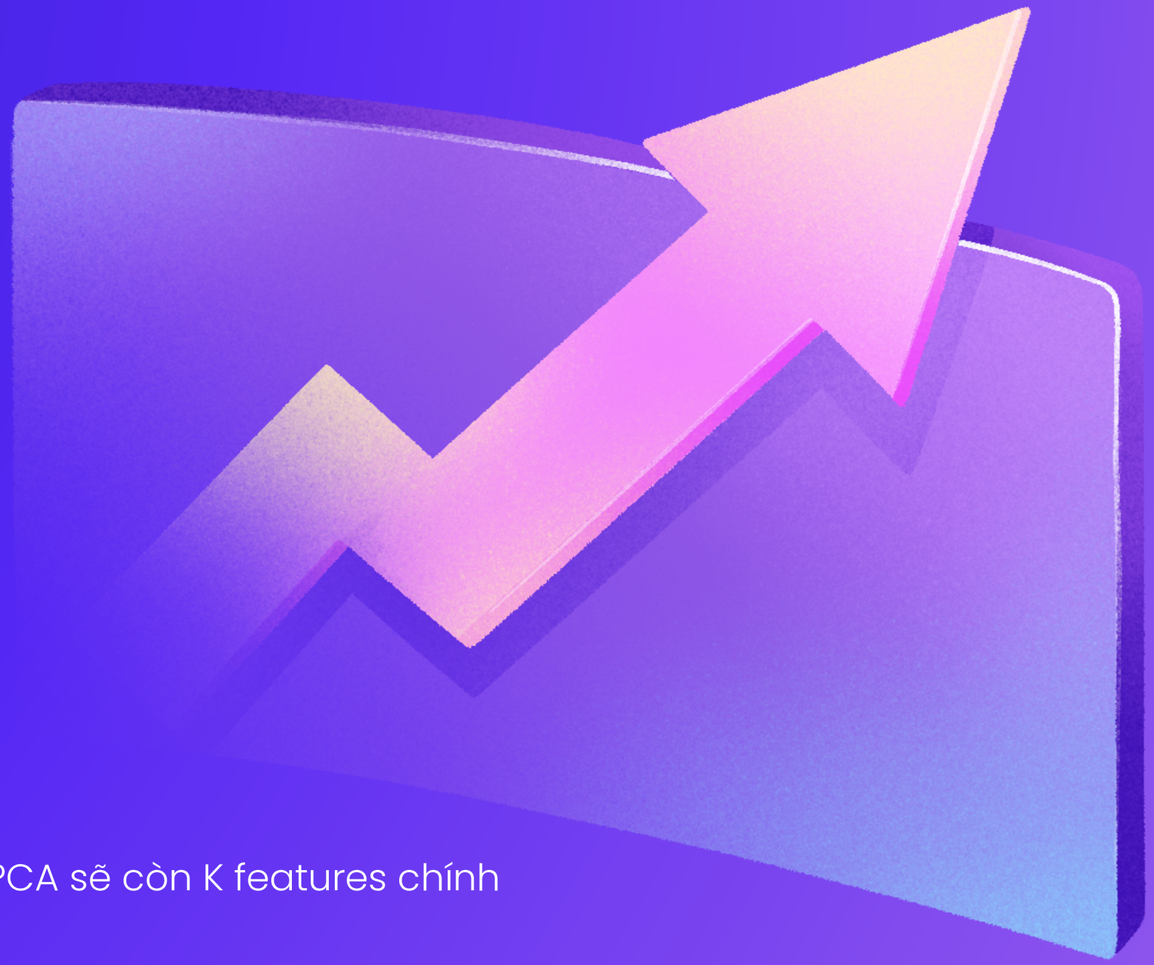
ARTIFICIAL INTELLIGENCE

PCA (PHÂN TÍCH THÀNH PHẦN CHÍNH)

PCA LÀ GÌ?

- Phân tích thành phần chính hoặc PCA là một kỹ dùng để giảm kích thước của tập dữ liệu lớn.

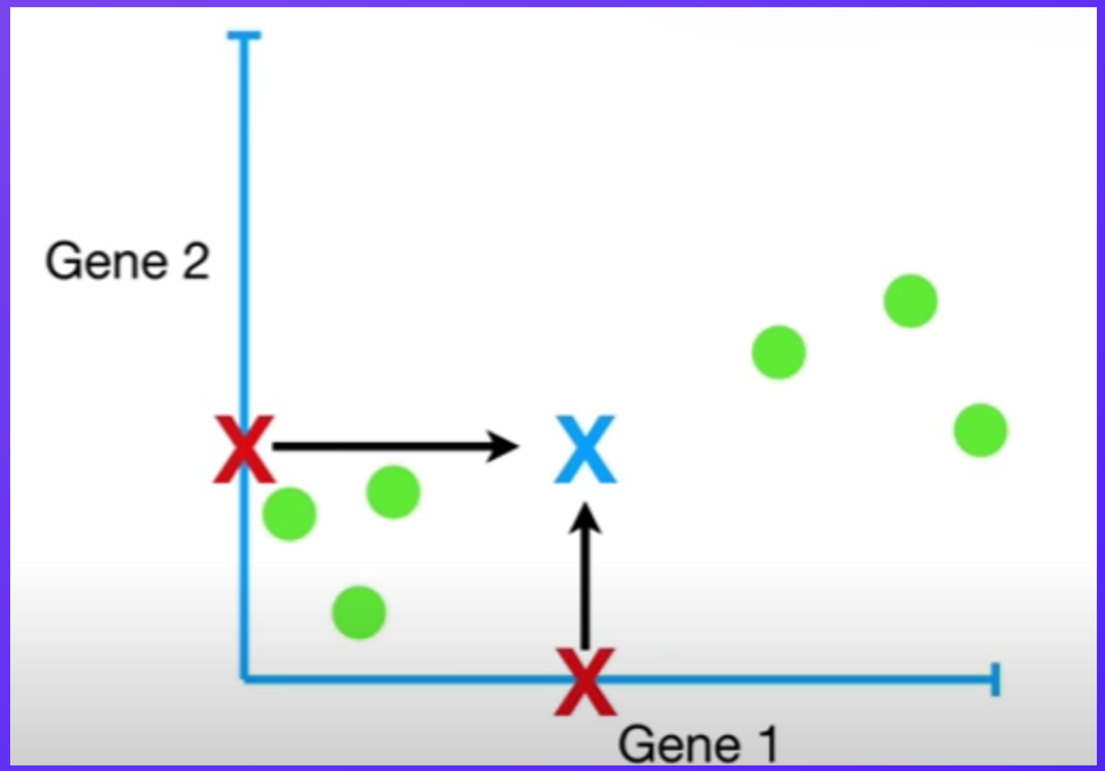
- Việc làm như trên sẽ giúp cho chúng ta: giảm chiều dữ liệu mà vẫn giữ được đặc trưng chính, chỉ mất đi "chút ít" đặc trưng, tiết kiệm thời gian, chi phí tính toán, dễ dàng visualize dữ liệu hơn

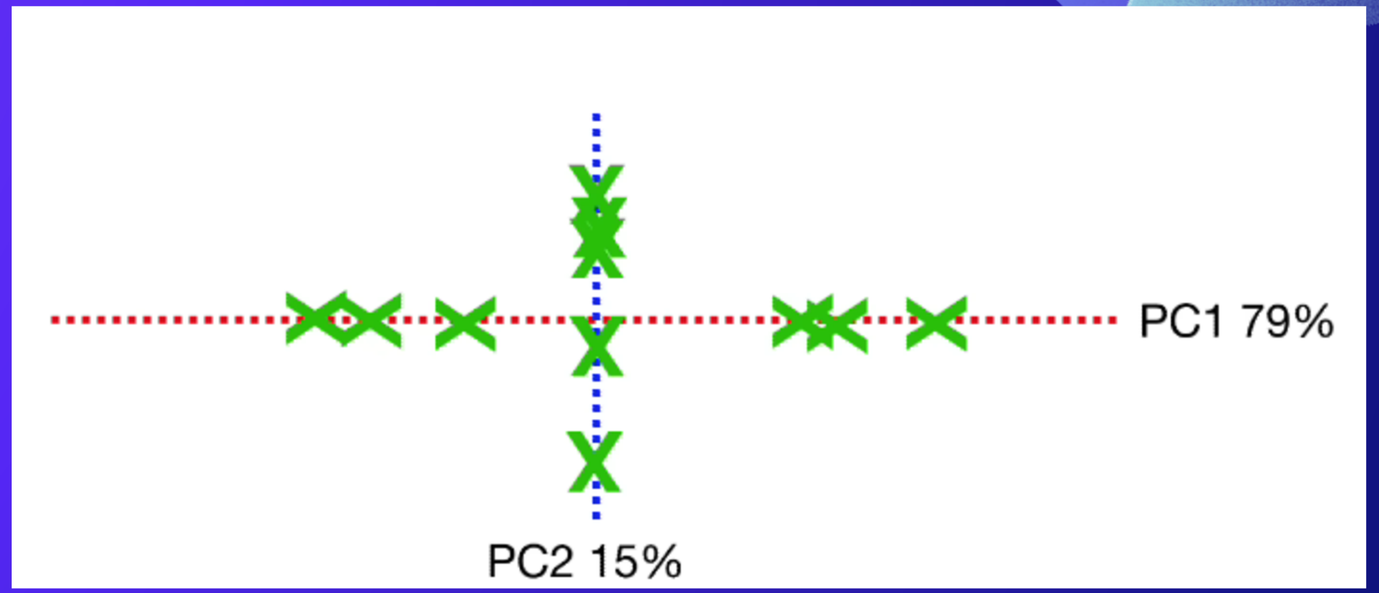


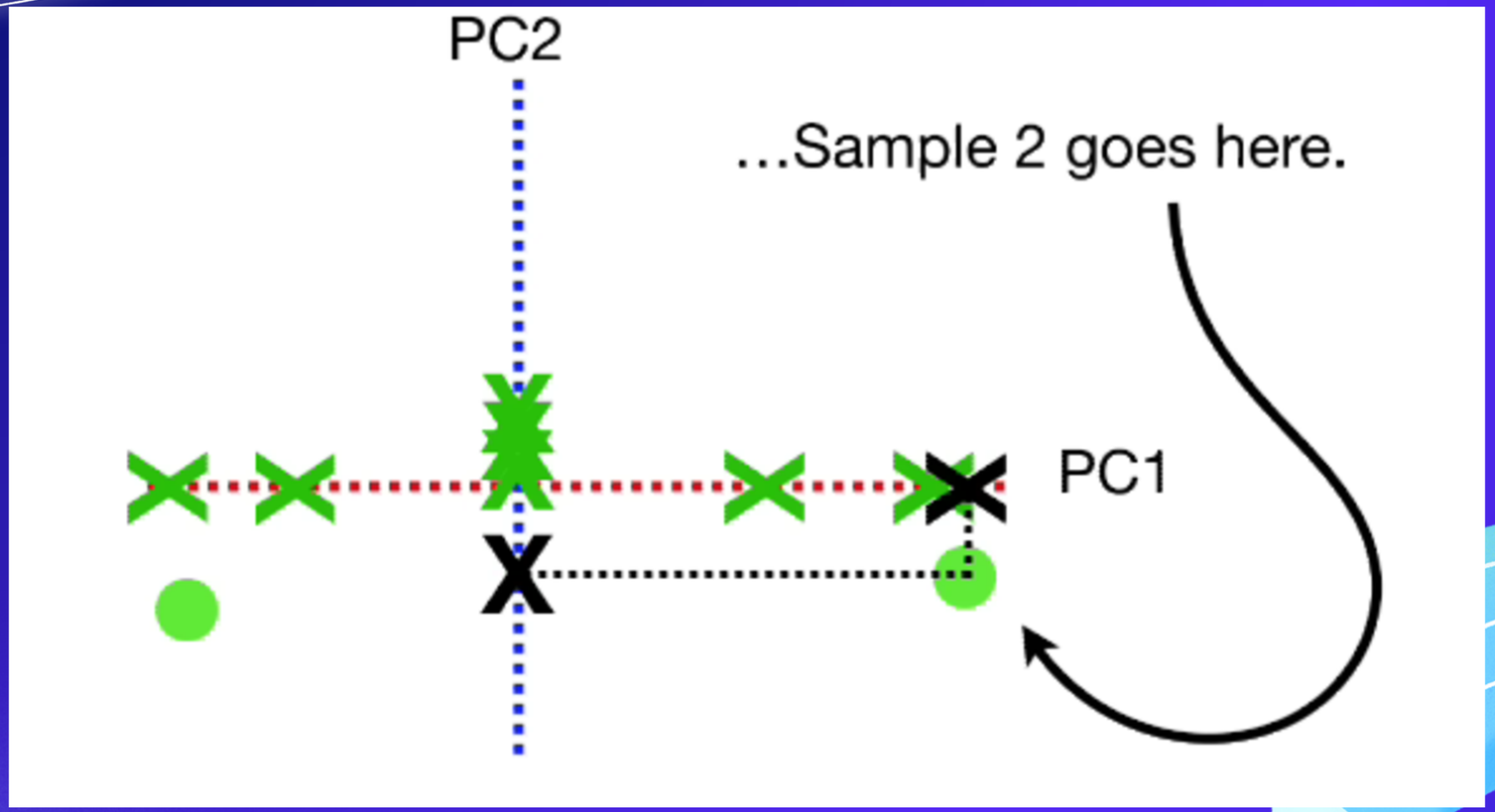
Ví dụ

Dữ liệu của chúng ta có N features thì sau khi áp dụng PCA sẽ còn K features chính mà thôi (K<N)

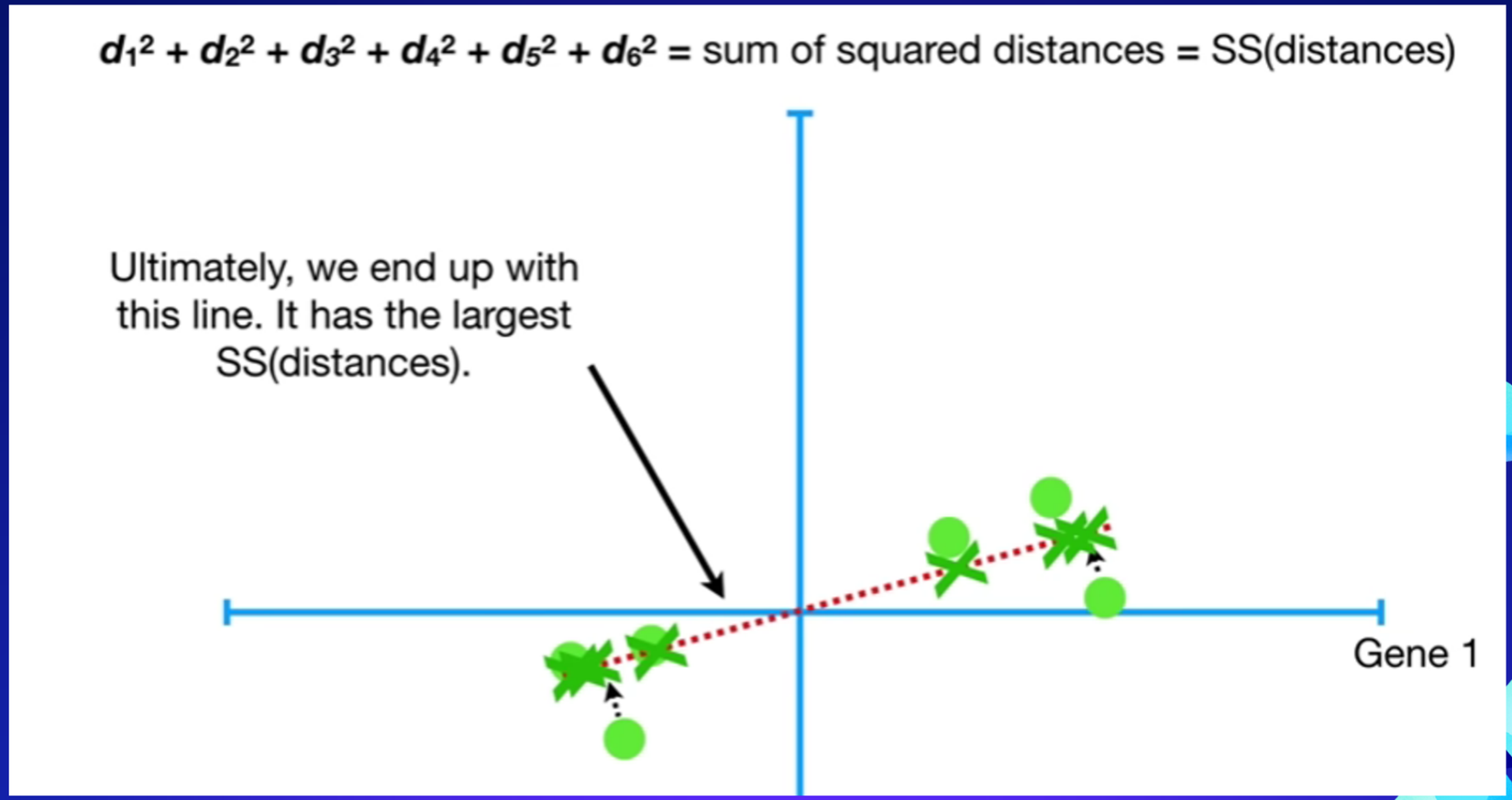
MỤC TIÊU PCA



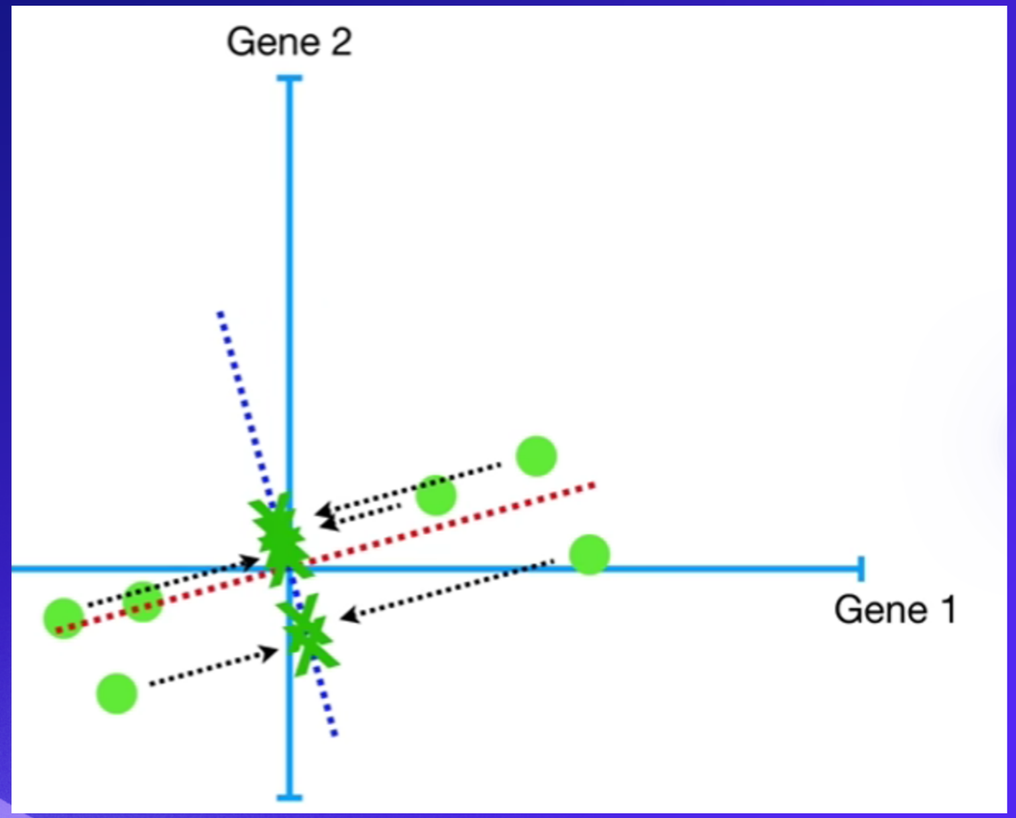



PC2
…Sample 2 goes here.
PC1

CÁCH XÁC ĐỊNH PCA

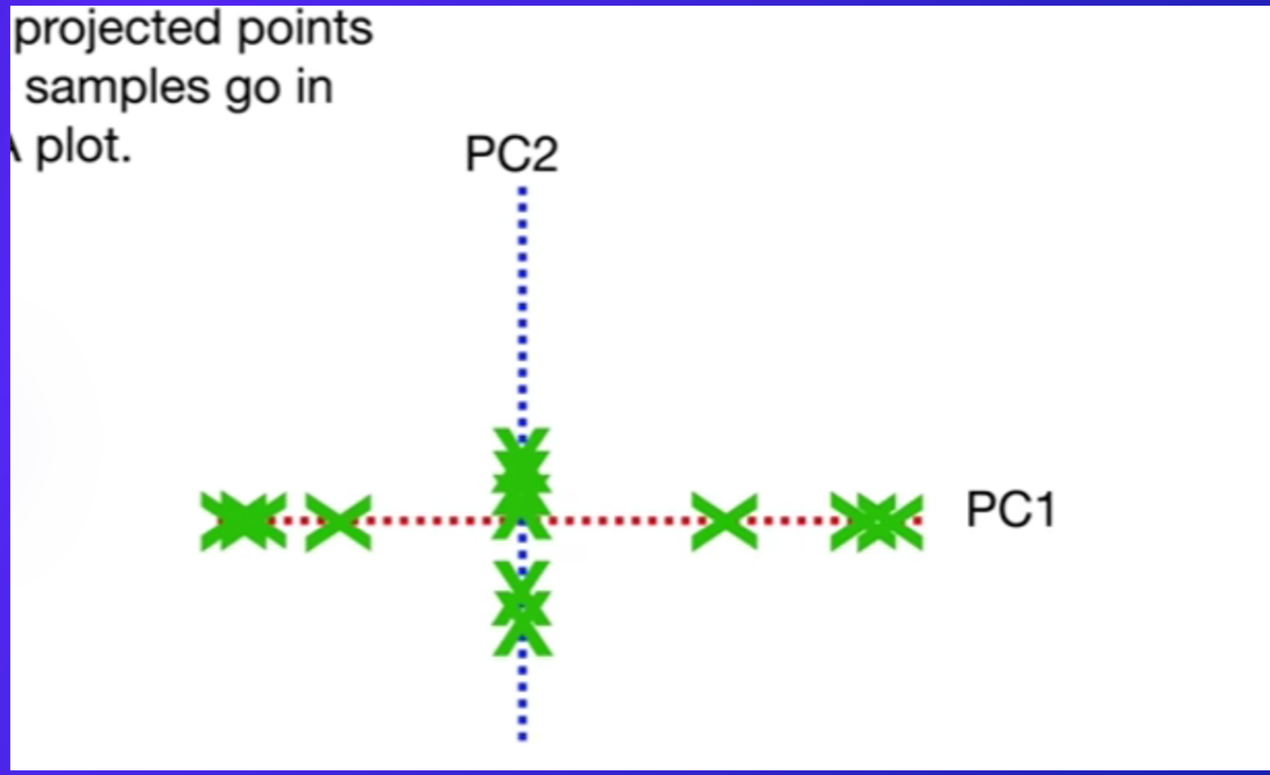


SAO CHO SS(DISTANCE) LỚN NHẤT



==>>

projected points
samples go in
plot.



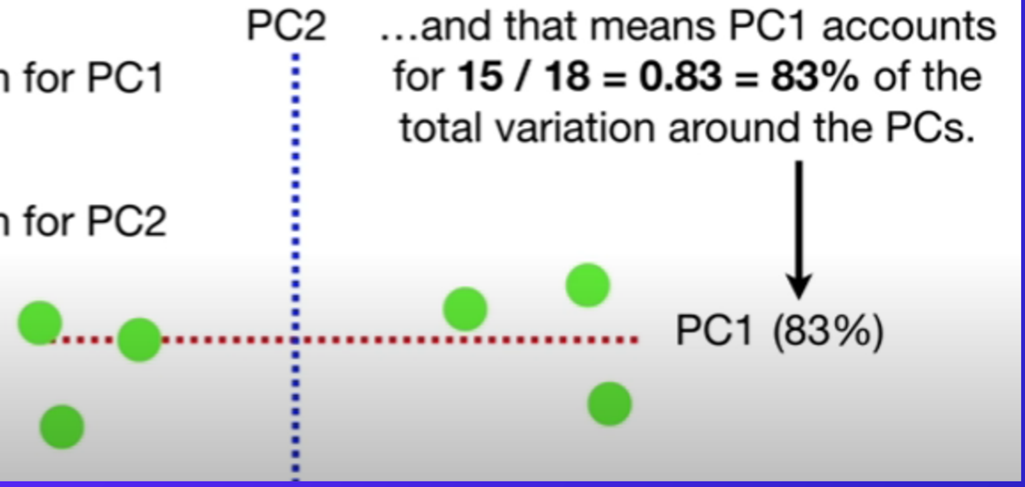
CÁCH TÍNH TỈ LỆ (%) CỦA CÁC PCA

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

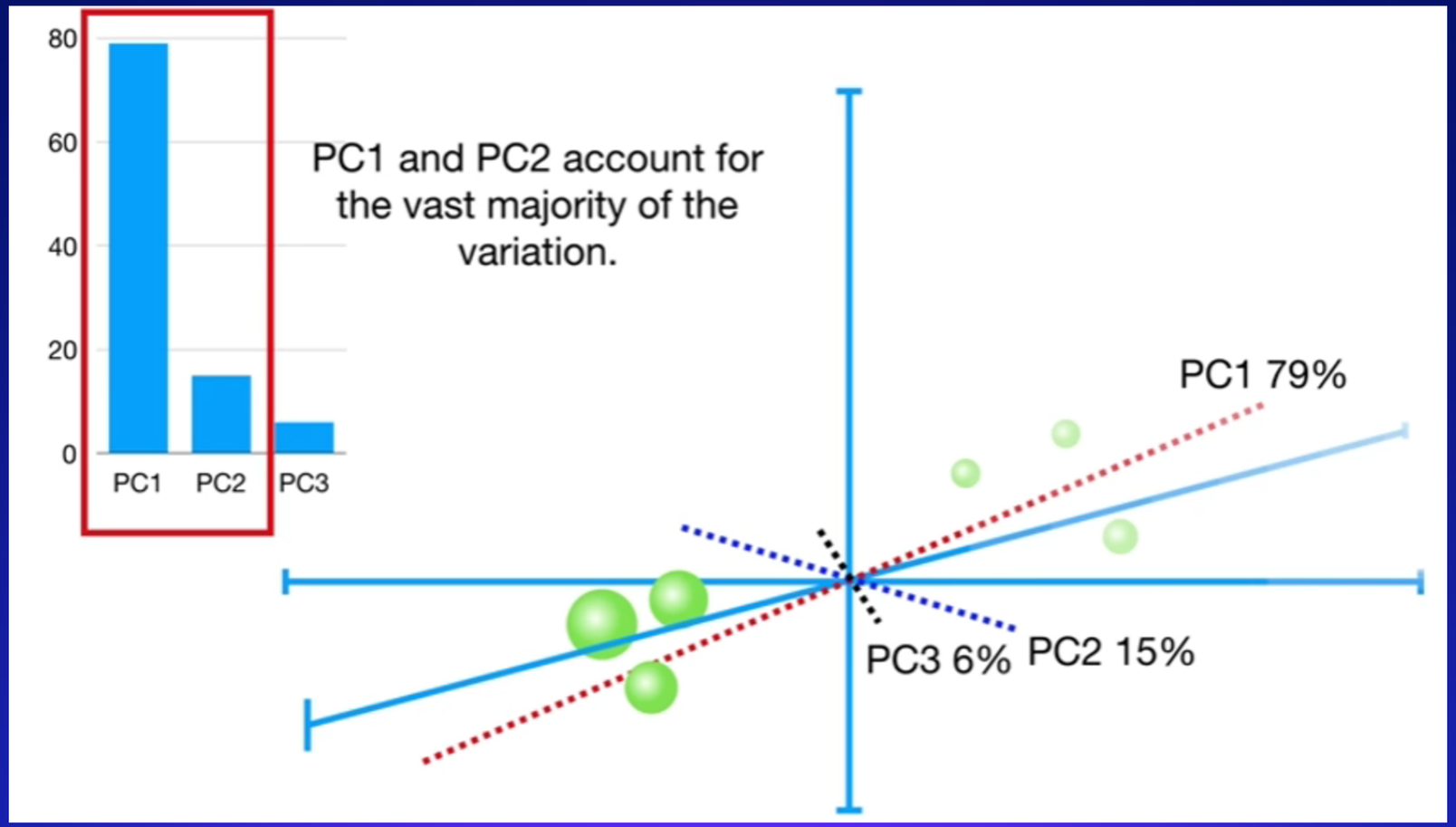
That means that the total variation around both PCs is **15 + 3 = 18**...

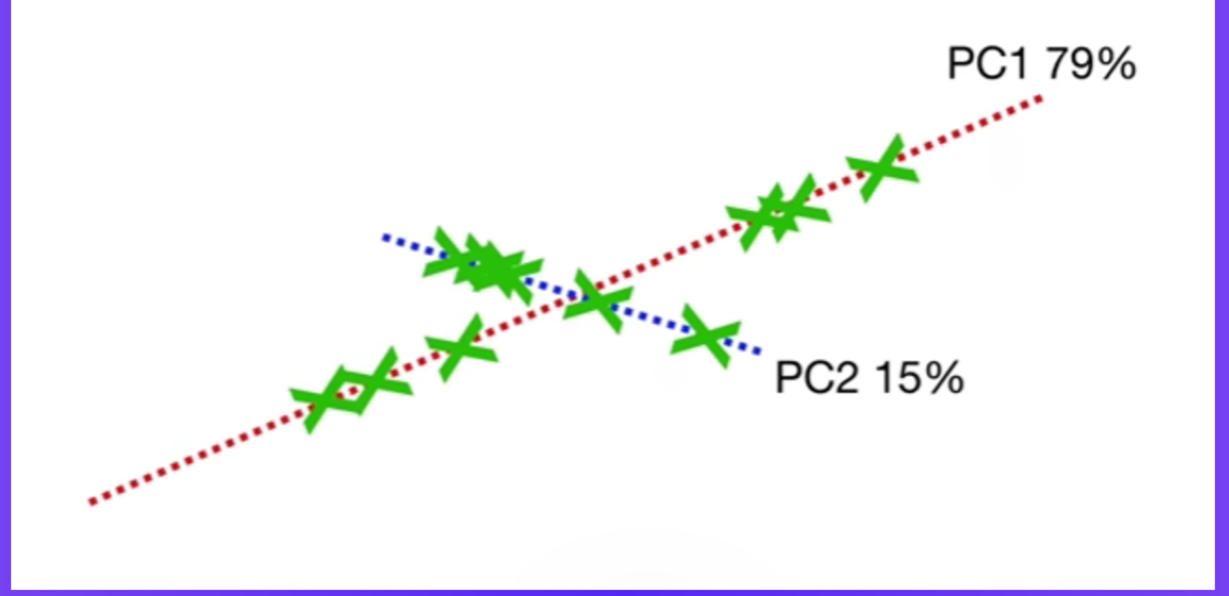
$$\frac{SS(\text{distances for PC1})}{n - 1} = \text{Variation for PC1}$$

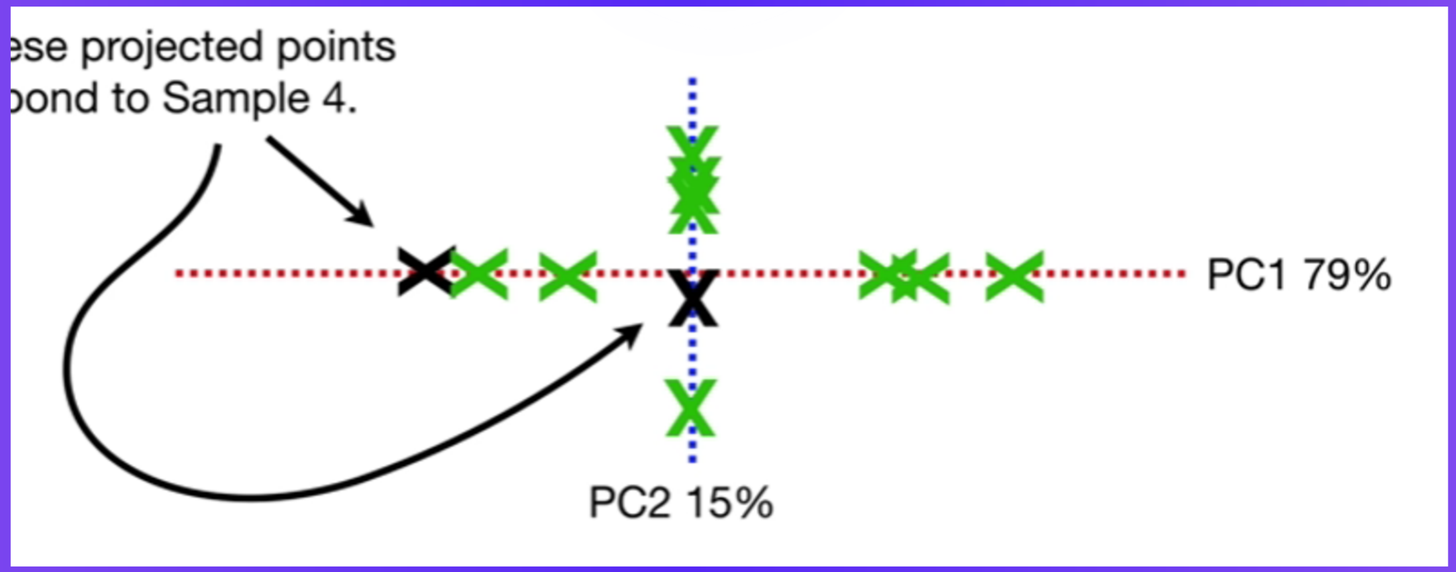
$$\frac{SS(\text{distances for PC2})}{n - 1} = \text{Variation for PC2}$$



ĐỐI VỚI DỮ LIỆU NHIỀU CHIỀU HƠN

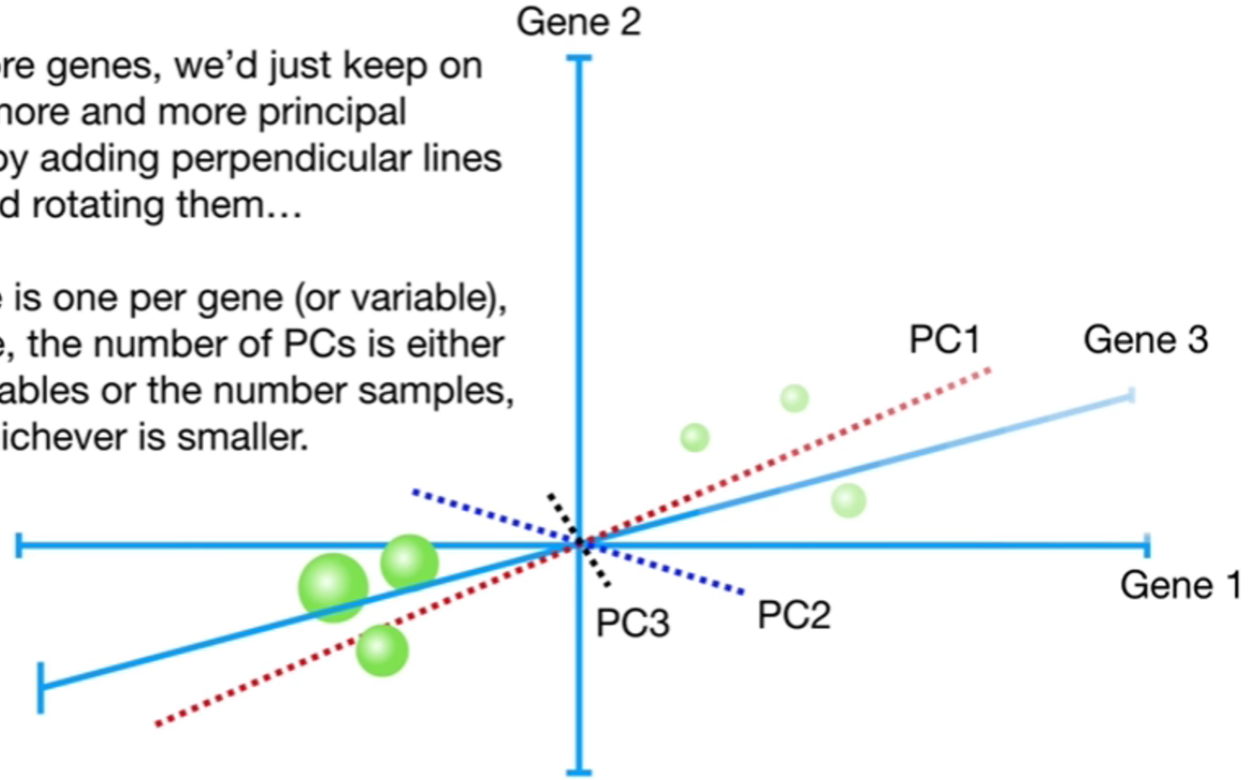



PC1 79%
PC2 15%


ese projected points
oond to Sample 4.
PC1 79%
PC2 15%

e had more genes, we'd just keep on
finding more and more principal
ponents by adding perpendicular lines
and rotating them…

ory there is one per gene (or variable),
practice, the number of PCs is either
er of variables or the number samples,
whichever is smaller.



THANK YOU!