## Exercise 1: SVD and its properties

**Data:**

Step 1: Generate two random matrices: A = randn(1000,2); B = randn(200,2);

Step 2: Construct matrix X = A*B’ + noise, i.e., where noise is =0.1*randn(1000,200)

- What is the rank of X?
- Plot the singular values of X
- Is X low-rank?
- What is the best rank-2 approximation of X?
- Compute what percent of the “energy” is explained using the best rank-2 approx.
- Plot the significant left and right singular vectors.

# Exercise 2: Recommender Systems

Given the following user-movie rating matrix, can you group users based on their interests in movies? Which groups of users are interested in what type of movies?

$$\begin{array}{c|cccccc|} & \text{Aloha} & \text{Star Wars} & \text{American Pie} & \text{Hunger Games} & \text{Silver Linings} & \text{Maze Runner} \\ \hline \text{User 1} & 5 & 2 & 4 & 1 & 3 & 2 \\ \text{User 2} & 1 & 4 & 1 & 4 & 1 & 4 \\ \text{User 3} & 3 & 1 & 5 & 2 & 5 & 2 \\ \text{User 4} & 1 & 4 & 1 & 5 & 3 & 4 \\ \text{User 5} & 3 & 1 & 4 & 2 & 3 & 2 \\ \text{User 6} & 2 & 4 & 2 & 4 & 2 & 4 \end{array}$$

**load exercise2_usermovie.mat**
**Compute its SVD**

# Exercise 3: Text Mining

Given the following document-term matrix, find the related documents and what they are about.

**Document 1: Tiger stopped playing golf**

**Document 2: News** about **Tiger** and his **golf career**

**Document 3: Golf career** of **Tiger** in **jeopardy**

**Document 4: Tiger** and his **wife** in the **news**

**Document 5:** The **new zoo featuring** the **big cat family**: **tigers** and **lions**

**Document 6: Tigers** – the **big cats** – in the **new zoo**

**Document 7: Tigers** and **lions**, **which** are the **biggest cats**?

$$
\begin{array}{c|cccccccccccccccc}
 & \text{tiger} & \text{stop} & \text{play} & \text{golf} & \text{news} & \text{career} & \text{jeopardy} & \text{wife} & \text{zoo} & \text{featuring} & \text{new} & \text{big} & \text{cat} & \text{family} & \text{lions} & \text{which} \\
\hline
\text{Doc1} & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Doc2} & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Doc3} & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Doc4} & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Doc5} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
\text{Doc6} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
\text{Doc7} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1
\end{array}
$$

**load exercise3_docterm.mat**
**Compute its SVD**

Which group does this document belong to?

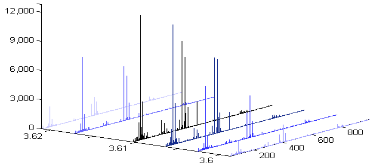
**Document 8: Tiger** back to **golf**

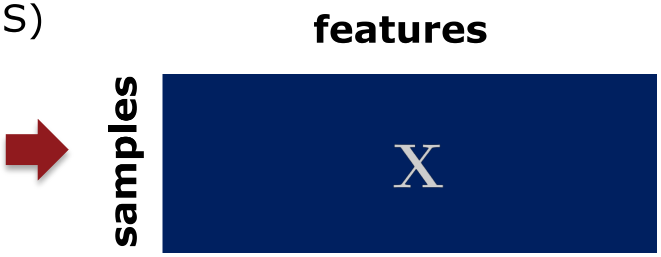
$$
q^{\mathsf{T}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
$$

# Exercise 4: Metabolomics

Some are fed with 10g apple while
some are controls (no apple)

Liquid Chromatography-
Mass Spectrometry (LC-MS)

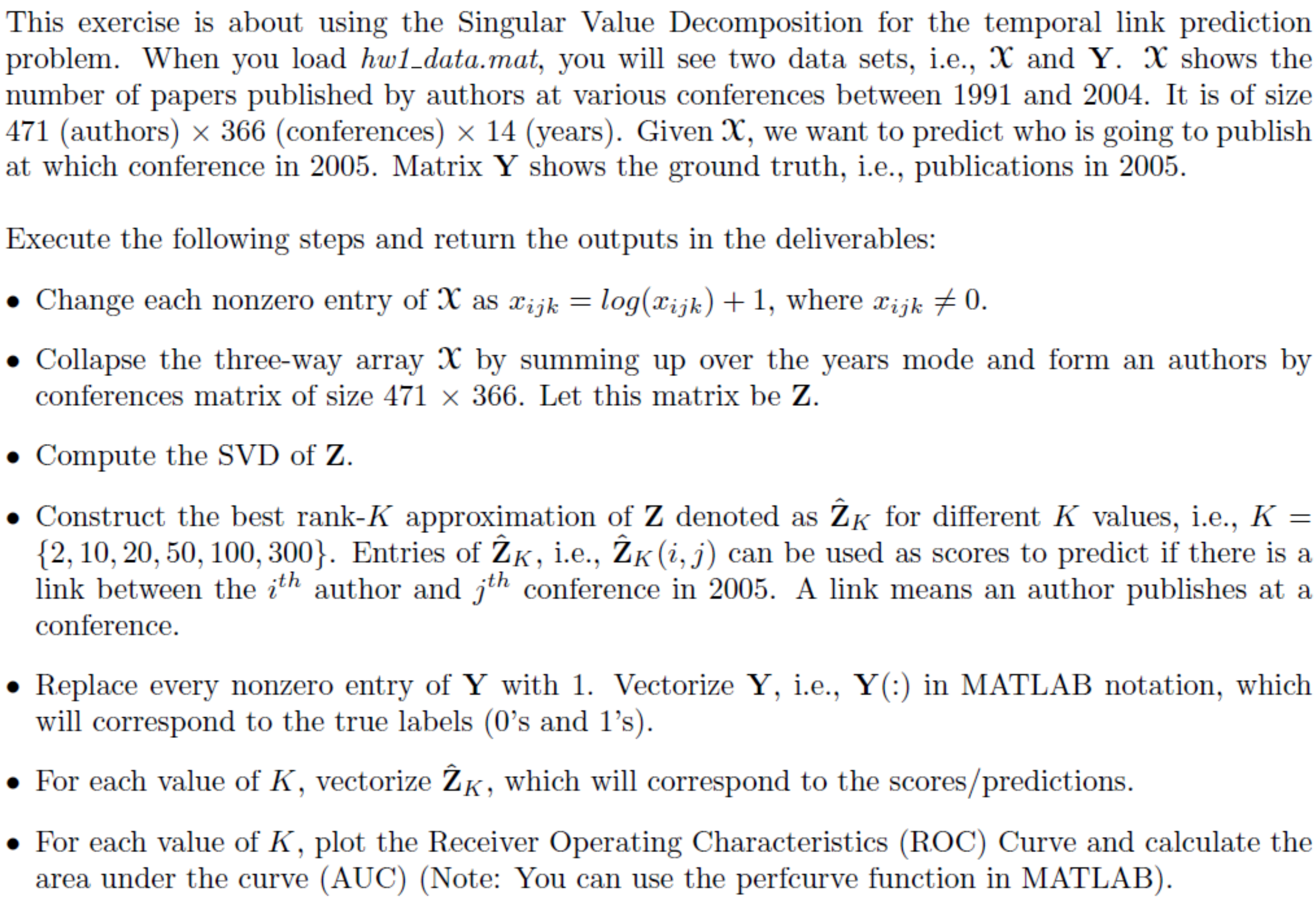




**load exercise4_metabolomics.mat**

- Preprocess the data
- Plot singular values
- Compute the best rank-r approximation that represents the 80% of the "energy"
- Can you group the samples based on apple consumption (Class1:10g apple, Class2: no apple)?

# Exercise 5: Link Prediction using the DBLP data



This exercise is about using the Singular Value Decomposition for the temporal link prediction problem. When you load *hw1_data.mat*, you will see two data sets, i.e., $\mathcal{X}$ and $\mathbf{Y}$. $\mathcal{X}$ shows the number of papers published by authors at various conferences between 1991 and 2004. It is of size 471 (authors) × 366 (conferences) × 14 (years). Given $\mathcal{X}$, we want to predict who is going to publish at which conference in 2005. Matrix $\mathbf{Y}$ shows the ground truth, i.e., publications in 2005.

Execute the following steps and return the outputs in the deliverables:

- Change each nonzero entry of $\mathcal{X}$ as $x_{ijk} = log(x_{ijk}) + 1$, where $x_{ijk} \neq 0$.
- Collapse the three-way array $\mathcal{X}$ by summing up over the years mode and form an authors by conferences matrix of size 471 × 366. Let this matrix be $\mathbf{Z}$.
- Compute the SVD of $\mathbf{Z}$.
- Construct the best rank-$K$ approximation of $\mathbf{Z}$ denoted as $\hat{\mathbf{Z}}_K$ for different $K$ values, i.e., $K = \{2, 10, 20, 50, 100, 300\}$. Entries of $\hat{\mathbf{Z}}_K$, i.e., $\hat{\mathbf{Z}}_K(i, j)$ can be used as scores to predict if there is a link between the $i^{th}$ author and $j^{th}$ conference in 2005. A link means an author publishes at a conference.
- Replace every nonzero entry of $\mathbf{Y}$ with 1. Vectorize $\mathbf{Y}$, i.e., $\mathbf{Y}(:)$ in MATLAB notation, which will correspond to the true labels (0's and 1's).
- For each value of $K$, vectorize $\hat{\mathbf{Z}}_K$, which will correspond to the scores/predictions.
- For each value of $K$, plot the Receiver Operating Characteristics (ROC) Curve and calculate the area under the curve (AUC) (Note: You can use the perfcurve function in MATLAB).

## Exercise 6: Solving Least Squares using SVD