Let $P$ be a set of $n$ points in $d$-dimensional space, where $d$ is a very large value. Informally, the goal of **dimensionality reduction** is to convert $P$ into a set $P'$ of points in a $k$-dimensional space where $k < d$, such that $P'$ loses as little information about $P$ as possible.

Today, we will learn a popular method of dimensionality reduction called **principled component analysis** (PCA).
A vector $\mathbf{v}$ is a $d \times 1$ matrix: $\mathbf{v} = (v[1], \ldots, v[d])^T$.

A point can be represented as a vector.

A vector $\mathbf{v}$ is a **unit vector** if $\sum_{i=1}^{d} v[i]^2 = 1$.

Dot product $\mathbf{v}_1 \cdot \mathbf{v}_2 = \sum_{i=1}^{d} (v_1[i]v_2[i])$.

If two vectors $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal, $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$.

Let $\mathbf{p}$ be a point and $\mathbf{v}$ a unit vector. Then, $\mathbf{p} \cdot \mathbf{v}$ gives the distance from the origin to the projection of $\mathbf{p}$ on $\mathbf{v}$. 
Let $S$ be a set of real numbers $r_1, ..., r_m$. The mean of $S$ equals:

$$\text{mean}(S) = \frac{1}{m} \sum_{i=1}^{m} r_i.$$ 

The variance of $S$ equals:

$$\text{var}(S) = \frac{1}{m} \sum_{i=1}^{m} (r_i - \text{mean}(S))^2.$$
Let $P$ be a set of $n d$-dimensional points $p_1, \ldots, p_n$. Its co-variance between dimensions $i$ and $j$ (where $1 \leq i \leq j \leq d$) equals

$$
\frac{1}{n} \sum_{k=1}^{n} (p_k[i] - mean_i)(p_k[j] - mean_j)
$$

where $mean_i$ (resp., $mean_j$) is the mean of the coordinates in $P$ along dimension $i$ (resp., $j$).
The **co-variance matrix** $A$ of point set $P$ is a $d \times d$ matrix whose value at the $i$-th row and $j$-th column ($i, j \in [1, d]$) is the co-variance of $P$ between dimensions $i$ and $j$.

Note that $A$ is symmetric, namely, $A = A^T$. 
Let $A$ be a $d \times d$ matrix. If

$$A\mathbf{v} = \lambda \mathbf{v}$$

for some $d \times 1$ unit vector $\mathbf{v}$ and some real value $\lambda$, then $\mathbf{v}$ is called a unit eigenvector of $A$ and $\lambda$ is called an eigenvalue of $A$. 
algorithm \((P, k)\)
/* input: \(P\) is a set of \(d\)-dimensional points and \(k\) is an integer in \([1, d]\) */
/* output: a subspace defined by \(k\) orthogonal vectors */
1. shift \(P\) such that its geometric mean is at the origin of the data space
2. \(A \leftarrow \) the co-variance matrix of \(P\)
3. compute all the \(d\) unit eigenvectors
4. arrange the eigenvectors in descending order of their eigenvalues
5. return the first \(k\) eigenvectors \(\mathbf{v}_1, \ldots, \mathbf{v}_k\)

Note

Each point \(p\) is then converted to a \(k\)-dimensional point whose \(i\)-th \((1 \leq i \leq k)\) coordinate is \(\mathbf{v}_i \cdot p\).
Here is a key property of PCA.

$v_1$ is the direction along which the projections of $P$ have the largest variance. In general, $v_i$ ($i > 1$) is the direction along which $P$ has the largest variance, among all directions orthogonal to all of $v_1, ..., v_{i-1}$.

Next we will prove the above for $v_1$ and $v_2$. Then, the cases with $v_3, ..., v_i$ follow the same idea.
Formally, redefine $P$ be a set of $n$ $d$-dimensional points with zero mean on all dimensions. Let $w$ be a unit vector. We can project $P$ onto $w$ to obtain a set of 1d values: $S = \{ p \cdot w \mid p \in P \}$. Define the quality of $w$ be $\text{var}(S)$.

**Theorem 1**

The first eigenvector output by PCA has the highest quality.
Proof of Theorem 1

Let $X$ be the $n \times d$ matrix where each row lists the coordinates of a point in $P$. Thus, we can view $S$ as a vector $Xw$. Thus:

$$\text{var}(S) = \frac{1}{n} (Xw)^T (Xw)$$

$$= w^T X^T X w$$

$$= w^T A w$$

where $A$ is the covariance matrix of $P$. Hence, we want to maximize the above subject to the constraint that $w^T w = 1.$
Proof of Theorem 1 (Cont.)

Now we apply the method of Lagrange multipliers to find the maximum. Introduce a real value $\lambda$, and now consider the objective function

$$f(w, \lambda) = w^T A w - \lambda (w^T w - 1) \implies$$

$$\frac{\partial f}{\partial w} = 2 A w - 2 \lambda w$$

Equating the above 0 gives $A w = \lambda w$. In other words, $w$ needs to be an eigenvector, and $\lambda$ the corresponding eigenvalue.
Proof of Theorem 1 (Cont.)

Now it remains to check which eigenvector gives the largest variance. Observe that:

\[
\text{var}(S) = w^T A w = w^T \lambda w = \lambda
\]

In other words, when we choose eigenvector \(w\) as our solution, its quality is exactly the eigenvalue \(\lambda\). Hence, the eigenvector with the maximum eigenvalue is what we are looking for. □
Theorem 2

The second eigenvector output by PCA has the highest quality, among all the vectors $w$ orthogonal to the first eigenvector $v_1$. 
Proof of Theorem 2

Let $A$ be the covariance matrix of $P$. As shown in the proof of Theorem 1, we proved that

$$\text{var}(S) = w^T A w.$$ 

Hence, we want to maximize the above subject to the constraints $w^T w = 1$ and $w^T v_1 = 0$.

Now we apply the method of Lagrange multipliers to find the maximum. Introduce real values $\lambda$ and $\phi$, and now consider the objective function

$$f(w, \lambda, \phi) = w^T A w - \lambda (w^T w - 1) - \phi w^T v_1 \Rightarrow$$

$$\frac{\partial f}{\partial w} = 2 A w - 2 \lambda w - \phi v_1.$$
Proof of Theorem 2 (Cont.)

The optimal $\mathbf{w}$ needs to satisfy $\frac{\partial f}{\partial \mathbf{w}} = 0$, namely:

$$2\mathbf{Aw} - 2\lambda \mathbf{w} - \phi \mathbf{v}_1 = 0. \quad (1)$$

Next we prove that $\phi$ must be 0. To see this, multiplying both sides of (1) by $\mathbf{v}_1^T$, we get:

$$2\mathbf{v}_1^T \mathbf{Aw} - 2\lambda \mathbf{v}_1^T \mathbf{w} + \phi \mathbf{v}_1^T \mathbf{v}_1 = 0. \quad (2)$$

We know that $\mathbf{v}_1^T \mathbf{w} = 0$, and $\mathbf{v}_1^T \mathbf{v}_1 = 1$. Furthermore,

$$\mathbf{v}_1^T \mathbf{Aw} = \mathbf{w}^T \mathbf{A}^T \mathbf{v}_1 = \mathbf{w}^T \mathbf{Av}_1 = \mathbf{w}^T (\mathbf{Av}_1) = \mathbf{w}^T \mathbf{v}_1 = 0.$$  

Hence, from (2), we get $\phi = 0$. 

Proof of Theorem 2 (Cont.)

Therefore, from (1), we know:

\[ 2Aw - 2\lambda w = 0 \]

namely, \( w \) must also be an eigenvector.

From the proof of Theorem 1, we know that \( \text{var}(S) \) equals the eigenvalue corresponding to \( w \). This thus indicates that \( w \) is the eigenvector of \( A \) with the second largest eigenvalue.