Roadmap

- Motivation
- Matrix tools
- Tensor tools
- Case studies

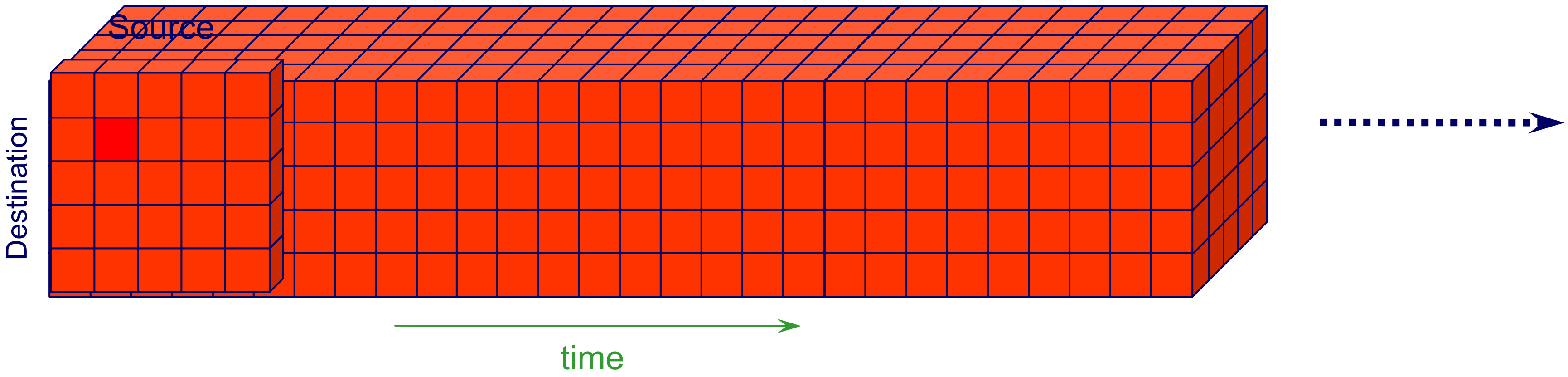
- Basics
- Tucker
- PARAFAC
- Incrementalization

Incrementalization

Incremental Tensor Decomposition

- Dynamic data model
  - Tensor Streams
- Dynamic Tensor Decomposition (DTA)
- Streaming Tensor Decomposition (STA)
- Window-based Tensor Decomposition (WTA)

Dynamic Tensor Stream



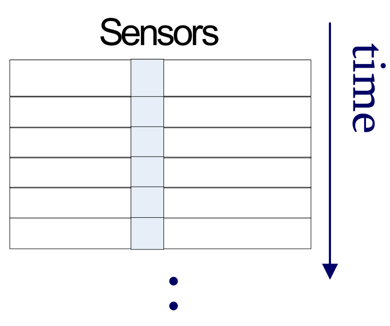
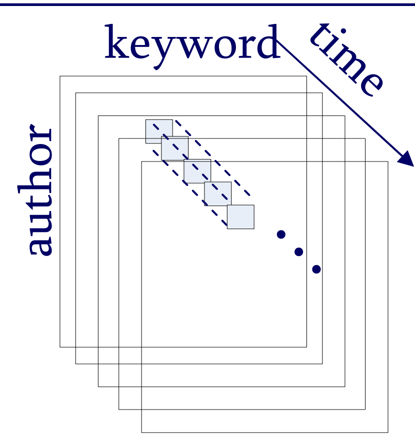
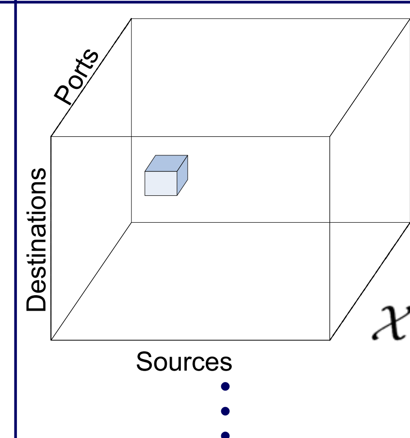
- Streams come with structure
  - (time, source, destination, port)
  - (time, author, keyword)
- How to summarize tensor streams effectively and incrementally?

Dynamic Data model

- Tensor Streams
  - A sequence of Mth order tensor

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

n is increasing over time

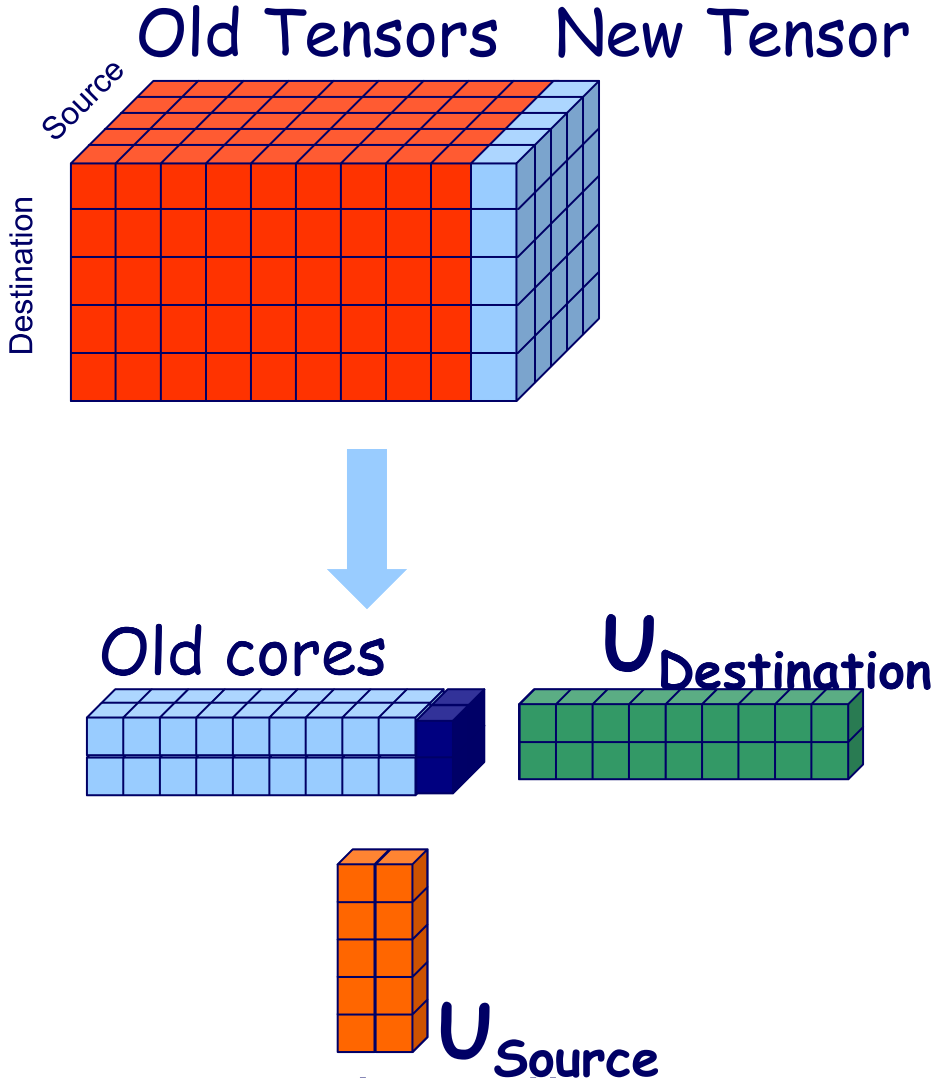
Order	1st	2nd	3rd
Correspondence	Multiple streams	Time evolving graphs	3D arrays
Example			

Incremental Tensor Decomposition

☺ Dynamic data model
  - Tensor Streams
- Dynamic Tensor Decomposition (DTA)
- Streaming Tensor Decomposition (STA)
- Window-based Tensor Decomposition (WTA)

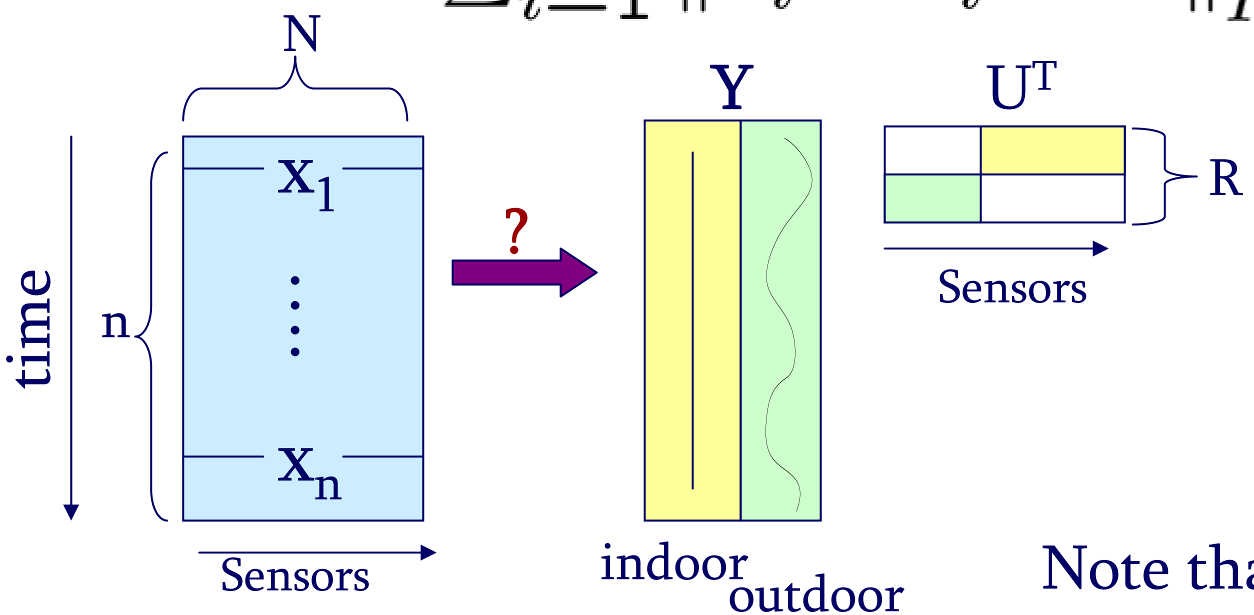
1. Jimeng Sun, Spiros Papadimitriou, Philip Yu. Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams, *ICDM 2006*
2. Jimeng Sun, Dacheng Tao, Christos Faloutsos. Beyond Streams and Graphs: Dynamic Tensor Analysis, *KDD 2006*

Incremental Tensor Decomposition



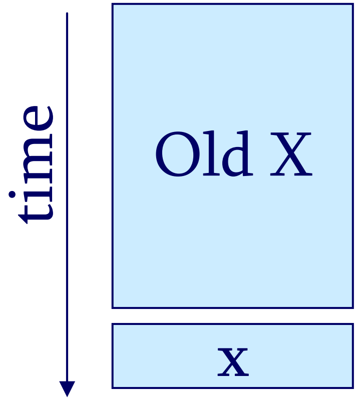
1st order DTA - problem

Given $x_1 \ldots x_n$ where each $x_i \in R^N$, find $U \in R^{N\times R}$ such that the error e is small: $e = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_i U U^T\|_F^2$



Note that Y = XU

1^{st} order Dynamic Tensor Analysis

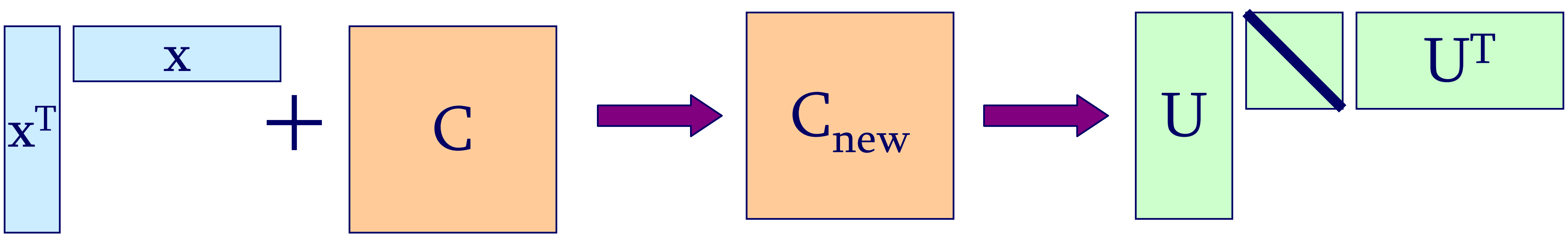


Input: new data vector $x \in R^N$, old variance matrix $C \in R^{N\times N}$

Output: new projection matrix $U \in R^{N\times R}$

Algorithm:

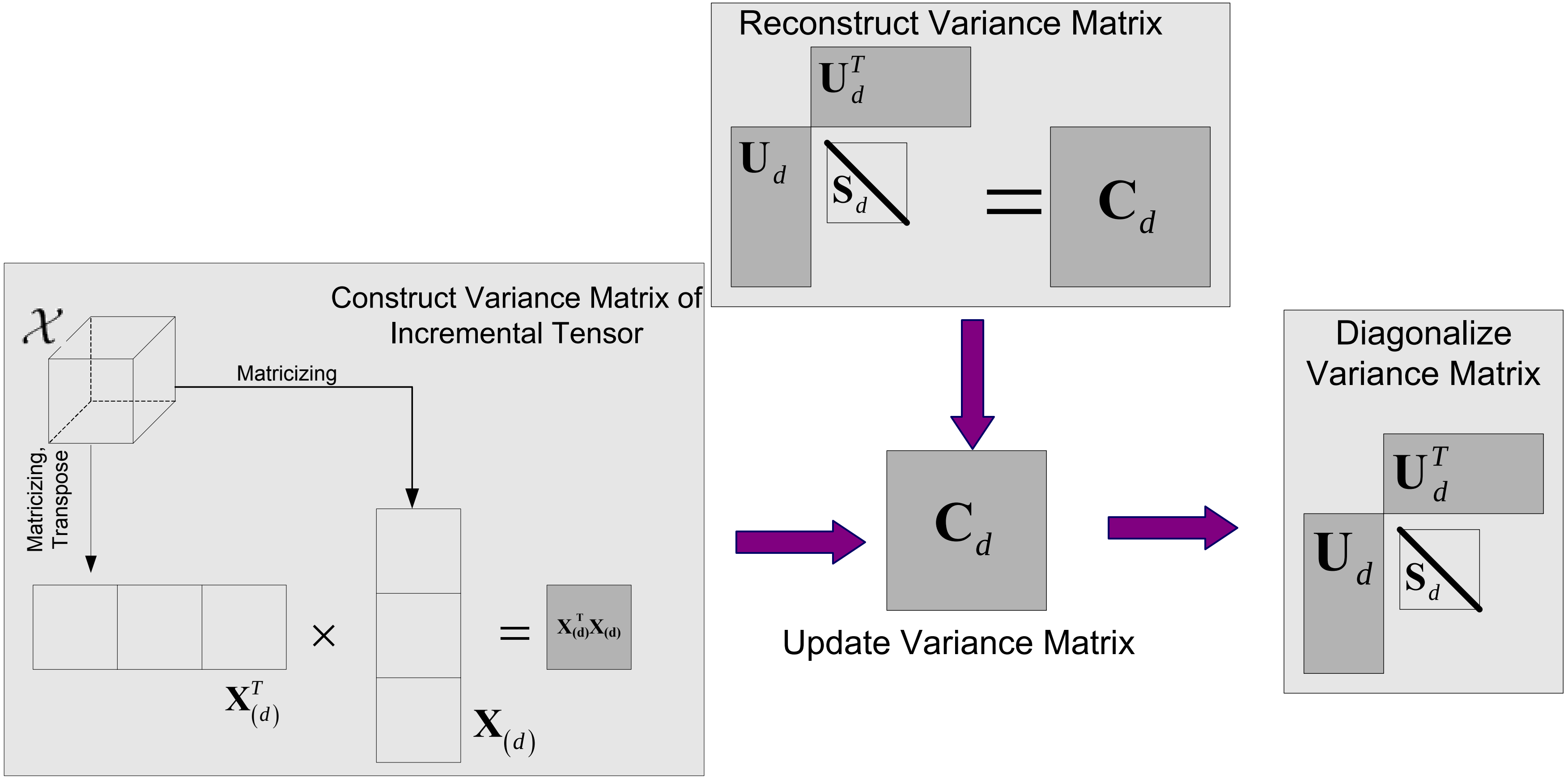
1. update variance matrix $C_{new} = x^T x + C$
2. Diagonalize $U\Lambda U^T = C_{new}$
3. Determine the rank R and return U



Diagonalization has to be done for every new **x**!

details

M^{th} order DTA



M^{th} order DTA – complexity

Storage:

$O(\prod N_i)$, i.e., size of an input tensor at a single timestamp

Computation:

$\sum N_i^3$ (or $\sum N_i^2$) diagonalization of C

$+ \sum N_i \prod N_i$ matrix multiplication $X_{(d)}^T X_{(d)}$

For low order tensor(<3), diagonalization is the main cost

For high order tensor, matrix multiplication is the main cost

Incremental Tensor Decomposition

- ☺ Dynamic data model
  - Tensor Streams
- ☺ Dynamic Tensor Decomposition (DTA)
- Streaming Tensor Decomposition (STA)
- Window-based Tensor Decomposition (WTA)

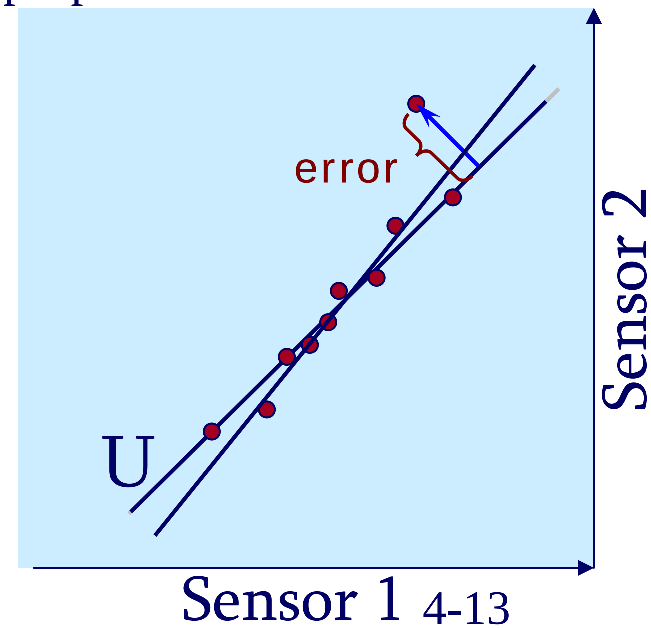
1. Jimeng Sun, Spiros Papadimitriou, Philip Yu. Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams, *ICDM 2006*
2. Jimeng Sun, Dacheng Tao, Christos Faloutsos. Beyond Streams and Graphs: Dynamic Tensor Analysis, *KDD 2006*

1^{st} order Streaming Tensor Analysis (STA)

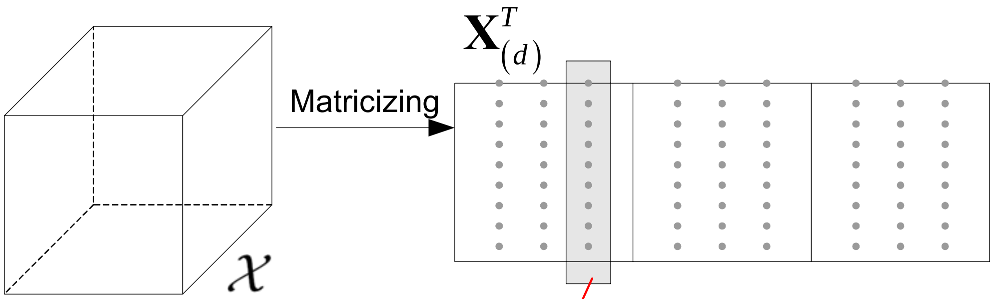
- Adjust U smoothly when new data arrive without diagonalization [VLDB05]
- For each new point x
  - Project onto current line
  - Estimate error
  - Rotate line in the direction of the error and in proportion to its magnitude

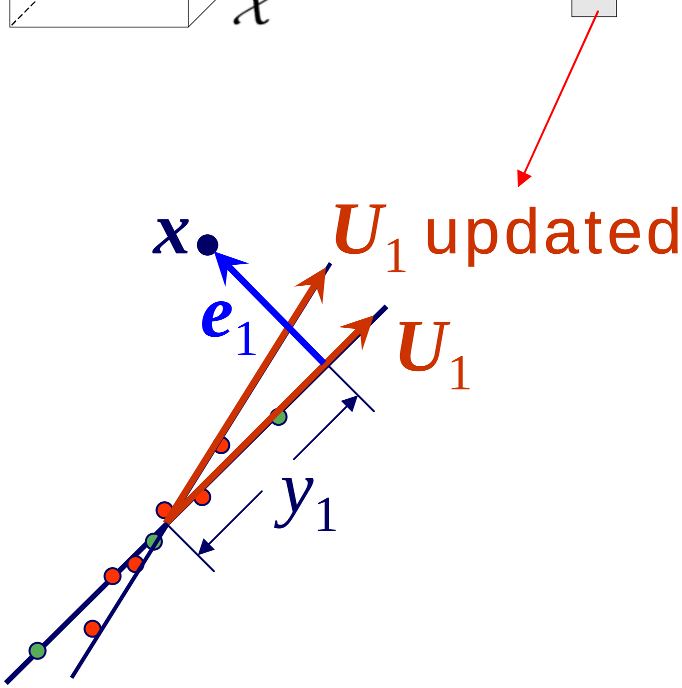
For each new point $\boldsymbol{x}$ and for $i = 1, \ldots, k$:

- $y_i := \boldsymbol{U}_i^T \boldsymbol{x}$ (proj. onto $\boldsymbol{U}_i$)
- $d_i \leftarrow \lambda d_i + y_i^2$ (energy $\propto$ i-th eigenval.)
- $\boldsymbol{e}_i := \boldsymbol{x} - y_i \boldsymbol{U}_i$ (error)
- $\boldsymbol{U}_i \leftarrow \boldsymbol{U}_i + (1/d_i)\, y_i \boldsymbol{e}_i$ (update estimate)
- $\boldsymbol{x} \leftarrow \boldsymbol{x} - y_i \boldsymbol{U}_i$ (repeat with remainder)



M^{th} order STA





- Run 1st order STA along each mode
- Complexity:
  - Storage: $O(\prod N_i)$
  - Computation: $\sum R_i \prod N_i$ which is smaller than DTA

Incremental Tensor Decomposition

☺ Dynamic data model
  - Tensor Streams

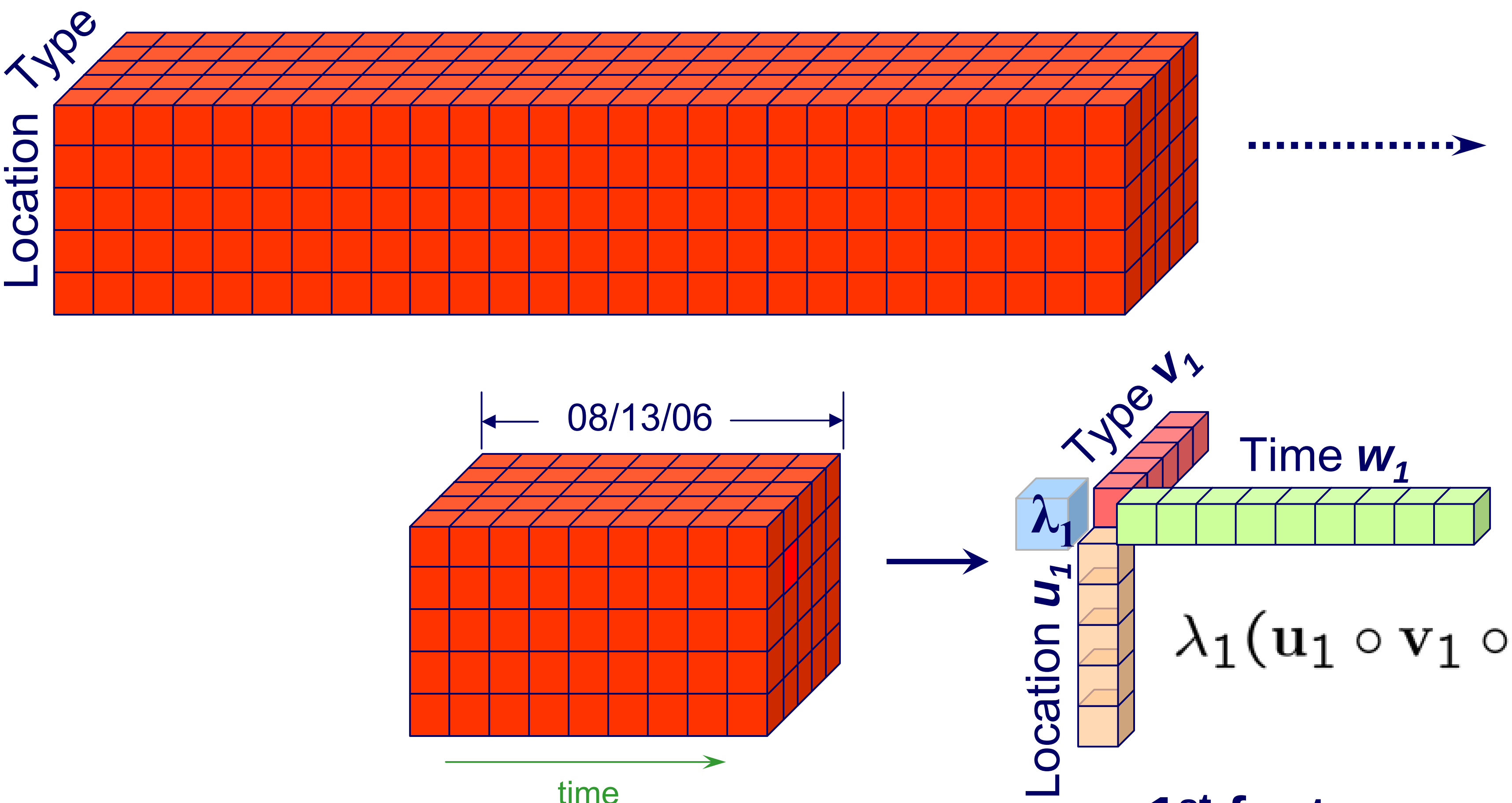
☺ Dynamic Tensor Decomposition (DTA)

☺ Streaming Tensor Decomposition (STA)

- Window-based Tensor Decomposition (WTA)

1. Jimeng Sun, Spiros Papadimitriou, Philip Yu. Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams, *ICDM 2006*
2. Jimeng Sun, Dacheng Tao, Christos Faloutsos. Beyond Streams and Graphs: Dynamic Tensor Analysis, *KDD 2006*

Window-based Tensor Analysis (WTA)



$$\lambda_1(\mathbf{u}_1 \circ \mathbf{v}_1 \circ \mathbf{w}_1)$$

1st factor

Meta-algorithm for window-based tensor analysis

Input:

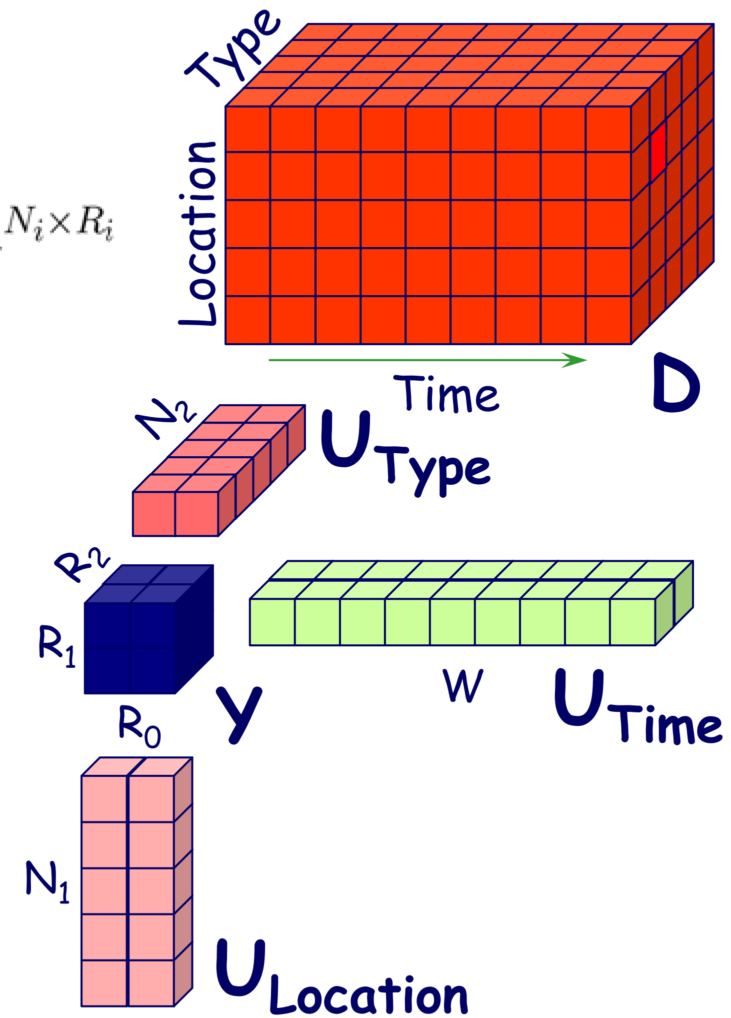
The tensor window $\mathcal{D} \in \mathbb{R}^{W \times N_1 \times \cdots \times N_M}$

Output:

The projection matrix $\mathbf{U}_0 \in \mathbb{R}^{W \times R_0}$, $\mathbf{U}_i|_{i=1}^{M} \in \mathbb{R}^{N_i \times R_i}$ and the core tensor $\mathcal{Y}$.

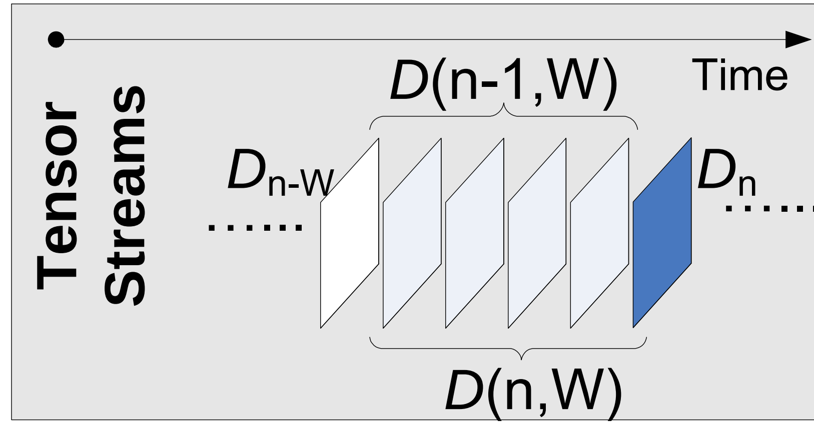
Algorithm:

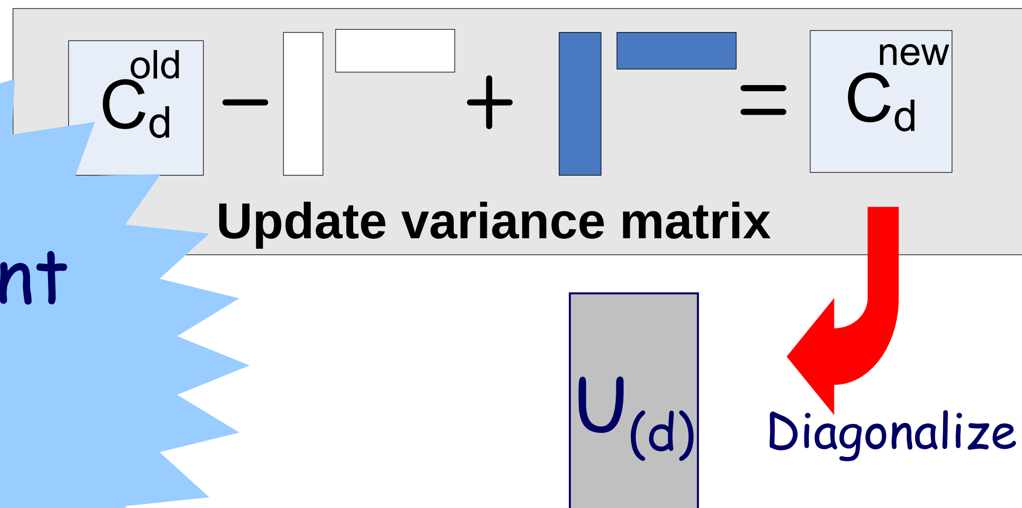
1. Initialize $\mathbf{U}_i|_{i=0}^{M}$
2. Conduct 3 - 5 iteratively
3. For $k = 0$ to M
4. Fix $\mathbf{U}_i$ for $i \neq k$ and find the $\mathbf{U}_k$ that minimizes $d(\mathcal{D}, \mathcal{D} \prod_{i=0}^{M} \times_i (\mathbf{U}_i \mathbf{U}_i^T))$
5. Check convergence
6. Calculate the core tensor $\mathcal{Y} = \mathcal{D} \prod_{i=0}^{M} \times_i \mathbf{U}_i$



Moving Window scheme (MW)

- Update the variance matrix $\mathbf{C}_{(\mathbf{i})}$ **incrementally**
- Diagonalize **C(i) to find U(i)**



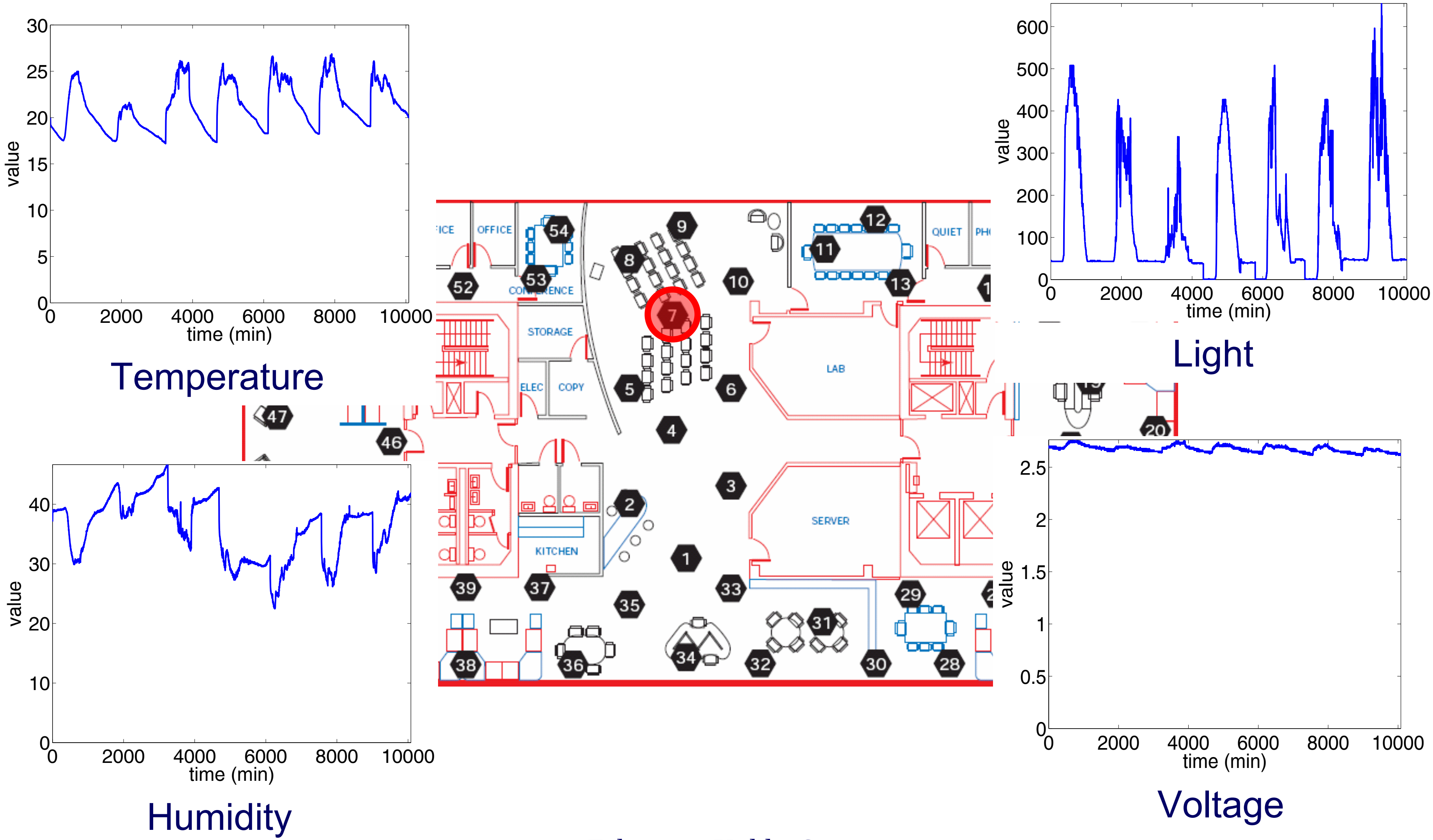


A good and efficient initialization

Roadmap

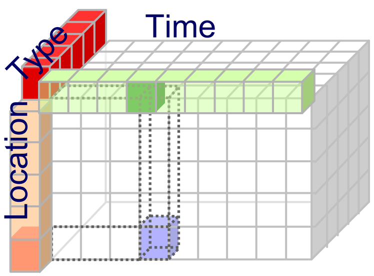
- Motivation
- Matrix tools
- Tensor tools
- Case studies
  - Sensors
  - social networks
  - web mining

P1: Environmental sensor monitoring

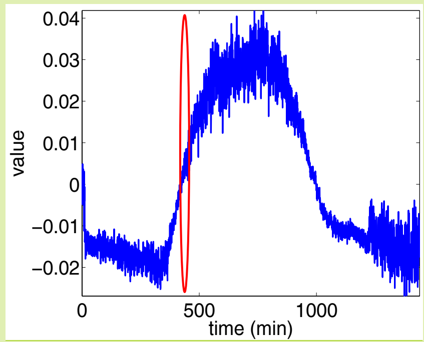


P1: sensor monitoring

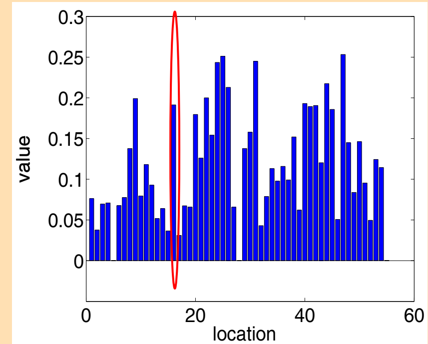
1st factor
Scaling factor 250



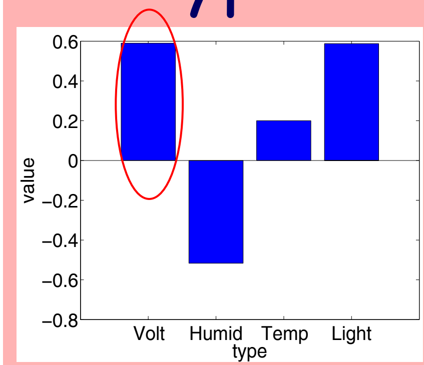
time



location



type



voltage light

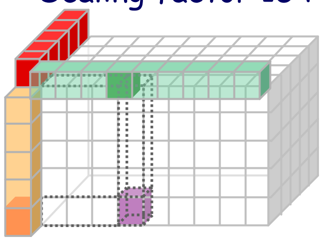
hum.
temp.

- 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

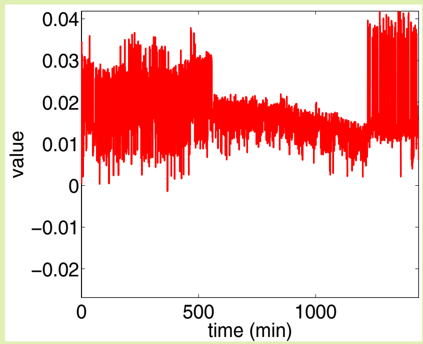
P1: sensor monitoring

2nd factor

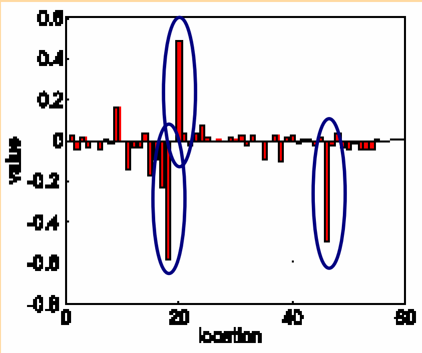
Scaling factor 154



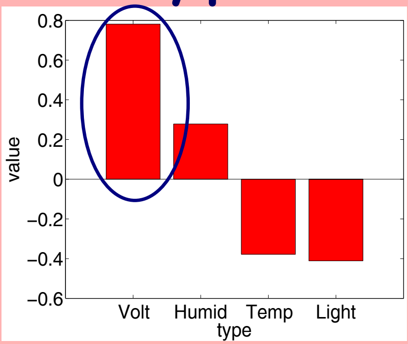
time



location



type



- 2nd factor captures an atypical trend:
  - Uniformly across all time
  - Concentrating on 3 locations
  - Mainly due to voltage

voltage light

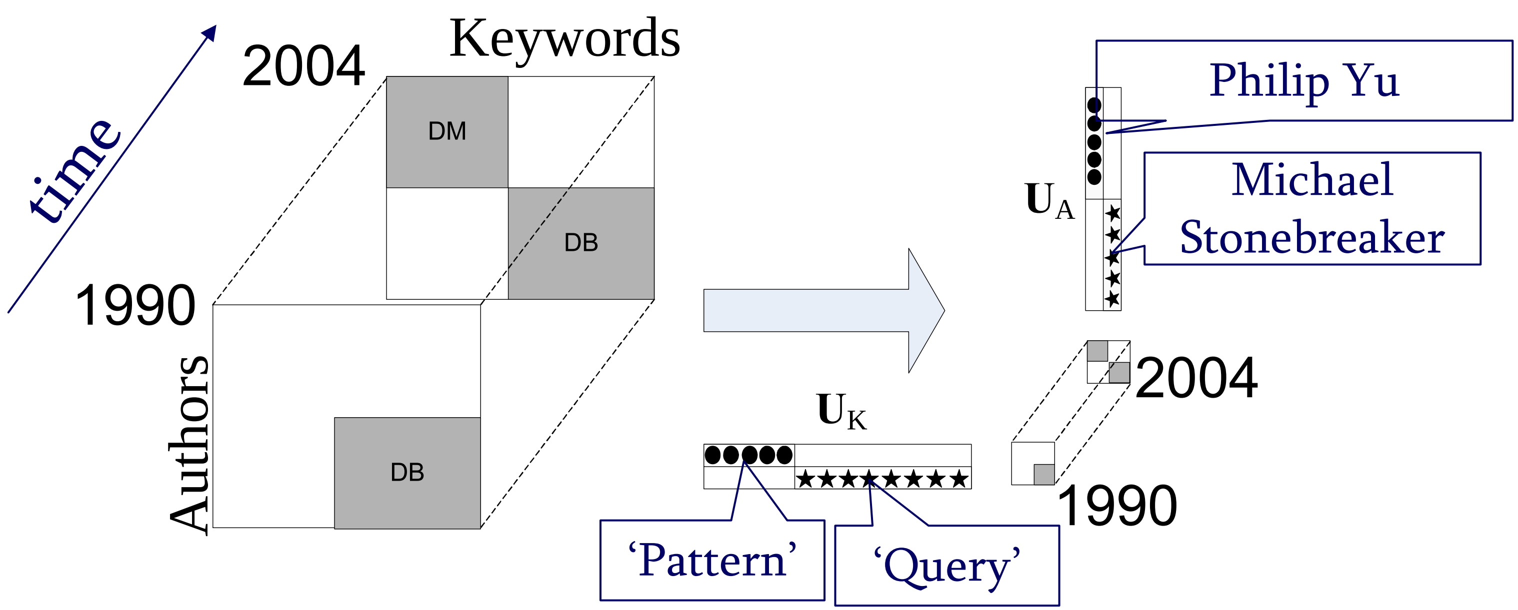
hum.

temp.

- Interpretation: two sensors have low battery, and the other one has high battery.

P3: Social network analysis

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



P3: Social network analysis (cont.)

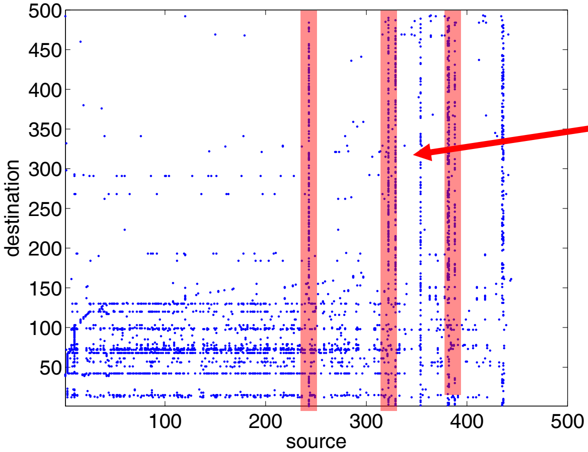
Authors	Keywords	Year
michael carey, michael stonebreaker, h. jagadish, hector garcia-molina	ueri,parallel,optimization,concurr, ctorient	1995
surajit chaudhuri,mitch cherniack,michael stonebreaker,ugur etintemel	distribut,systems,view,storage,servic,process, cache	2004
jiawei han,jian pei,philip s.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

DB

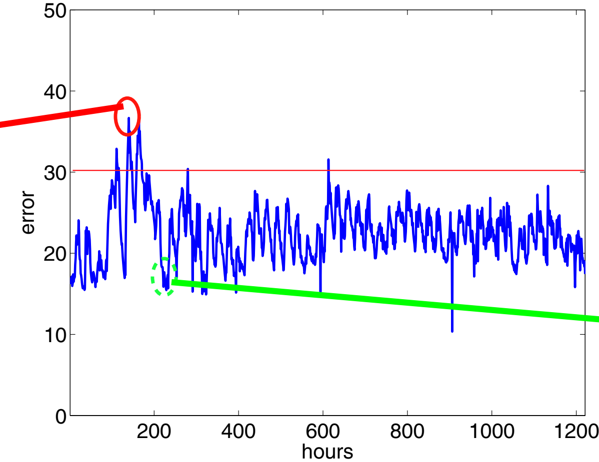
DM

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

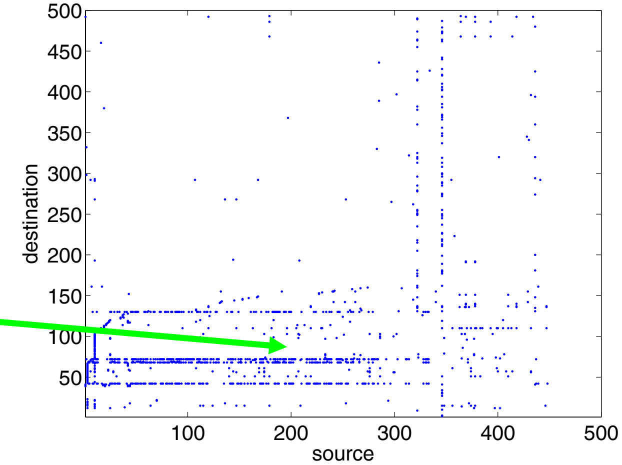
P4: Network anomaly detection



Abnormal traffic



Reconstruction error over time

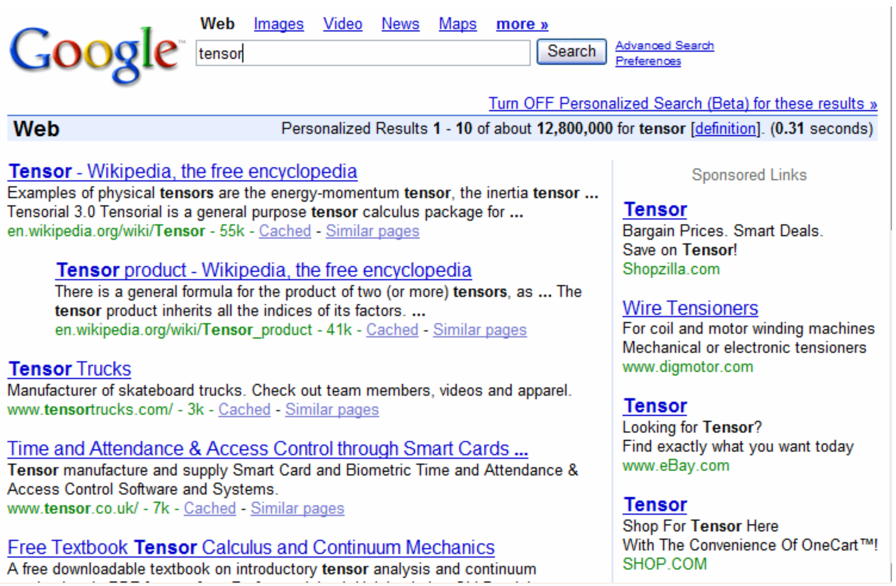


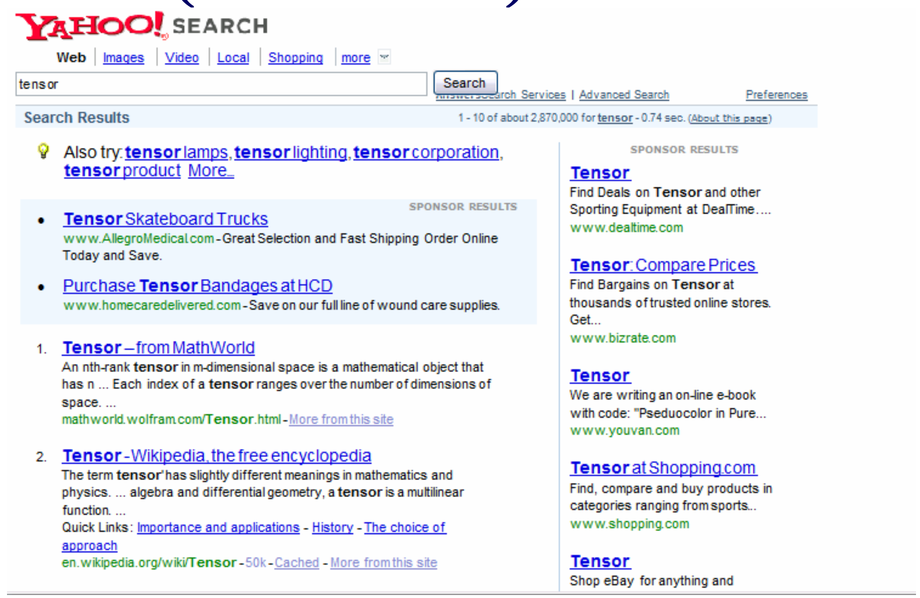
Normal traffic

- 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).

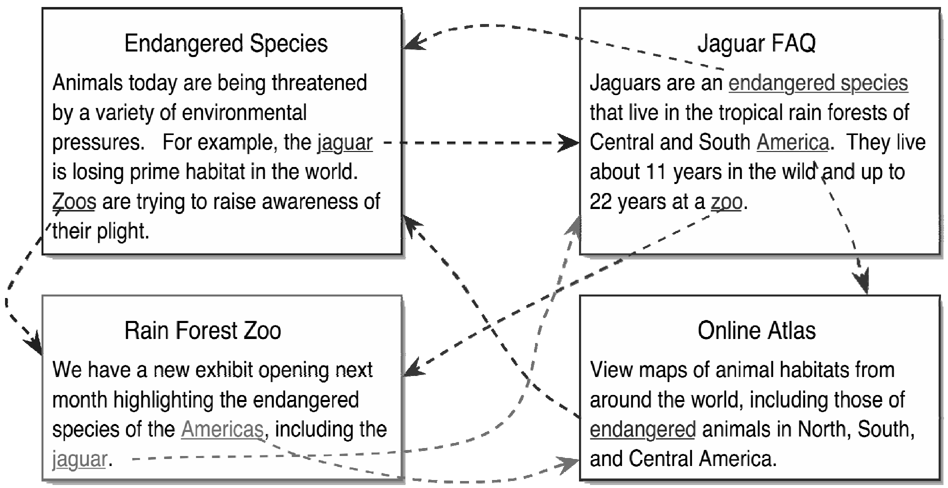
P5: Web graph mining

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





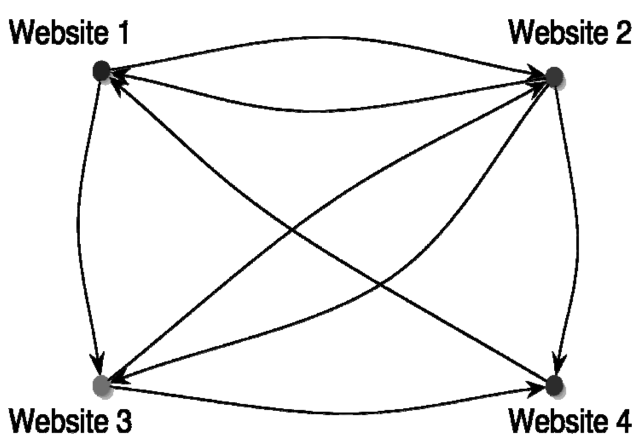
Kleinberg's Hubs and Authorities (the HITS method)

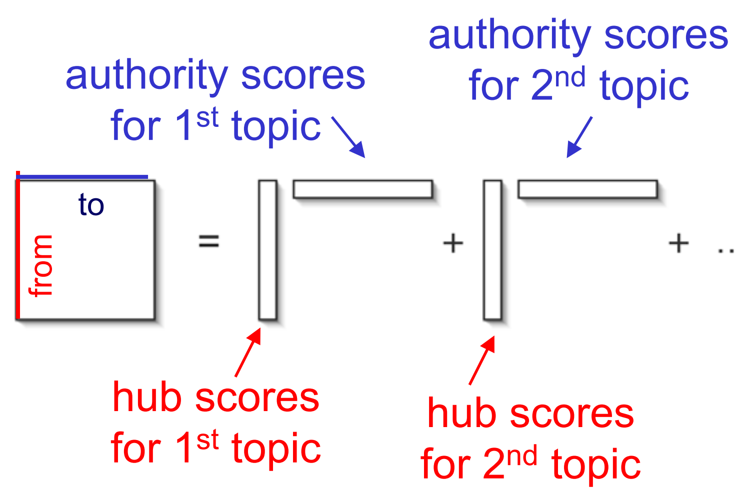


Sparse adjacency matrix and its SVD:

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

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





Kleinberg, JACM, 1999

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.co

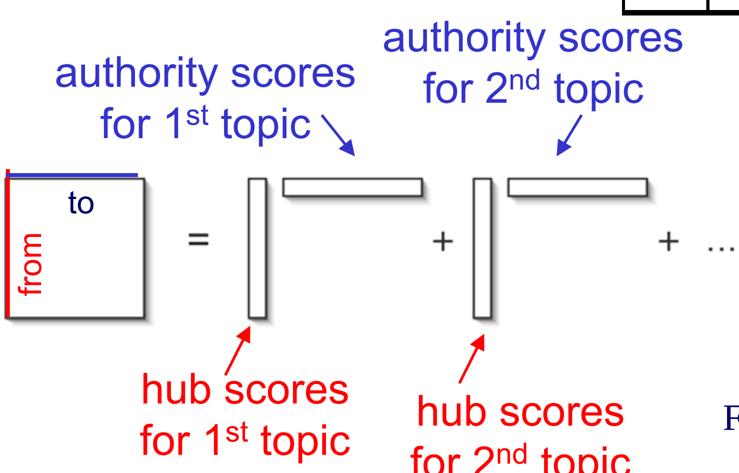
2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh
.06	www.lehigha
.06	www.lehighs
.02	www.bethleh
.02	www.adobe.c
.02	lewisweb.cc.
.02	www.leo.lehi
.02	www.distanc
.02	fp1.cc.lehigh

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.co
.08	www.sun-catalo
.08	news.com.com

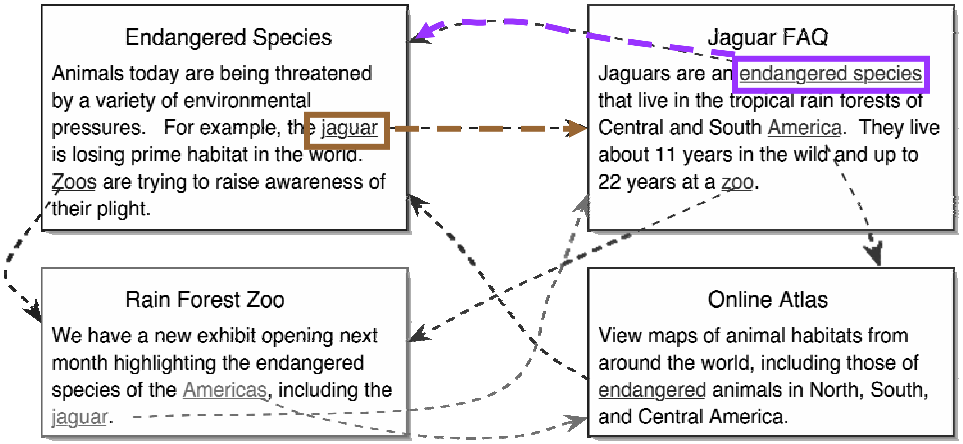
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state
.22	www.gsa.g
.20	www.ssa.g
.16	www.censu
.14	www.govbe
.13	www.kids.g
.13	www.usdoj

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

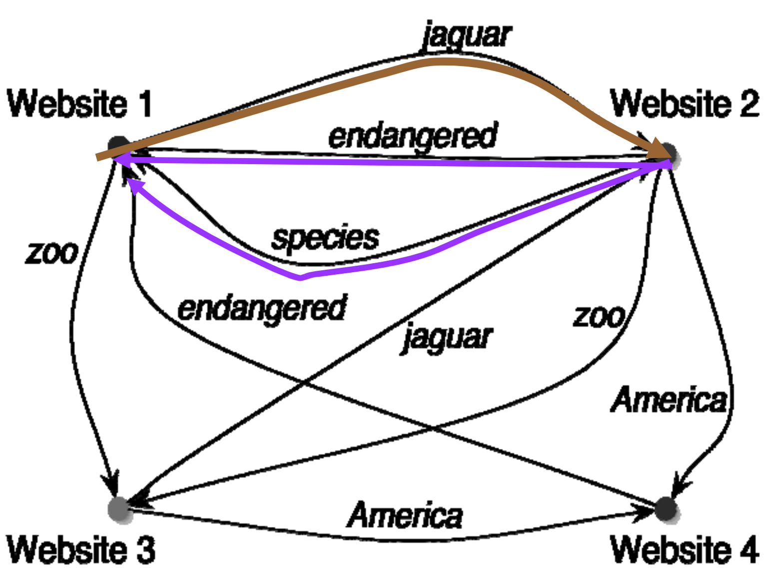
We started our crawl from http://www-neos.mcs.anl.gov/neos, and crawled 4700 pages, resulting in 560 cross-linked hosts.



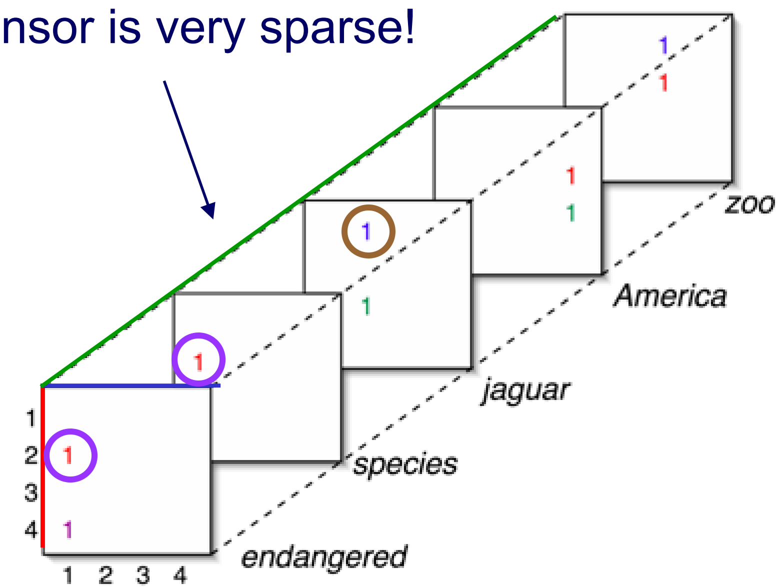
Three-Dimensional View of the Web



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



Observe that this tensor is very sparse!

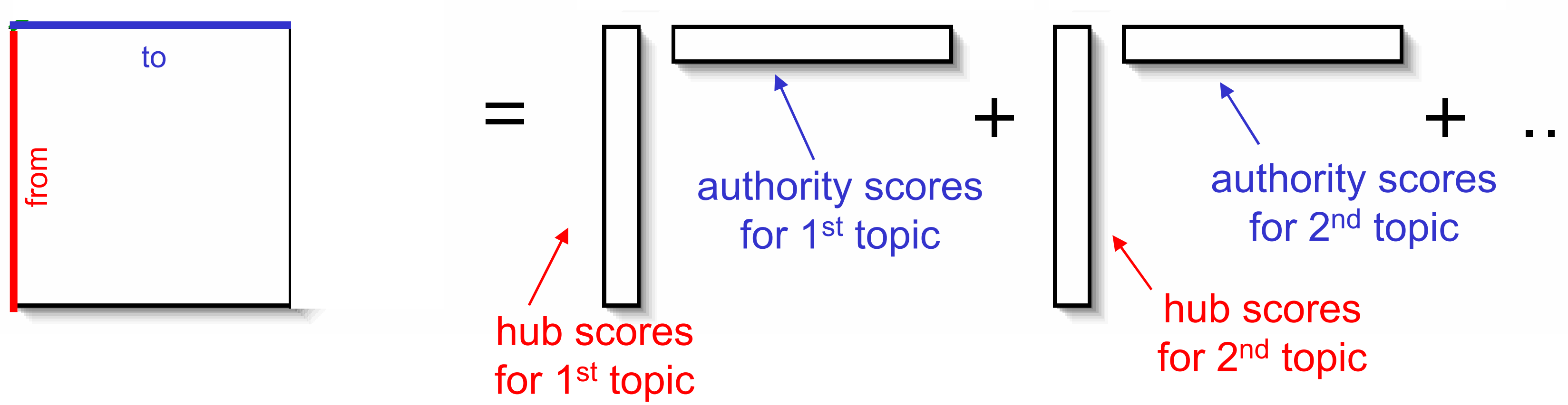


Kolda, Bader, Kenny, ICDM05

Topical HITS (TOPHITS)

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

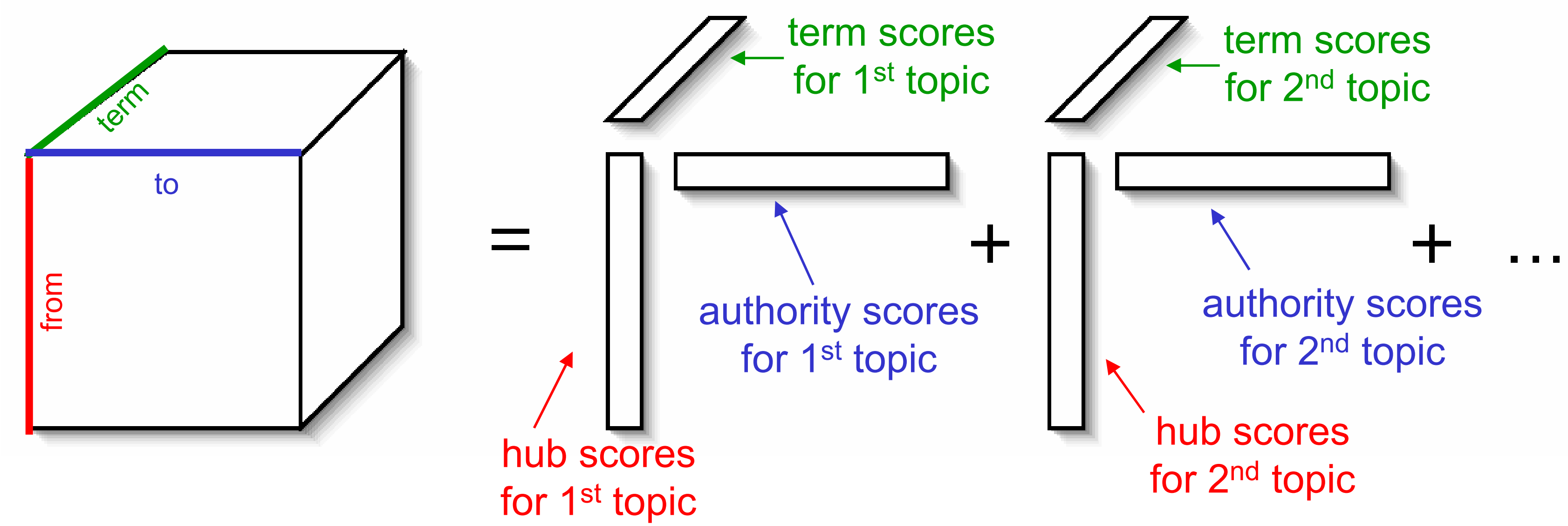
$$\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{h}_r \circ \mathbf{a}_r$$



Topical HITS (TOPHITS)

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

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



TOPHITS Terms & Authorities on Sample Data

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

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

w_k = # unique links using term k

1st Principal Factor

.23	JAVA	.86	java.sun.com
.18	SUN		
.17	PLATF		
.16	SOLAR		
.16	DEVEL		
.15	EDITIO		
.15	DOWN		
.14	INFO		
.12	SOFTW		
.12	NO-RE		

2nd Principal Factor

.20	NO-READABLE-TEXT	.99	www.lehigh.edu
.16	FACUL		
.16	SEARC		
.16	NEWS		
.16	LIBRAR		
.16	COMP		
.12	LEHIG		

3rd Principal Factor

.15	NO-READABLE-TEXT	.97	www.ibm.com
.15	IBM	.18	www.alphaworks.ibm.com
.12	SERVI		
.12	WEBS		
.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	OTHER		
.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	WHIT		
.16	U.S		
.15	HOUS		
.13	BUDG		
.13	PRES		
.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	CONTR		
.05	ILOG		

13th Principal Factor

.46	ADOBE	.99	www.adobe.com
.45	READER		
.45	ACROB		
.30	FREE		
.30	NO-RE		
.29	HERE		
.29	COPY		
.05	DOWN		

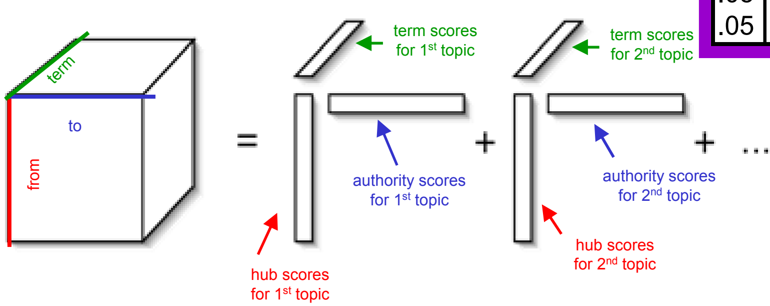
16th Principal Factor

.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	SEVER		
.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

Tensor PARAFAC



Faloutsos, Kolda, Sun

Summary

Methods	Pros	Cons	Applications
SVD, PCA	Optimal in L2 and Frobenius	Dense representation, Negative entries	LSI, PageRank, HITS
CUR, CMD	Interpretability, sparse bases	Not optimal like SVD, dense core	DNA SNP data, network forensics
Co-clustering	Interpretability	Local minimum	Social networks, microarray data
Tucker	Flexible representation	Interpretability, non-uniqueness, dense core	TensorFaces
PARAFAC	Interpretability, efficient parse computation	Slow convergence	TOPHISTS
Incrementalization	Efficiency	Non-optimal	Tensor Streams

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

Acknowledgements

- Cross-association foils:
  - from Dr. Deepayan Chakrabarti (Yahoo!)

- Funding:

National Science Foundation
WHERE DISCOVERIES BEGIN

LAWRENCE LIVERMORE NATIONAL LABORATORY
Science in the National Interest

IBM

PITA (PA Inf. Tech. Alliance)

Thank you!

- Christos Faloutsos
  www.cs.cmu.edu/~christos

- Tamara Kolda
  csmr.ca.sandia.gov/~tgkolda

- Jimeng Sun
  www.cs.cmu.edu/~jimeng

www.cs.cmu.edu/~christos/TALKS/SIGMOD-07-tutorial/

3h version: www.cs.cmu.edu/~christos/TALKS/SDM-tut-07/