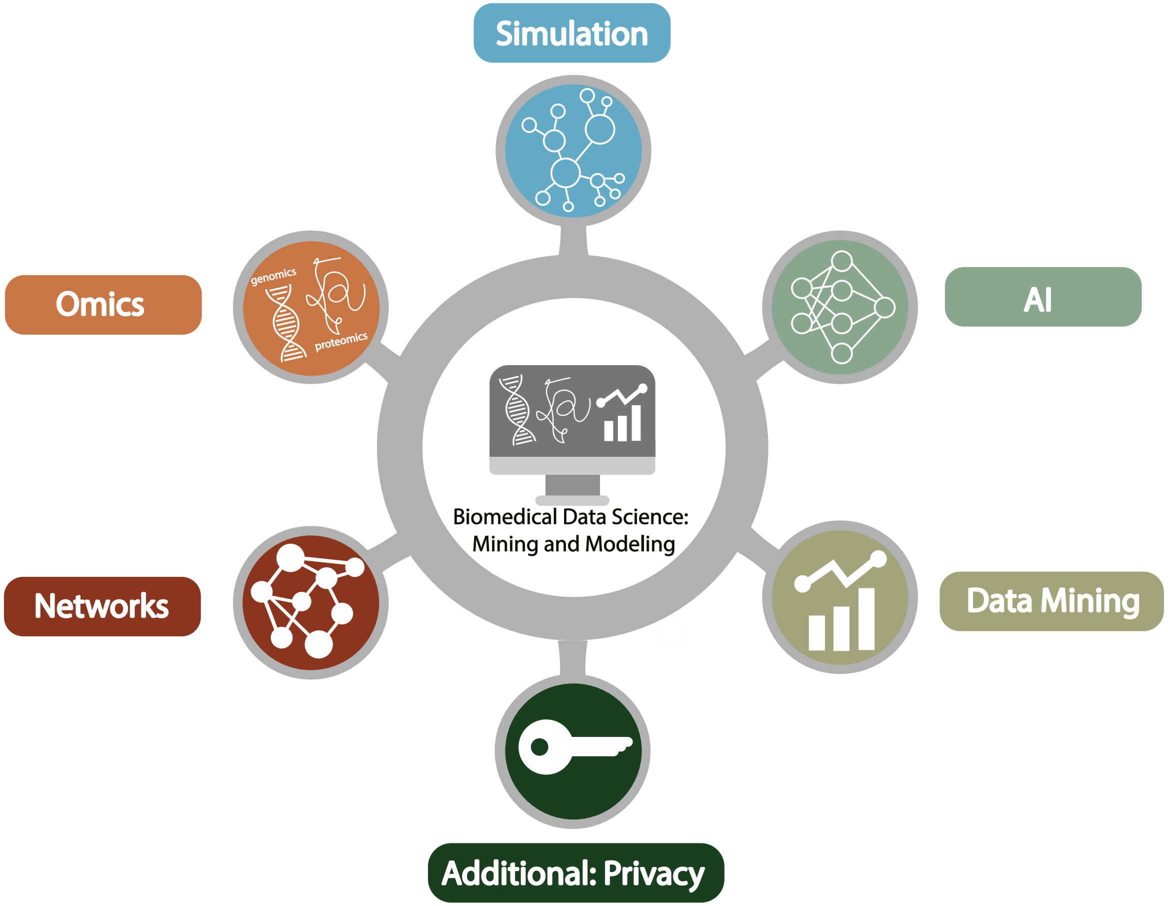
Biomedical Data Science (GersteinLab.org/courses/452)

Unsupervised Datamining – SVD (25m9c)



Last edit in spring ’25. Condensing by ~3 slide deletions from 2022’s 22m9c, which is similar to 2021’s M9c [which has a video].

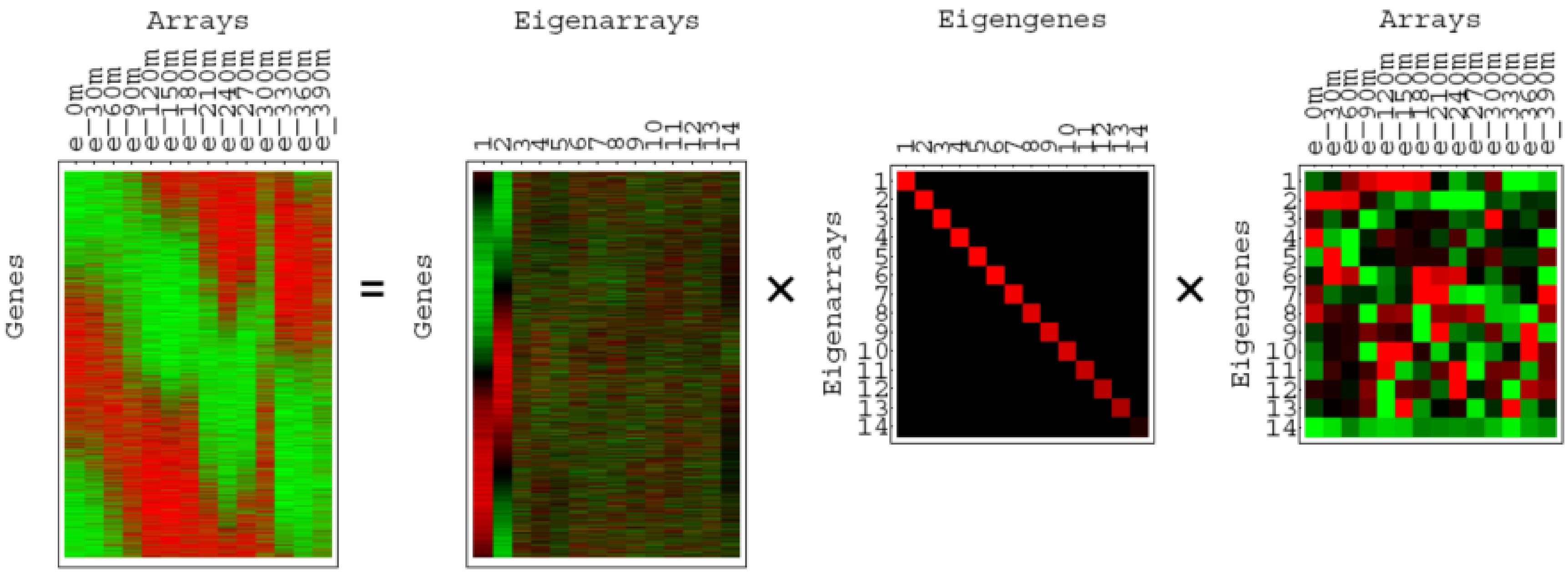
Mark Gerstein
Yale U.

Unsupervised Mining

SVD

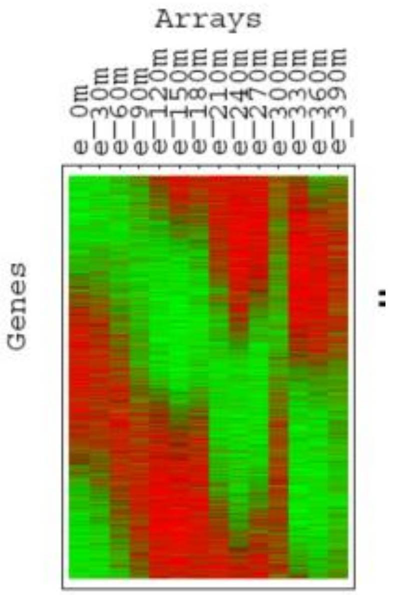
Puts together slides prepared by
Brandon Xia with images from
Alter et al. papers

SVD for microarray data
(Alter et al, PNAS 2000)



$$\color{red}{A} = USV^T$$

- A is any rectangular matrix (m ≥ n)
- Row space: vector subspace generated by the row vectors of A
- Column space: vector subspace generated by the column vectors of A
  - The dimension of the row & column space is the rank of the matrix A: r (≤ n)
- A is a linear transformation that maps vector x in row space into vector Ax in column space

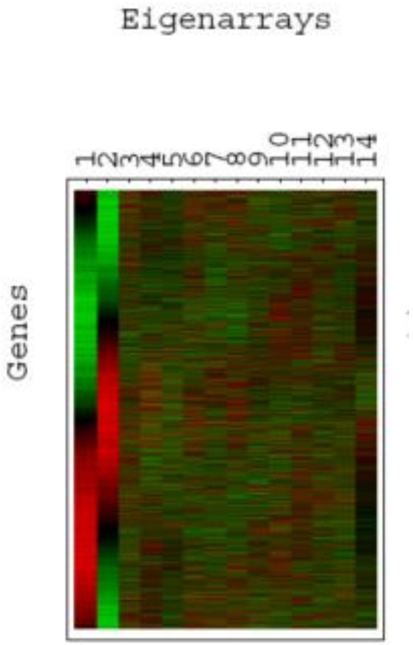


$A = USV^T$

- U is an “orthogonal” matrix (m ≥ n)
- Column vectors of U form an orthonormal basis for the column space of A: $U^TU=I$

$$U = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathrm{L} & \mathbf{u}_n \\ | & | & & | \end{pmatrix}$$

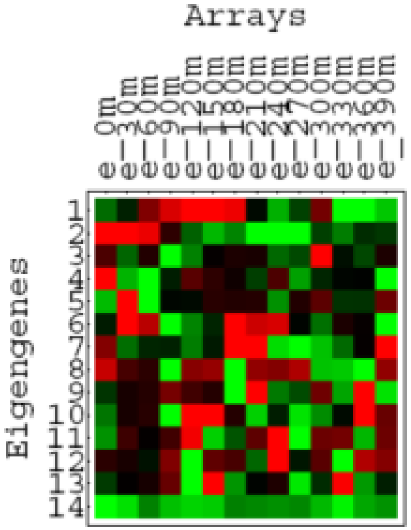
- $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ in U are eigenvectors of AA^T
  - $AA^T = USV^T VSU^T = US^2 U^T$
  - “Left singular vectors”



$A = USV^T$

- V is an orthogonal matrix (n by n)
- Column vectors of V form an orthonormal basis for the row space of A: $V^TV=VV^T=I$

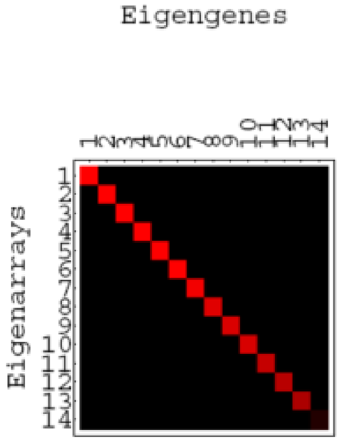
$$V = \begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathrm{L} & \mathbf{v}_n \\ | & | & & | \end{pmatrix}$$



- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ in V are eigenvectors of A^TA
  - $A^TA = VSU^T USV^T = VS^2 V^T$
  - "Right singular vectors"

$$A = USV^T$$

- S is a diagonal matrix (n by n) of non-negative singular values
- Typically sorted from largest to smallest
- Singular values are the non-negative square root of corresponding eigenvalues of A^TA and AA^T



$$AV = US$$

- Means each $A\boldsymbol{v}_i = s_i\boldsymbol{u}_i$
- Remember A is a linear map from row space to column space
- Here, A maps an orthonormal basis $\{\boldsymbol{v}_i\}$ in row space into an orthonormal basis $\{\boldsymbol{u}_i\}$ in column space
- Each component of u_i is the projection of a row of the data matrix A onto the vector v_i

SVD as sum of rank-1 matrices

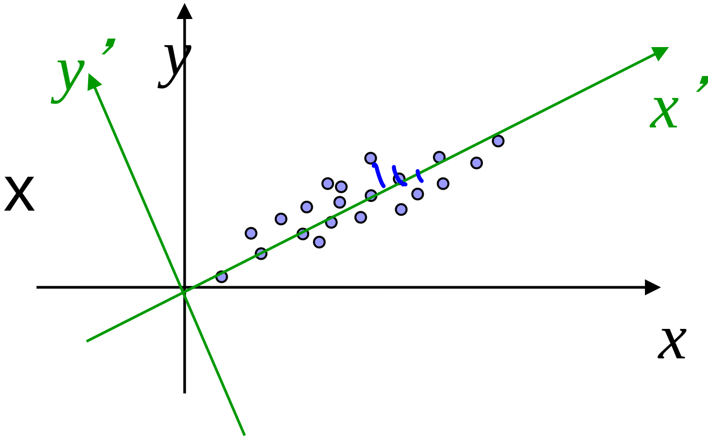
- $A = USV^T$
- $A = s_1\boldsymbol{u}_1\boldsymbol{v}_1{}^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2{}^T + \ldots + s_n\boldsymbol{u}_n\boldsymbol{v}_n{}^T$
- $s_1 \geq s_2 \geq \ldots \geq s_n \geq 0$
- What is the rank-r matrix $\hat{A}$ that best approximates A ?
  - Minimize $\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\hat{A}_{ij} - A_{ij}\right)^2$
- $\hat{A} = s_1\boldsymbol{u}_1\boldsymbol{v}_1{}^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2{}^T + \ldots + s_r\boldsymbol{u}_r\boldsymbol{v}_r{}^T$
- Very useful for matrix approximation

an outer product (uv^T) giving a matrix rather than the scalar of the inner product

LSQ approx. If r=1, this amounts to a line fit.

Geometry of SVD in row space

- A as a collection of m row vectors (points) in the row space of A
- $s_1\boldsymbol{u}_1\boldsymbol{v}_1^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2^T$ is the best rank-2 matrix approximation for A
- Geometrically: $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are the directions of the best approximating rank-2 subspace that goes through origin
- $s_1\boldsymbol{u}_1$ and $s_2\boldsymbol{u}_2$ gives coordinates for row vectors in rank-2 subspace
- $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ gives coordinates for row space basis vectors in rank-2 subspace



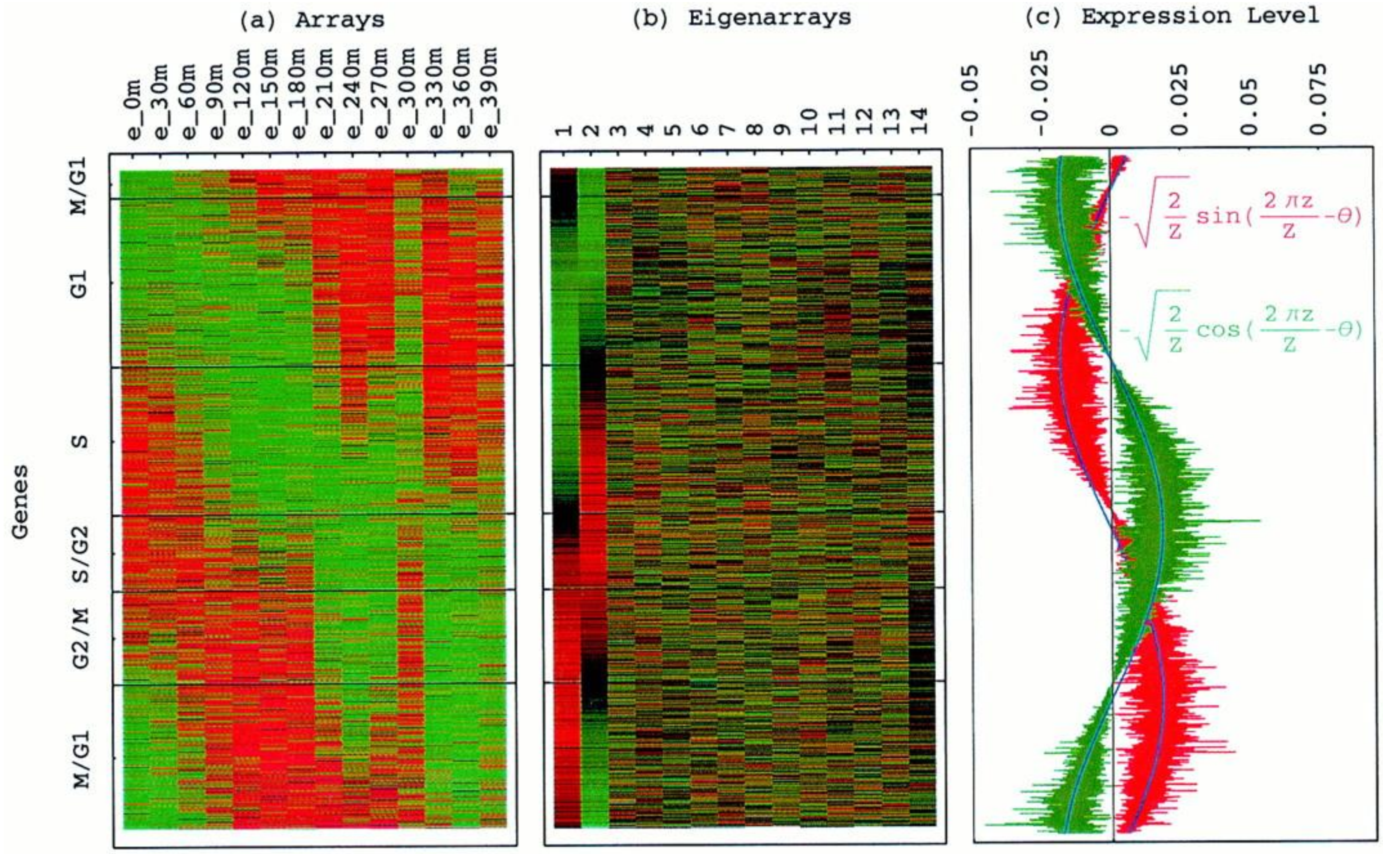
$$\boxed{A\mathbf{v_i} = s_i\mathbf{u_i}}$$

$$\boxed{I\,\mathbf{v_i} = \mathbf{v_i}}$$

Unsupervised Mining

Intuition on interpretation of SVD
in terms of genes and conditions

Genes sorted by correlation with top 2 eigengenes



Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Fig. 3. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized elutriation. (*a*) Normalized elutriation expression of the sorted 5,981 genes in the 14 arrays, showing traveling wave of expression. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z \equiv N - 1 = 5{,}980$ and phase $\theta \approx 2\pi/13$ (blue), respectively.

PNAS

Normalized elutriation expression in the subspace associated with the cell cycle

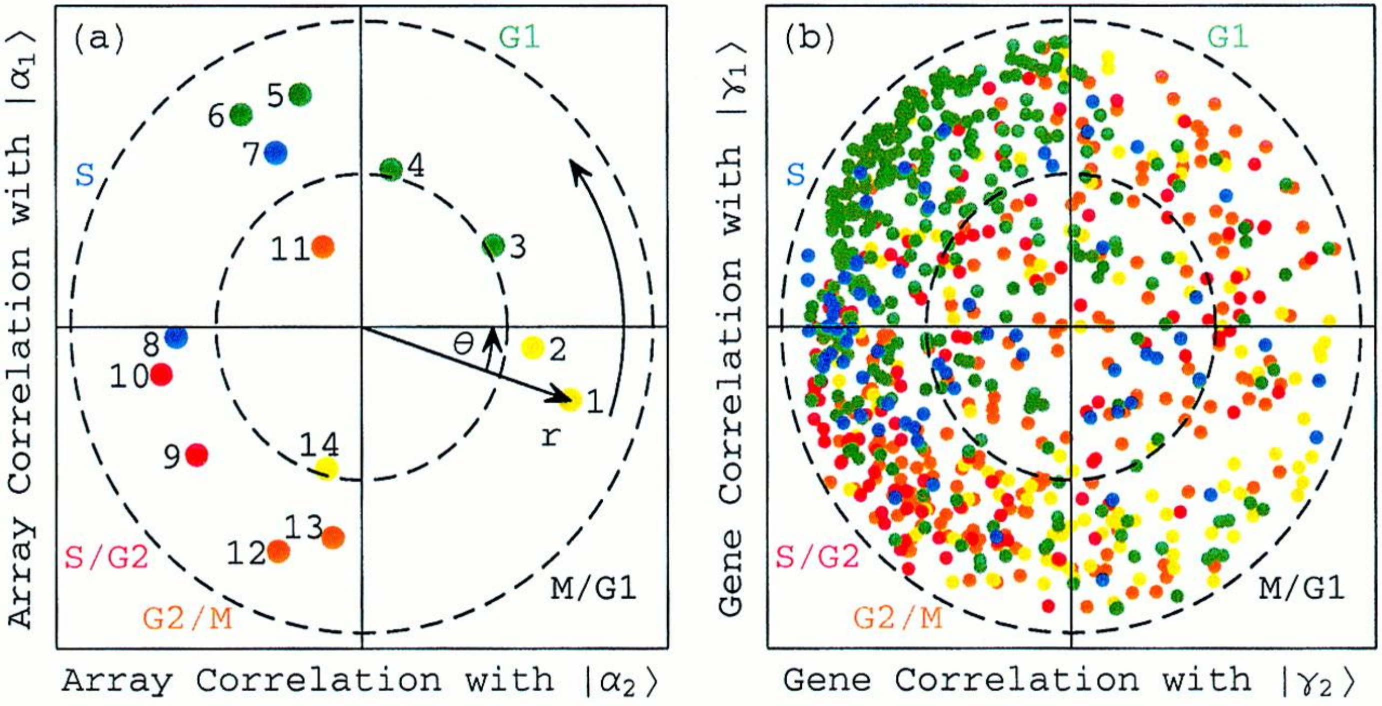


Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_N$ along the *y*-axis vs. that with $|\alpha_2\rangle_N$ along the *x*-axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 784 cell cycle regulated genes, color-coded according to the classification by Spellman *et al.* (3).

Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

References

- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert
  An Introduction to Statistical Learning: with Applications in R
  [ISLR (2nd edition)]
  https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1071614177/ + https://www.statlearning.com
  (Chapters 6.3.1 [up to section on "The Principal Components Regression Approach"] and 12.2 gives background on PCA/SVD.)

- Alter, O., Brown, P. O., & Botstein, D. (2000). PNAS, 97(18), 10101–10106.
  Singular value decomposition for genome-wide expression data processing and modeling.
  https://doi.org/10.1073/pnas.97.18.10101
  (Example discussed in class.)