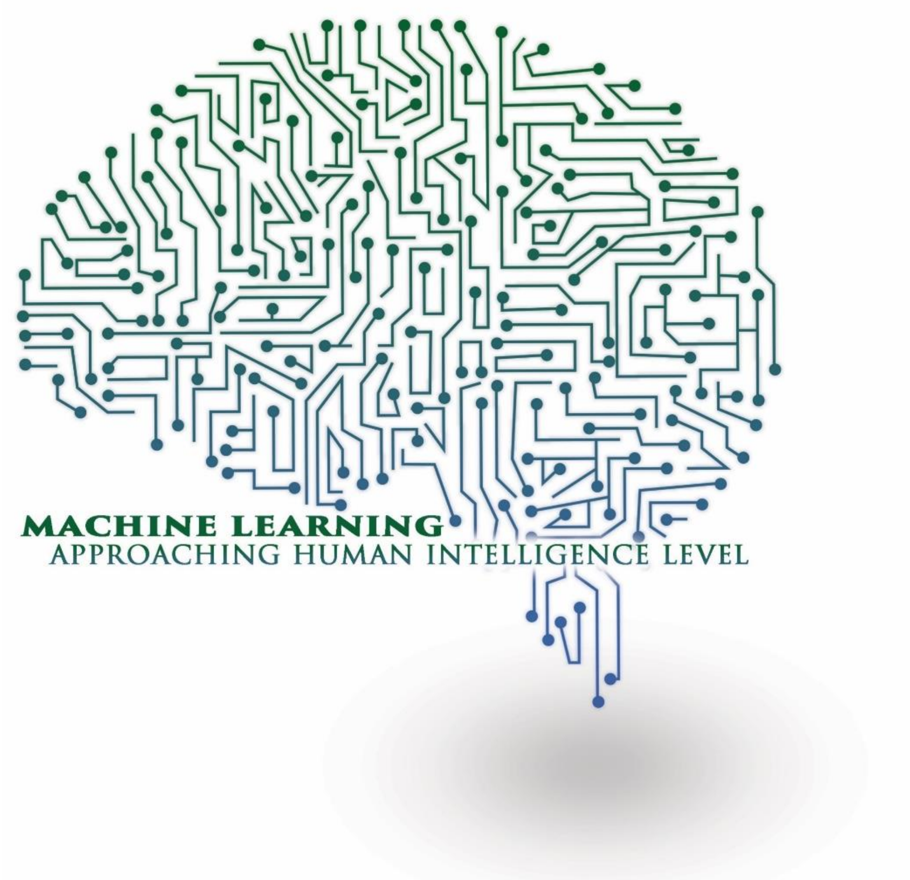
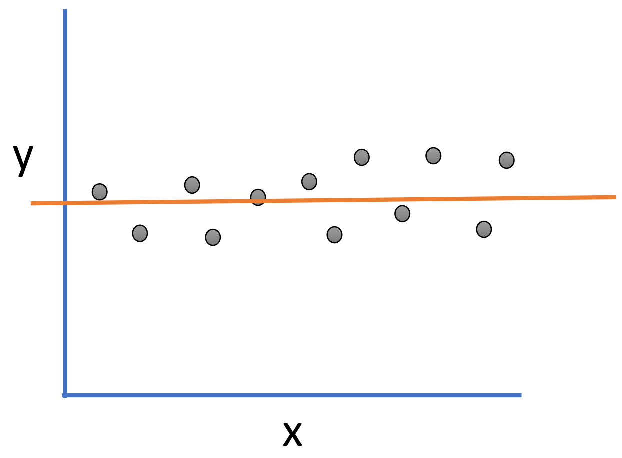
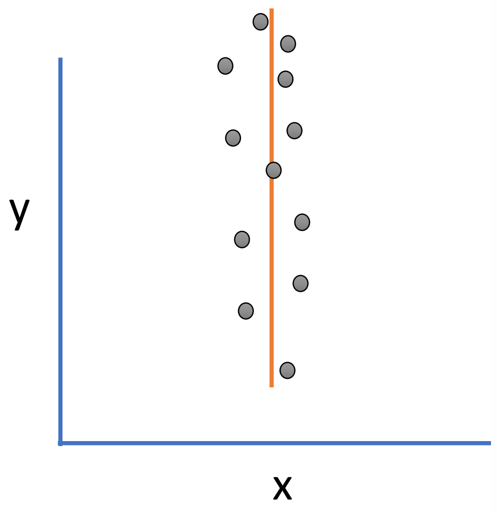


Dimentionality Reduction

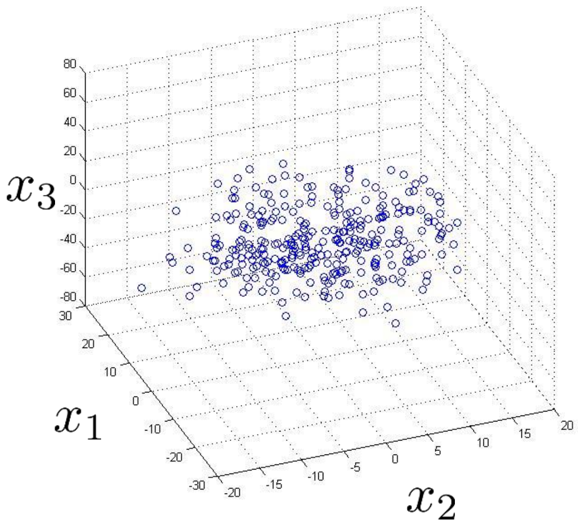
Motivation

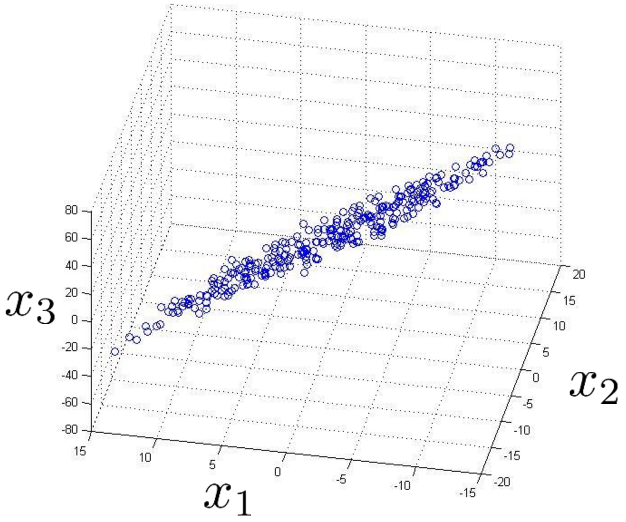
I. Data Compression

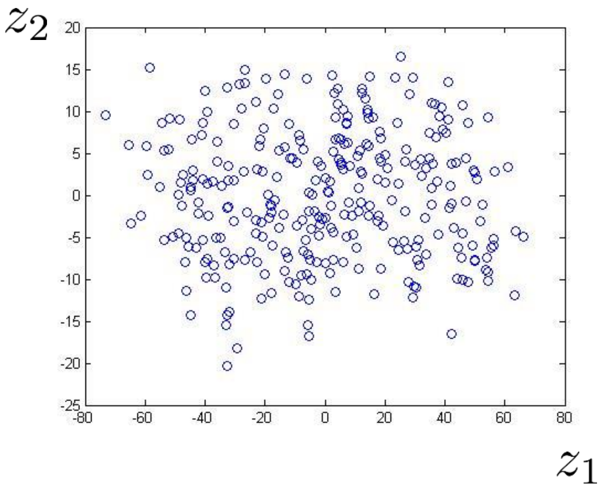




$$from\ 2D\ to\ 1D$$
$$\mathcal{R}^2 \longrightarrow \mathcal{R}$$







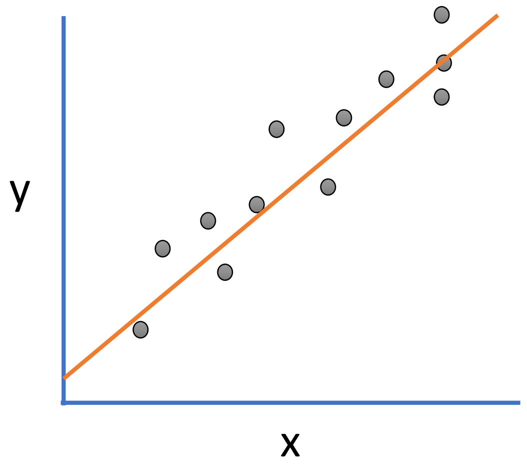
$from\ 3D\ to\ 2D$

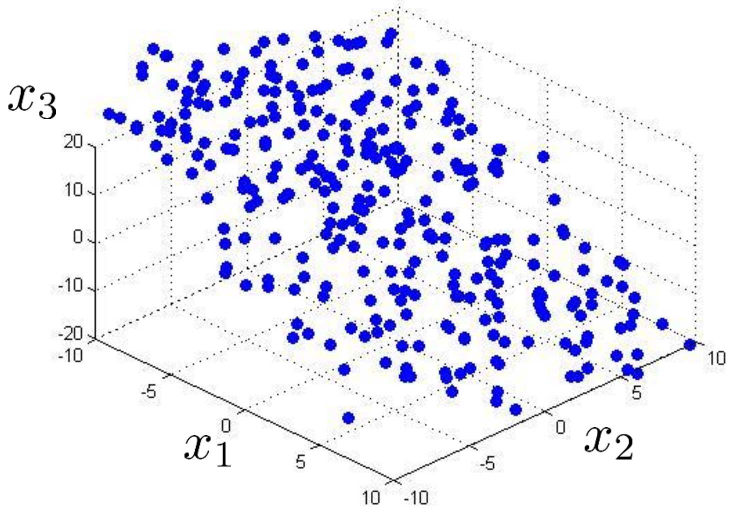
$\mathcal{R}^3 \longrightarrow \mathcal{R}^2$

II. Data Visualization

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY
-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY
-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY
-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY
-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY
-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY
-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY
-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY
-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY
-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY

Principal Component Analysis problem formulation





Reduce from 2-dimension to 1-dimension: Find a direction (a vector $u^{(1)} \in \mathbb{R}^n$) onto which to project the data so as to minimize the projection error.

In general:
Reduce from n-dimension to k-dimension: Find vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ onto which to project the data, so as to minimize the projection error.

Data preprocessing

Why?!

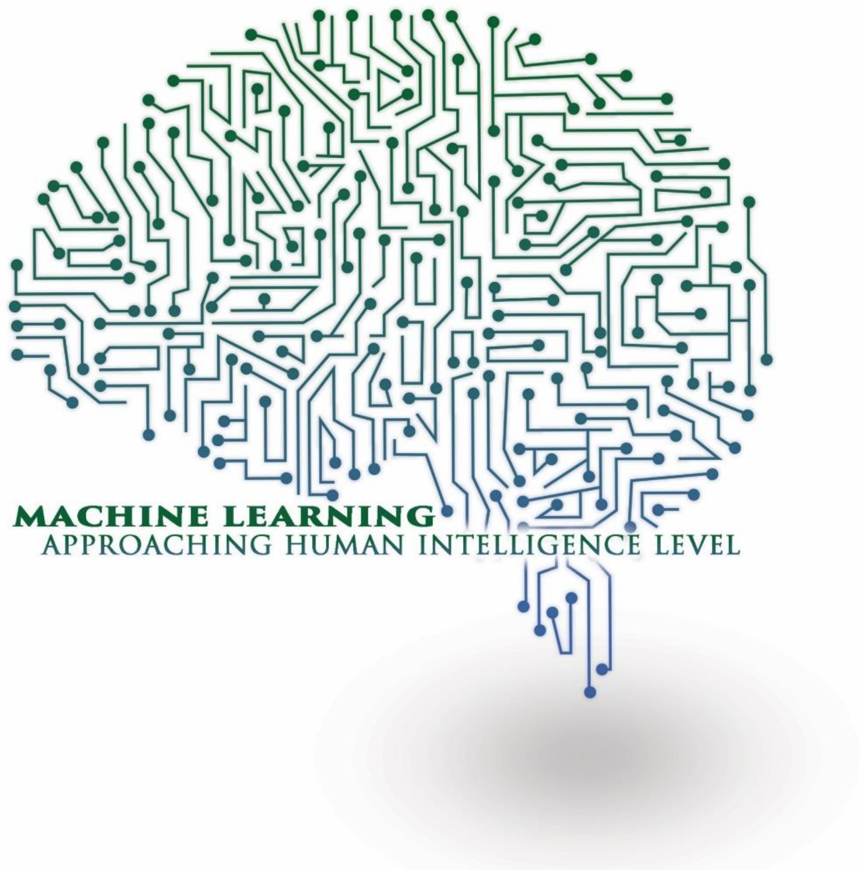
Training set: $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$

Preprocessing (feature scaling/mean normalization):

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

Replace each $x_j^{(i)}$ with $x_j - \mu_j$

If different features on different scales (e.g., $x_1 =$ median age of house, $x_2 =$ number of bedrooms), scale features to have comparable range of values.



SVD

Linear algebra revisited

Orthogonal matrix

In linear algebra, an orthogonal matrix or real orthogonal matrix is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e., orthonormal vectors), i.e.

$$QQ^T = Q^TQ = I$$

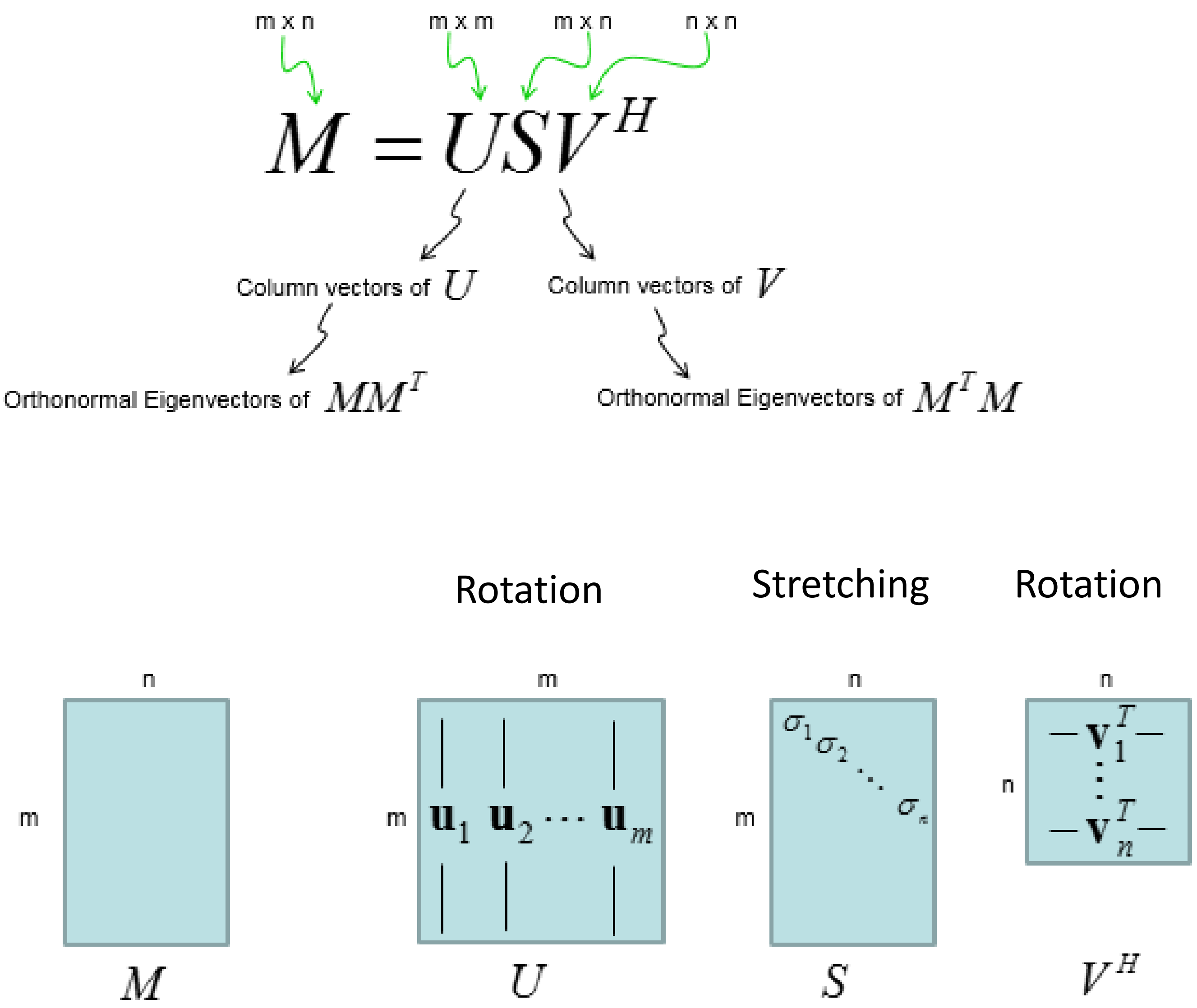
$$Q^T = Q^{-1}$$

Diagonal matrix

Is a matrix in which the entries outside the main diagonal are all zero.

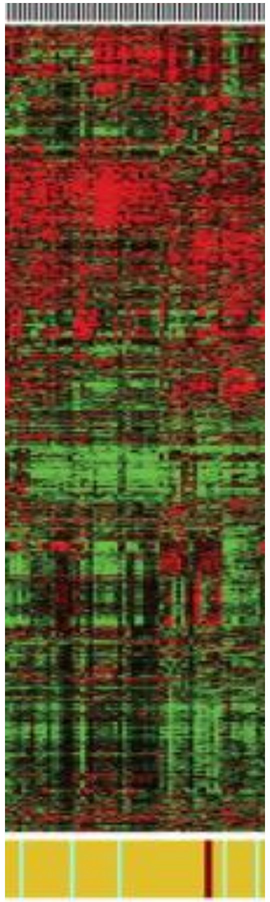
$$\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

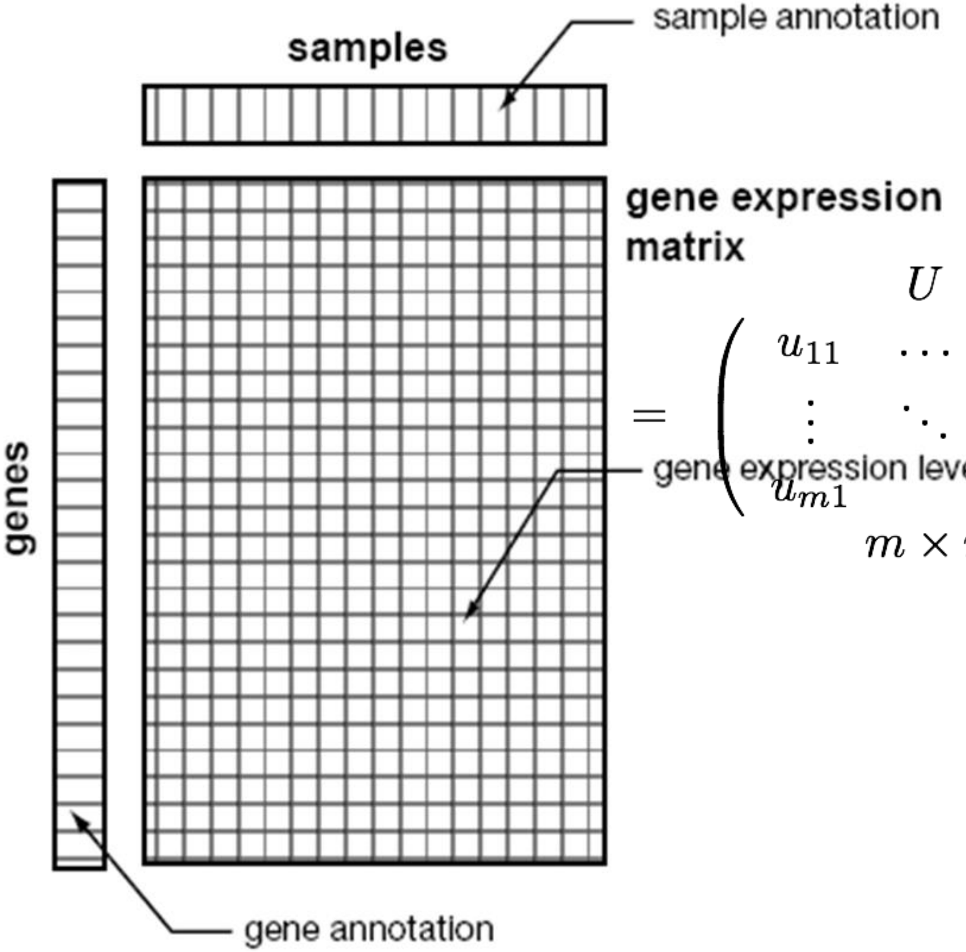
SVD



m x n
m x m
m x n
n x n
$M = USV^H$
Column vectors of U
Column vectors of V
Orthonormal Eigenvectors of MM^T
Orthonormal Eigenvectors of M^TM
Rotation
Stretching
Rotation
n
m
m
m
n
m
n
n
$\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_m$
$\sigma_1\ \sigma_2 \ \ddots\ \sigma_n$
$\mathbf{v}_1^T$
$\mathbf{v}_n^T$
M
U
S
V^H

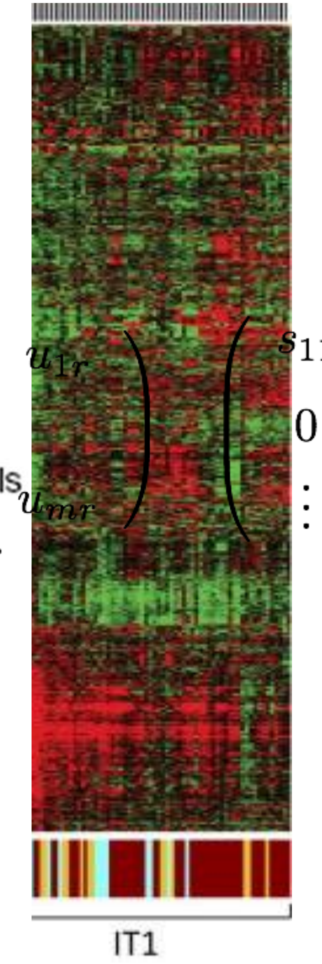
Appling SVD





Gene Expression Matrix

$$= \underset{m \times r}{\overset{U}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \underset{r \times r}{\overset{S}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \underset{r \times n}{\overset{V^{\mathsf{T}}}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}$$



Principal component analysis

$u_1 \, \sigma_1 \, v_1^T$

Principal Component Analysis (PCA) algorithm

Reduce data from n-dimensions to k-dimensions

Compute "covariance matrix":

$$\Sigma = \frac{1}{m} \sum_{i=1}^{n} (x^{(i)})(x^{(i)})^T$$

Compute "eigenvectors" of matrix Σ:

python implementation