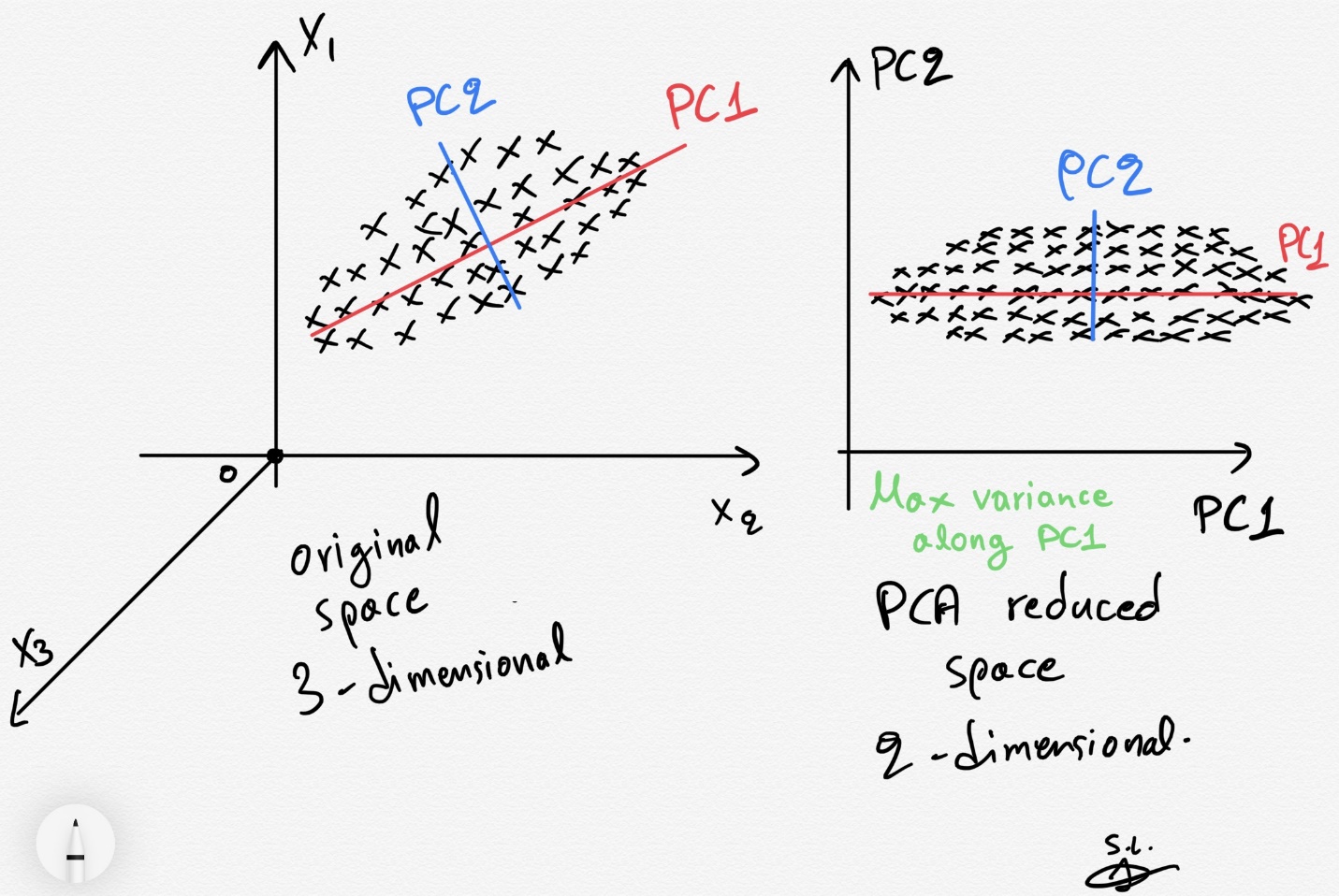
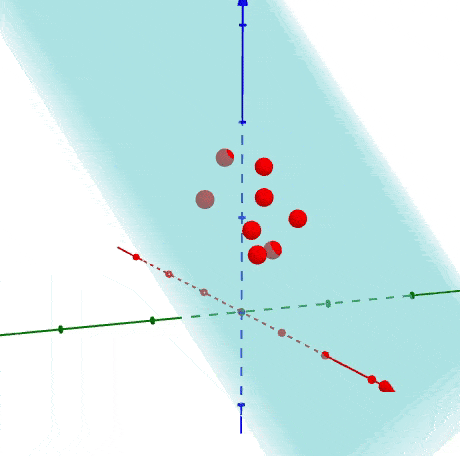
**Phép phân tích thành phần chính** (Principal Components Analysis - PCA)





Cách đơn giản nhất để giảm chiều dữ liệu từ **D** về **K<D** là chỉ giữ lại **K** phần tử quan trọng nhất. Tuy nhiên, việc làm này chắc chắn chưa phải tốt nhất vì chúng ta chưa biết xác định thành phần nào là quan trọng hơn. Hoặc trong trường hợp xấu nhất, lượng thông tin mà mỗi thành phần mang là như nhau, bỏ đi thành phần nào cũng dẫn đến việc mất một lượng thông tin lớn.

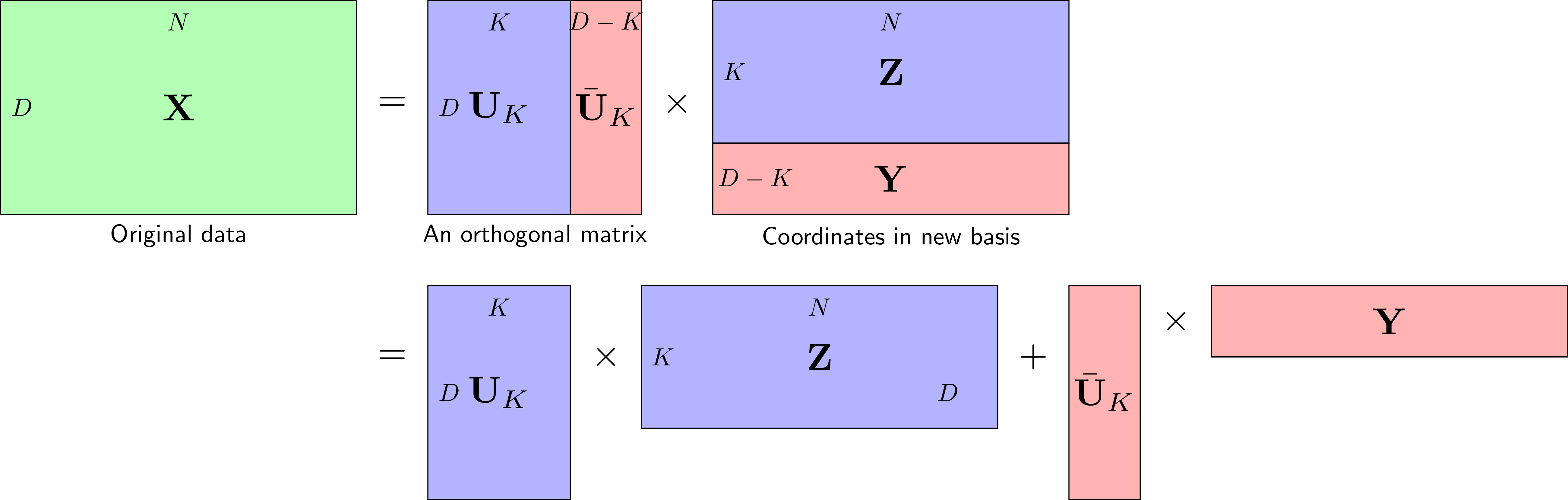
Tuy nhiên, nếu chúng ta có thể biểu diễn các vector dữ liệu ban đầu trong một hệ cơ sở mới mà trong hệ cơ sở mới đó, tầm quan trọng giữa các thành phần là khác nhau rõ rệt, thì chúng ta có thể bỏ qua những thành phần ít quan trọng nhất.

Lấy một ví dụ về việc có hai camera đặt dùng để chụp một con người, một camera đặt phía trước người và một camera đặt trên đầu. Rõ ràng là hình ảnh thu được từ camera đặt phía trước người mang nhiều thông tin hơn so với hình ảnh nhìn từ phía trên đầu. Vì vậy, bức ảnh chụp từ phía trên đầu có thể được bỏ qua mà không có quá nhiều thông tin về hình dáng của người đó bị mất.

**PCA** chính là phương pháp đi tìm một hệ cơ sở mới sao cho thông tin của dữ liệu chủ yếu tập trung ở một vài toạ độ, phần còn lại chỉ mang một lượng nhỏ thông tin. Và để cho đơn giản trong tính toán, **PCA** sẽ tìm một [hệ trực chuẩn](https://machinelearningcoban.com/2017/06/07/svd/#-he-truc-giao-va-truc-chuan) để làm cơ sở mới.

Giả sử hệ cơ sở trực chuẩn mới là **U** và chúng ta muốn giữ lại **K** toạ độ trong hệ cơ sở mới này. Không mất tính tổng quát, giả sử đó là **K** thành phần đầu tiên. Quan sát Hình 3 dưới đây:

Giả sử hệ cơ sở trực chuẩn mới là **U** và chúng ta muốn giữ lại **K** toạ độ trong hệ cơ sở mới này. Không mất tính tổng quát, giả sử đó là **K** thành phần đầu tiên. Quan sát hình dưới đây:



Ý tưởng chính của **PCA**: Tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong **K** thành phần đầu tiên.

Mục đích của **PCA** là đi tìm ma trận trực giao **U** sao cho phần lớn thông tin được giữ lại ở phần màu xanh **UKZ** và phần màu đỏ **¯UKY** sẽ được lược bỏ và thay bằng một ma trận không phụ thuộc vào từng điểm dữ liệu. Nói cách khác, ta sẽ xấp xỉ **Y** bởi một ma trận có toàn bộ các cột là như nhau. Chú ý rằng các cột này có thể phụ thuộc vào dữ liệu training nhưng không phụ thuộc vào dữ liệu test.