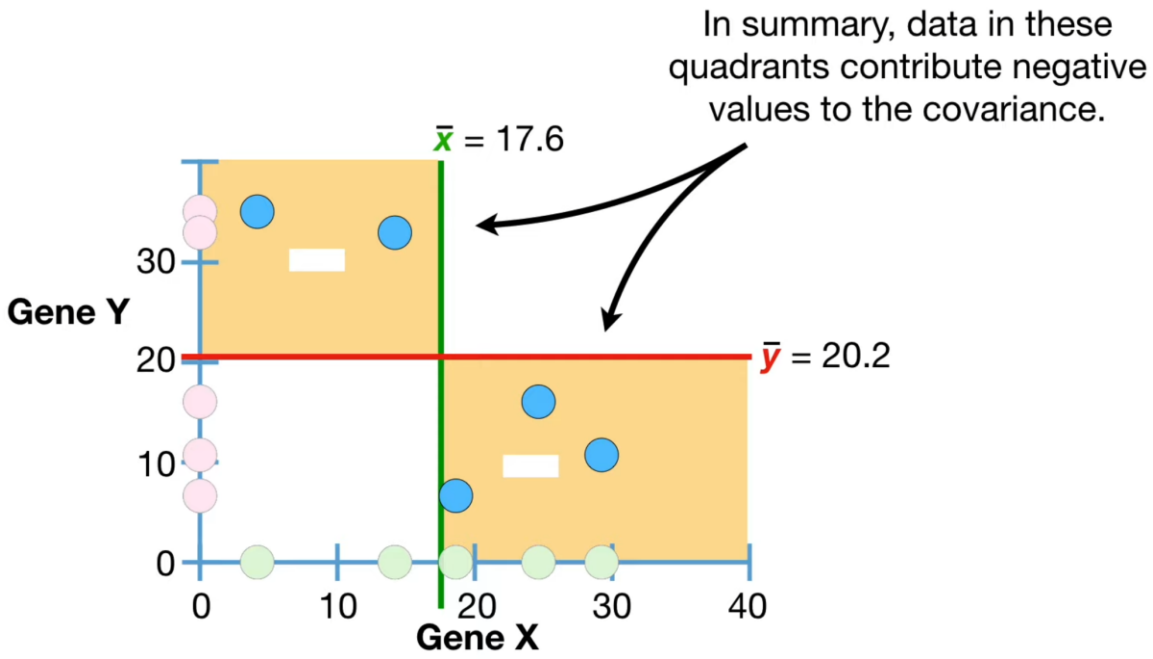

- No relationship




## Formular of covariance

$$Cov = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

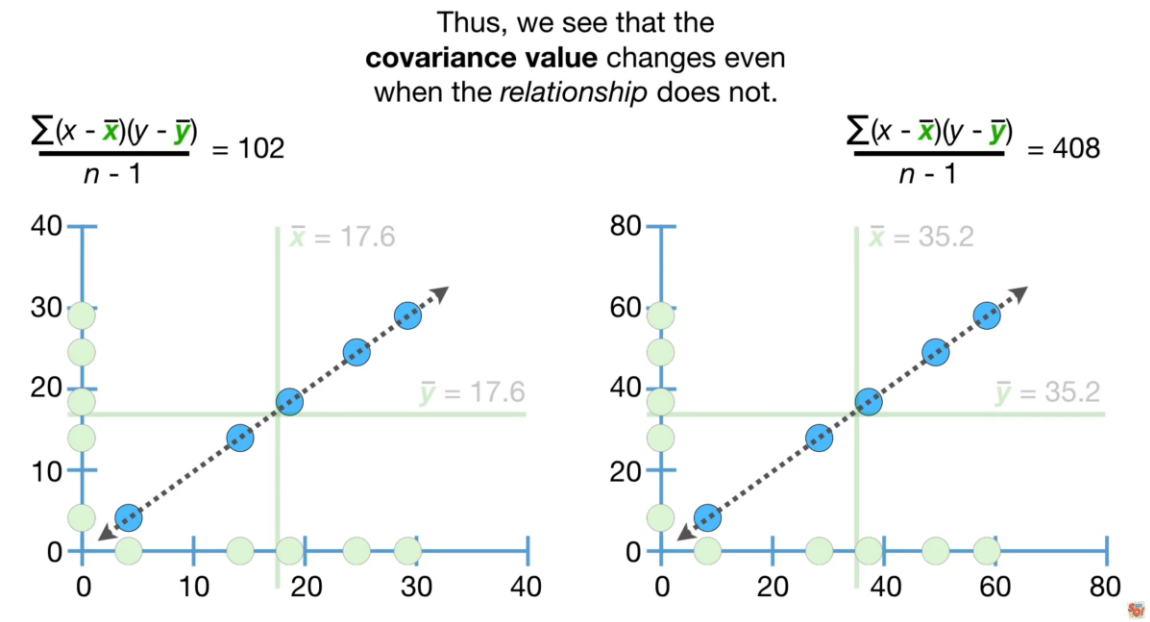
You could imagine, it is like, move the origin to the point $(\bar{x}, \bar{y})$, and calculate the product of points' new coordinates, which indicates what quadrant they are located.




The value of co-variance do not indicate the slop is steeper or not, it only tell us whether the relationship is positive or negative.

The covariance is hard to interpret, so we consider it as the computational steppingstone to more interesting things.

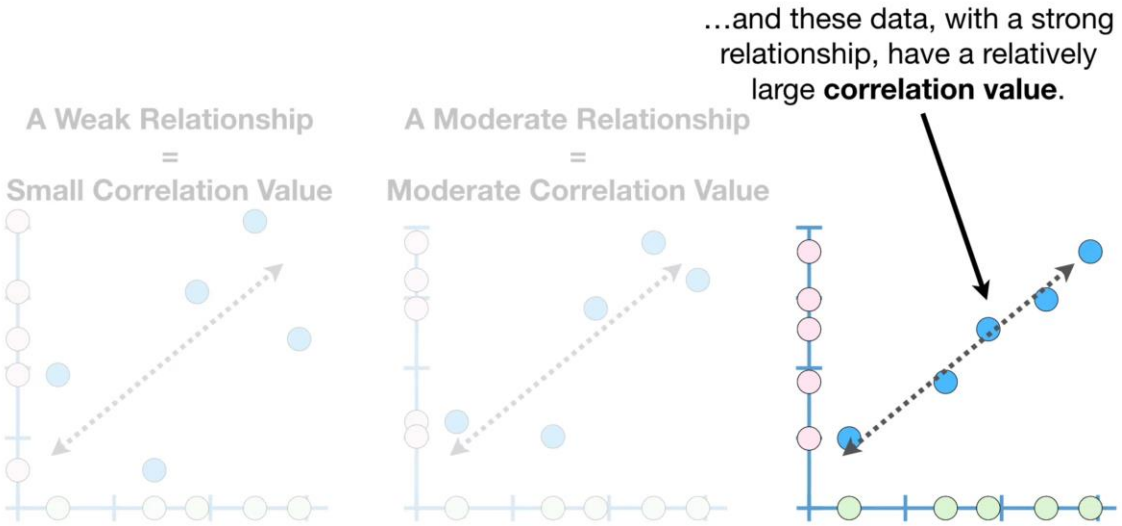
The reason why covariance is hard to interpret, is because it comes from variance. The value of variance depends on the gratitude of original inputs. Image we multiply 2 to each input, the value of variance would enlarge to 4 times, but the slop would not be changed.

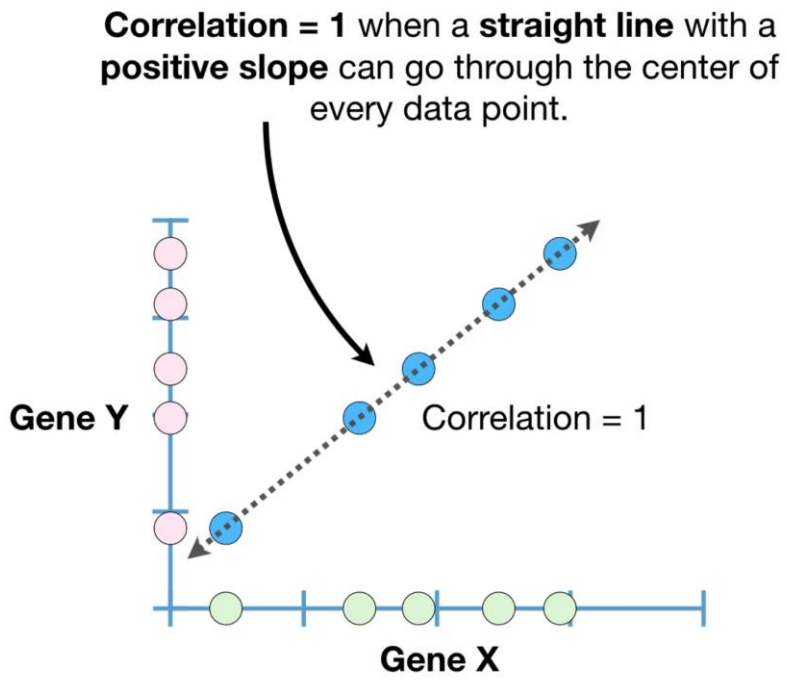



So, we come up the concept of correlation.

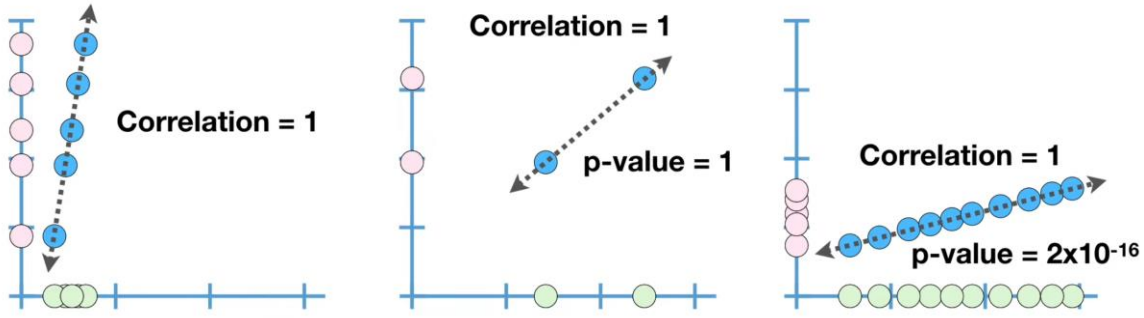
# Correlation

## Purpose of correlation





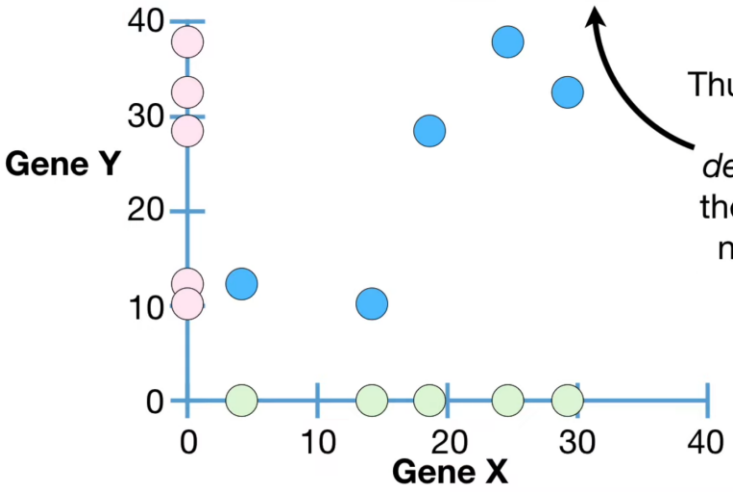

Although the value of correlation does not depend on the gratitude of inputs, but the number of observations would impact the confidence of correlation.




## Formular of correlation

$$Correlation = \frac{Covariance(X,Y)}{\sqrt{Variance(X)}\sqrt{Variance(Y)}}$$

$$\text{Correlation} = \frac{\text{Covariance(Gene X, Gene Y)}}{\sqrt{\text{Variance(Gene X)}}\ \sqrt{\text{Variance(Gene Y)}}}$$




Thus, when we calculate **correlation**, the *denominator* squeezes the **covariance** to be a number from **-1** to **1**.