

INTRO to DATA SCIENCE

LECTURE 12: DIMENSIONALITY REDUCTION

I. DIMENSIONALITY REDUCTION

II. SINGULAR VALUE DECOMPOSITION

III. PRINCIPAL COMPONENTS ANALYSIS

EXERCISE:

IV. DIMENSIONALITY REDUCTION IN SCIKIT-LEARN

INTRO TO DATA SCIENCE

K-MEANS CLUSTERING

Friday, April 4, 14

- 1) *choose k initial centroids (note that k is an input)*
- 2) *for each point:*
 - *find distance to each centroid*
 - *assign point to nearest centroid*
- 3) *recalculate centroid positions*
- 4) *repeat steps 2-3 until stopping criteria met*

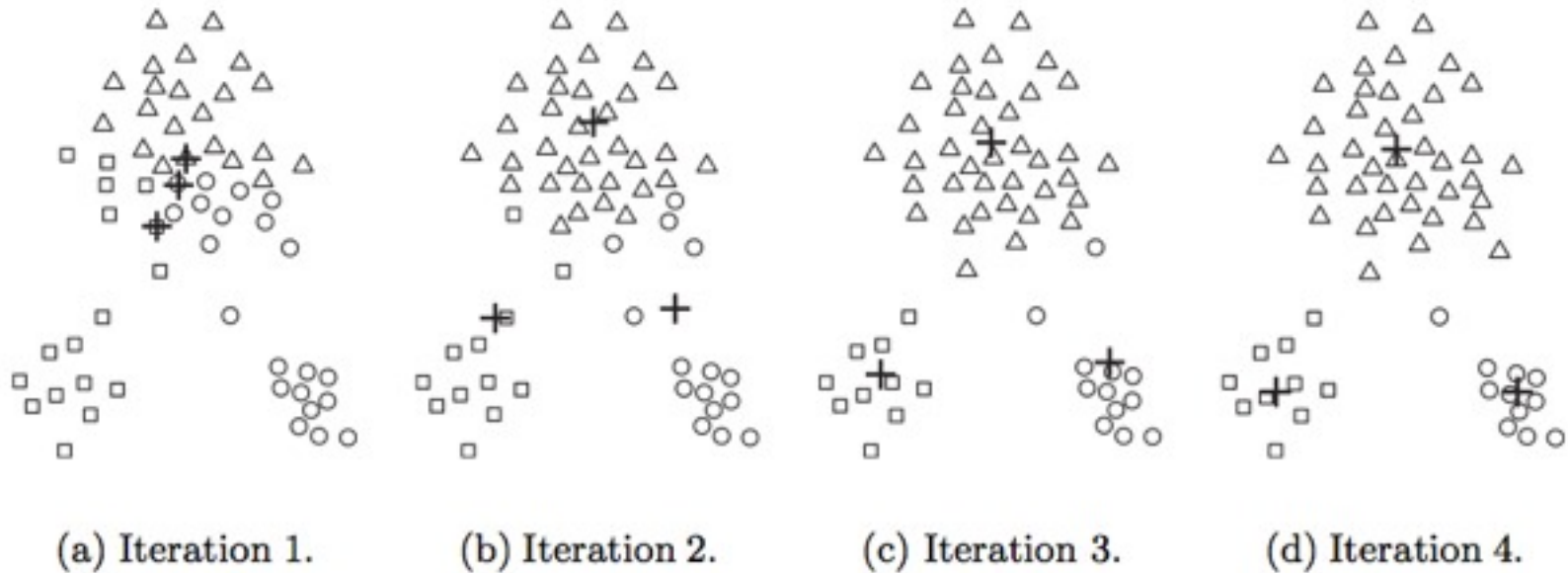


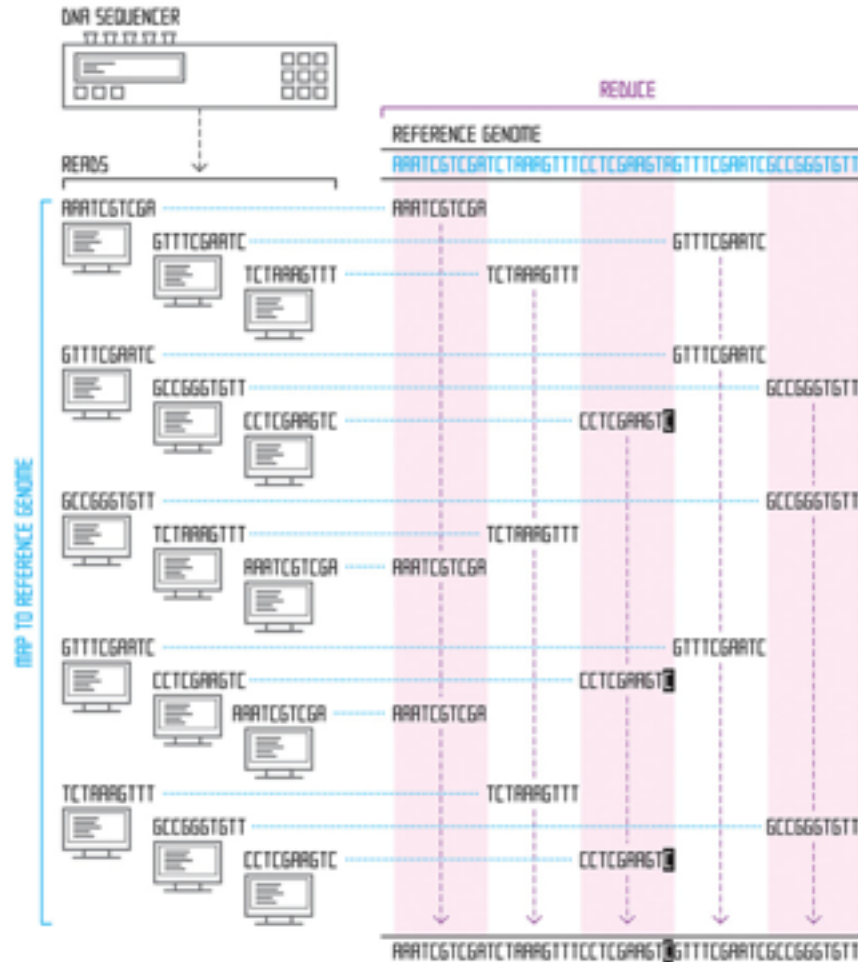
Figure 8.3. Using the K-means algorithm to find three clusters in sample data.

source: <http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf>

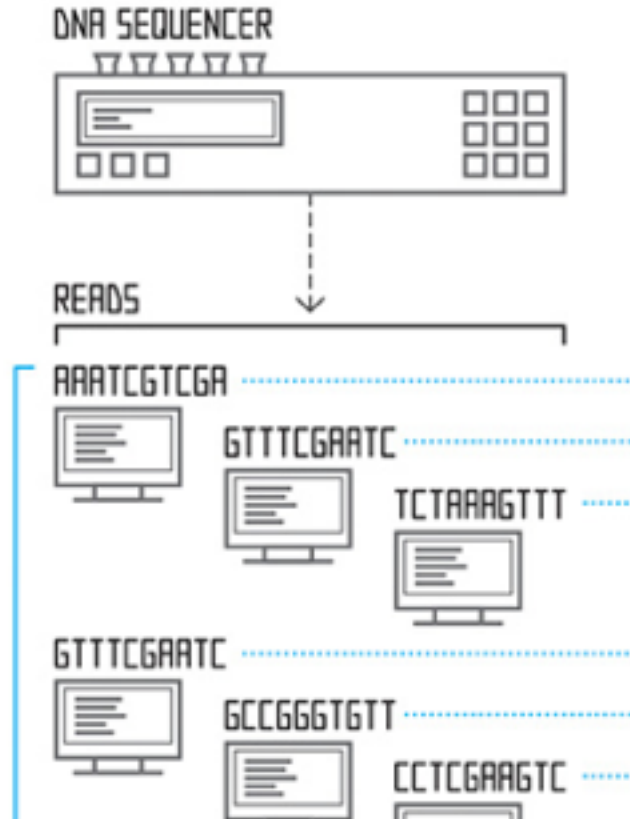
INTRO TO DATA SCIENCE

GENOMICS DATA PROCESSING

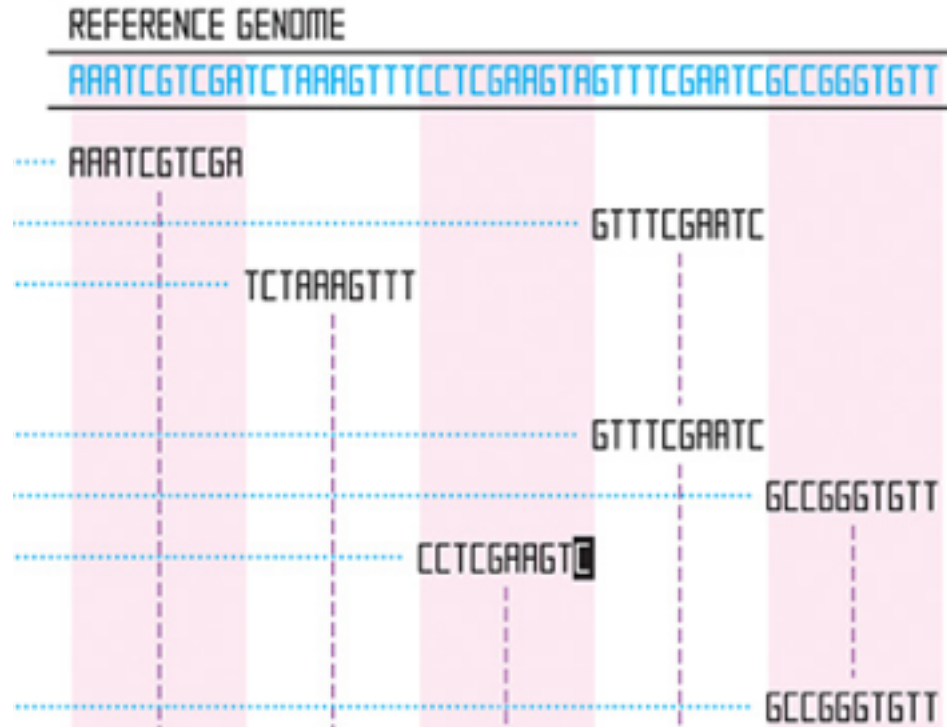
Friday, April 4, 14



Source : <http://spectrum.ieee.org/biomedical/devices/the-dna-data-deluge>



Source : <http://spectrum.ieee.org/biomedical/devices/the-dna-data-deluge>



Source : <http://spectrum.ieee.org/biomedical/devices/the-dna-data-deluge>





- **FASTQ** files contain the raw read sequences and a quality score



- **FASTQ** files contain the raw read sequences and a quality score
- **BAM** files contain read sequences aligned to some reference genome



- **FASTQ** files contain the raw read sequences and a quality score
- **BAM** files contain read sequences aligned to some reference genome
- **VCF** files contain the suspected differences (mutations) from the reference genome

Variant Calling is the process of deciding where a sample differs from the reference genome based on the sequence/read data

Variant Calling is the process of deciding where a sample differs from the reference genome based on the sequence/read data

chr12	25398284	.	C	T	.	.	GE=KRAS
chr6	159653970	.	G	A	.	.	GE=FNDC1

Variant Calling is the process of deciding where a sample differs from the reference genome based on the sequence/read data

What is a **disease**: genetic variants/mutations that give rise to irregular biological process

Variant Calling is the process of deciding where a sample differs from the reference genome based on the sequence/read data

What is a **cancer**: genetic variants/mutations that give rise to irregular biological process ... mainly proliferation of cell growth

Variant Calling is the process of deciding where a sample differs from the reference genome based on the sequence/read data

chr12	25392284	.	C	T	.	.	GE=KRAS
chr6	159653870	.	G	A	.	.	GE=FNDC1

KRAS is a common gene affected in pancreatic cancer

I. DIMENSIONALITY REDUCTION

Q: What is dimensionality reduction?

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

In general, the idea is to regard the dataset as a matrix and to decompose the matrix into simpler, meaningful pieces.

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

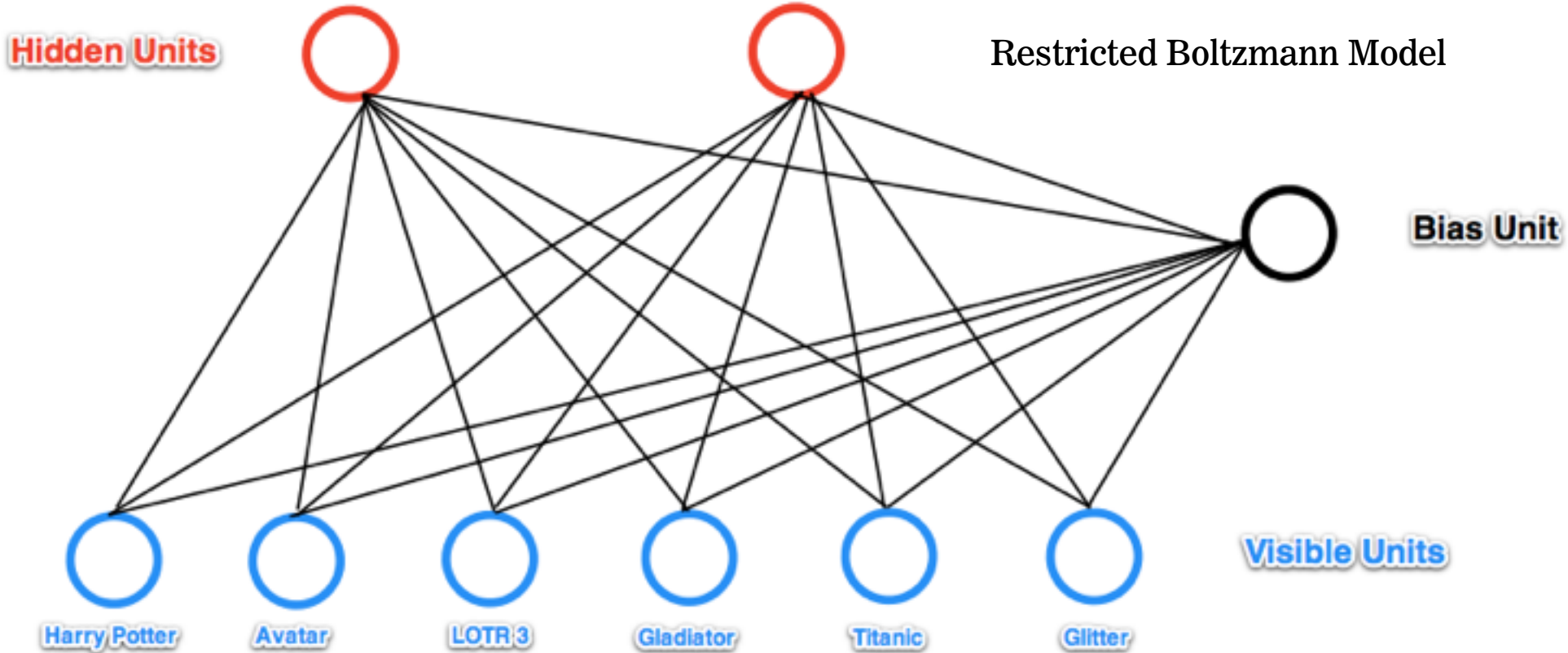
In general, the idea is to regard the dataset as a matrix and to decompose the matrix into simpler, meaningful pieces.

Dimensionality reduction is frequently performed as a pre-processing step before another learning algorithm is applied.

Q: What are the motivations for dimensionality reduction?

Q: What are the motivations for dimensionality reduction?

The number of features in our dataset can be difficult to manage, or even misleading (eg, if the relationships are actually simpler than they appear).



Fantasy? Oscars?

Restricted Boltzmann Model

Hidden Units

Bias Unit

Visible Units



Q: What is the goal of dimensionality reduction?

Q: What is the goal of dimensionality reduction?

- reduce computational expense*
- reduce susceptibility to overfitting*
- reduce noise in the dataset*
- enhance our intuition*

Q: How is dimensionality reduction performed?

Q: How is dimensionality reduction performed?

A: There are two approaches: feature selection and feature extraction.

Q: How is dimensionality reduction performed?

A: There are two approaches: feature selection and feature extraction.

feature selection – *selecting a subset of features using an external criterion (filter) or the learning algo accuracy itself (wrapper)*

feature extraction – *mapping the features to a lower dimensional space*

Feature selection is important, but typically when people say dimensionality reduction, they are referring to feature extraction.

The goal of feature extraction is to create a new set of coordinates that simplify the representation of the data.

Q: What are some applications of dimensionality reduction?

Q: What are some applications of dimensionality reduction?

- topic models (document clustering)*
- image recognition/computer vision*
- recommender systems*

II. SINGULAR VALUE DECOMPOSITION

Consider a matrix A with n rows and d features.

Consider a matrix A with n rows and d features.

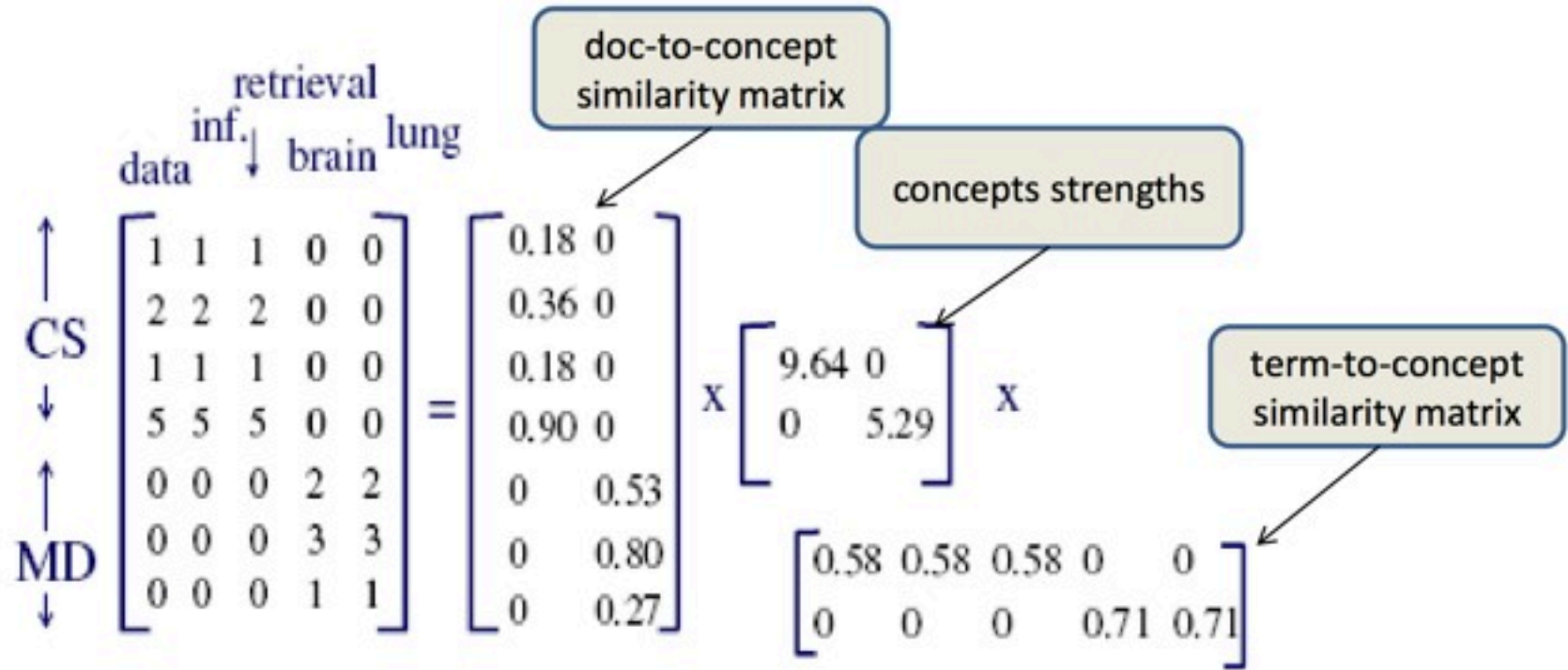
The singular value decomposition of A is given by:

$$A = U \Sigma V^T$$

Consider a matrix A with n rows and d features.

The singular value decomposition of A is given by:

$$\underset{(n \times d)}{A} = \underset{(n \times k)}{U} \underset{(k \times k)}{\Sigma} \underset{(k \times d)}{V^T}$$



Consider a matrix A with n rows and d features.

The singular value decomposition of A is given by:

$$\underset{(n \times d)}{A} = \underset{(n \times k)}{U} \underset{(k \times k)}{\Sigma} \underset{(k \times d)}{V^T}$$

st. U , V are orthogonal matrices and Σ is a diagonal matrix.

Consider a matrix A with n rows and d features.

The singular value decomposition of A is given by:

$$\underset{(n \times d)}{A} = \underset{(n \times k)}{U} \underset{(k \times k)}{\Sigma} \underset{(k \times d)}{V^T}$$

st. U, V are orthogonal matrices and Σ is a diagonal matrix.

$$\rightarrow UU^T = I_n, \quad VV^T = I_d \qquad \rightarrow \Sigma_{ij} = 0 \quad (i \neq j)$$

The singular value decomposition of A is given by:

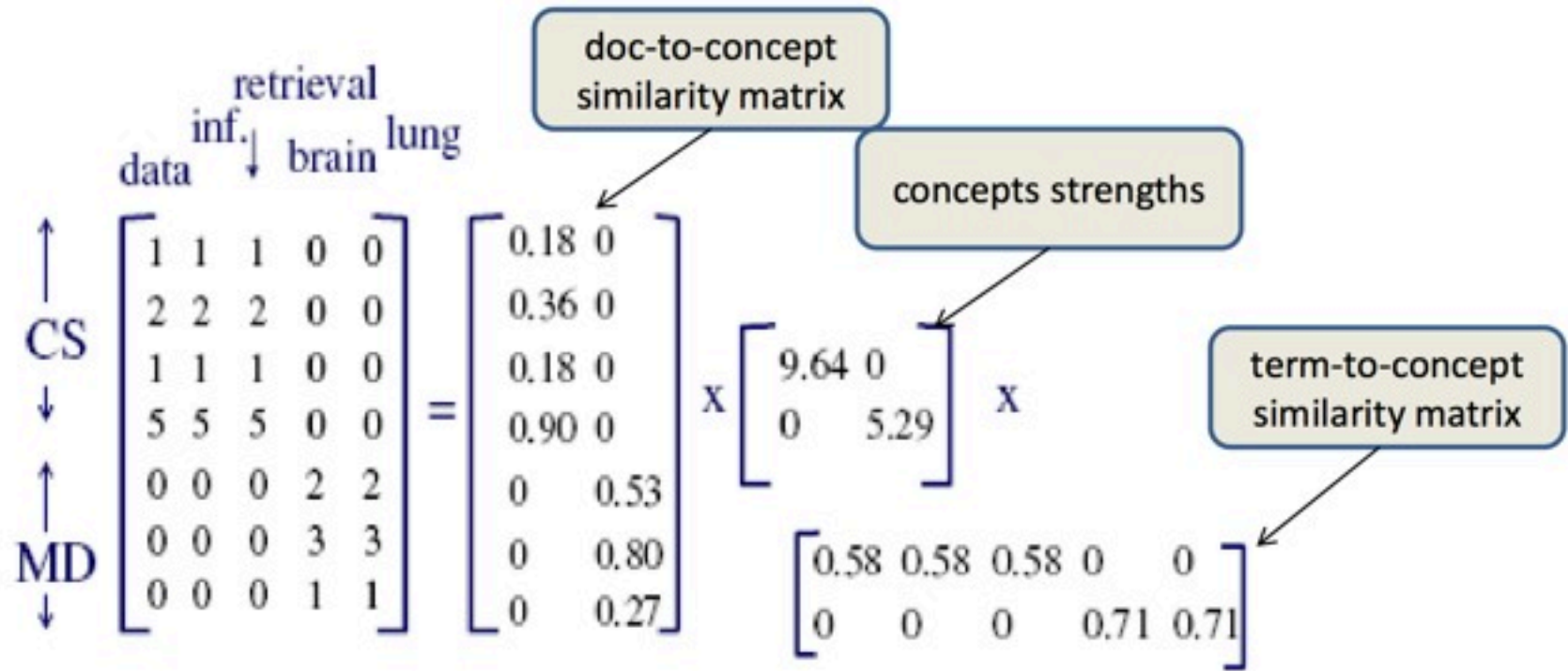
$$\begin{matrix} A & = & U & \Sigma & V^T \\ (n \times d) & & (n \times k) & (k \times k) & (k \times d) \end{matrix}$$

*The nonzero entries of Σ are the **singular values** of A . These are real, nonnegative, and rank-ordered (decreasing from left to right).*

Q: How do you interpret the SVD?

Q: How do you interpret the SVD?

A: Recall that given a set of n points in d -dimensional space (eg, a matrix A), we want to find the best $k < d$ dimensional subspace to represent the data.



III. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

Principal component analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

This procedure produces a new basis, each of whose components retain as much variance from the original data as possible.

Principal component analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

This procedure produces a new basis, each of whose components retain as much variance from the original data as possible.

*The PCA of a matrix A boils down to the **eigenvalue decomposition** of the **covariance matrix** of A .*

The covariance matrix C of a matrix A is always square:

$$C = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

off-diagonal elements C_{ij} give the covariance between X_i, X_j ($i \neq j$)

diagonal elements C_{ii} give the variance of X_i

The eigenvalue decomposition of a square matrix A is given by:

$$A = Q \Lambda Q^{-1}$$

The eigenvalue decomposition of a square matrix A is given by:

$$A = Q \Lambda Q^{-1}$$

*The columns of Q are the **eigenvectors** of A , and the values in Λ are the associated **eigenvalues** of A .*

The eigenvalue decomposition of a square matrix A is given by:

$$A = Q \Lambda Q^{-1}$$

*The columns of Q are the **eigenvectors** of A , and the values in Λ are the associated **eigenvalues** of A .*

For an eigenvector v of A and its eigenvalue λ , we have the important relation:

$$Av = \lambda v$$

The eigenvalue decomposition of a square matrix A is given by:

$$A = Q \Lambda Q^{-1}$$

*The columns of Q are the **eigenvectors** of A , and the values on the diagonal of Λ are the associated **eigenvalues** of A .*

NOTE

This relationship defines what it means to be an eigenvector of A .

For an eigenvector v of A and its eigenvalue λ , we have the important relation:

$$Av = \lambda v$$

The eigenvectors form a basis of the vector space on which A acts (eg, they are orthogonal).

The eigenvectors form a basis of the vector space on which A acts (eg, they are orthogonal).

Furthermore the basis elements are ordered by their eigenvalues (from largest to smallest), and these eigenvalues represent the amount of variance explained by each basis element.

The singular value decomposition of A is given by:

$$\underset{(n \times d)}{A} = \underset{(n \times k)}{U} \underset{(k \times k)}{\Sigma} \underset{(k \times d)}{V^T}$$

*The nonzero entries of Σ are the **singular values** of A . These are real, nonnegative, and rank-ordered (decreasing from left to right).*

In any case, the key difficulties with dimensionality reduction are time/space complexity, randomness (eg different results for different runs), and selecting the number of dimensions in the lower-dim subspace.

