Tools for large graph mining

WWW 2008 tutorial

Part 3: Matrix tools for graph mining

Jure Leskovec and Christos Faloutsos

Machine Learning Department

Carnegie Mellon

Joint work with: Deepay Chakrabarti, Tamara Kolda and Jimeng Sun.

Tutorial outline

- Part 1: Structure and models for networks
  - What are properties of large graphs?
  - How do we model them?
- Part 2: Dynamics of networks
  - Diffusion and cascading behavior
  - How do viruses and information propagate?
- Part 3: Matrix tools for mining graphs
  - Singular value decomposition (SVD)
  - Random walks
- Part 4: Case studies
  - 240 million MSN instant messenger network
  - Graph projections: how does the web look like

About part 3

- Introduce **matrix and tensor tools** through **real mining applications**
- **Goal:** find **patterns, rules, clusters, outliers, …**
  - in matrices and
  - in tensors

What is this part about?

- Connection of matrix tools and networks
- Matrix tools
  - Singular Value Decomposition (SVD)
  - Principal Component Analysis (PCA)
  - Webpage ranking algorithms: HITS, PageRank
  - CUR decomposition
  - Co-clustering (in part 4 of the tutorial)
- Tensor tools
  - Tucker decomposition
- Applications

Why matrices? Examples

- Social networks
- Documents and terms
- Authors and terms

	John	Peter	Mary	Nick	…
John	0	11	22	55	…
Peter	5	0	6	7	…
Mary	…	…	…	…	…
Nick	…	…	…	…	…
…	…	…	…	…	…

Why tensors? Example

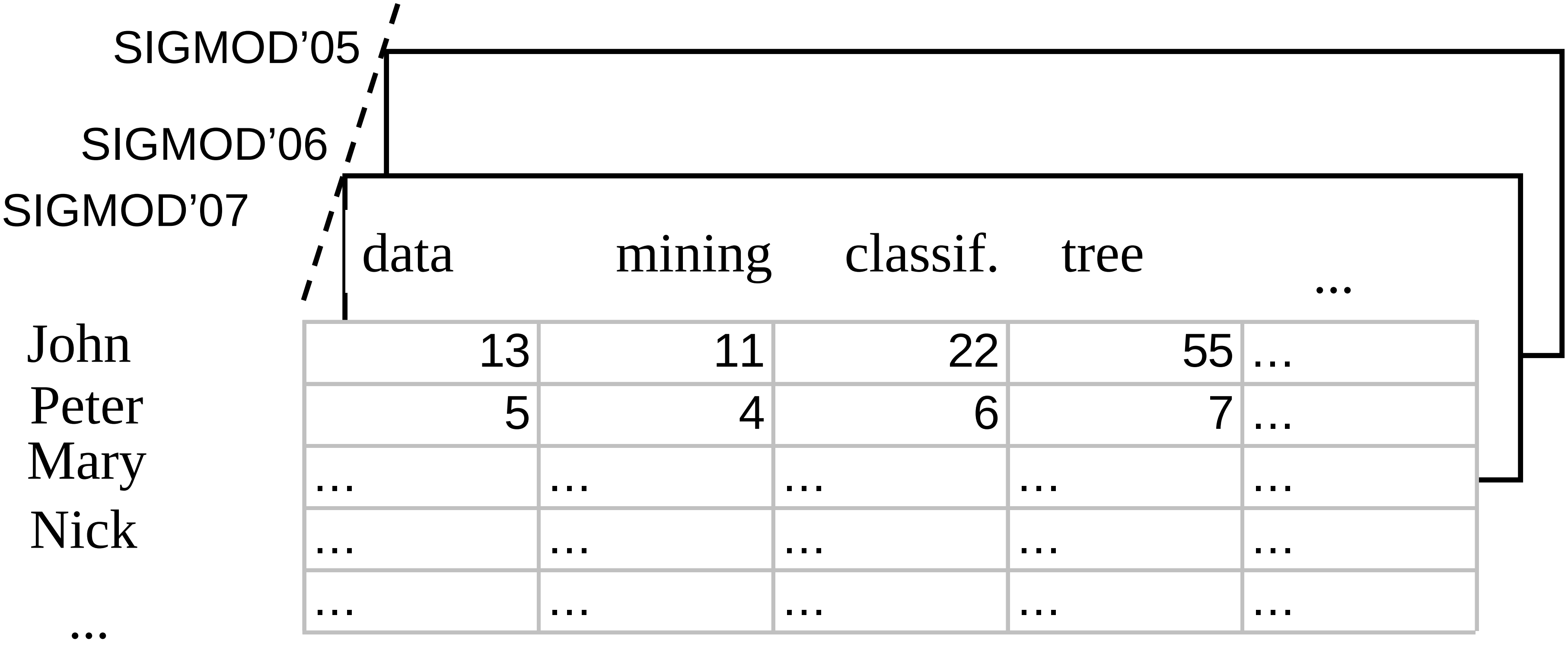
- Tensor:
  - n-dimensional generalization of matrix

SIGMOD’07

	data	mining	classif.	tree	…
John	13	11	22	55	…
Peter	5	4	6	7	…
Mary	…	…	…	…	…
Nick	…	…	…	…	…
…	…	…	…	…	…

Why tensors? Example

- Tensor:
  - n-dimensional generalization of matrix



SIGMOD’05

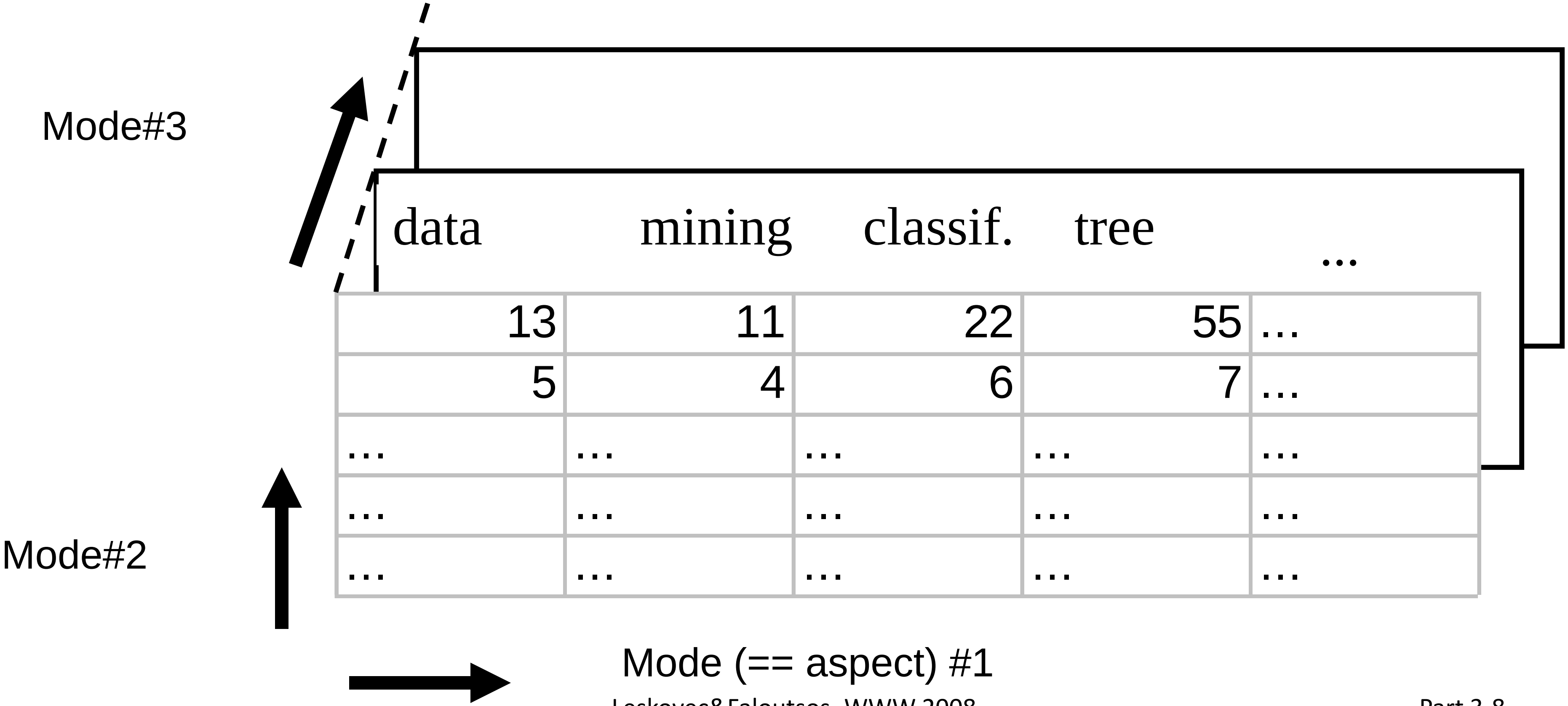
SIGMOD’06

SIGMOD’07

	data	mining	classif.	tree	...
John	13	11	22	55	…
Peter	5	4	6	7	…
Mary	…	…	…	…	…
Nick	…	…	…	…	…
…	…	…	…	…	…

Tensors are useful for 3 or more modes

Terminology: ‘mode’ (or ‘aspect’):



data	mining	classif.	tree	...
13	11	22	55	…
5	4	6	7	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

Mode#3

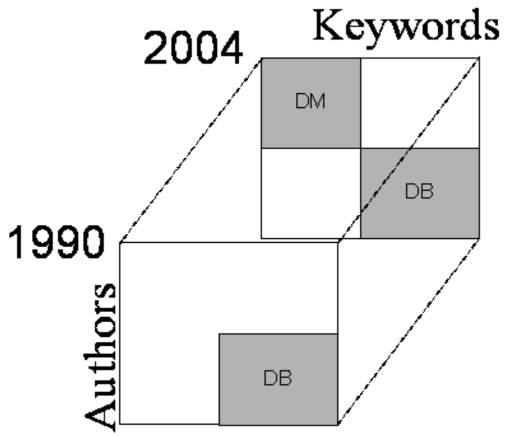
Mode#2

Mode (== aspect) #1

Motivating applications

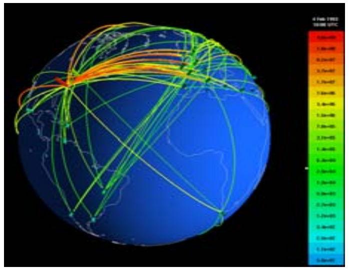
- Why matrices are important?
- Why tensors are useful?
    - P1: social networks
    - P2: web & text mining
    - P3: network forensics
    - P4: sensor networks

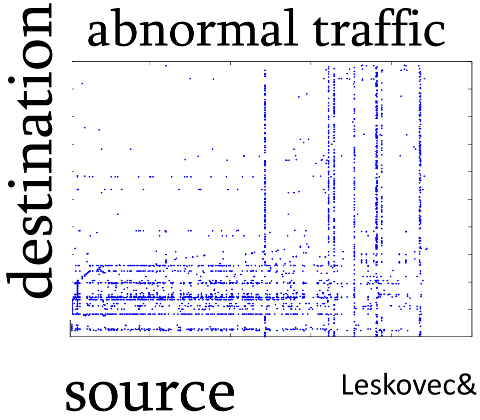
Social networks

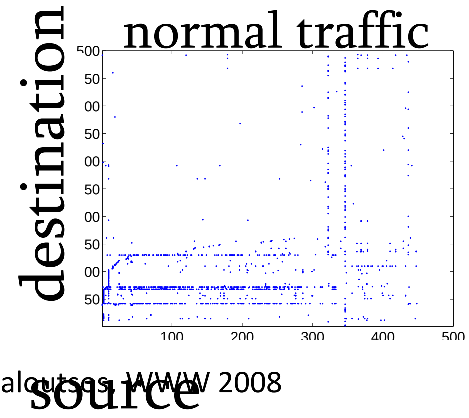


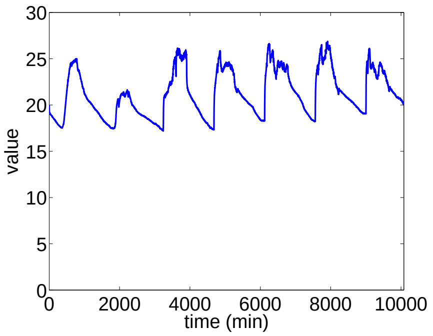
Sensor networks

Network forensics





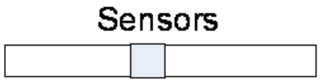
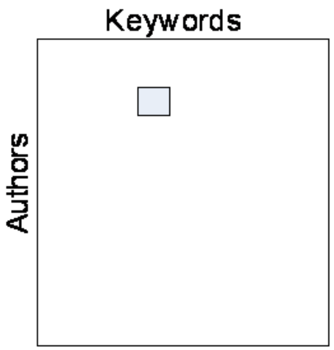
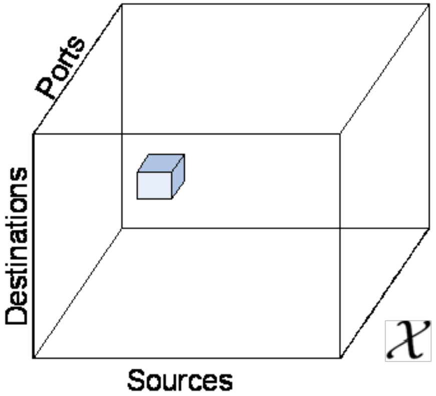




Temperature

Static Data model

- Tensor
  - Formally, $\mathcal{X} \in \mathbf{R}^{N_1 \times \ldots \times N_M}$
  - Generalization of matrices
  - Represented as multi-array, (~ data cube).

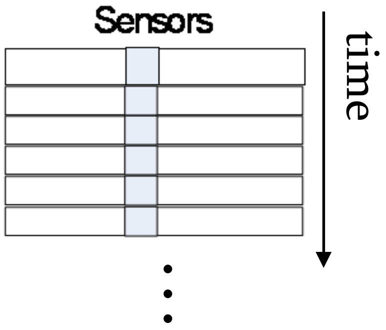
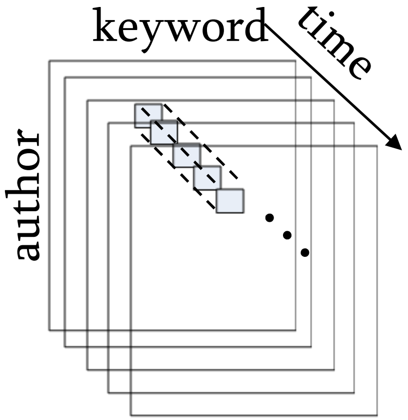
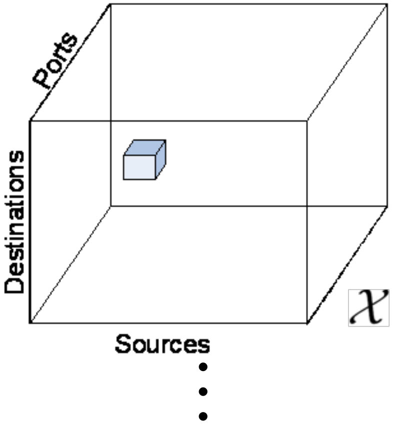
Order	1st	2^{nd}	3^{rd}
Correspondence	Vector	Matrix	3D array
Example			

Dynamic Data model

- Tensor Streams
  - A sequence of Mth order tensors

$$\mathcal{X}_1 \ldots \mathcal{X}_t \text{ where } \mathcal{X}_i \in \mathbf{R}^{N_1 \times \ldots \times N_M}$$

t is increasing over time

Order	1st	2nd	3rd
Correspondence	Multiple streams	Time evolving graphs	3D arrays
Example			

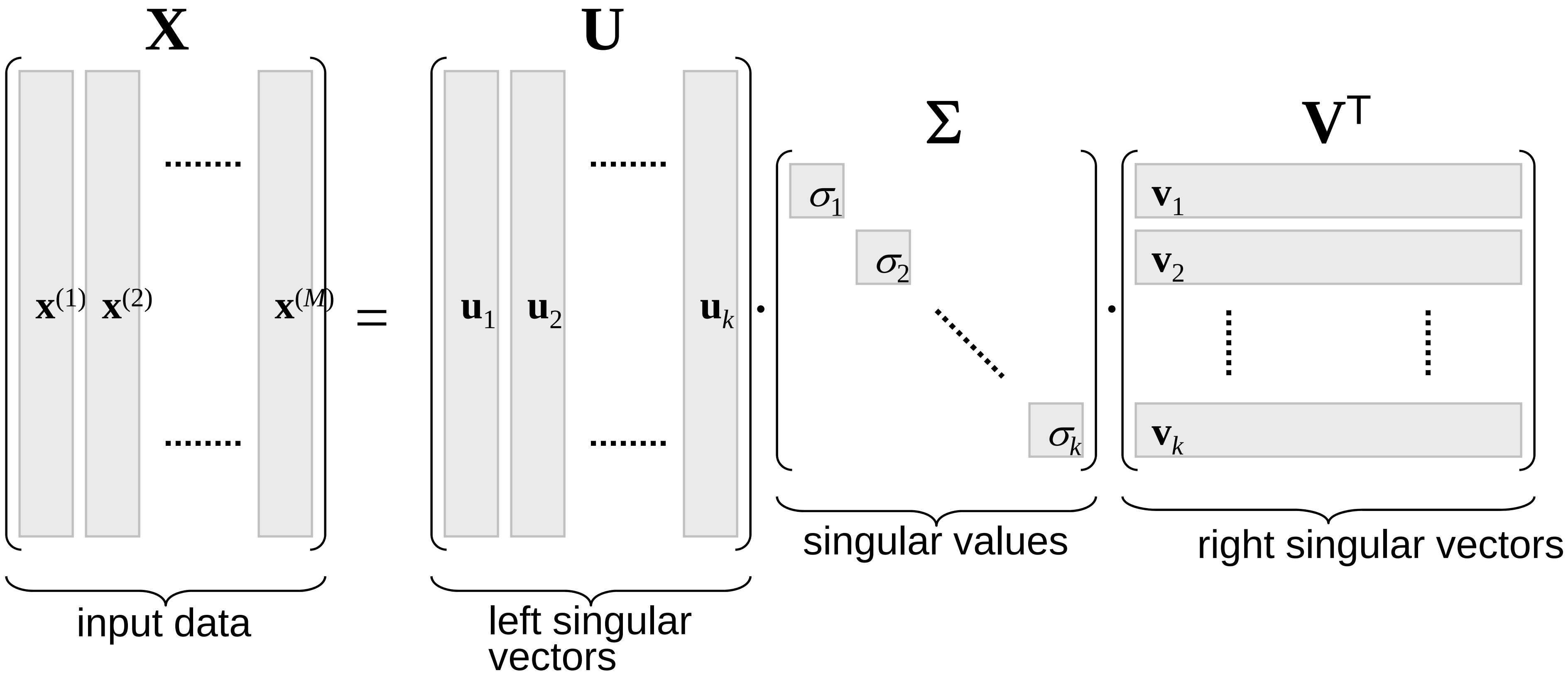
SVD: Examples of Matrices

- Example/Intuition: Documents and terms
- Find patterns, groups, concepts

	data	mining	classif.	tree	…
Paper#1	13	11	22	55	…
Paper#2	5	4	6	7	…
Paper#3	…	…	…	…	…
Paper#4	…	…	…	…	…
…	…	…	…	…	…

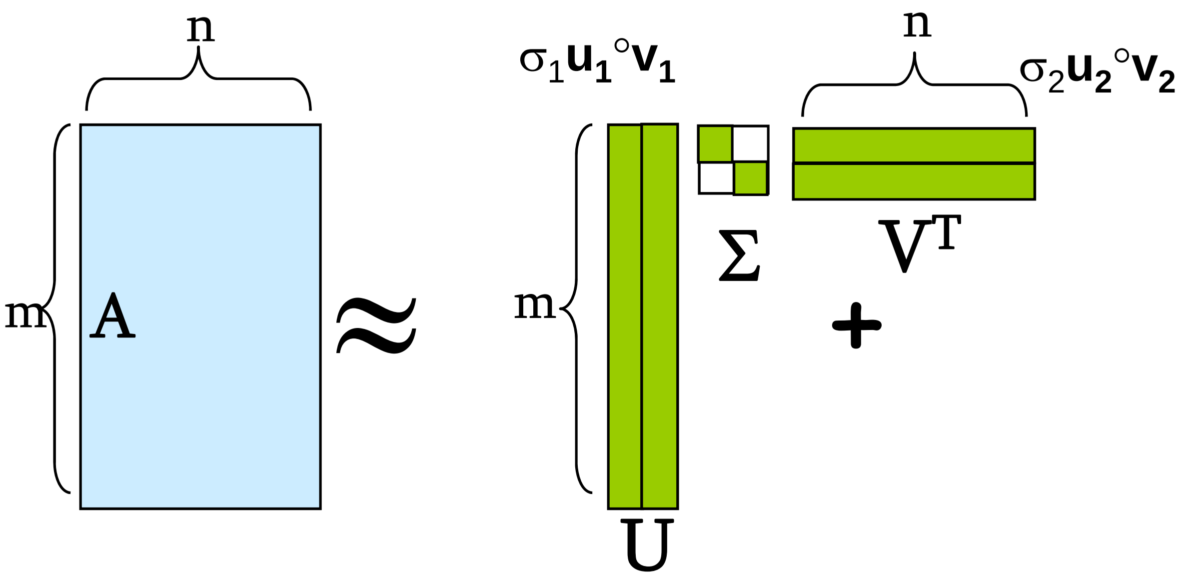
Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}$$



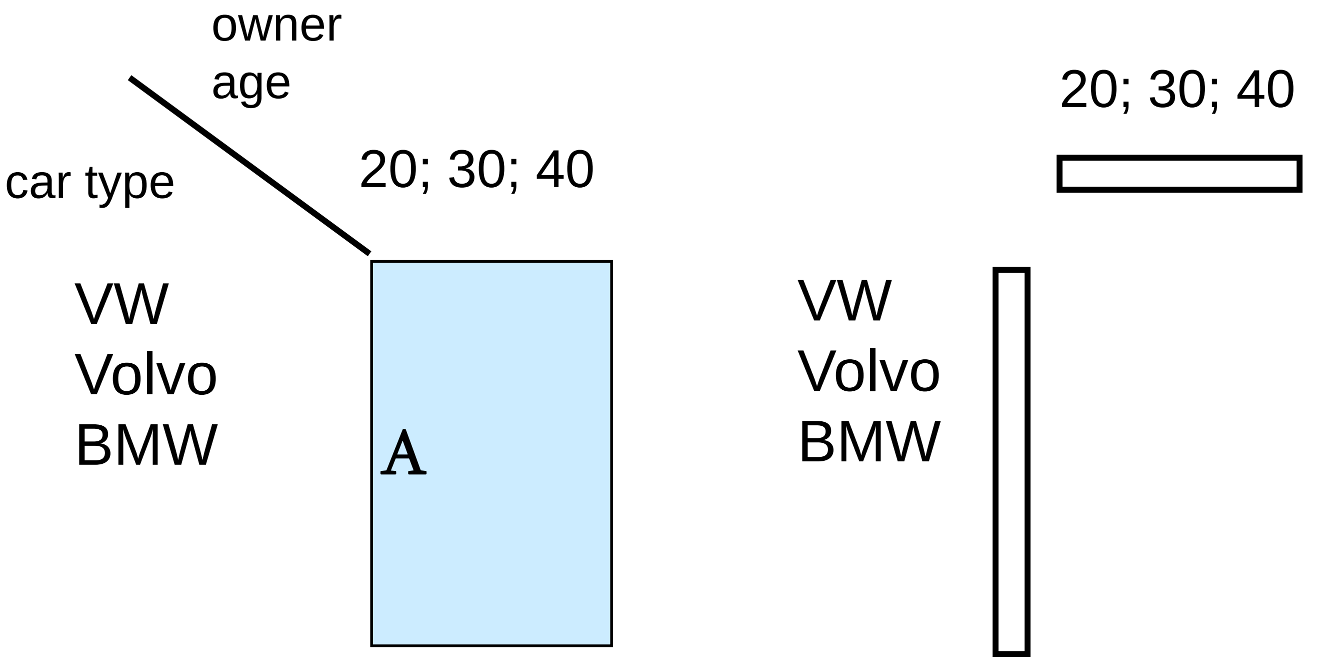
SVD as spectral decomposition

$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2 and Frobenius
- SVD only works for static matrices (a single 2nd order tensor)

Vector outer product – intuition:



owner
age

car type

20; 30; 40

VW
Volvo
BMW

A

20; 30; 40

VW
Volvo
BMW

2-d histogram

1-d histograms +
independence assumption

SVD - Example

- $\mathbf{A} = \mathbf{U}\, \Sigma\, \mathbf{V}^T$ - example:

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

SVD - Example

- **A = U Σ V**^{T} - example:

$$\begin{array}{c} \\ \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \\ \end{array} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \\ \end{array}$$

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

SVD - Example

- **A = U Σ V**^{T} - example:

doc-to-concept
similarity matrix

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

SVD - Example

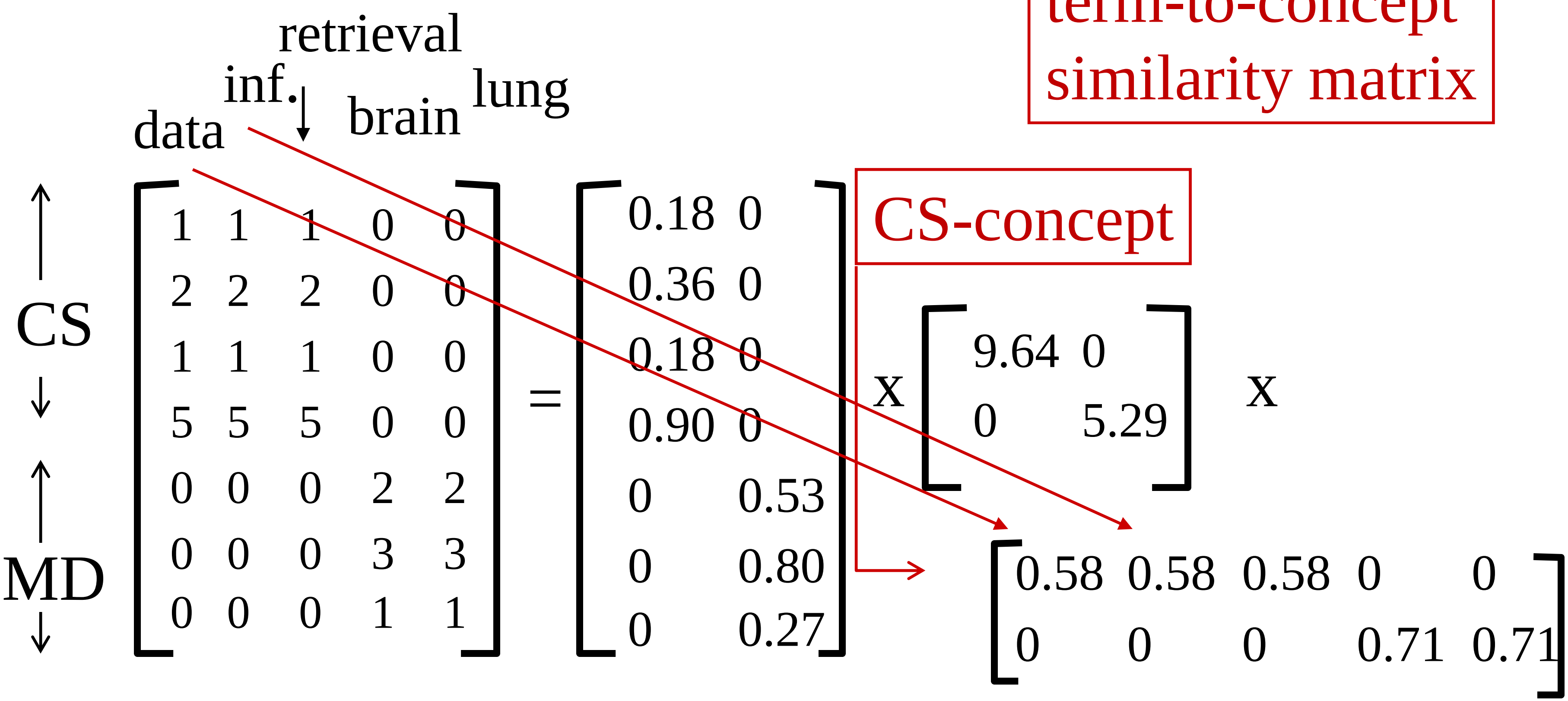
- $\mathbf{A} = \mathbf{U}\, \Sigma\, \mathbf{V}^T$ - example:

‘strength’ of CS-concept

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

SVD - Example

- **A** = **U** **Σ** **V**^{T} - example:



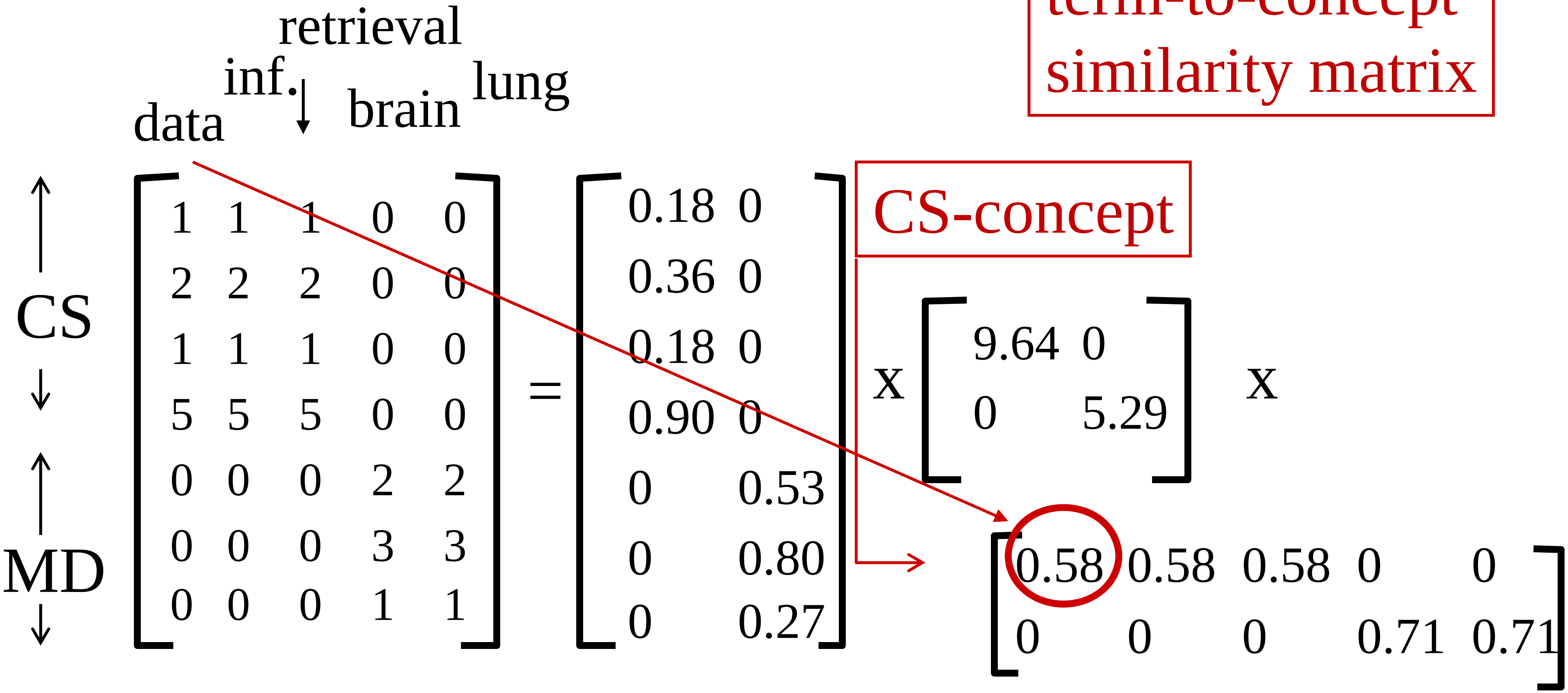
term-to-concept similarity matrix

CS-concept

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

SVD - Example

- $\mathbf{A} = \mathbf{U}\, \Sigma\, \mathbf{V}^T$ - example:



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

$$\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

term-to-concept similarity matrix

CS-concept

SVD - Interpretation

‘documents’, ‘terms’ and ‘concepts’:

Q: if **A** is the document-to-term matrix, what is $\mathbf{A}^T \mathbf{A}$?

A: term-to-term ([m x m]) similarity matrix

Q: $\mathbf{A} \mathbf{A}^T$?

A: document-to-document ([n x n]) similarity matrix

SVD properties

- **V** are the eigenvectors of the *covariance matrix* $\mathbf{A}^T\mathbf{A}$

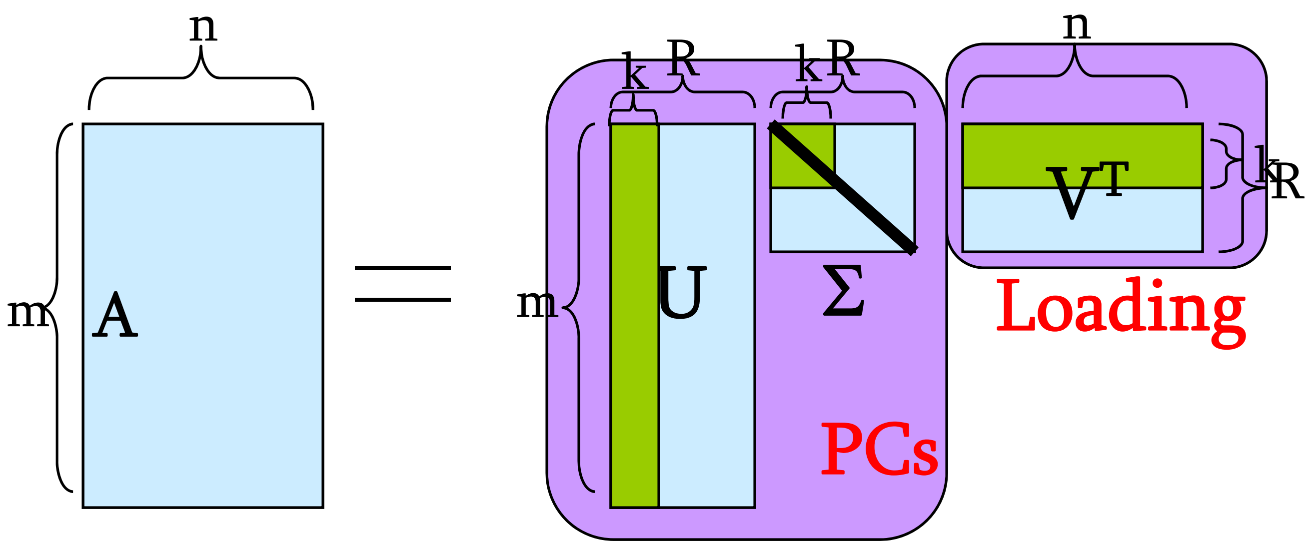
- **U** are the eigenvectors of the *Gram (inner-product) matrix* $\mathbf{A}\mathbf{A}^T$

Further reading:

1. Ian T. Jolliffe, *Principal Component Analysis* (2nd ed), Springer, 2002.
2. Gilbert Strang, *Linear Algebra and Its Applications* (4th ed), Brooks Cole, 2005.

Principal Component Analysis (PCA)

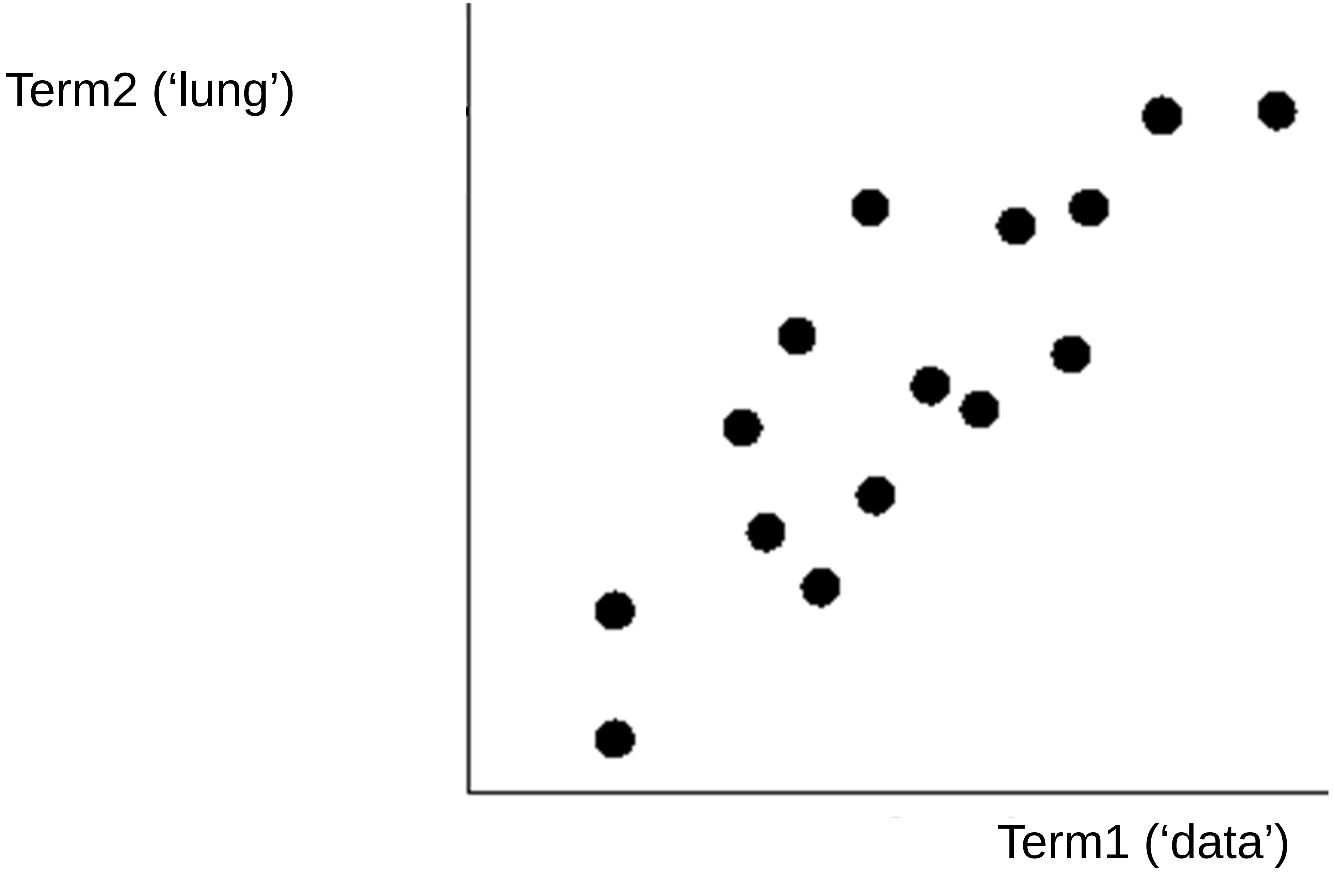
- SVD $$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$$



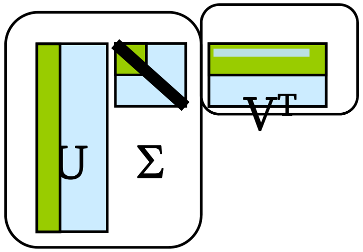
- PCA is an important application of SVD
- Note that U and V are dense and may have negative entries

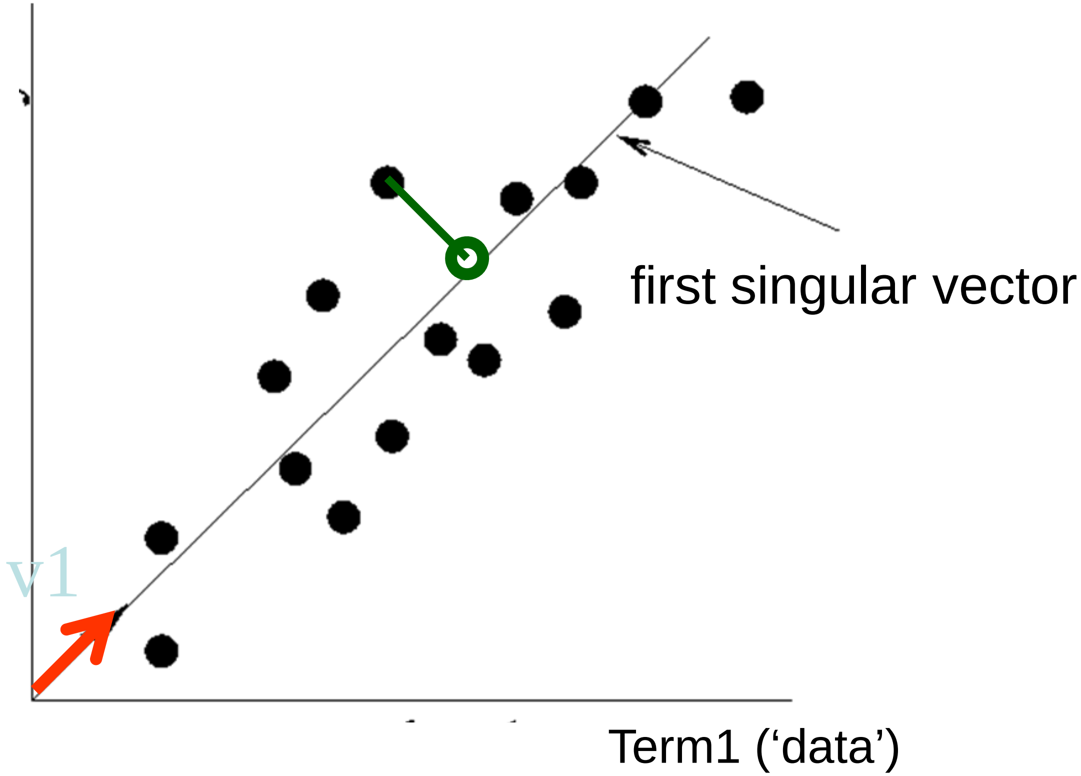
PCA interpretation

- best axis to project on: (‘best’ = min sum of squares of projection errors)



PCA - interpretation





PCA projects points
Onto the “best” axis

- minimum RMS error

Kleinberg’s algorithm HITS

- Problem definition:
  - given the web and a query
  - find the most ‘authoritative’ web pages for this query

Step 0: find all pages containing the query terms

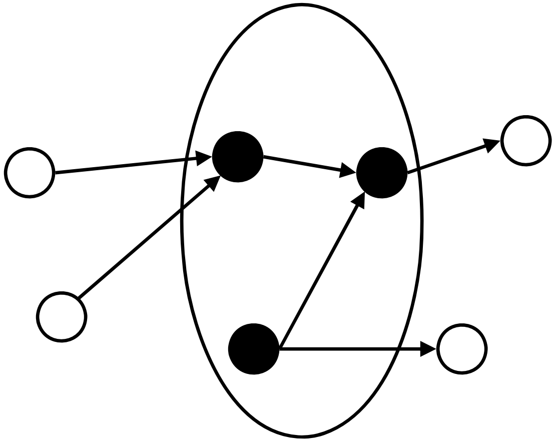
Step 1: expand by one move forward and backward

Further reading:

1. J. Kleinberg. Authoritative sources in a hyperlinked environment. SODA 1998

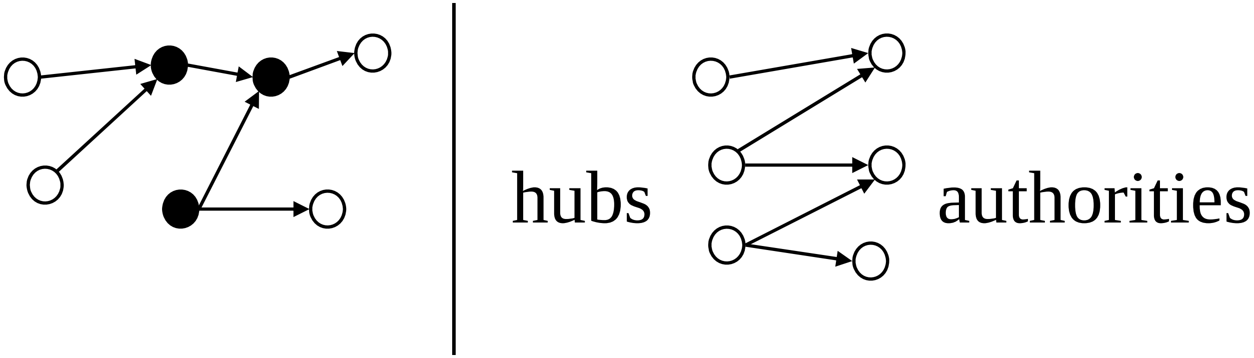
Kleinberg’s algorithm HITS

- Step 1: expand by one move forward and backward



Kleinberg’s algorithm HITS

- on the resulting graph, give high score (= ‘authorities’) to nodes that many important nodes point to
- give high importance score (‘hubs’) to nodes that point to good ‘authorities’



Kleinberg’s algorithm HITS

observations

- recursive definition!
- each node (say, ‘*i*’-th node) has both an authoritativeness score a_i and a hubness score h_i

Kleinberg’s algorithm: HITS

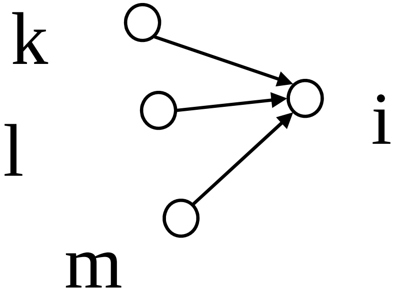
Let **A** be the adjacency matrix:

the (*i,j*) entry is 1 if the edge from *i* to *j* exists

Let **h** and **a** be [n x 1] vectors with the ‘hubness’ and ‘authoritativiness’ scores.

Then:

Kleinberg's algorithm: HITS



Then:

$$a_i = h_k + h_l + h_m$$

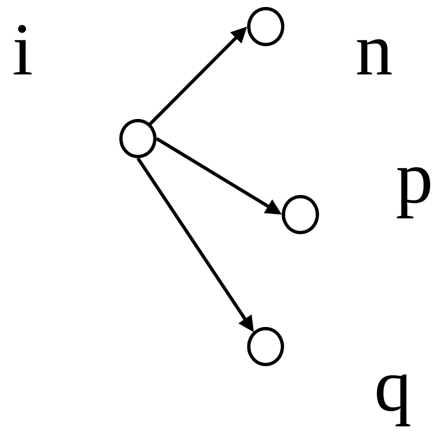
that is

a_i = Sum (h_j) over all j that (j,i) edge exists

or

$\mathbf{a} = \mathbf{A}^T \mathbf{h}$

Kleinberg's algorithm: HITS



symmetrically, for the 'hubness':

$$h_i = a_n + a_p + a_q$$

that is

h_i = Sum (q_j) over all j that (i,j) edge exists

or

$\mathbf{h} = \mathbf{A}\ \mathbf{a}$

Kleinberg's algorithm: HITS

In conclusion, we want vectors **h** and **a** such that:

$$\mathbf{h} = \mathbf{A}\,\mathbf{a}$$

$$\mathbf{a} = \mathbf{A}^T\,\mathbf{h}$$

That is:

$$\mathbf{a} = \mathbf{A}^T\mathbf{A}\,\mathbf{a}$$

Kleinberg's algorithm: HITS

a is a <u>right singular vector</u> of the adjacency matrix **A** (by dfn!), a.k.a the <u>eigenvector</u> of $\mathbf{A}^T\mathbf{A}$

Starting from random **a'** and iterating, we'll eventually converge

Q: to which of all the eigenvectors? why?

A: to the one of the strongest eigenvalue,

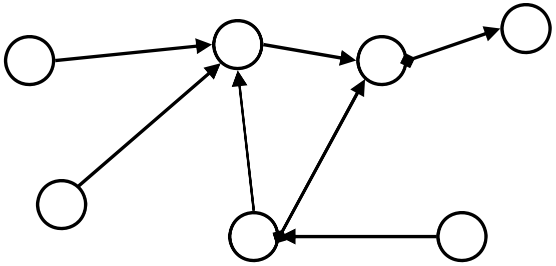
$$(\mathbf{A}^T \mathbf{A})^k \mathbf{a} = \lambda_1^k \mathbf{a}$$

Kleinberg’s algorithm - discussion

- ‘authority’ score can be used to find ‘similar pages’ (how?)
- closely related to ‘citation analysis’, social networks / ‘small world’ phenomena

Motivating problem: PageRank

Given a directed graph, find its most interesting/central node

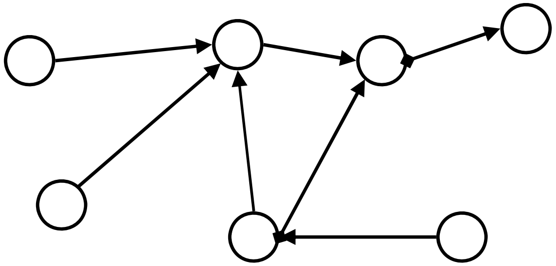


A node is important, if it is connected with important nodes (recursive, but OK!)

Motivating problem – PageRank solution

Given a directed graph, find its most interesting/central node

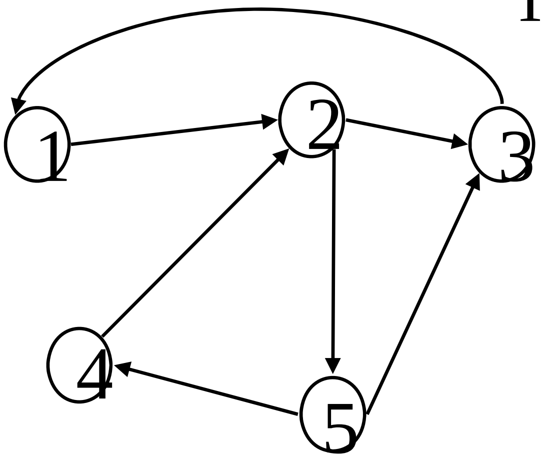
Proposed solution: Random walk; spot most 'popular' node (-> steady state prob. (ssp))



A node has high ssp, if it is connected with high ssp nodes (recursive, but OK!)

(Simplified) PageRank algorithm

- Let **A** be the transition matrix (= adjacency matrix); let **B** be the transpose, column-normalized - then



From

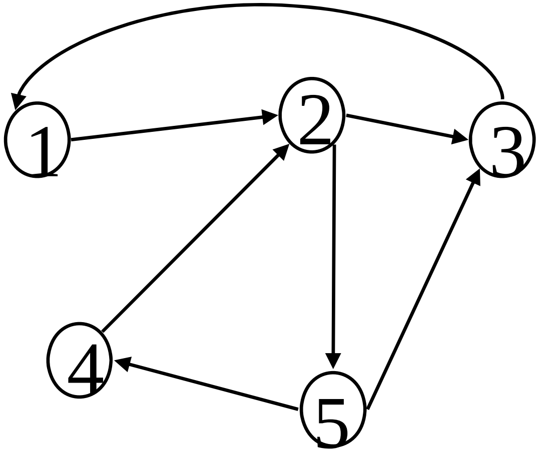
To

B

$$\begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix} \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix} = \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}$$

(Simplified) PageRank algorithm

- **B p = p**



$$\mathbf{B} \quad \mathbf{p} = \mathbf{p}$$

$$\begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix} \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix} = \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}$$

(Simplified) PageRank algorithm

- **B p =** 1 * **p**
- thus, **p** is the **eigenvector** that corresponds to the highest eigenvalue (=1, since the matrix is column-normalized)
- Why does such a **p** exist?
    - **p** exists if **B** is nxn, nonnegative, irreducible [Perron–Frobenius theorem]

(Simplified) PageRank algorithm

- In short: imagine a particle randomly moving along the edges
- compute its steady-state probabilities (ssp)

Full version of algo: with occasional random jumps

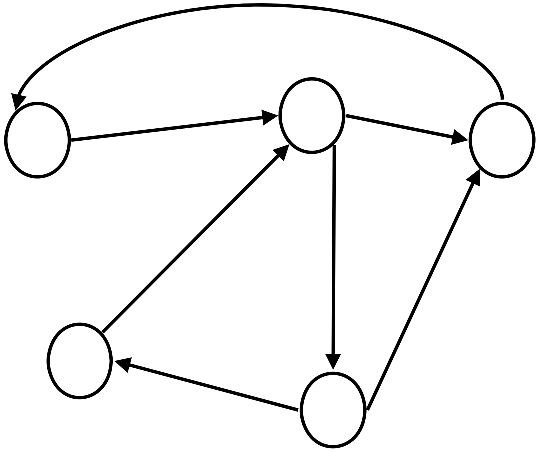
Why? To make the matrix irreducible

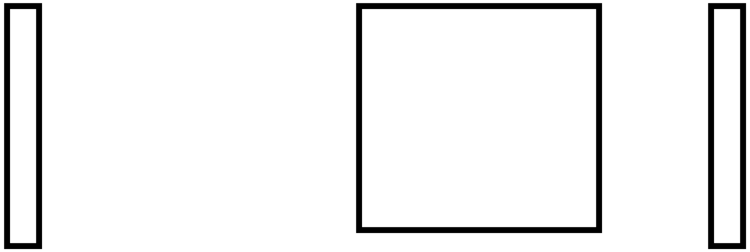
Full Algorithm

- With probability *1-c*, fly-out to a random node
- Then, we have

$$\mathbf{p} = c\,\mathbf{B}\,\mathbf{p} + (1-c)/n\ \mathbf{1} \Rightarrow$$

$$\mathbf{p} = (1-c)/n\ [\mathbf{I} - c\,\mathbf{B}]^{-1}\ \mathbf{1}$$

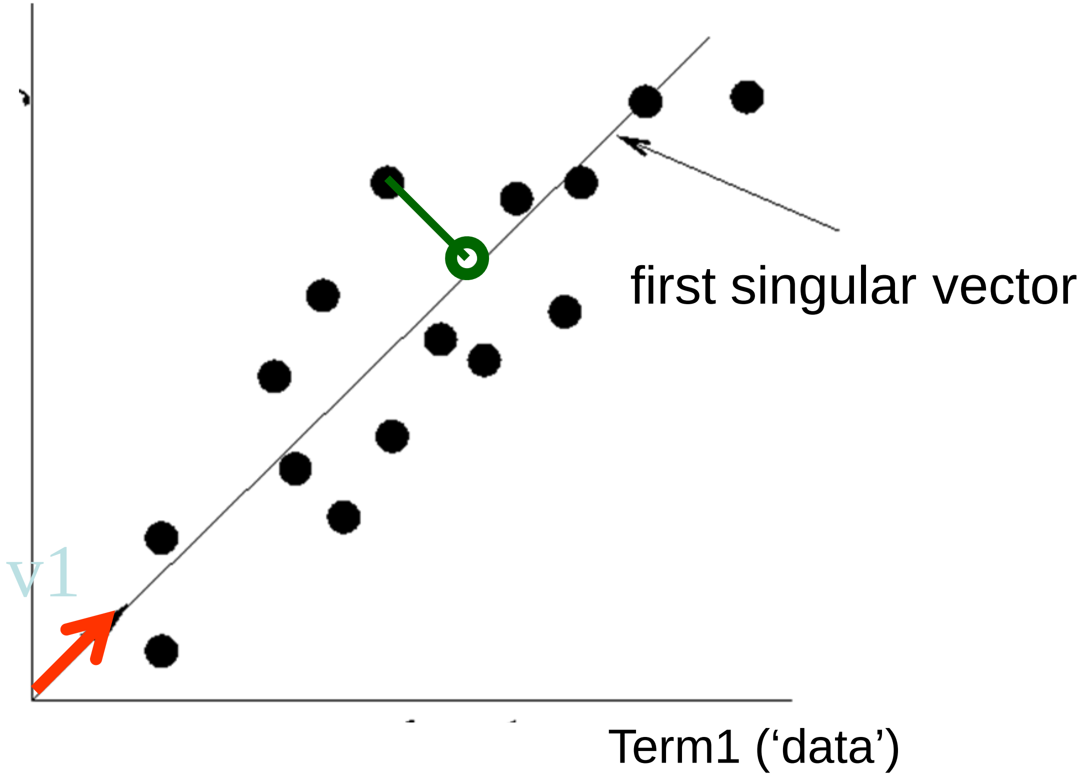




Motivation of CUR or CMD

- SVD, PCA all transform data into some abstract space (specified by a set basis)
    - Interpretability problem
    - Loss of sparsity

PCA - interpretation

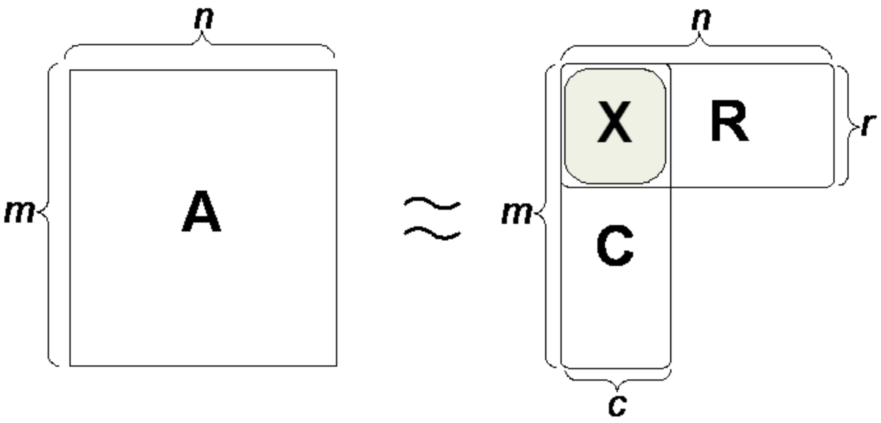


PCA projects points
Onto the "best" axis

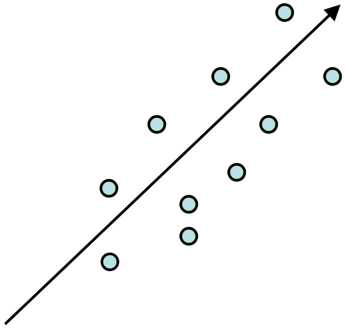
- minimum RMS error

CUR

- Example-based projection: use actual rows and columns to specify the subspace
- Given a matrix $A \in R^{m \times n}$, find three matrices $C \in R^{m \times c}$, $U \in R^{c \times r}$, $R \in R^{r \times n}$, such that $||A\text{-}CUR||$ is small



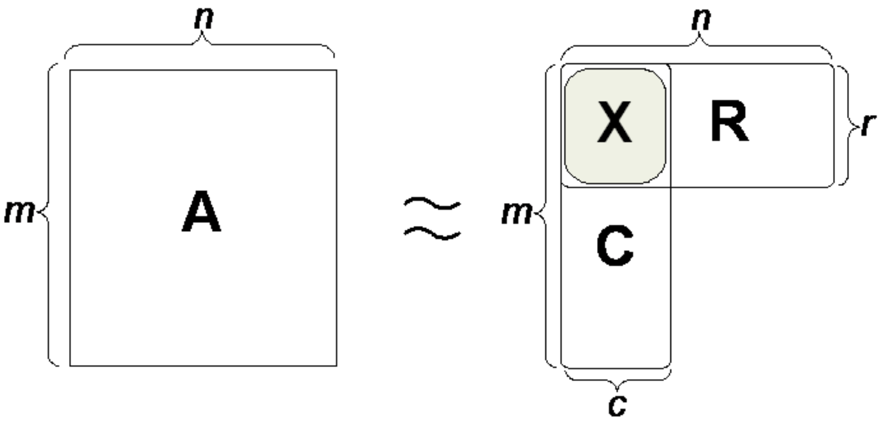
U is the pseudo-inverse of X

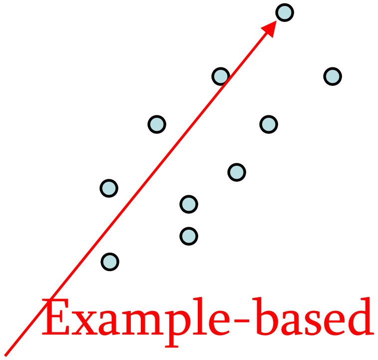


Orthogonal projection

CUR

- Example-based projection: use actual rows and columns to specify the subspace
- Given a matrix A∈$R^{m\times n}$, find three matrices C∈ $R^{m\times c}$, U∈ $R^{c\times r}$, R∈ $R^{r\times n}$, such that ||A-CUR|| is small





U is the pseudo-inverse of X:

$U = X^{\dagger} = (U^T U)^{-1} U^T$

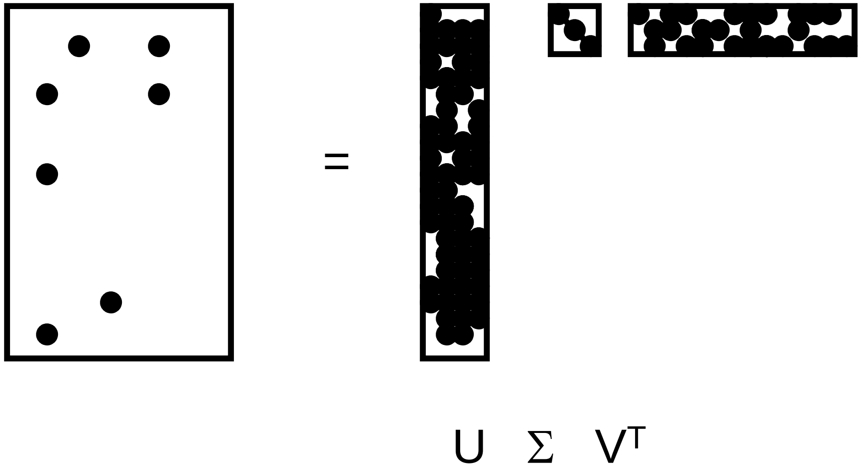
CUR (cont.)

- Key question:
  - How to select/sample the columns and rows?
- Uniform sampling
- Biased sampling
  - CUR w/ absolute error bound
  - CUR w/ relative error bound

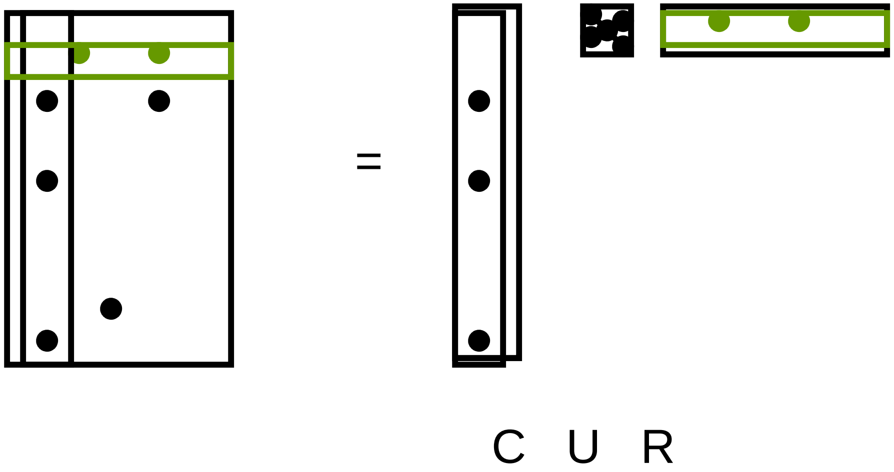
Reference:

1. Tutorial: Randomized Algorithms for Matrices and Massive Datasets, SDM'06
2. Drineas et al. Subspace Sampling and Relative-error Matrix Approximation: Column-Row-Based Methods, ESA2006
3. Drineas et al., Fast Monte Carlo Algorithms for Matrices III: Computing a Compressed Approximate Matrix Decomposition, SIAM Journal on Computing, 2006.

The sparsity property – pictorially:

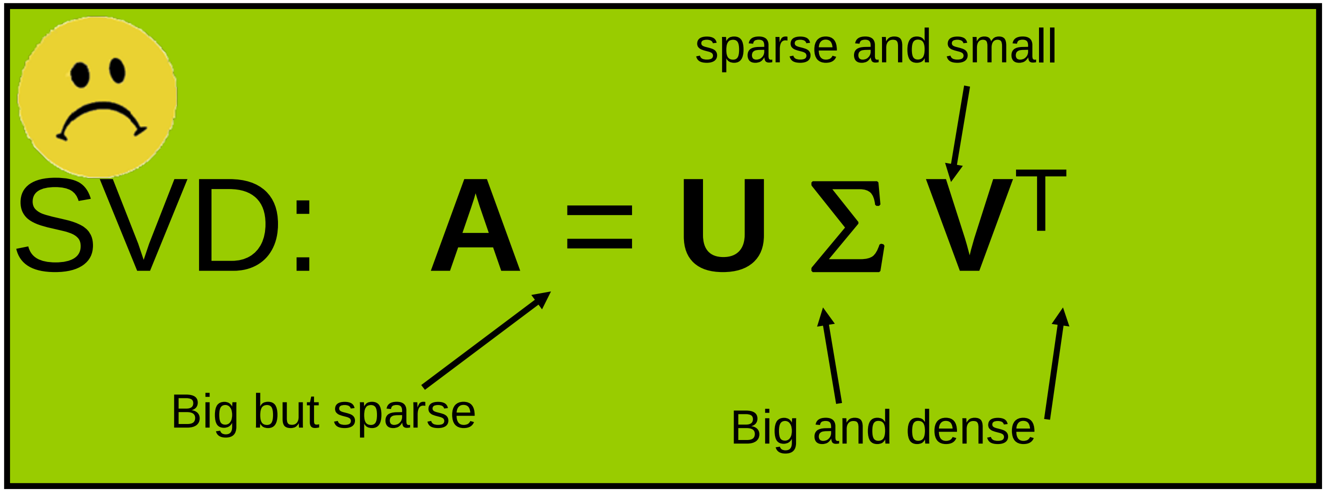


SVD/PCA:
Destroys sparsity

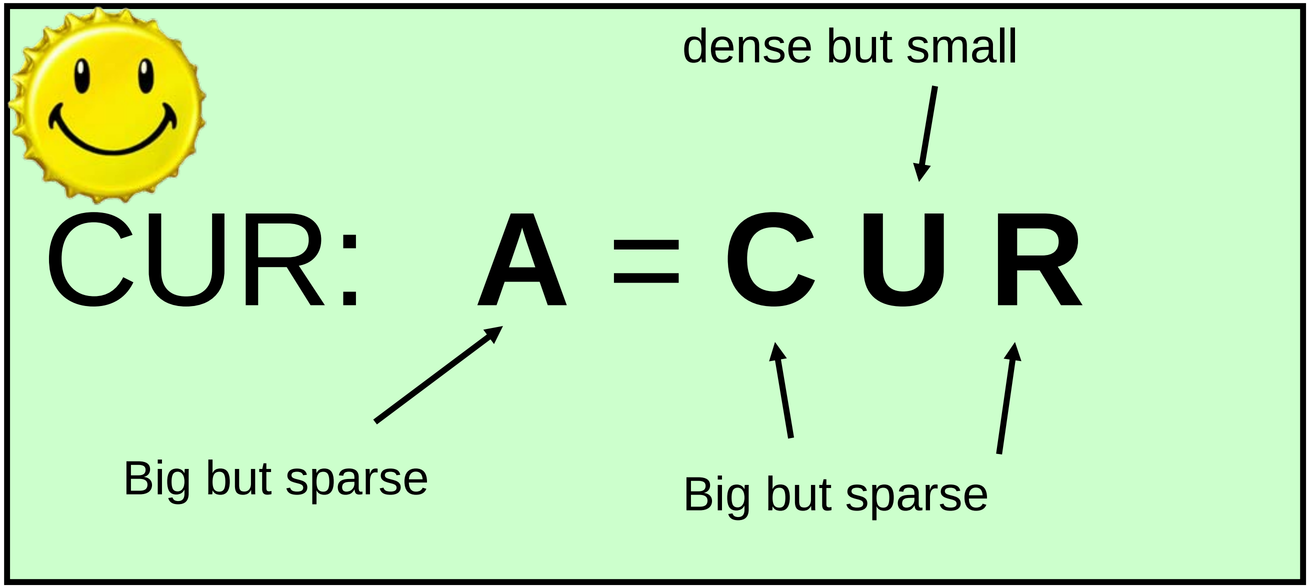


CUR: maintains sparsity

The sparsity property



SVD: $\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^T$



CUR: $\mathbf{A} = \mathbf{C}\,\mathbf{U}\,\mathbf{R}$

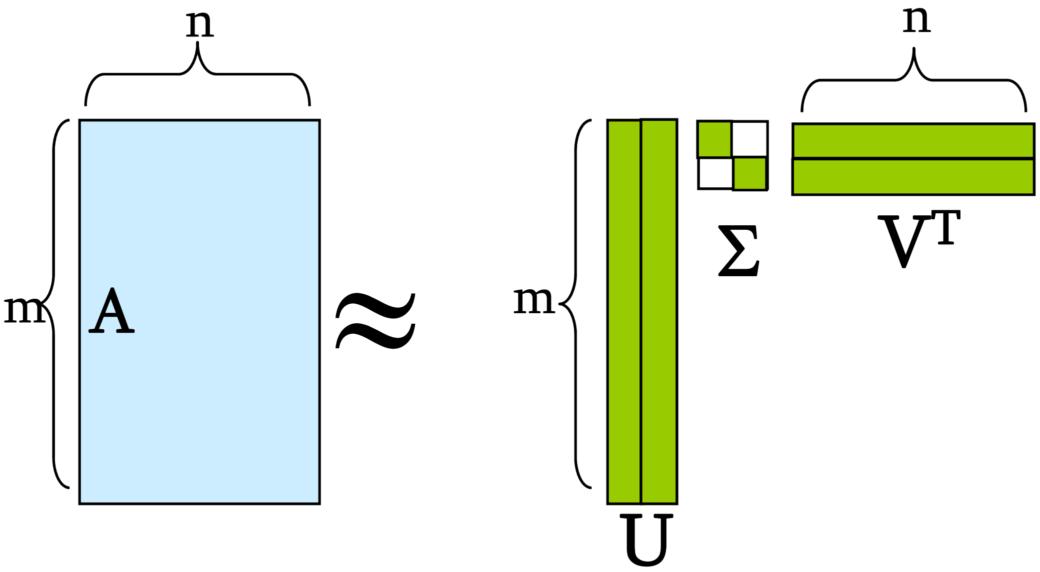
Matrix tools - summary

- SVD:
  - optimal for L2 – VERY popular (HITS, PageRank, Karhunen-Loeve, Latent Semantic Indexing, PCA, etc etc)
- C-U-R (CMD etc)
  - near-optimal; sparsity; interpretability

TENSORS

Reminder: SVD

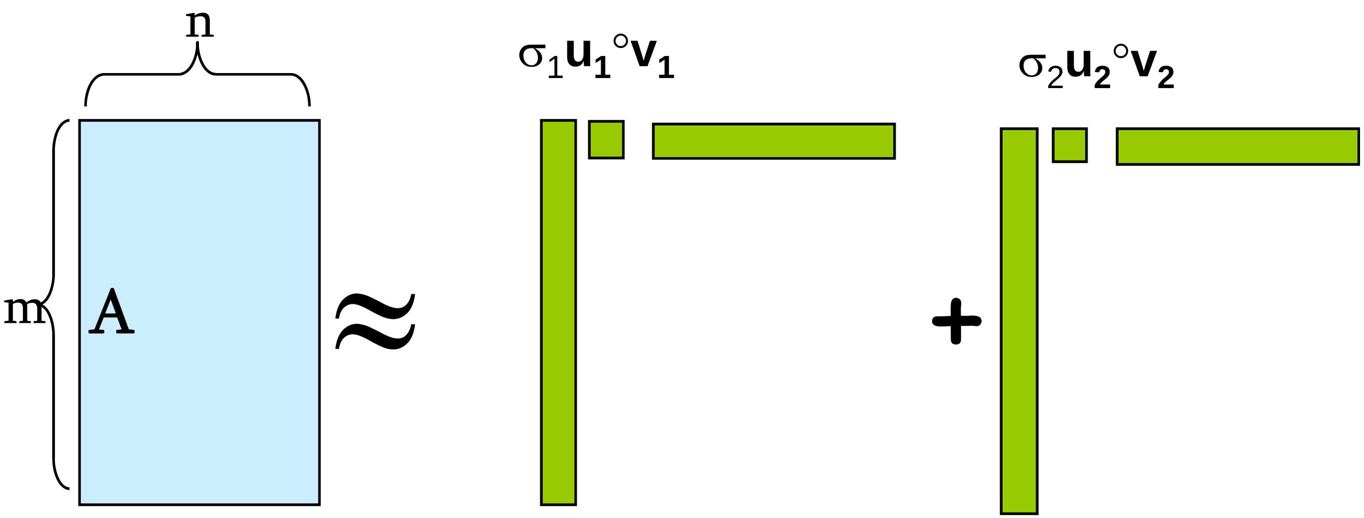
$$\mathbf{A} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

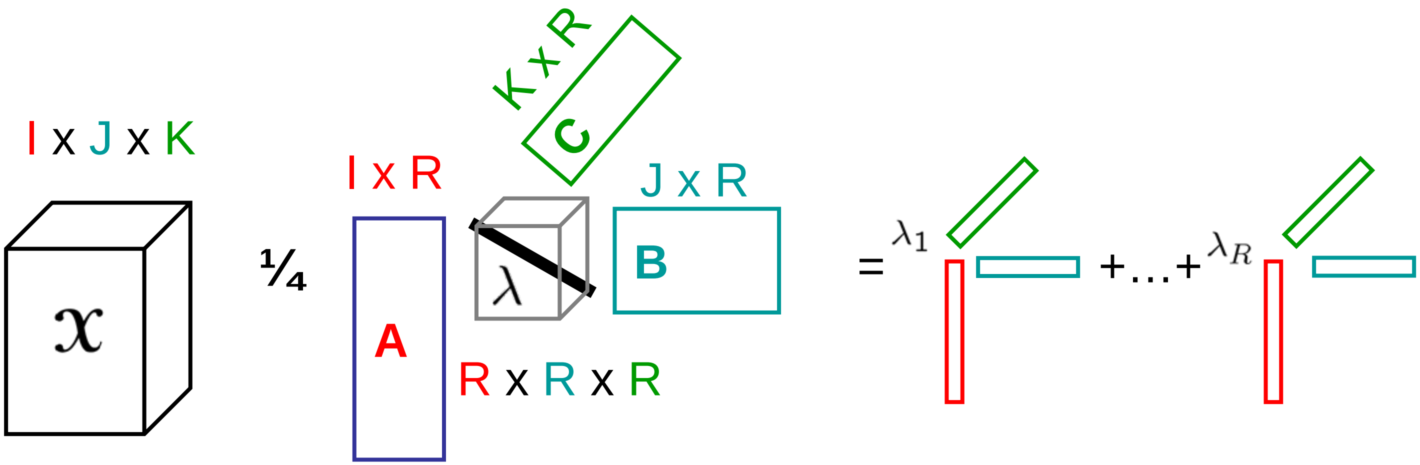
Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

Goal: extension to >=3 modes

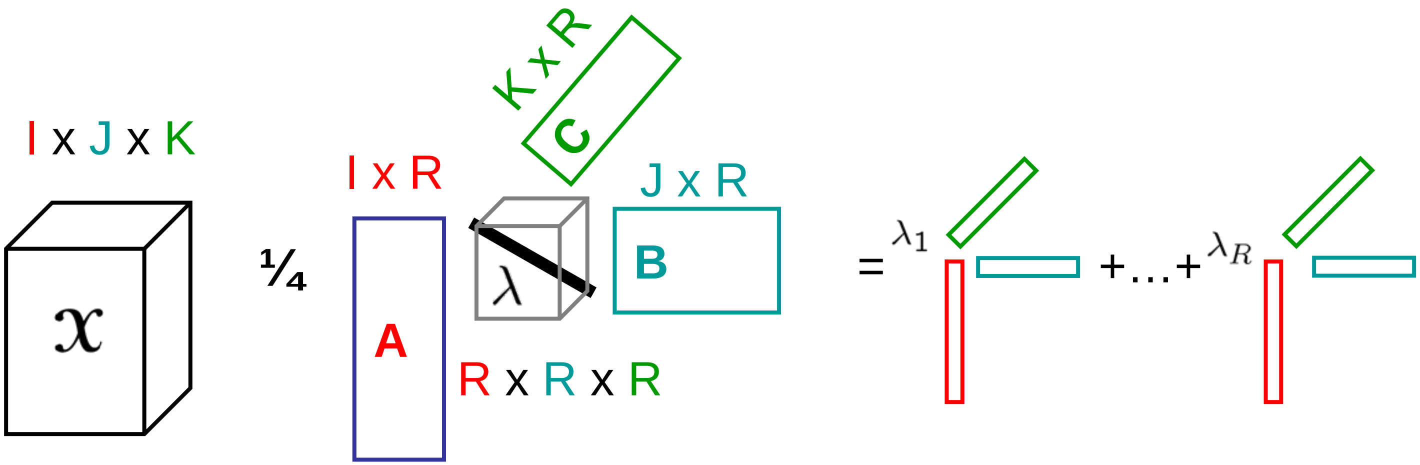


$$\mathcal{X} \approx [\![\lambda \,;\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_r \lambda_r \; \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Tensors: Main points

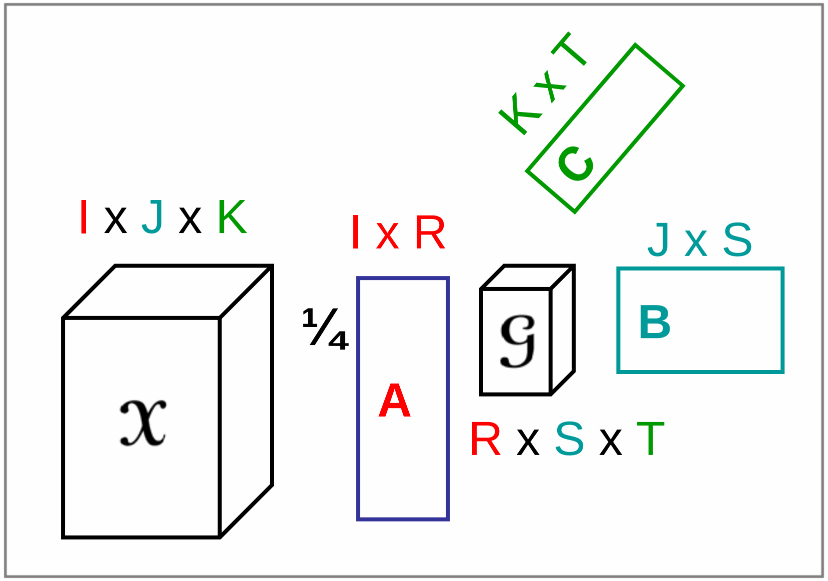
- 2 major types of tensor decompositions: Kruskal and Tucker
- both can be solved with ``alternating least squares'' (ALS)
- Details follow – we start with terminology:

Kruskal’s Decomposion - intuition



$$\mathcal{X} \approx [\![\lambda \,; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: conf. x conf-group
- **G**: how groups relate to each other

2-d analog of Tucker decomposition

e.g., terms x documents

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \end{array}$$

$$\begin{array}{cc} & k \\ m & \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \end{array} \begin{array}{cc} & l \\ k & \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \end{array} \begin{array}{cc} & n \\ l & \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} \end{array} = \left[\begin{array}{ccc|ccc} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ \hline .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{array}\right]$$

med. doc

cs doc

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

med. terms

cs terms

common terms

term group x
doc. group

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \left[\begin{array}{ccc|ccc} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ \hline .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{array}\right]$$

doc x
doc group

term x
term-group

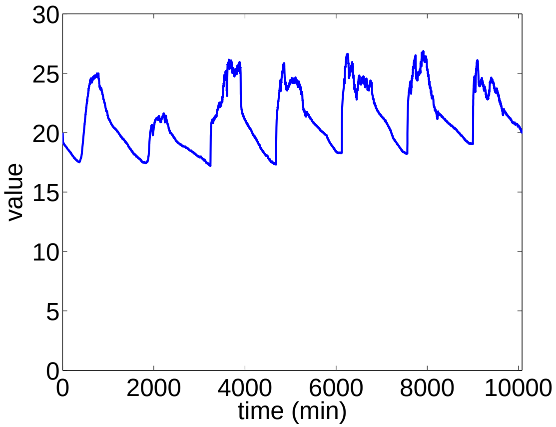
Tensor tools - summary

- Two main tools
  - PARAFAC
  - Tucker
- Both find row-, column-, tube-groups
  - but in PARAFAC the three groups are identical
- To solve: Alternating Least Squares

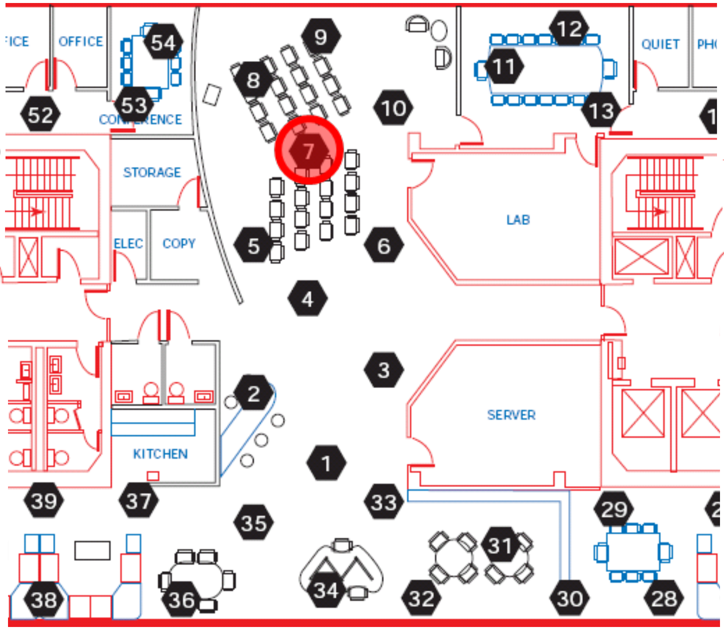
- Toolbox: from Tamara Kolda:

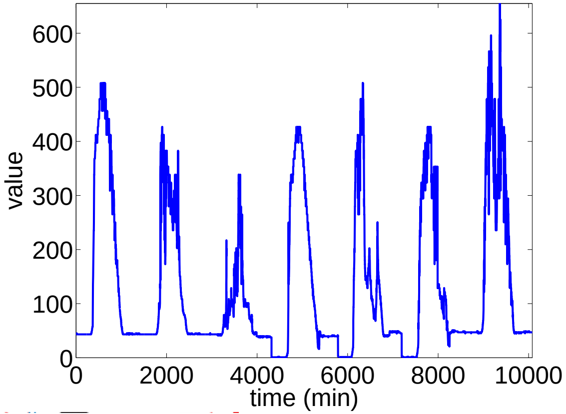
http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/

P1: Environmental sensor monitoring

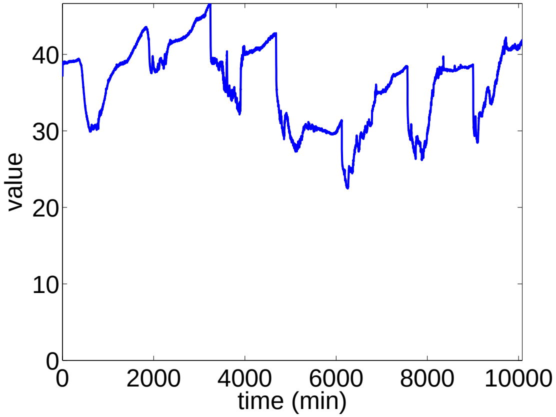


Temperature

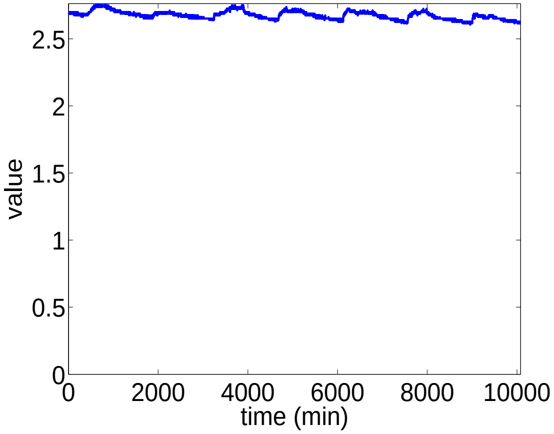




Light



Humidity

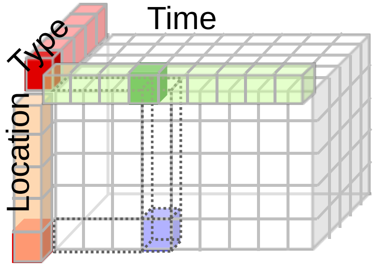


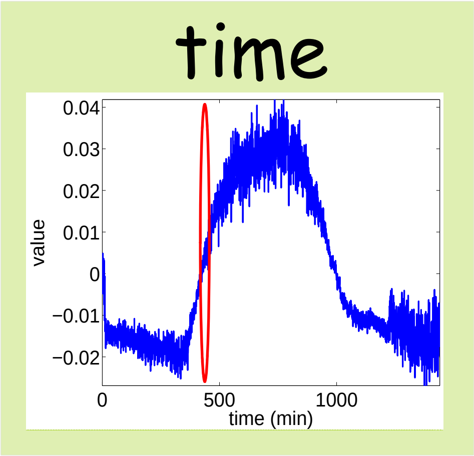
Voltage

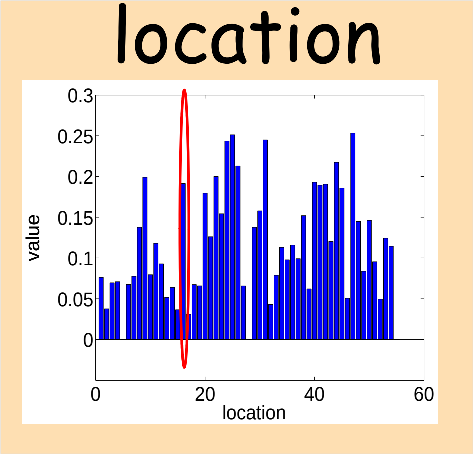
P1: sensor monitoring

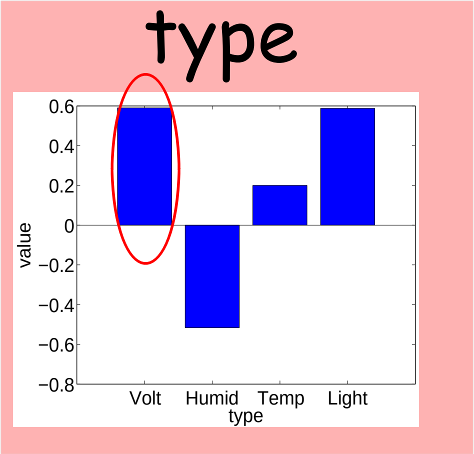
1st factor

Scaling factor 250









- 1st factor consists of the main trends:
    - Daily periodicity on time
    - Uniform on all locations
    - Temp, Light and Volt are positively correlated while negatively correlated with Humid

voltage

light

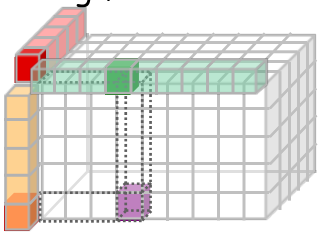
hum.

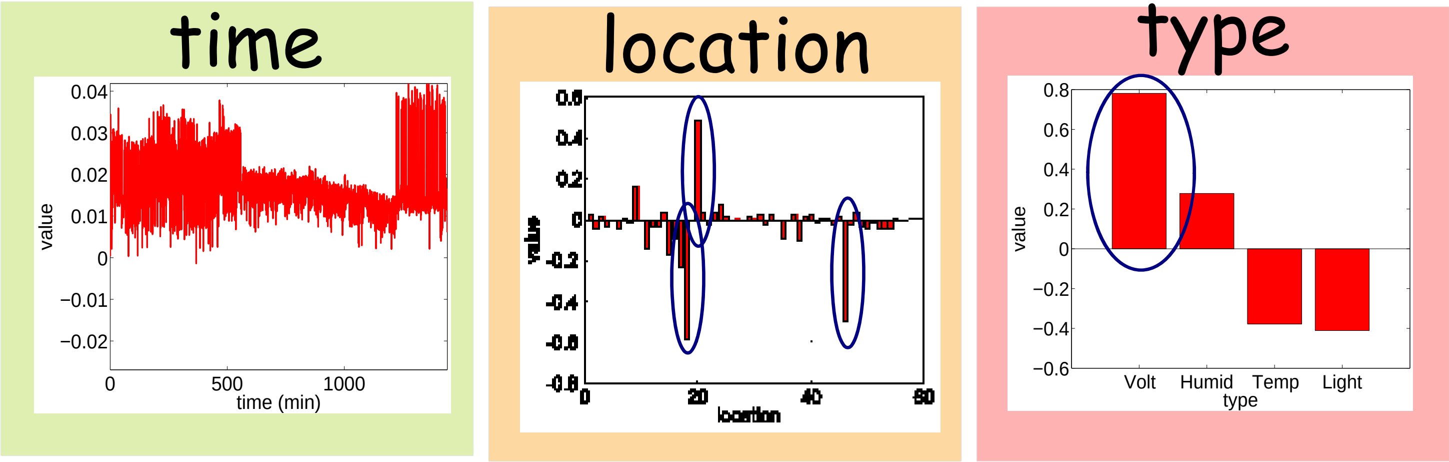
temp.

P1: sensor monitoring

2nd factor

Scaling factor 154





- 2nd factor captures an atypical trend:
  - Uniformly across all time
  - Concentrating on 3 locations
  - Mainly due to voltage
- Interpretation: two sensors have low battery, and the other one has high battery.

voltage

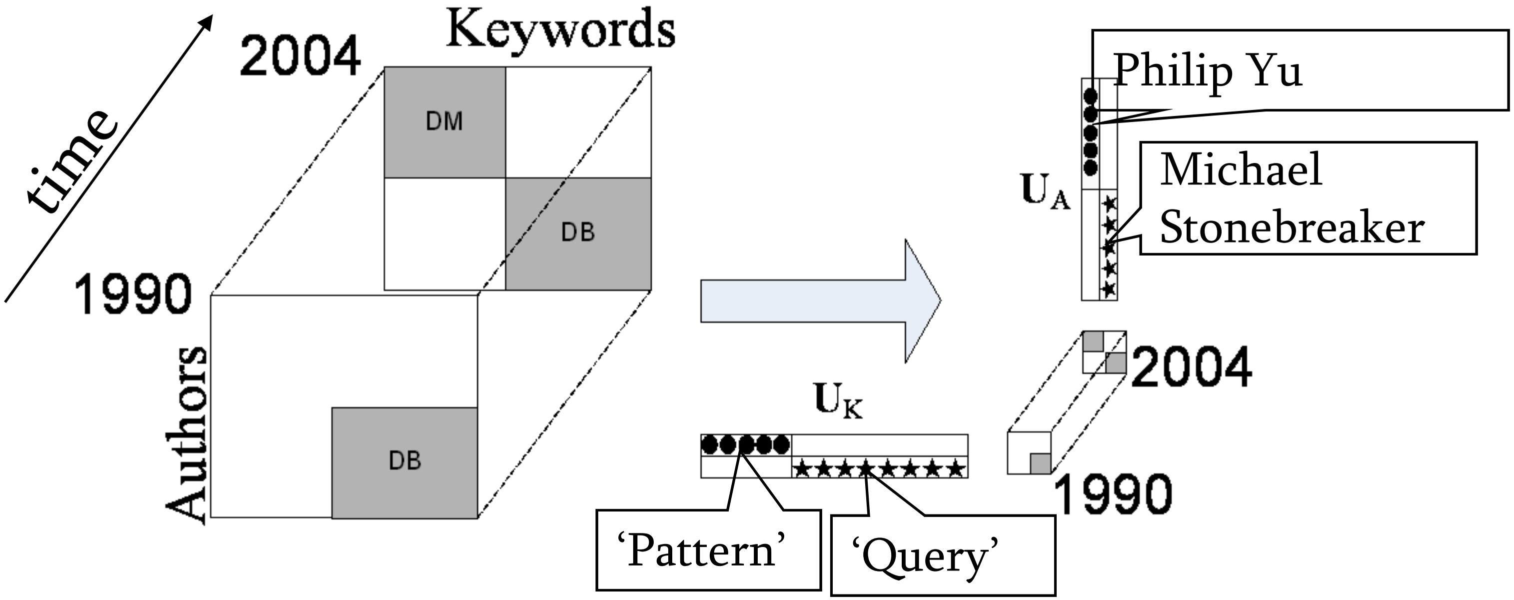
light

hum.

temp.

P3: Social network analysis

- Multiway latent semantic indexing (LSI)
  - Monitor the change of the community structure over time



P3: Social network analysis (cont.)

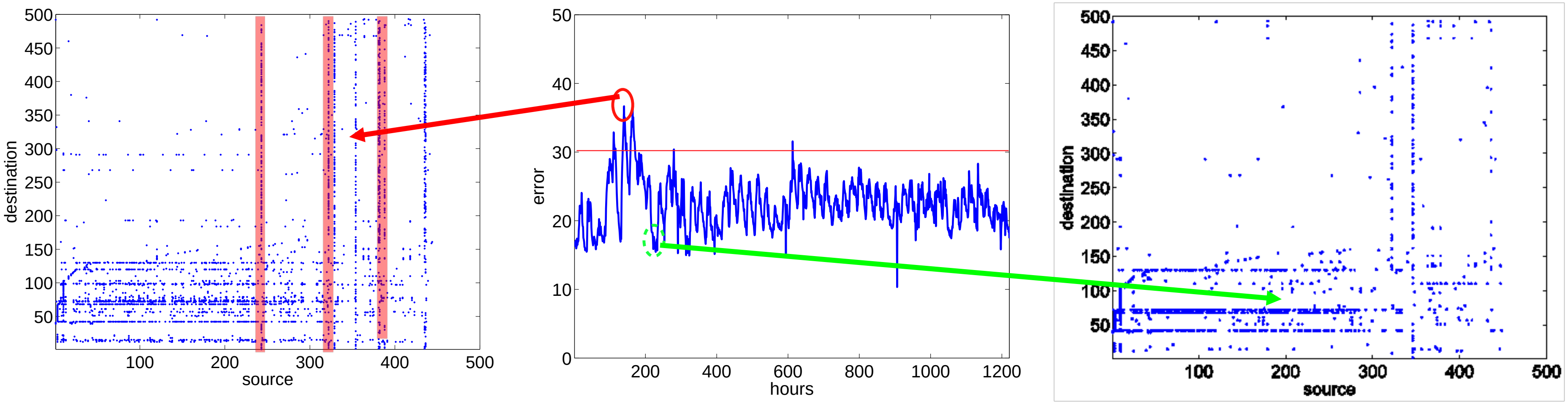
Authors	Keywords	Year
michael carey, michael stonebreaker, h. jagadish, hector garcia-molina	ueri,parallel,optimization,concurr, ient	1995
surajit chaudhuri,mitch cherniack,michael stonebreaker,ugur etintemel	distribut,systems,view,storage,servic,process, cache	2004
jiawei han,jian pei,philip s. yu, jianyong wang,charu c. aggarwal	pattern,support, cluster, ner,queri	2004

DB

DM

- Two groups are correctly identified: Databases and Data mining
- People and concepts are drifting over time

P4: Network anomaly detection



Abnormal traffic

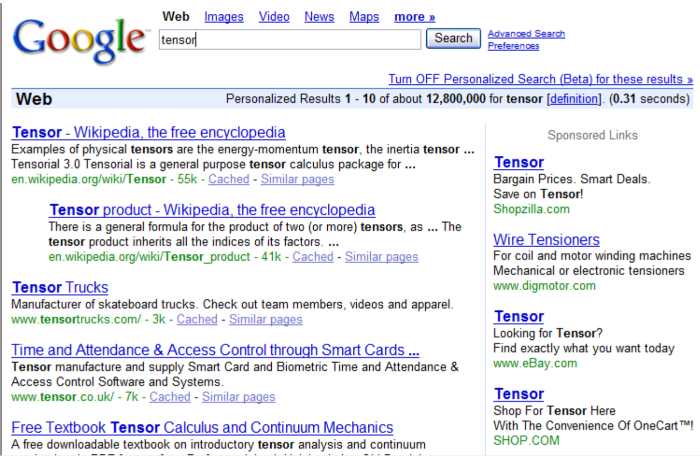
Reconstruction error over time

Normal traffic

- Reconstruction error gives indication of anomalies.
- Prominent difference between normal and abnormal ones is mainly due to the unusual scanning activity (confirmed by the campus admin).

P5: Web graph mining

- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)



Turn OFF Personalized Search (Beta) for these results »

Web Personalized Results **1** - **10** of about **12,800,000** for **tensor** [definition]. (**0.31** seconds)

Tensor - Wikipedia, the free encyclopedia
Examples of physical **tensors** are the energy-momentum **tensor**, the inertia **tensor** ... Tensorial 3.0 Tensorial is a general purpose **tensor** calculus package for ...
en.wikipedia.org/wiki/**Tensor** - 55k - Cached - Similar pages

Tensor product - Wikipedia, the free encyclopedia
There is a general formula for the product of two (or more) **tensors**, as ... The **tensor** product inherits all the indices of its factors. ...
en.wikipedia.org/wiki/**Tensor**_product - 41k - Cached - Similar pages

Tensor Trucks
Manufacturer of skateboard trucks. Check out team members, videos and apparel.
www.**tensor**trucks.com/ - 3k - Cached - Similar pages

Time and Attendance & Access Control through Smart Cards ...
Tensor manufacture and supply Smart Card and Biometric Time and Attendance & Access Control Software and Systems.
www.**tensor**.co.uk/ - 7k - Cached - Similar pages

Free Textbook **Tensor** Calculus and Continuum Mechanics
A free downloadable textbook on introductory **tensor** analysis and continuum

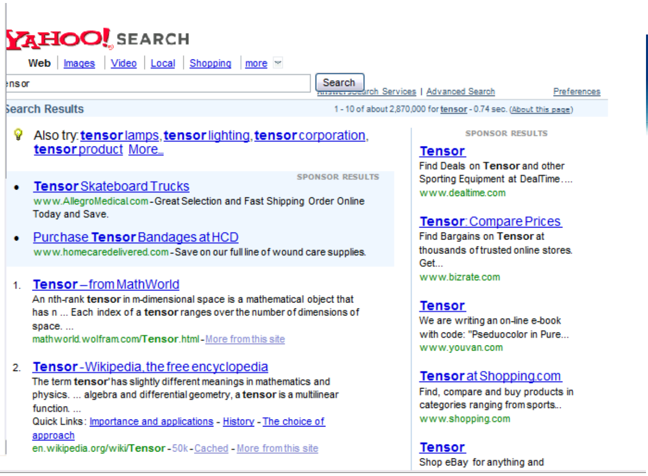
Sponsored Links

Tensor
Bargain Prices. Smart Deals.
Save on **Tensor**!
Shopzilla.com

Wire Tensioners
For coil and motor winding machines
Mechanical or electronic tensioners
www.digmotor.com

Tensor
Looking for **Tensor**?
Find exactly what you want today
www.eBay.com

Tensor
Shop For **Tensor** Here
With The Convenience Of OneCart™!
SHOP.COM



Search Results 1 - 10 of about 2,870,000 for tensor - 0.74 sec. (About this page)

Also try: **tensor** lamps, **tensor** lighting, **tensor** corporation, **tensor** product More...

SPONSOR RESULTS

- **Tensor** Skateboard Trucks
  www.AllegroMedical.com - Great Selection and Fast Shipping Order Online Today and Save.
- Purchase **Tensor** Bandages at HCD
  www.homecaredelivered.com - Save on our full line of wound care supplies.

1. **Tensor** – from MathWorld
   An nth-rank **tensor** in m-dimensional space is a mathematical object that has n ... Each index of a **tensor** ranges over the number of dimensions of space. ...
   mathworld.wolfram.com/**Tensor**.html - More from this site
2. **Tensor** - Wikipedia, the free encyclopedia
   The term **tensor**' has slightly different meanings in mathematics and physics. ... algebra and differential geometry, a **tensor** is a multilinear function. ...
   Quick Links: Importance and applications - History - The choice of approach
   en.wikipedia.org/wiki/**Tensor** - 50k - Cached - More from this site

SPONSOR RESULTS

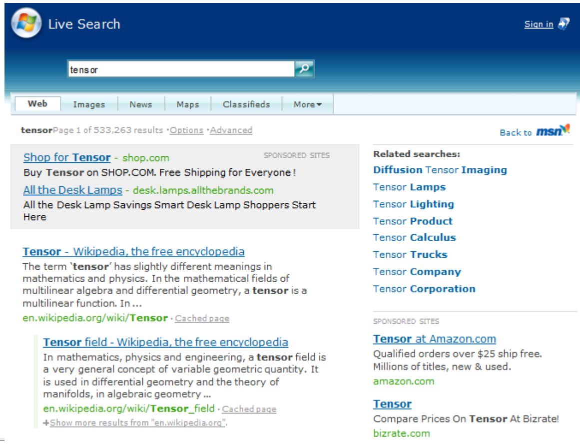
Tensor
Find Deals on **Tensor** and other Sporting Equipment at DealTime....
www.dealtime.com

Tensor: Compare Prices
Find Bargains on **Tensor** at thousands of trusted online stores. Get...
www.bizrate.com

Tensor
We are writing an on-line e-book with code: "Pseduocolor in Pure...
www.youvan.com

Tensor at Shopping.com
Find, compare and buy products in categories ranging from sports...
www.shopping.com

Tensor
Shop eBay for anything and



tensor Page 1 of 533,263 results · Options · Advanced

SPONSORED SITES

Shop for **Tensor** - shop.com
Buy **Tensor** on SHOP.COM. Free Shipping for Everyone!

All the Desk Lamps - desk.lamps.allthebrands.com
All the Desk Lamp Savings Smart Desk Lamp Shoppers Start Here

Tensor - Wikipedia, the free encyclopedia
The term '**tensor**' has slightly different meanings in mathematics and physics. In the mathematical fields of multilinear algebra and differential geometry, a **tensor** is a multilinear function. In ...
en.wikipedia.org/wiki/**Tensor** · Cached page

Tensor field - Wikipedia, the free encyclopedia
In mathematics, physics and engineering, a **tensor** field is a very general concept of variable geometric quantity. It is used in differential geometry and the theory of manifolds, in algebraic geometry ...
en.wikipedia.org/wiki/**Tensor**_field · Cached page
+Show more results from "en.wikipedia.org".

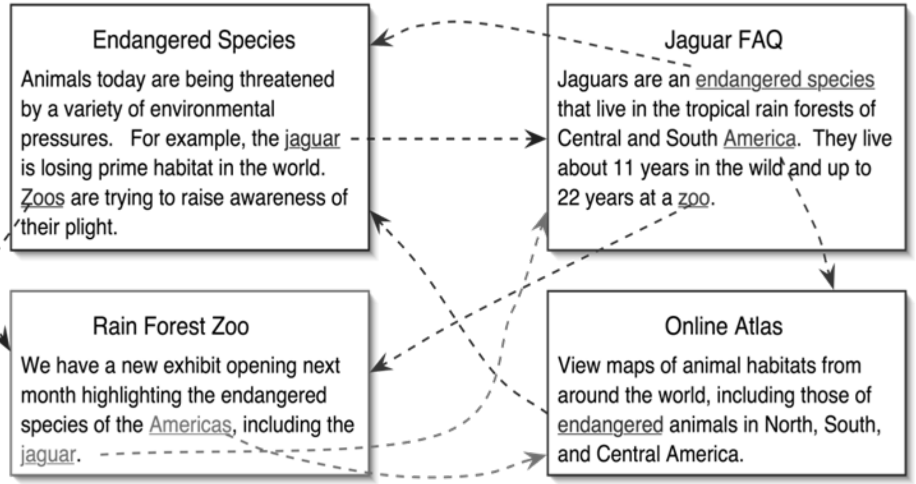
Related searches:
Diffusion Tensor **Imaging**
Tensor **Lamps**
Tensor **Lighting**
Tensor **Product**
Tensor **Calculus**
Tensor **Trucks**
Tensor **Company**
Tensor **Corporation**

SPONSORED SITES

Tensor at Amazon.com
Qualified orders over $25 ship free.
Millions of titles, new & used.
amazon.com

Tensor
Compare Prices On **Tensor** At Bizrate!
bizrate.com

Kleinberg's Hubs and Authorities (the HITS method)



Endangered Species

Animals today are being threatened by a variety of environmental pressures. For example, the jaguar is losing prime habitat in the world. Zoos are trying to raise awareness of their plight.

Jaguar FAQ

Jaguars are an endangered species that live in the tropical rain forests of Central and South America. They live about 11 years in the wild and up to 22 years at a zoo.

Rain Forest Zoo

We have a new exhibit opening next month highlighting the endangered species of the Americas, including the jaguar.

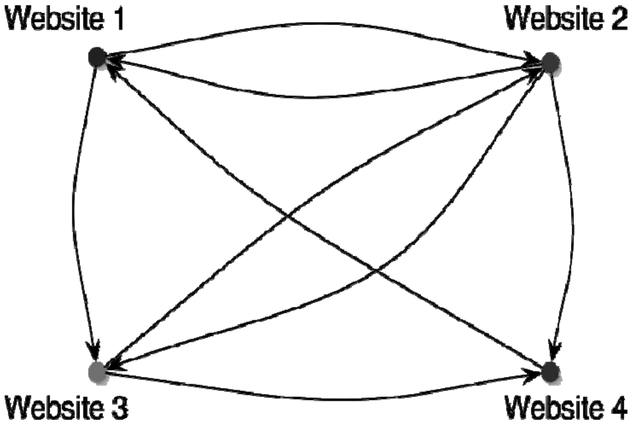
Online Atlas

View maps of animal habitats from around the world, including those of endangered animals in North, South, and Central America.

Sparse adjacency matrix and its SVD:

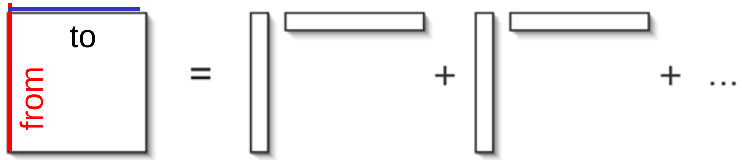
$$x_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{X} \approx \sum_r \sigma_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



authority scores for 1st topic

authority scores for 2nd topic



hub scores for 1st topic

hub scores for 2nd topic

Kleinberg, JACM, 1999 Leskovec&Faloutsos, WWW 2008

HITS Authorities on Sample Data

1st Principal Factor	
.97	www.ibm.com
.24	www.alphaw
.08	www-128.ibm
.05	www.develop
.02	www.researc
.01	www.redbook
.01	news.com.cc

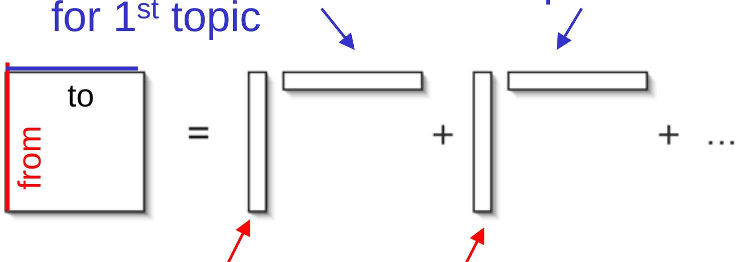
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh.edu
.06	www.lehighal
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.c
.02	lewisweb.cc.
.02	www.leo.lehi
.02	www.distanc
.02	fp1.cc.lehigh

3rd Principal Factor	
.75	java.sun.com
.38	www.sun.com
.36	developers.sun.
.24	see.sun.com
.16	www.samag.co
.13	docs.sun.com
.12	blogs.sun.com
.08	sunsolve.sun.c
.08	www.sun-catalo
.08	news.com.com

4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.g
.13	www.usdoj

6th Principal Factor	
.97	mathpost.asu.edu
.18	math.la.asu.edu
.17	www.asu.edu
.04	www.act.org
.03	www.eas.asu.edu
.02	archives.math.utk.edu
.02	www.geom.uiuc.edu
.02	www.fulton.asu.edu
.02	www.amstat.org
.02	www.maa.org

We started our crawl from
http://www-neos.mcs.anl.gov/neos,
and crawled 4700 pages,
resulting in 560
cross-linked hosts.



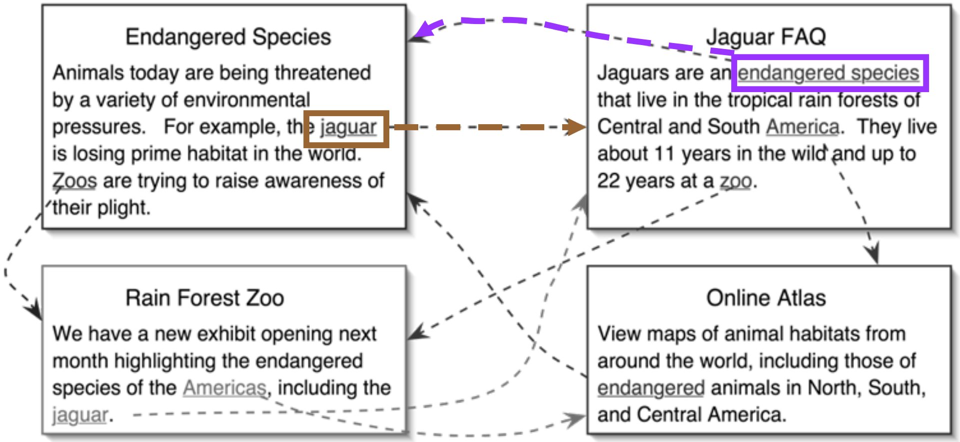
authority scores for 1st topic

authority scores for 2nd topic

hub scores for 1st topic

hub scores for 2nd topic

Three-Dimensional View of the Web



Endangered Species

Animals today are being threatened by a variety of environmental pressures. For example, the jaguar is losing prime habitat in the world. Zoos are trying to raise awareness of their plight.

Jaguar FAQ

Jaguars are an endangered species that live in the tropical rain forests of Central and South America. They live about 11 years in the wild and up to 22 years at a zoo.

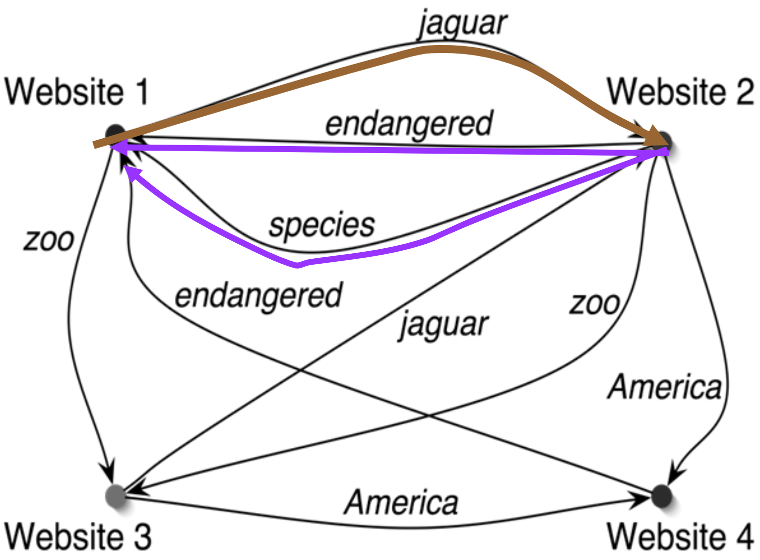
Rain Forest Zoo

We have a new exhibit opening next month highlighting the endangered species of the Americas, including the jaguar.

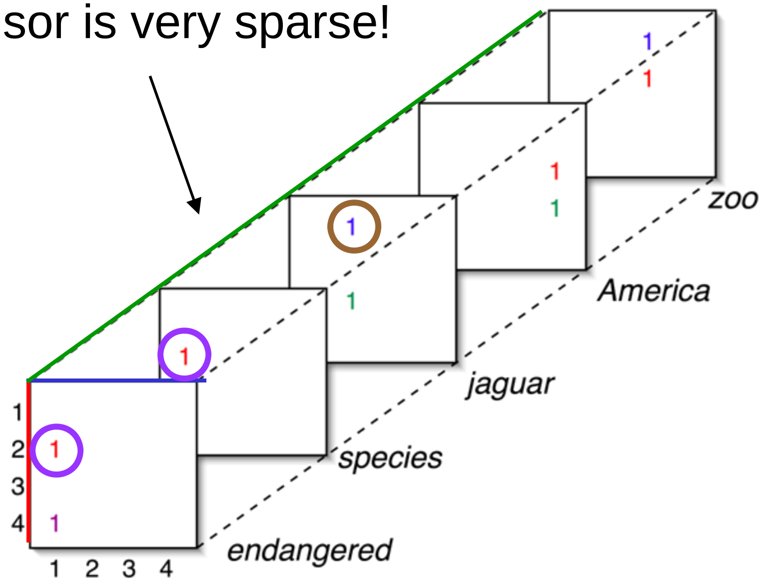
Online Atlas

View maps of animal habitats from around the world, including those of endangered animals in North, South, and Central America.

$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{page } j \\ & \text{with term } k \\ 0 & \text{otherwise} \end{cases}$$



Observe that this tensor is very sparse!

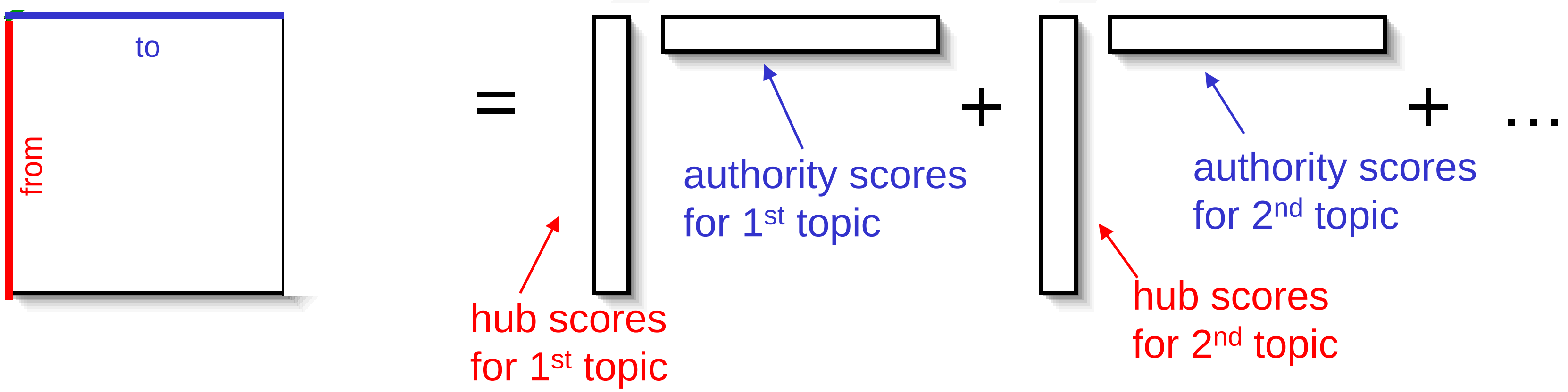


Kolda, Bader, Kenny, ICDM05

Topical HITS (TOPHITS)

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

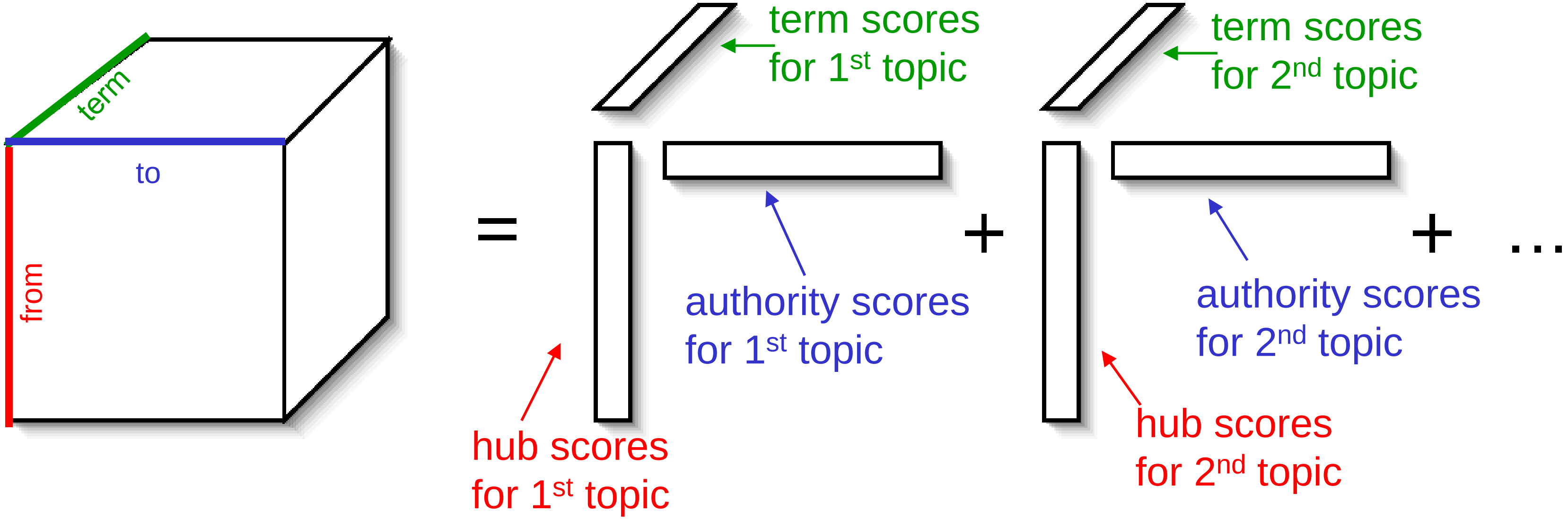
$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



Topical HITS (TOPHITS)

Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r \circ \mathbf{t}_r$$



term scores for 1st topic

authority scores for 1st topic

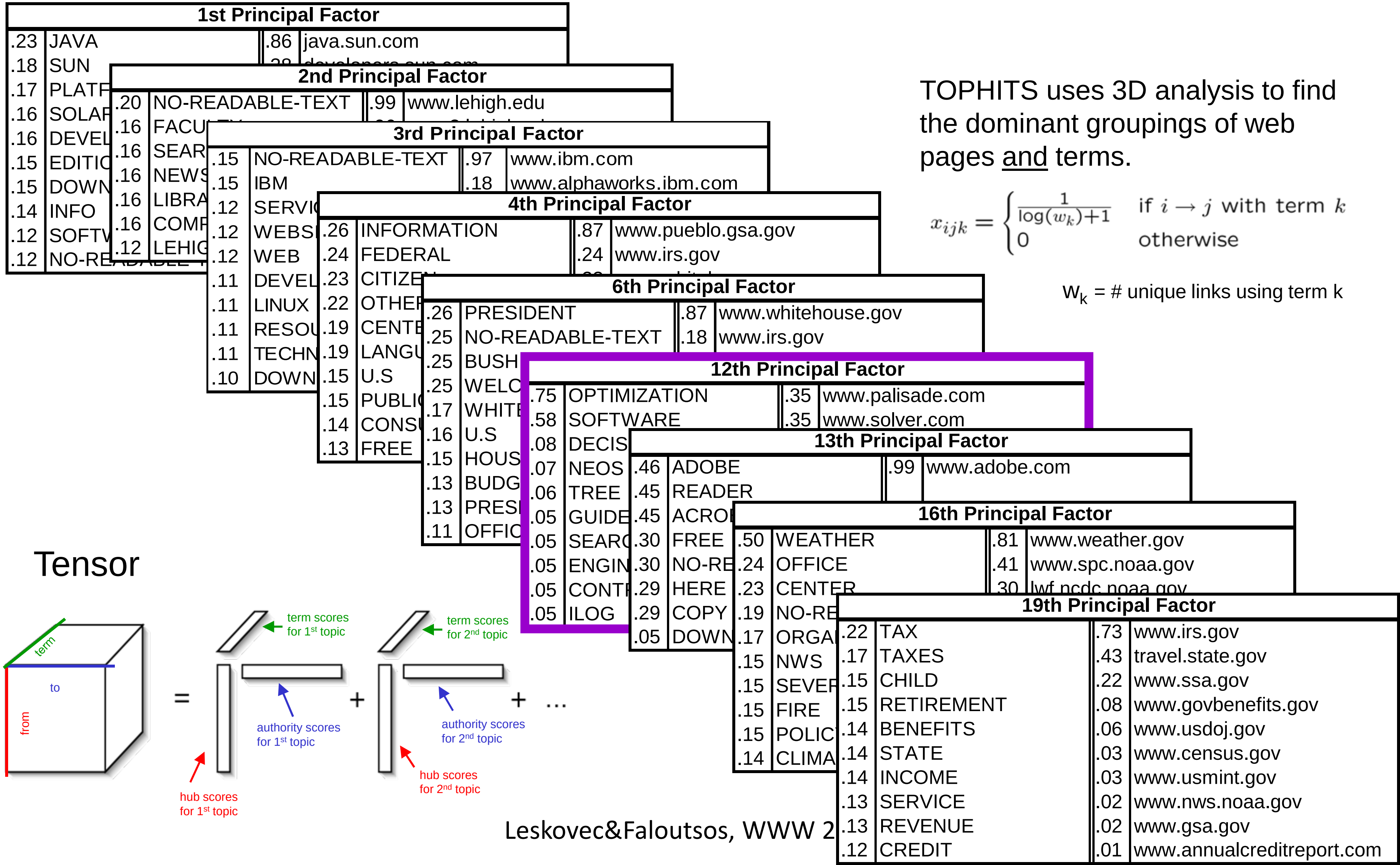
hub scores for 1st topic

term scores for 2nd topic

authority scores for 2nd topic

hub scores for 2nd topic

TOPHITS Terms & Authorities on Sample Data



1st Principal Factor			
.23	JAVA	.86	java.sun.com
.18	SUN	.38	developers.sun.com
.17	PLATF		
.16	SOLAF		
.16	DEVEL		
.15	EDITIC		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-RE		

2nd Principal Factor			
.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACU		
.16	SEAR		
.16	NEWS		
.16	LIBRA		
.16	COMP		
.12	LEHIG		

3rd Principal Factor			
.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM	.18	www.alphaworks.ibm.com
.12	SERVIC		
.12	WEBSI		
.12	WEB		
.11	DEVEL		
.11	LINUX		
.11	RESOU		
.11	TECHN		
.10	DOWN		

4th Principal Factor			
.26	INFORMATION	.87	www.pueblo.gsa.gov
.24	FEDERAL	.24	www.irs.gov
.23	CITIZE		
.22	OTHEF		
.19	CENTE		
.19	LANGU		
.15	U.S		
.15	PUBLIC		
.14	CONSU		
.13	FREE		

6th Principal Factor			
.26	PRESIDENT	.87	www.whitehouse.gov
.25	NO-READABLE-TEXT	.18	www.irs.gov
.25	BUSH		
.25	WELC		
.17	WHITE		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRESI		
.11	OFFIC		

12th Principal Factor			
.75	OPTIMIZATION	.35	www.palisade.com
.58	SOFTWARE	.35	www.solver.com
.08	DECIS		
.07	NEOS		
.06	TREE		
.05	GUIDE		
.05	SEARC		
.05	ENGIN		
.05	CONTI		
.05	ILOG		

13th Principal Factor			
.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACROE		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

16th Principal Factor			
.50	WEATHER	.81	www.weather.gov
.24	OFFICE	.41	www.spc.noaa.gov
.23	CENTER	.30	lwf.ncdc.noaa.gov
.19	NO-RE		
.17	ORGA		
.15	NWS		
.15	SEVEF		
.15	FIRE		
.15	POLIC		
.14	CLIMA		

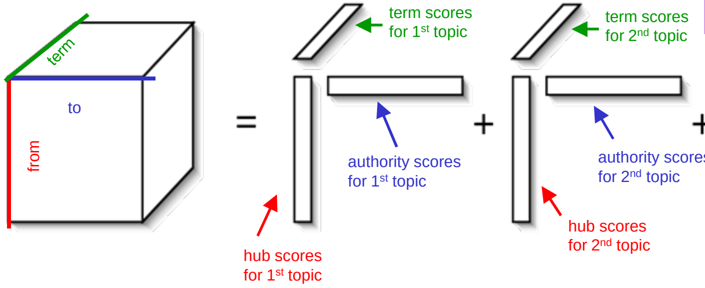
19th Principal Factor			
.22	TAX	.73	www.irs.gov
.17	TAXES	.43	travel.state.gov
.15	CHILD	.22	www.ssa.gov
.15	RETIREMENT	.08	www.govbenefits.gov
.14	BENEFITS	.06	www.usdoj.gov
.14	STATE	.03	www.census.gov
.14	INCOME	.03	www.usmint.gov
.13	SERVICE	.02	www.nws.noaa.gov
.13	REVENUE	.02	www.gsa.gov
.12	CREDIT	.01	www.annualcreditreport.com

TOPHITS uses 3D analysis to find the dominant groupings of web pages <u>and</u> terms.

$$x_{ijk} = \begin{cases} \frac{1}{\log(w_k)+1} & \text{if } i \to j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

w_k = # unique links using term k

Tensor



Leskovec&Faloutsos, WWW 2

Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Matrices and tensors provide elegant theory and algorithms
- Several research problems are still open
    - skewed distribution, anomaly detection, streaming algorithms, distributed/parallel algorithms, efficient out-of-core processing

References

- Slides borrowed from SIGMOD ‘07 tutorial by Falutsos, Kolda and Sun.