GENERAL ASSEMBLY

# INTRO TO DATA SCIENCE

## LECTURE 15: ADVANCED UNSUPERVISED LEARNING

# AGENDA

I. LDA

II. LDA EXERCISE WITH PYTHON AND GENSIM

# REVIEW: DIMENSIONALITY REDUCTION

# 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.*

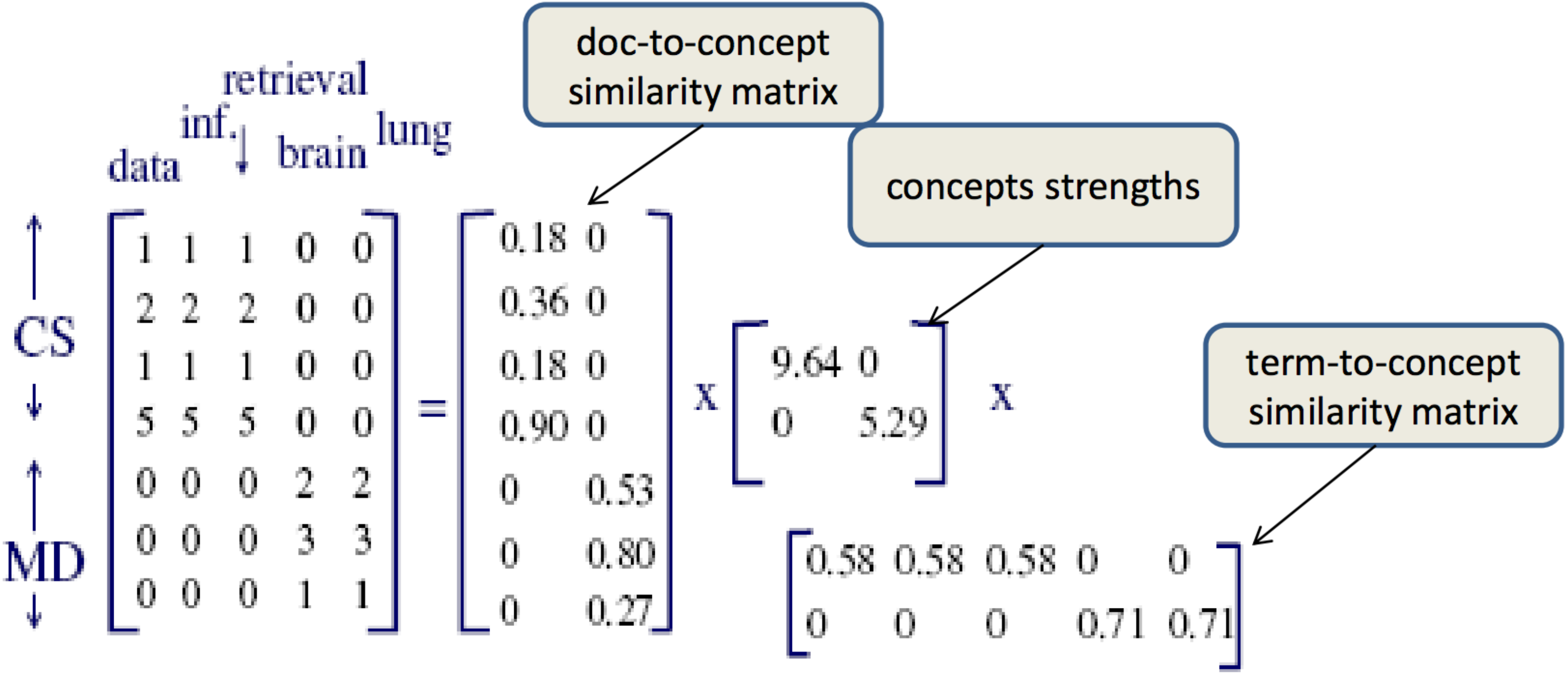
*In general, the idea is to regard the dataset as a matrix and to decompose the matrix into simpler, meaningful pieces.*

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

INTRO TO DATA SCIENCE

# REVIEW: SINGULAR VALUE DECOMPOSITION

# SINGULAR VALUE DECOMPOSITION



$$\begin{array}{c|ccccc} & \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \\ \hline \text{CS} & 1 & 1 & 1 & 0 & 0 \\ & 2 & 2 & 2 & 0 & 0 \\ & 1 & 1 & 1 & 0 & 0 \\ & 5 & 5 & 5 & 0 & 0 \\ \text{MD} & 0 & 0 & 0 & 2 & 2 \\ & 0 & 0 & 0 & 3 & 3 \\ & 0 & 0 & 0 & 1 & 1 \end{array} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \mathrm{x} \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \mathrm{x} \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

doc-to-concept similarity matrix

concepts strengths

term-to-concept similarity matrix

# SINGULAR VALUE DECOMPOSITION

*Consider a matrix* A *with* n *rows and* d *features.*

*The* **singular value decomposition** *of* A *is given by:*

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

# SINGULAR VALUE DECOMPOSITION

*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* $\Sigma$ *are the* **singular values** *of* A. *These are real, nonnegative, and rank-ordered (decreasing from left to right).*

# I. LATENT DIRICHLET ALLOCATION