



Roadmap

- Motivation
 - Matrix tools
 - **Tensor tools**
 - Case studies
- Basics
 - Tucker
 - PARAFAC
 - **Incrementalization**





Incrementalization

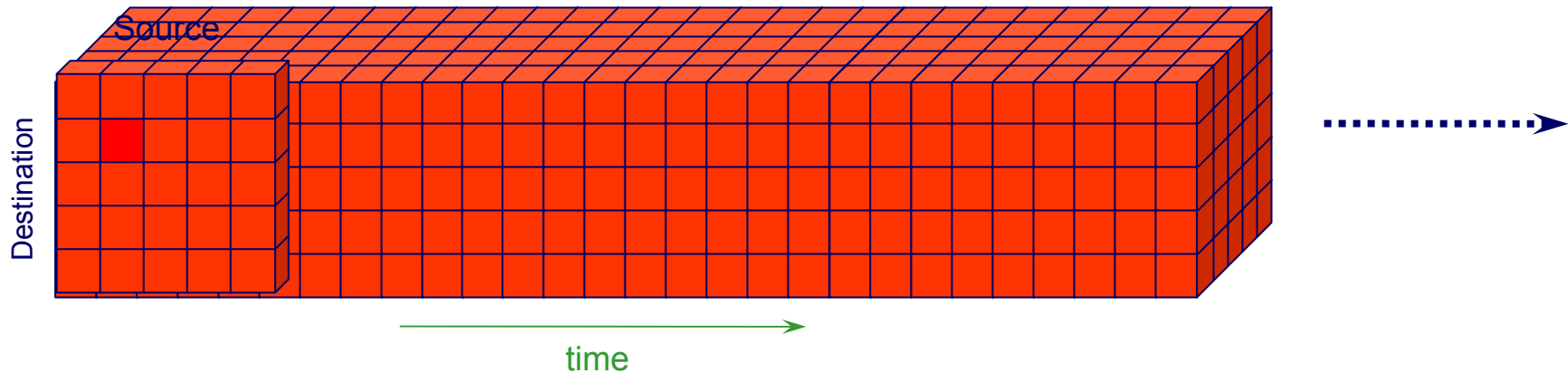


Incremental Tensor Decomposition

- 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**?

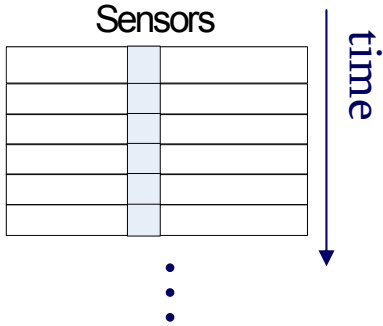
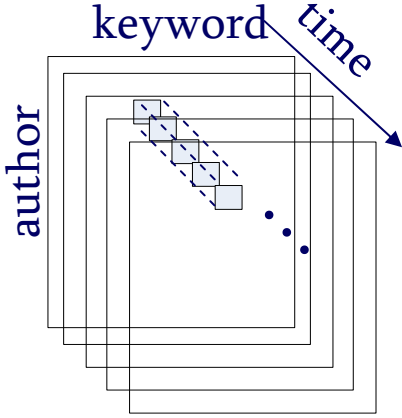
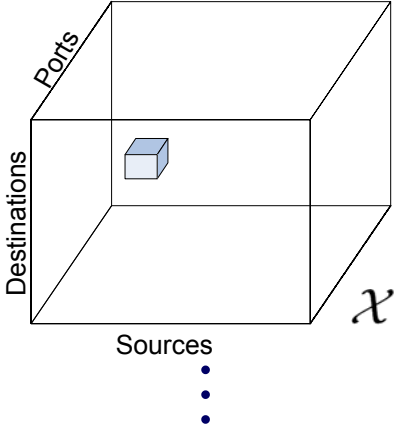


Dynamic Data model

- Tensor Streams
 - A sequence of Mth order tensor

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

n is increasing over time

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



Incremental Tensor Decomposition

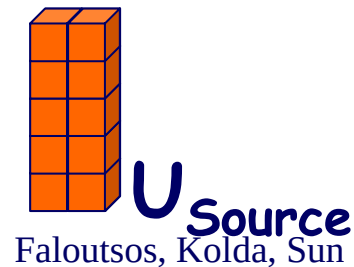
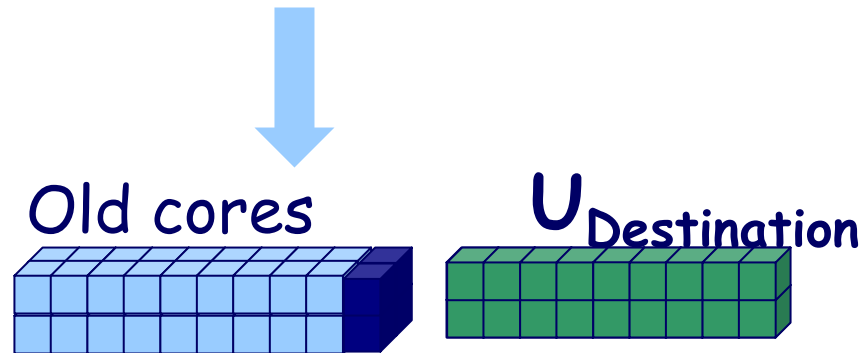
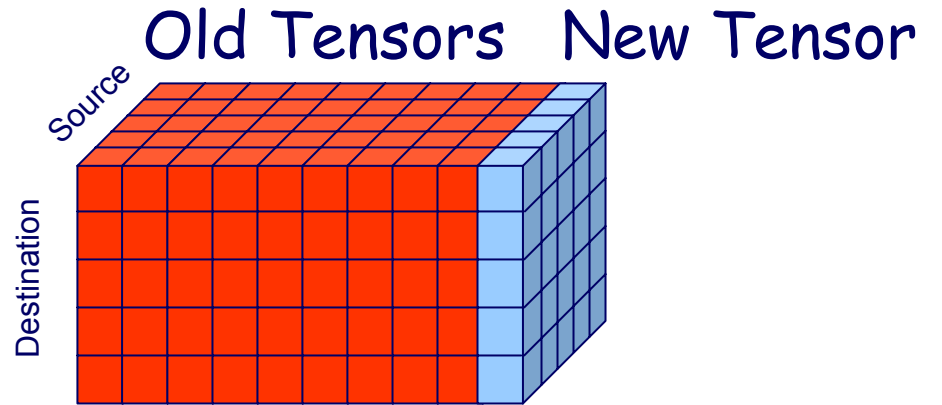
☺ Dynamic data model

- 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



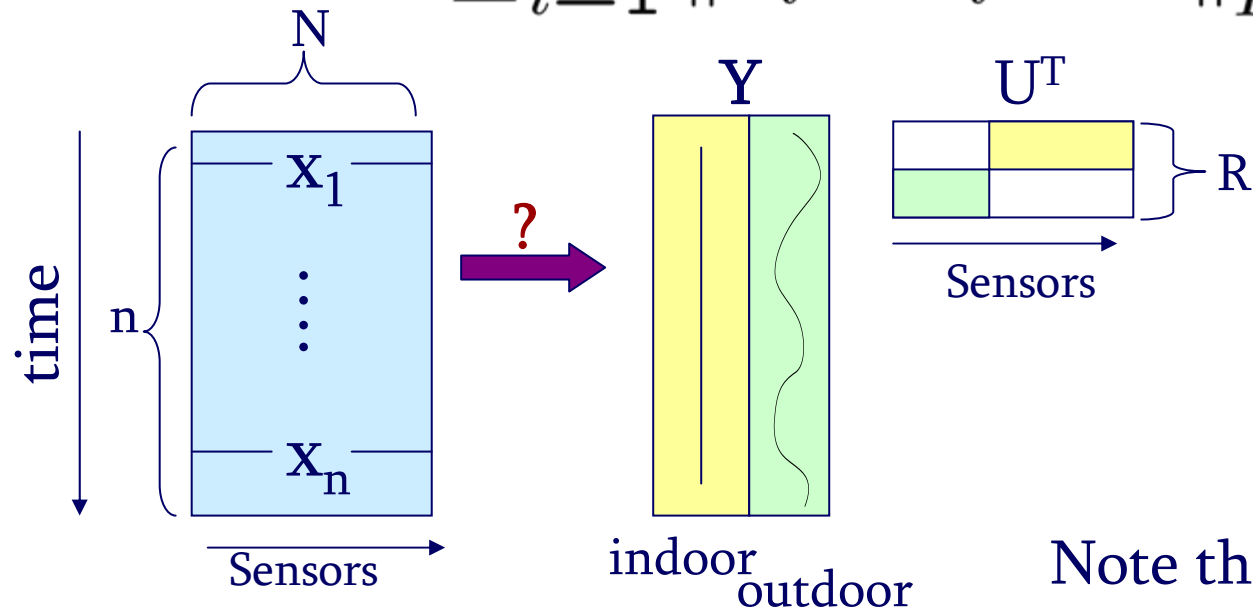


1st order DTA - problem

Given $x_1 \dots x_n$ where each $x_i \in \mathbb{R}^N$, find

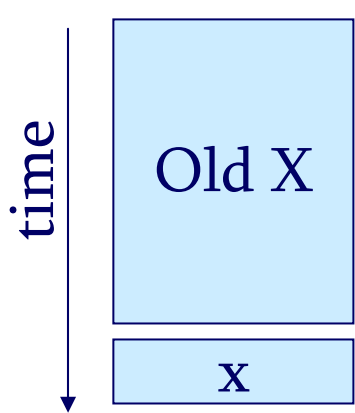
$U \in \mathbb{R}^{N \times R}$ such that the error e is

small: $e = \sum_{i=1}^n \|x_i - x_i U U^T\|_F^2$





1st order Dynamic Tensor Analysis

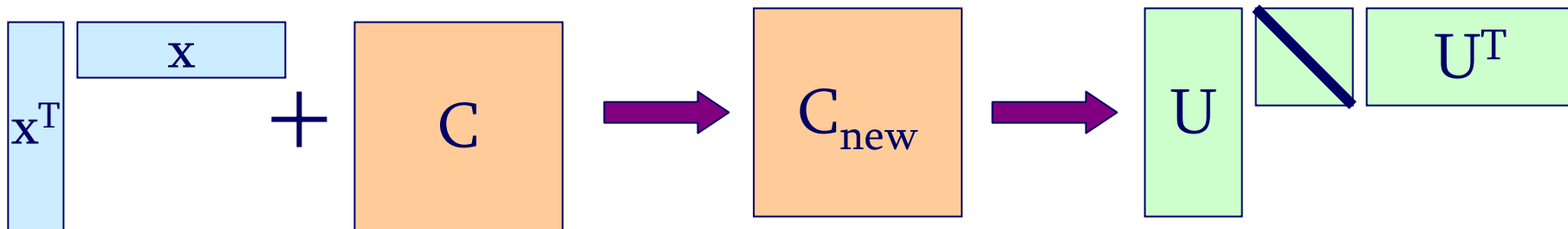


Input: new data vector $x \in \mathbb{R}^N$, old variance matrix $C \in \mathbb{R}^{N \times N}$

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

Algorithm:

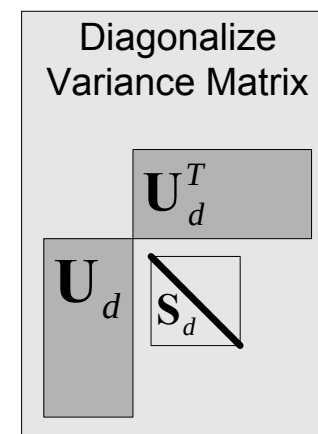
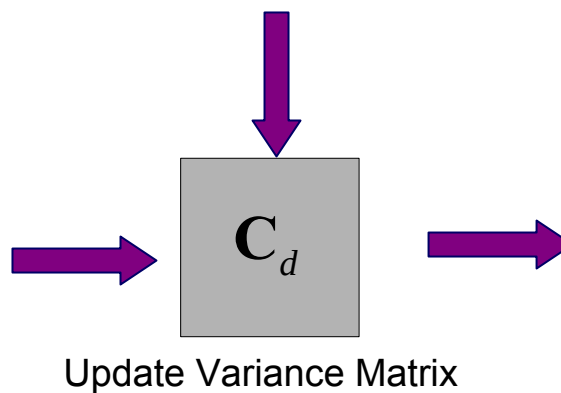
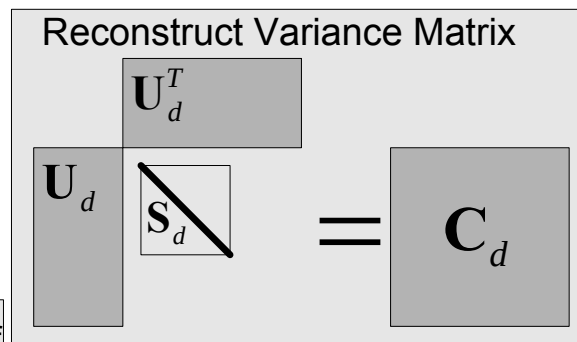
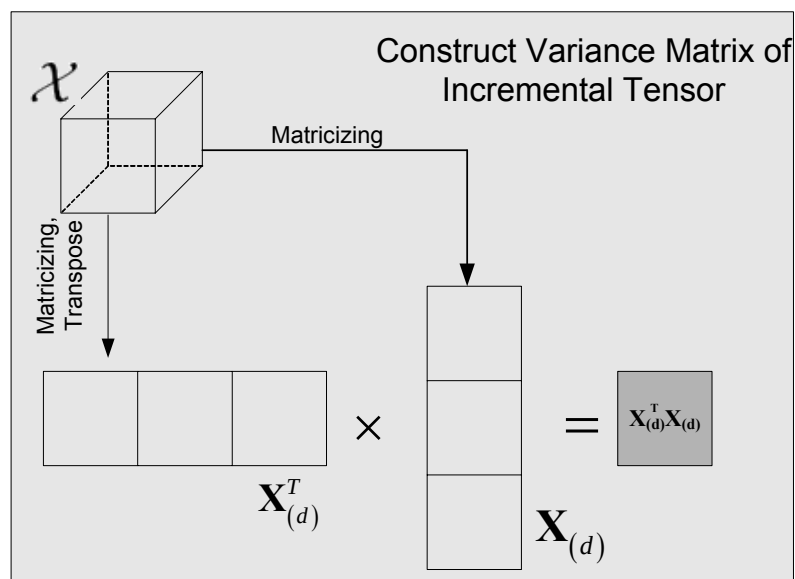
1. update variance matrix $C_{\text{new}} = x^T x + C$
2. Diagonalize $U \Lambda U^T = C_{\text{new}}$
3. Determine the rank R and return U



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



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*

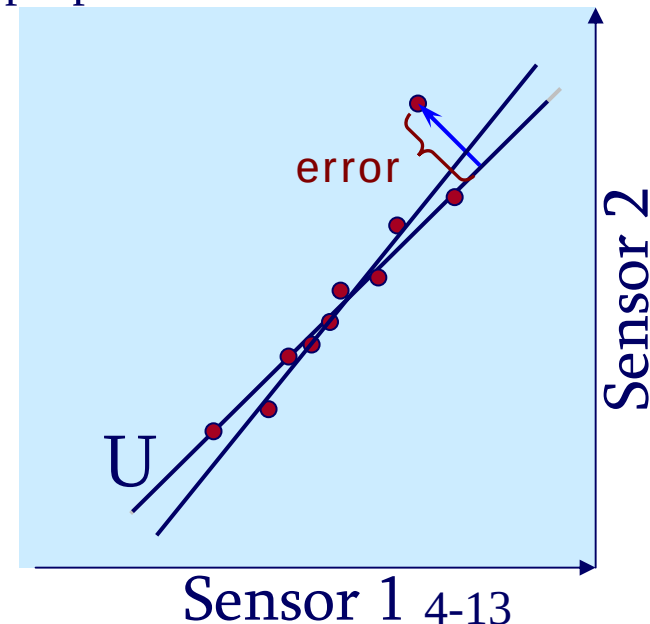


1st order Streaming Tensor Analysis (STA)

- 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

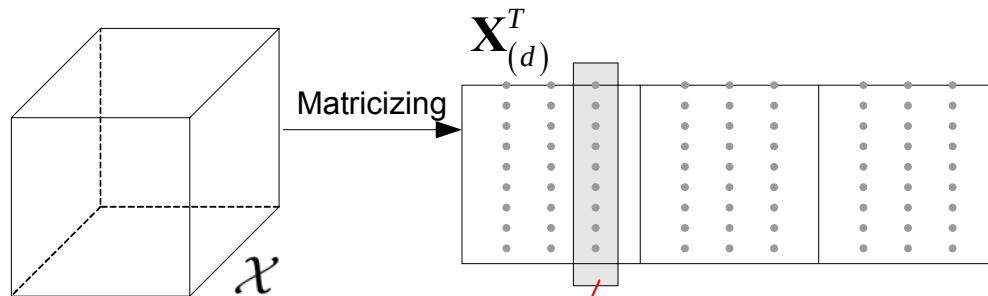
For each new point x and for $i = 1, \dots, k$:

- $y_i := U_i^T x$ (proj. onto U_i)
- $d_i \leftarrow \lambda d_i + y_i^2$ (energy $\propto i$ -th eigenval.)
- $e_i := x - y_i U_i$ (error)
- $U_i \leftarrow U_i + (1/d_i) y_i e_i$ (update estimate)
- $x \leftarrow x - y_i U_i$ (repeat with remainder)

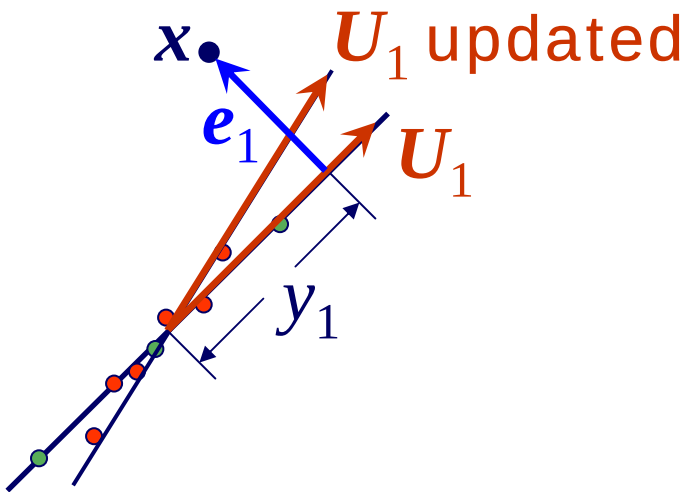




M^{th} order STA



- 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





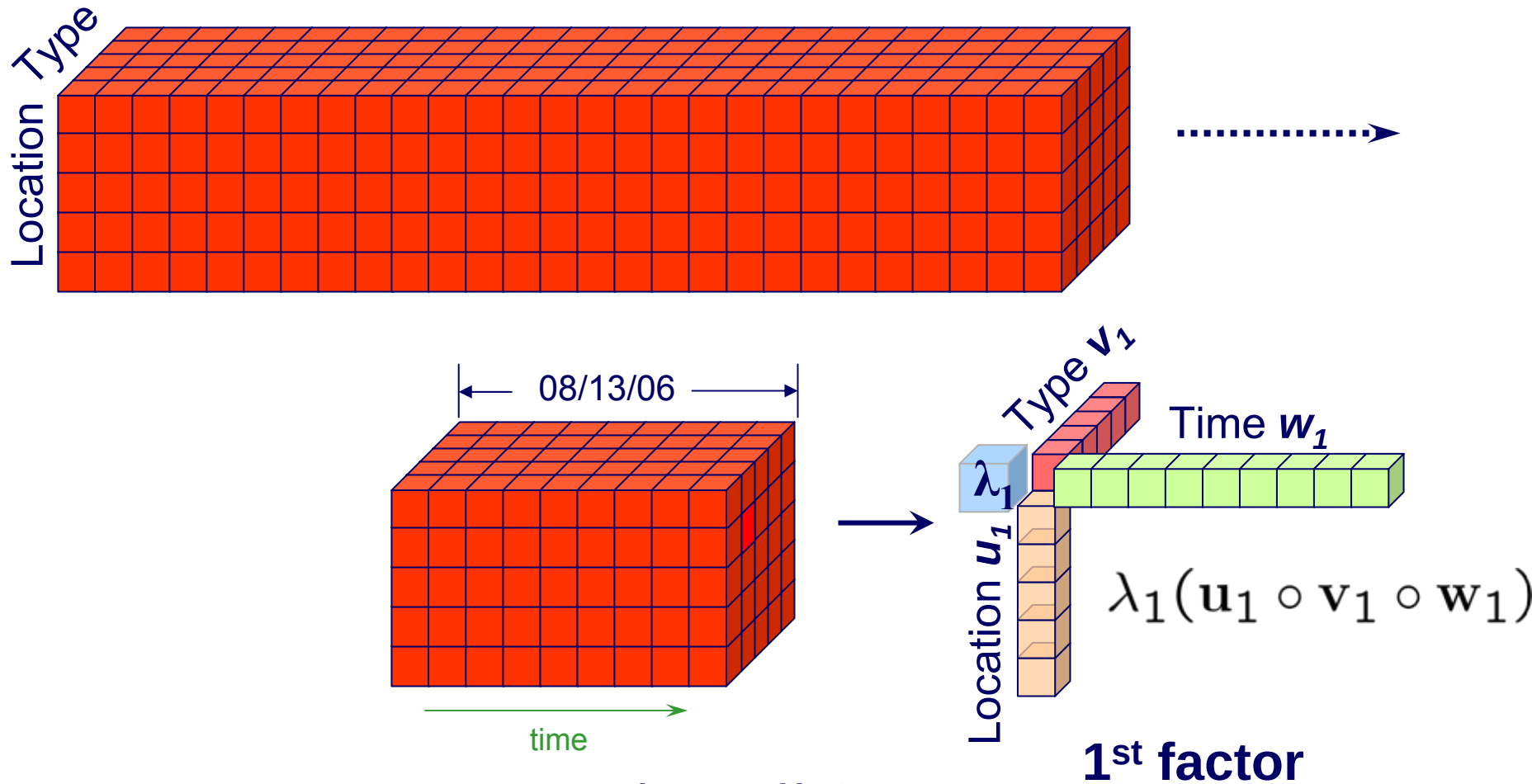
Incremental Tensor Decomposition

- ☺ 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)





Meta-algorithm for window-based tensor analysis

Input:

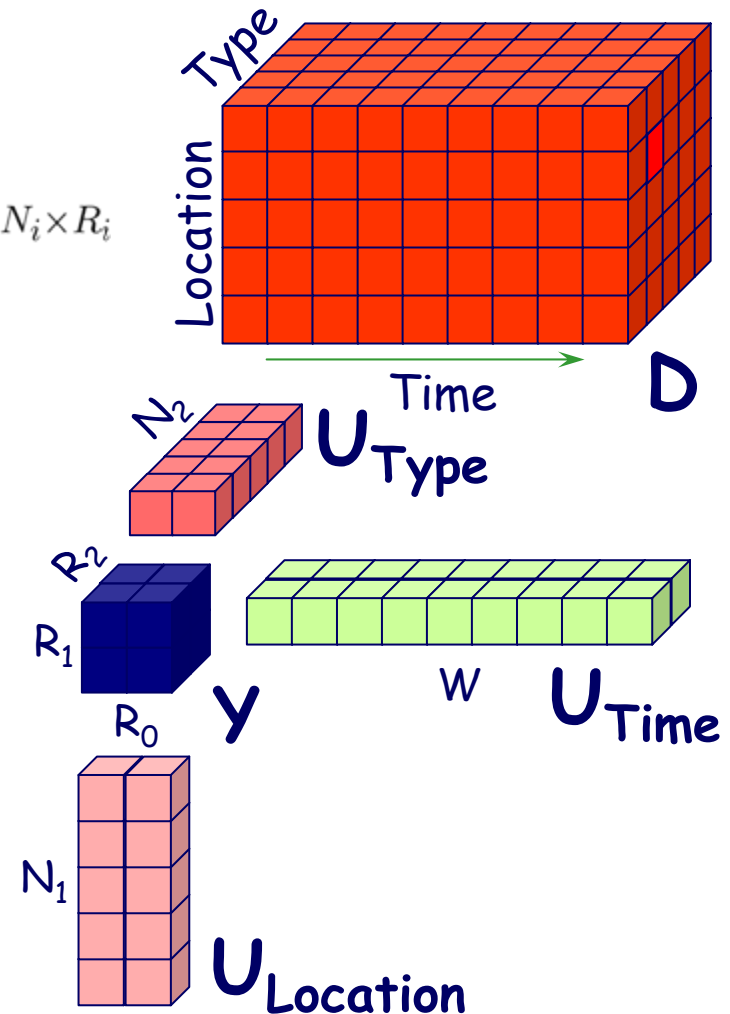
The tensor window $\mathcal{D} \in \mathbb{R}^{W \times N_1 \times \dots \times N_M}$

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:

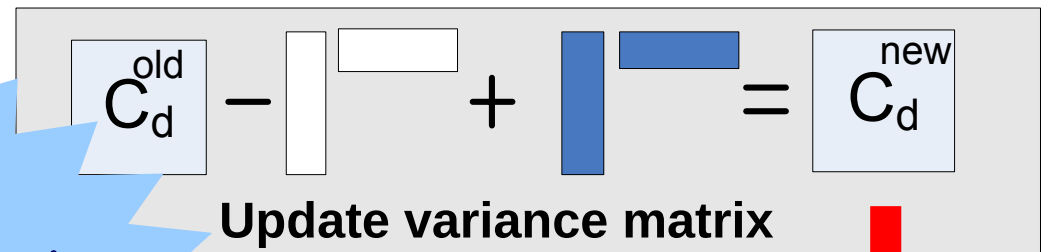
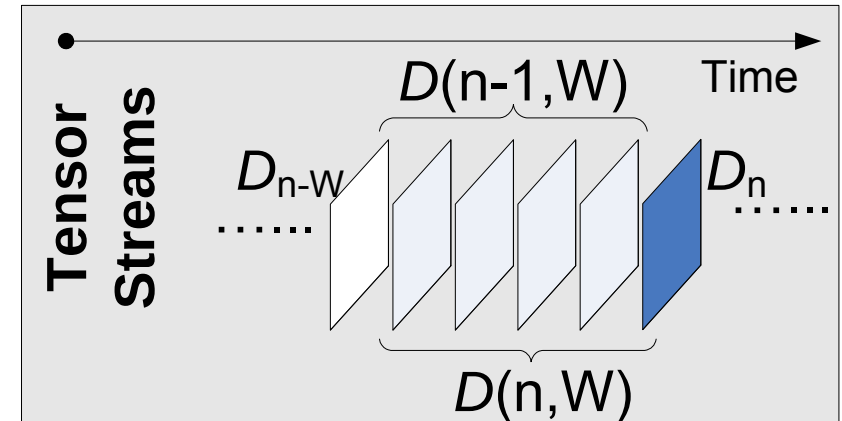
1. Initialize $\mathbf{U}_i|_{i=0}^M$
2. Conduct 3 - 5 iteratively
3. For $k = 0$ to M
4. Fix $\mathbf{U}_i$ for $i \neq k$ and find the $\mathbf{U}_k$ that minimizes $d(\mathcal{D}, \mathcal{D} \prod_{i=0}^M \times_i (\mathbf{U}_i \mathbf{U}_i^T))$
5. Check convergence
6. Calculate the core tensor $\mathcal{Y}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



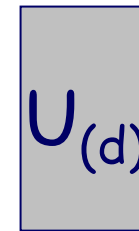


Moving Window scheme (MW)

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



A good and efficient initialization





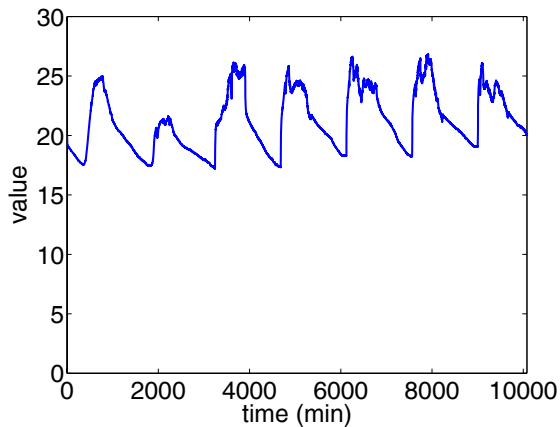
Roadmap

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

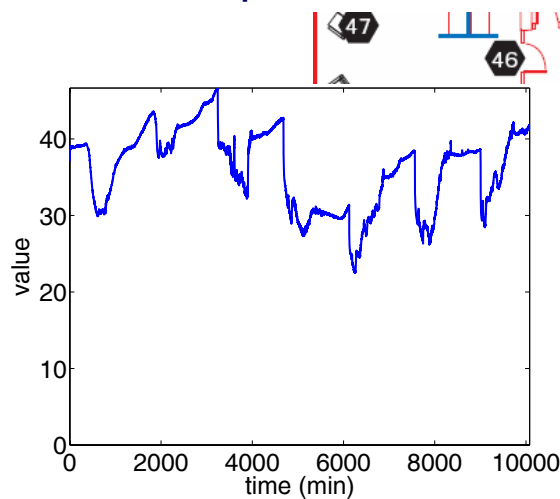




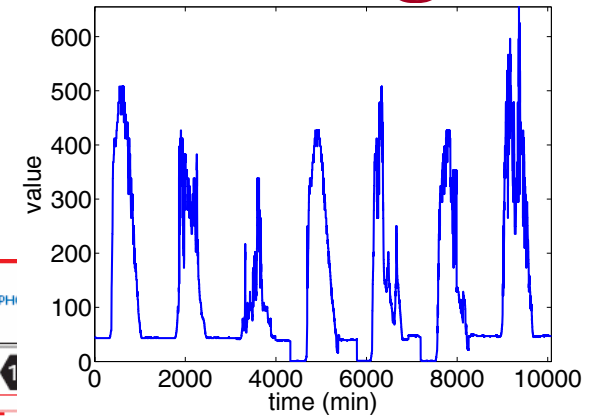
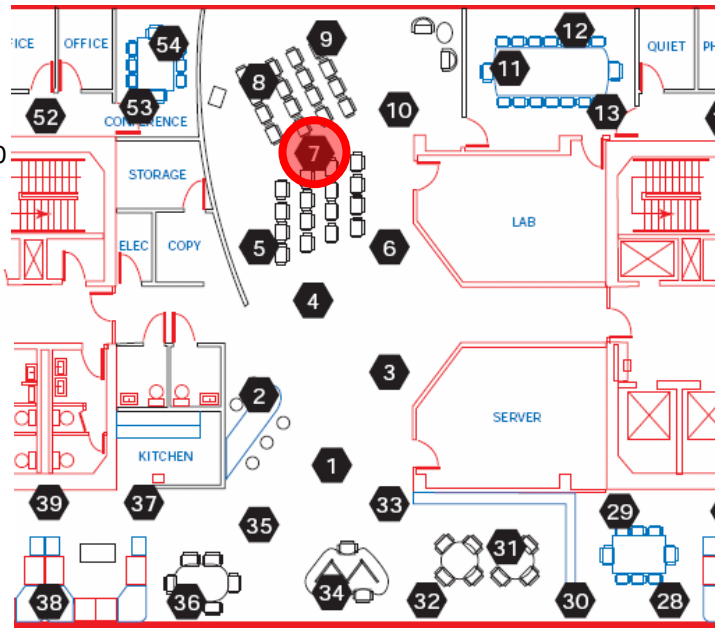
P1: Environmental sensor monitoring



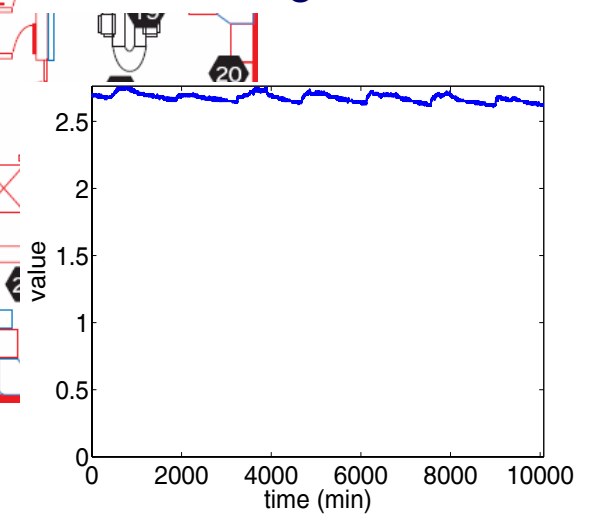
Temperature



Humidity



Light



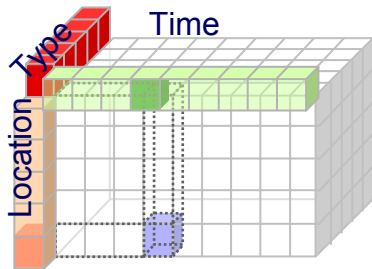
Voltage



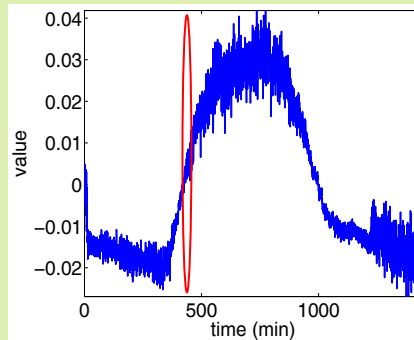
P1: sensor monitoring

1st factor

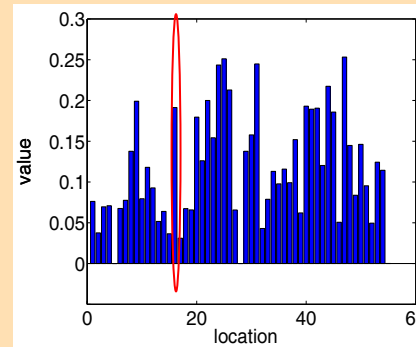
Scaling factor 250



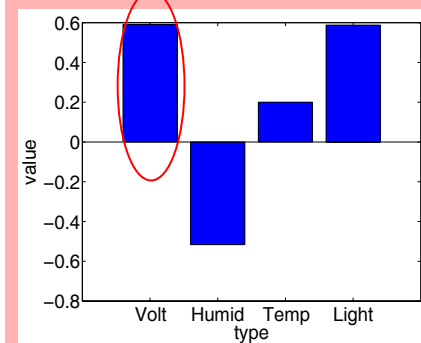
time



location



type



- 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

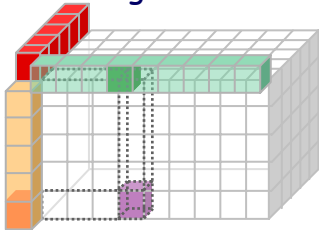
voltage light
hum.
temp.



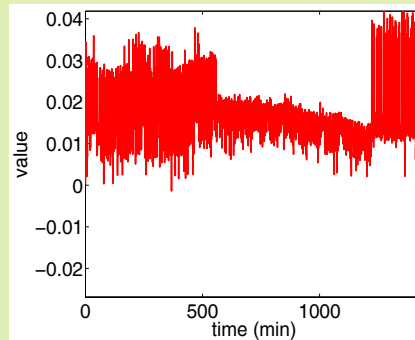
P1: sensor monitoring

2nd factor

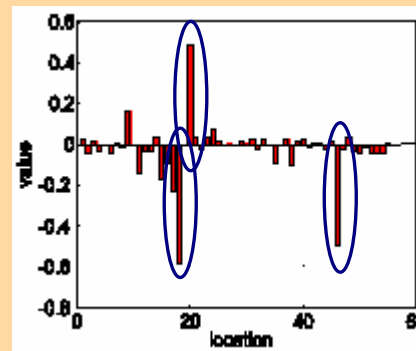
Scaling factor 154



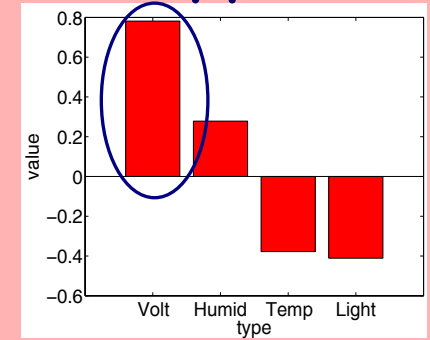
time



location



type

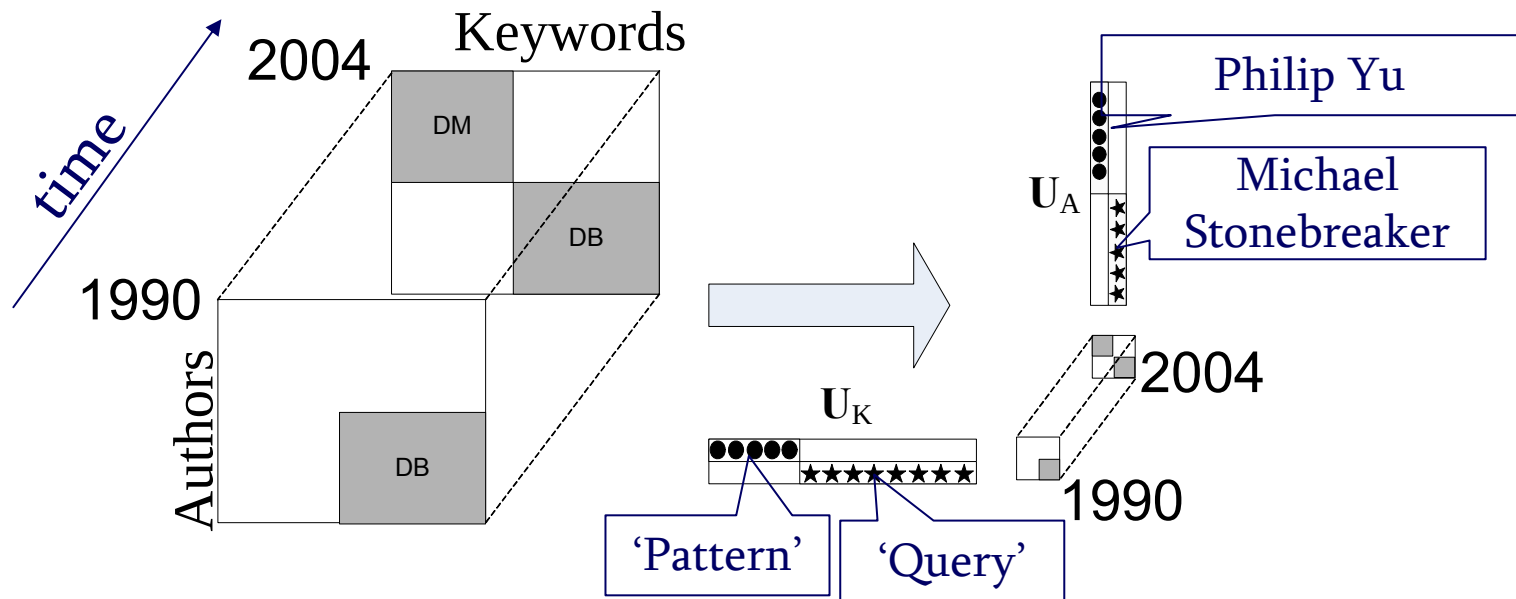


- 2nd factor captures an atypical trend:
 - Uniformly across all time
 - Concentrating on 3 locations
 - Mainly due to voltage
 - Interpretation: two sensors have low battery, and the other one has high battery.
- voltage light
hum.
temp.



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	query, parallel, optimization, concurr, distributed	1995
surajit chaudhuri, mitch cherniack, michael stonebreaker, ugur etintemel	distributed systems, view, storage, servic, process, cache	2004
jiawei han, jian pei, philip s. yu, jianyong wang, charu c. aggarwal	pattern, support, cluster, miner, query	2004

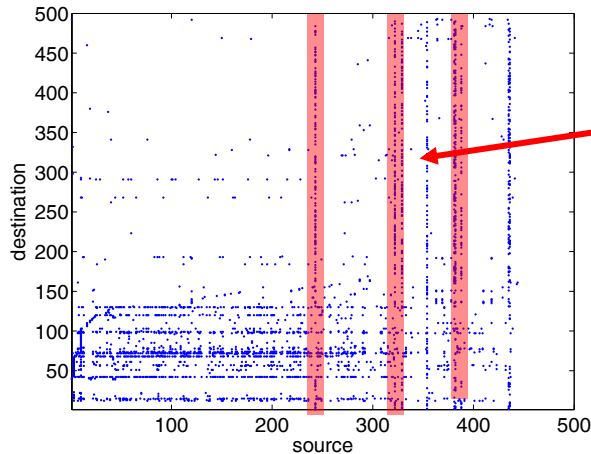
DB

DM

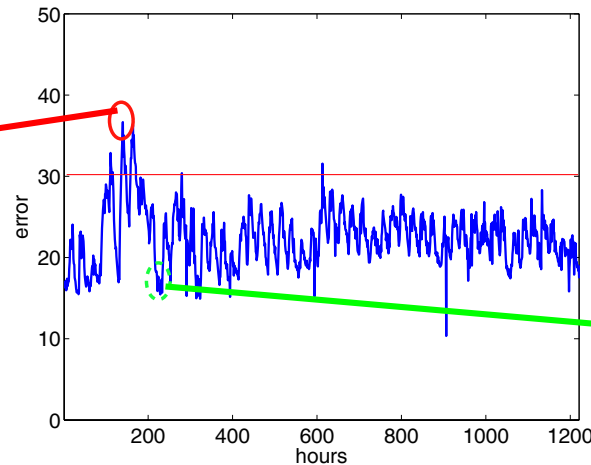
- Two groups are correctly identified: Databases and Data mining
- People and concepts are drifting over time



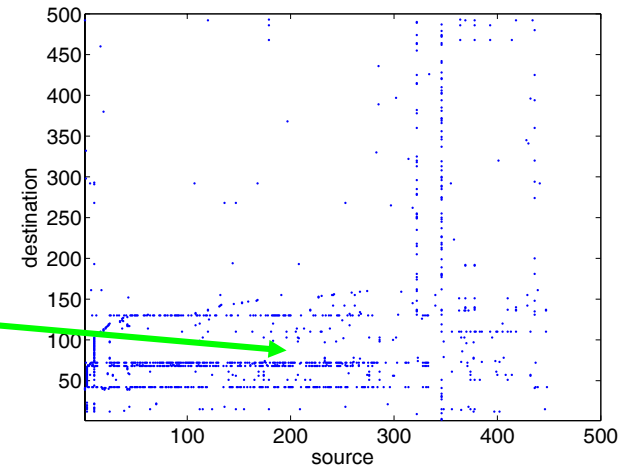
P4: Network anomaly detection



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)

Google Web Images Video News Maps more »

tensor Search Advanced Search Preferences

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.tensortrucks.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

YAHOO! SEARCH

Web Images Video Local Shopping more »

tensor Search

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

Also try: [tensorlamps](#), [tensorlighting](#), [tensorcorporation](#), [tensorproduct](#) [More...](#)

Tensor Skateboard Trucks
www.AllegroMedical.com - Great Selection and Fast Shipping Order Online Today and Save.

Purchase Tensor Bandages atHCD
www.homecaredelivered.com - Save on our full line of wound care supplies.

1. **Tensor--from MathWorld**
An n -rank **tensor** in m -dimensional space is a mathematical object that has $n \dots$ 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](#)

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

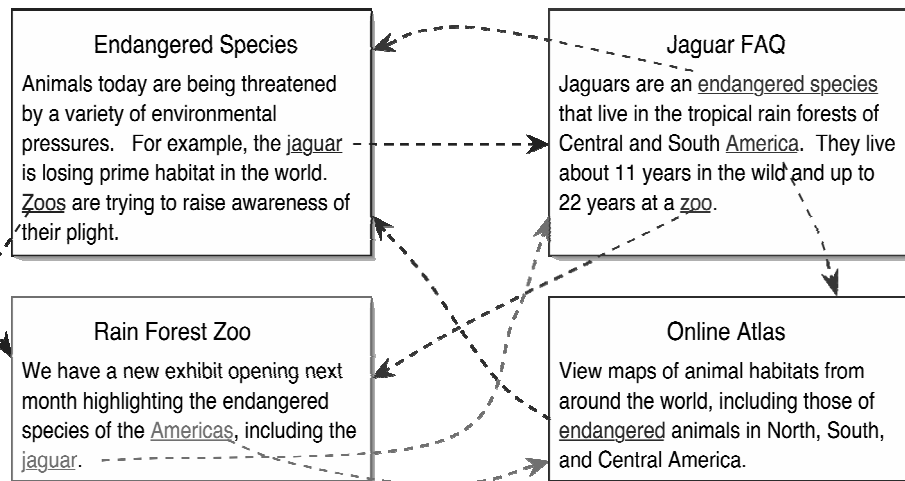
Tensor
We are writing an on-line e-book with code: "Pseduocolor in Pure...
www.youvan.com

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

Tensor
Shop eBay for anything and



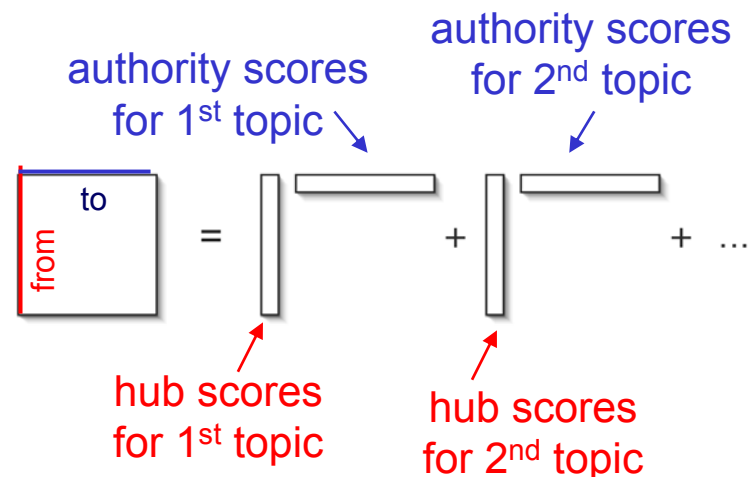
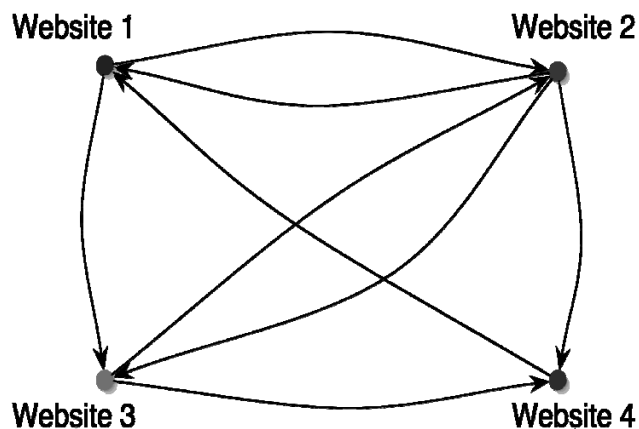
Kleinberg's Hubs and Authorities (the HITS method)



Sparse adjacency matrix and its SVD:

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

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





HITS Authorities on Sample Data

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

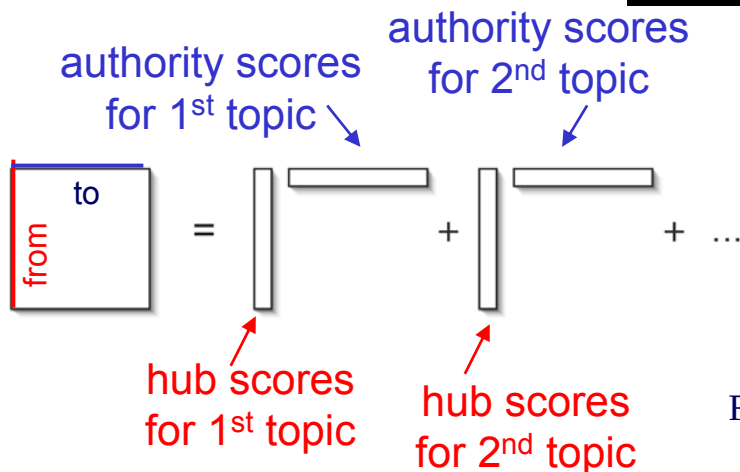
1st Principal Factor	
.97	www.ibm.com
.24	www.alphaw.com
.08	www-128.ibm.com
.05	www.develop.com
.02	www.research.com
.01	www.redbook.com
.01	news.com.com

2nd Principal Factor	
.99	www.lehigh.edu
.11	www2.lehigh.edu
.06	www.lehigh.edu
.06	www.lehigh.edu
.02	www.bethlehigh.edu
.02	www.adobe.com
.02	lewisweb.cc
.02	www.leo.lehigh.edu
.02	www.distance.edu
.02	fp1.cc.lehigh.edu

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

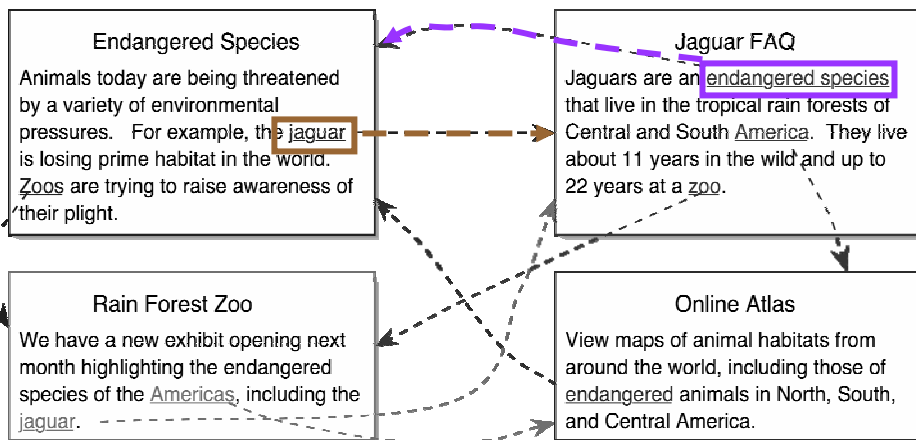
4th Principal Factor	
.60	www.pueblo.gsa.gov
.45	www.whitehouse.gov
.35	www.irs.gov
.31	travel.state.gov
.22	www.gsa.gov
.20	www.ssa.gov
.16	www.census.gov
.14	www.govbeat.com
.13	www.kids.gov
.13	www.usdoj.gov

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



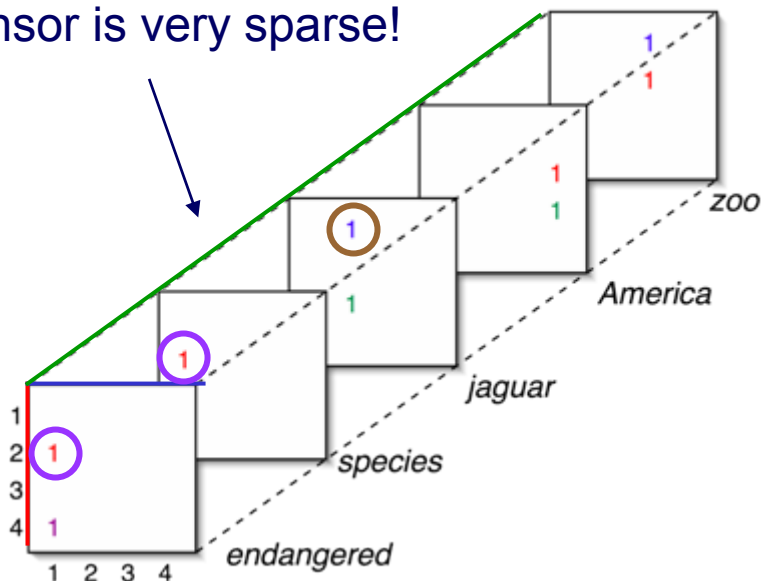
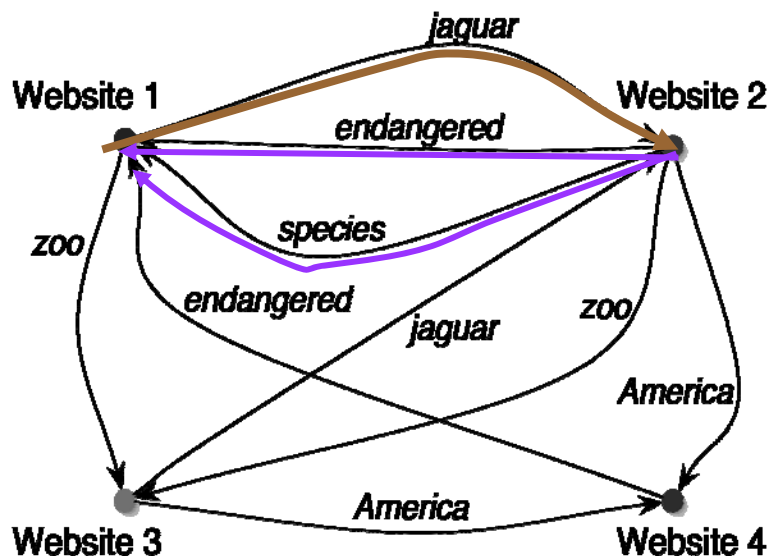


Three-Dimensional View of the Web



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

Observe that this tensor is very sparse!

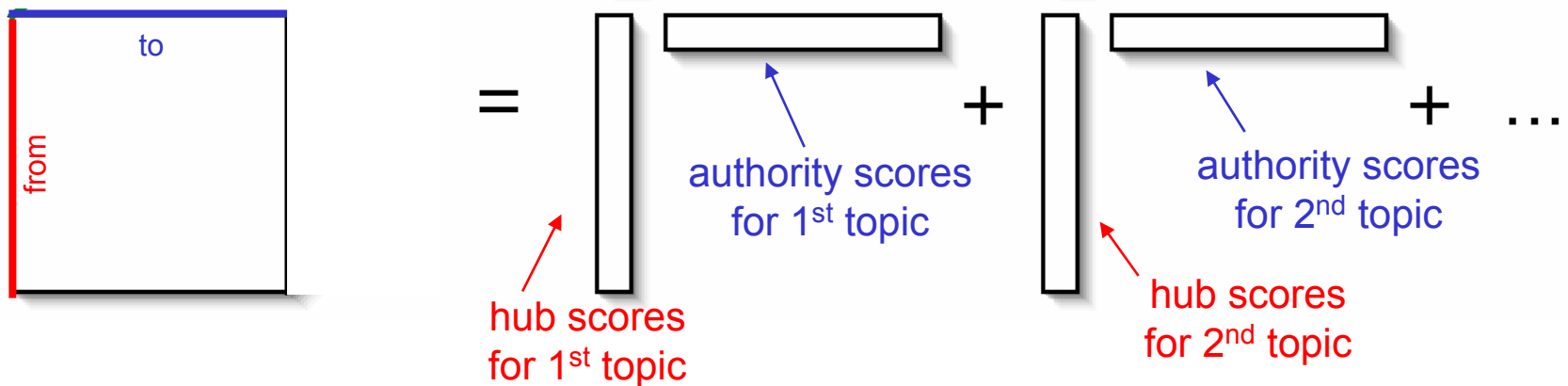




Topical HITS (TOPHITS)

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

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

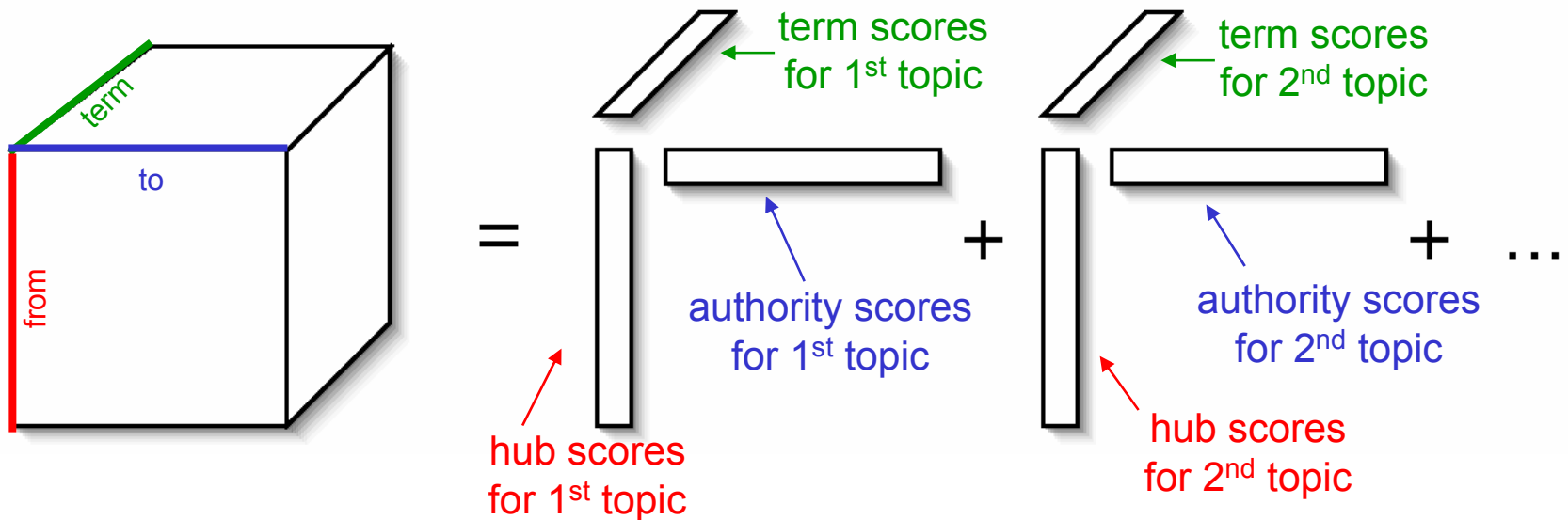




Topical HITS (TOPHITS)

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

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





TOPHITS Terms & Authorities

on Sample Data

1st Principal Factor			
.23	JAVA	.86	java.sun.com
.18	SUN	.28	development.sun.com

2nd Principal Factor			
.17	PLATF	.20	NO-READABLE-TEXT
.16	SOLAR	.99	www.lehigh.edu

3rd Principal Factor			
.16	DEVEL	.15	NO-READABLE-TEXT
.15	EDITIO	.97	www.ibm.com
.15	DOWN	.18	www.alphaworks.ibm.com

4th Principal Factor			
.14	INFO	.12	SERVI
.12	SOFTV	.26	INFORMATION
.12	NO-RE	.87	www.pueblo.gsa.gov
		.24	www.irs.gov

6th Principal Factor			
.11	DEVEL	.26	PRESIDENT
.11	LINUX	.87	www.whitehouse.gov
.11	RESOL	.25	NO-READABLE-TEXT
.11	TECHN	.18	www.irs.gov

12th Principal Factor			
.15	PUBLIC	.75	OPTIMIZATION
.14	CONS	.35	www.palisade.com
.13	FREE	.58	SOFTWARE
		.35	www.solver.com

13th Principal Factor			
.15	HOUS	.08	DECIS
.13	BUDG	.46	ADOBE
.13	PRES	.99	www.adobe.com
.11	OFFIC	.07	NEOS
		.45	READER

16th Principal Factor			
.05	SEARCH	.50	WEATHER
.05	ENGIN	.81	www.weather.gov
.05	CONT	.41	www.spc.noaa.gov
.05	ILOG	.30	NO-RE
		.24	OFFICE
		.23	CENTER
		.30	lwf.ncdc.noaa.gov

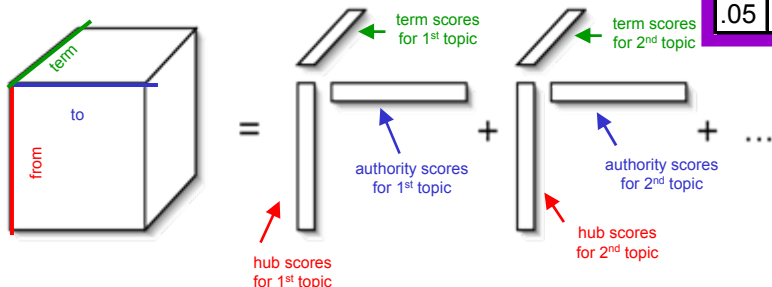
19th Principal Factor			
.17	ORGA	.22	TAX
.15	NWS	.73	www.irs.gov
.15	SEVER	.43	travel.state.gov
.15	FIRE	.22	www.ssa.gov
.15	POLIC	.08	www.govbenefits.gov
.14	CLIMA	.06	www.usdoj.gov
		.03	www.census.gov
		.03	www.usmint.gov
		.02	www.nws.noaa.gov
		.02	www.gsa.gov
		.01	www.annualcreditreport.com

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

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:



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/