Summer School On Mining Big And Complex Data
6 September 2016
Ohrid, Macedonia

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

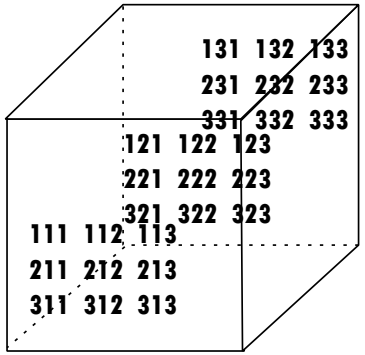
Mining Tensor Data

Hadi Fanaee-T
INESC TEC, Portugal
hadi.fanaee@fe.up.pt

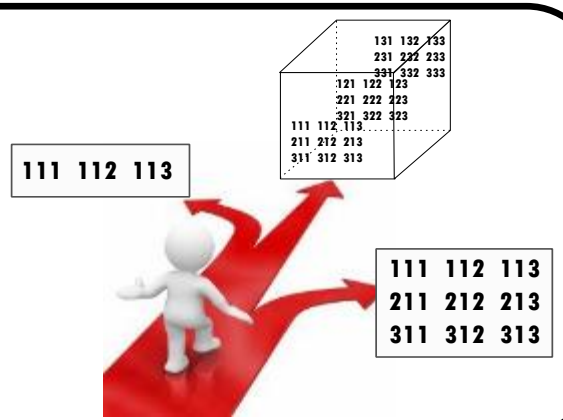
Level: ● Easy ○ Medium ○ Advanced
Mathematical Equations: No
Duration: 45 Min

LIAAD INESCTEC LABORATÓRIO ASSOCIADO
U.PORTO

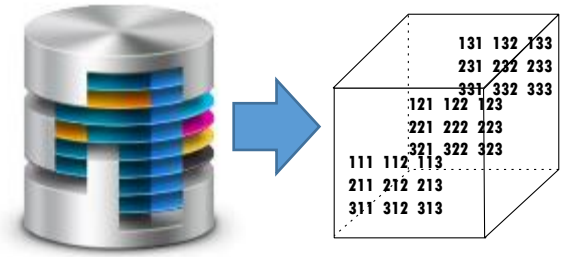
Tensor



Why Tensors?

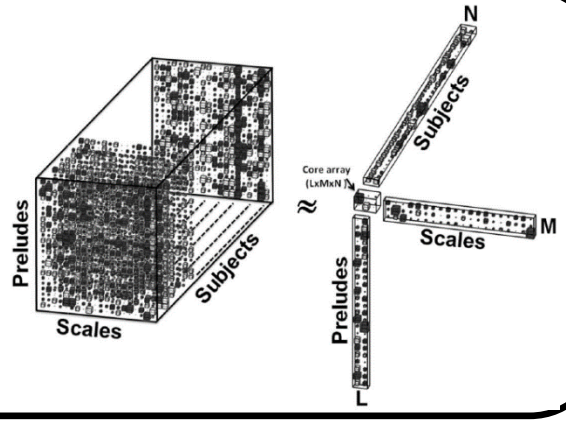


Convert DBs to Tensor



Tensor Datasets

Tensor Decomposition



Mining Applications

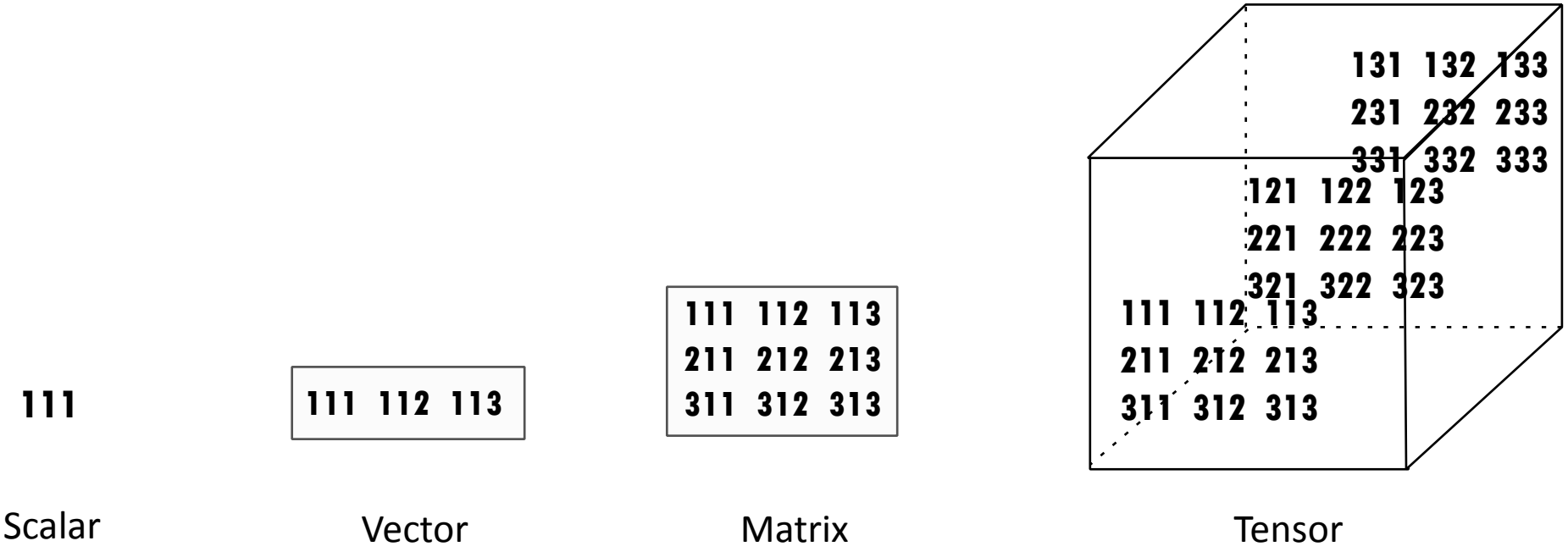
Software

Starting Points

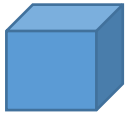
Interdisciplinary Showcases

Tensor

- Geometric object for extension of scalars, vectors and matrices to higher dimensions

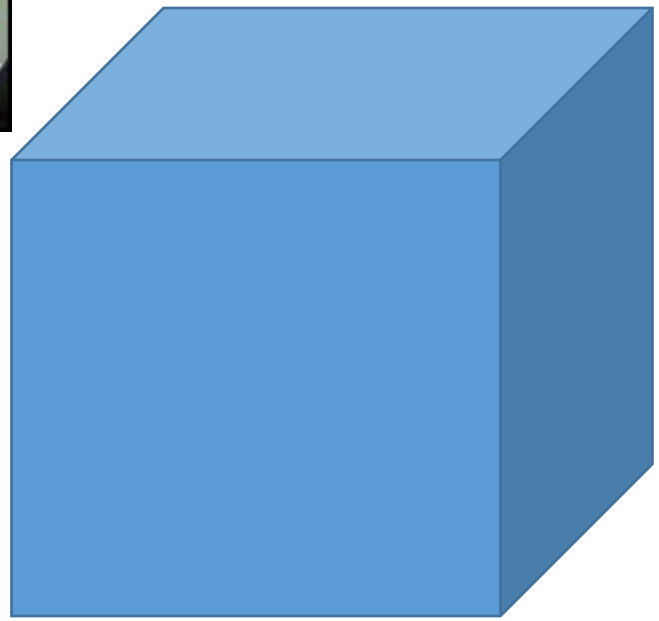


Why tensor analysis is emerging?



50 years ago

Now



Necessity of Tensors

Misleading knowledge!

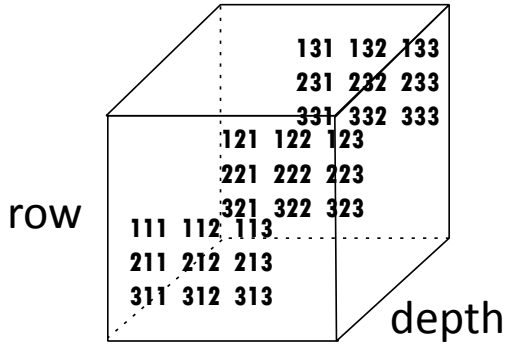
Camera

Human Eye

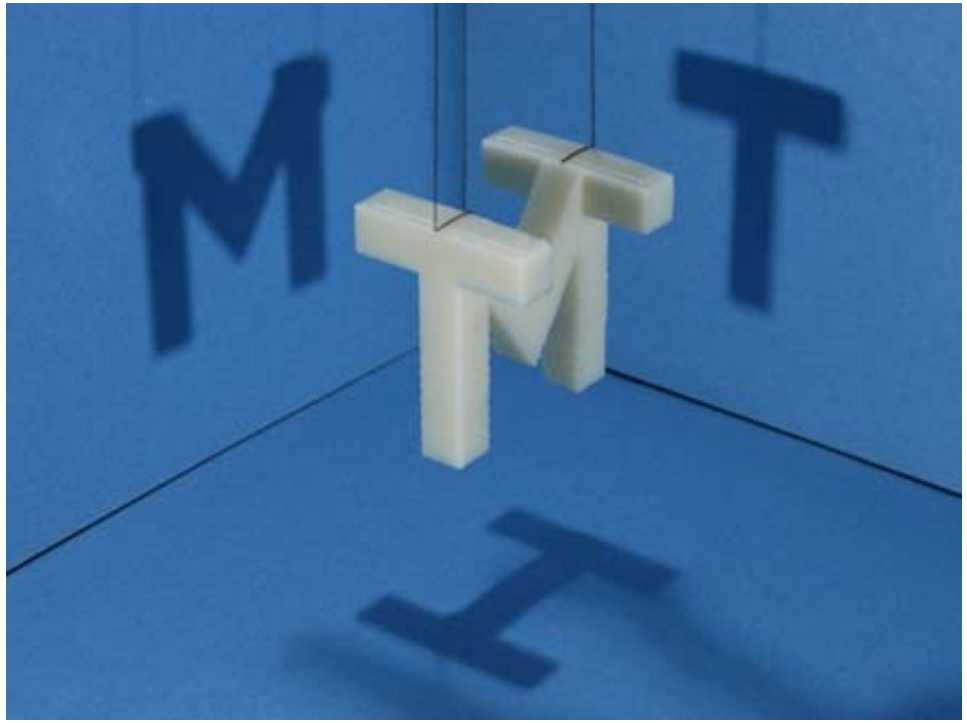
row

111	112	113
211	212	213
311	312	313

column



Necessity of Tensors

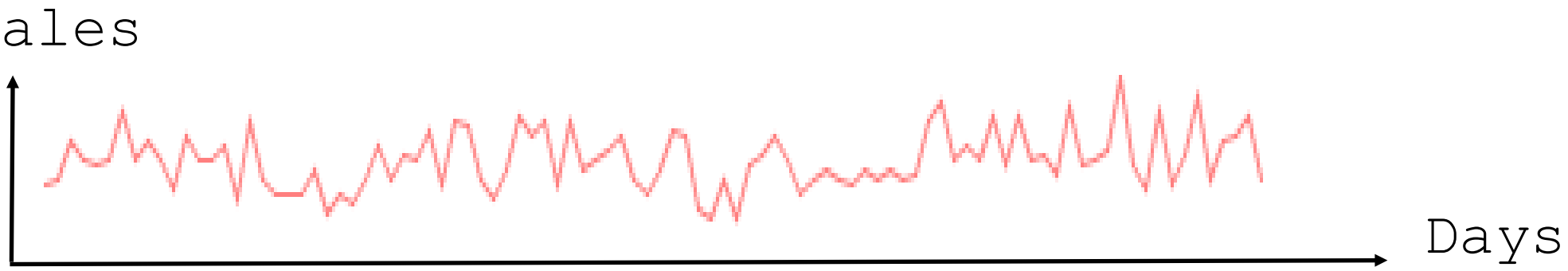


Misleading knowledge!

Example: Supermarket Sales Data

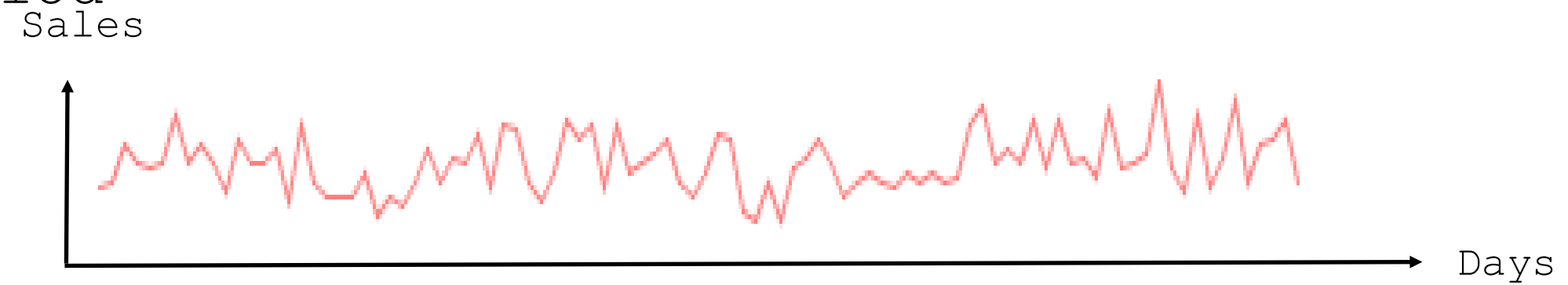
Univariate data

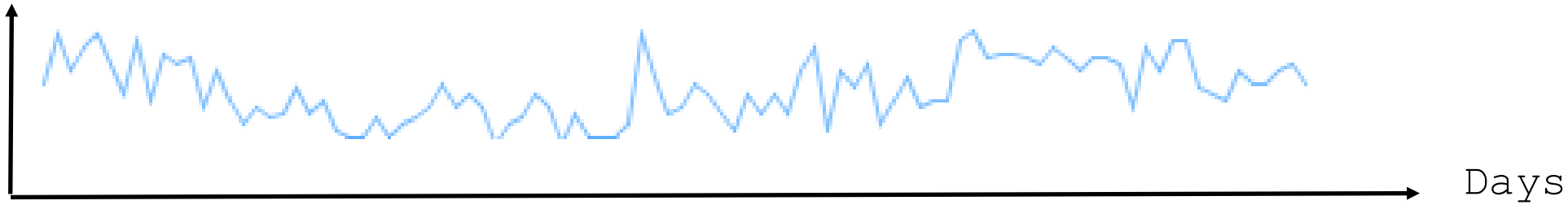
- Sales of Water in Porto during 1 month period



Multivariate data

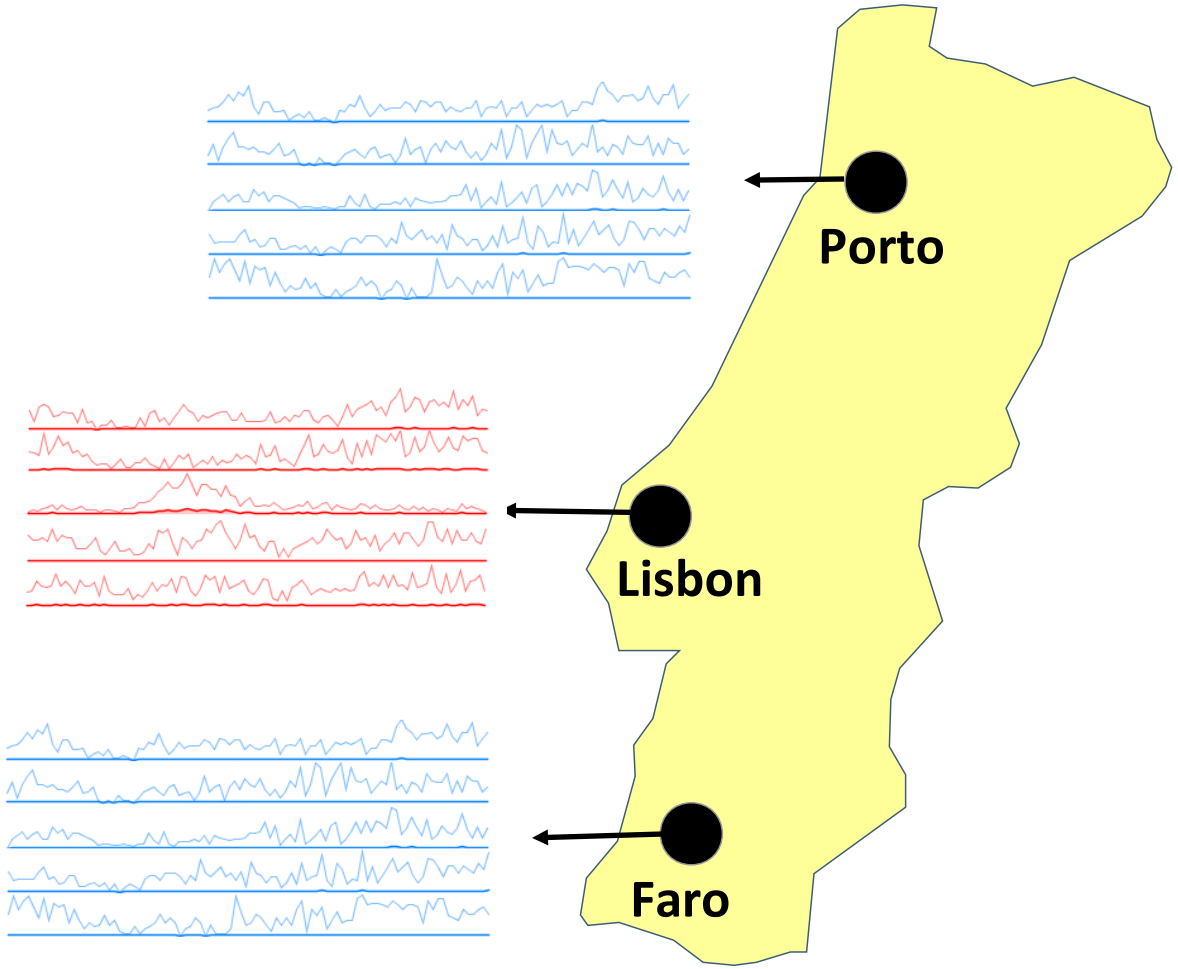
- Sales of Water and Cola in Porto during 1 month period





Tensor data

Sales of multiple products through multiple locations during 1 month period



Transforming data to tensor format

1) Indexing modes:

Dimension#1 Locations

Porto : 1
Lisbon: 2
Faro: 3

Dimension#2 Products

Water: 1
Cola: 2
Tea: 3
Cafe: 4
Bread: 5

Dimension#3 Time

2016/05/01: 1
2016/05/02: 2
....
2016/05/30: 30

2) Translating to index

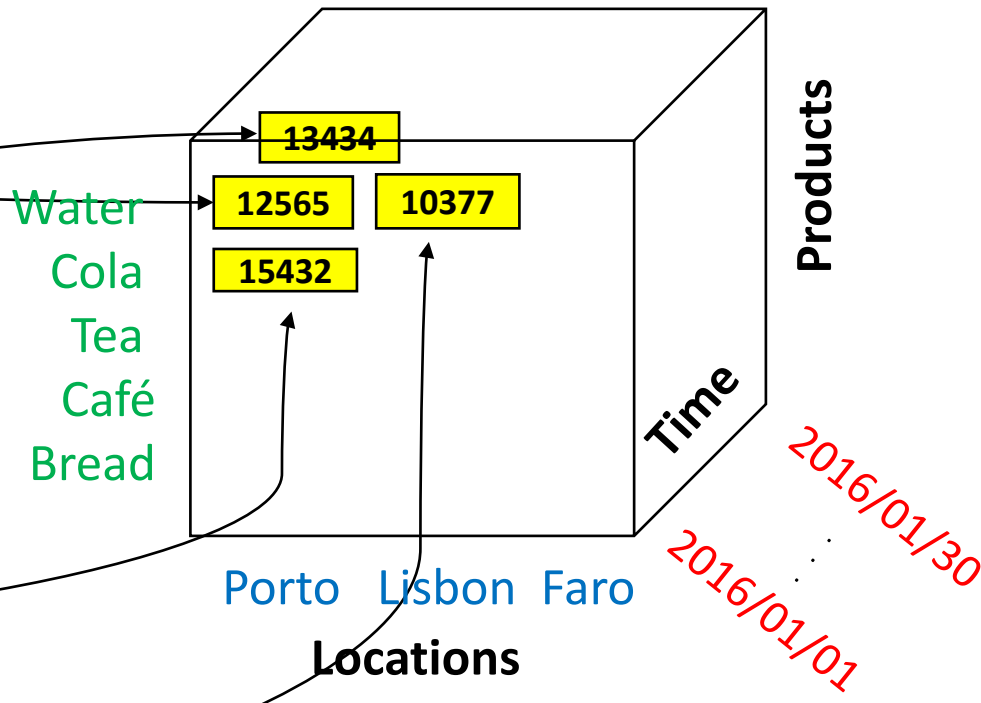
Water Sales in € at **Porto** in **2016/04/01** :
12565 € →
X(1,1,1)= 12565

Water Sales in € at **Porto** in **2016/04/02**:
13434 € →
X(1,1,2)=13434

Cola sales in € at **Porto** in **2016/04/01**:
15432 € →
X(1,2,1)=15432

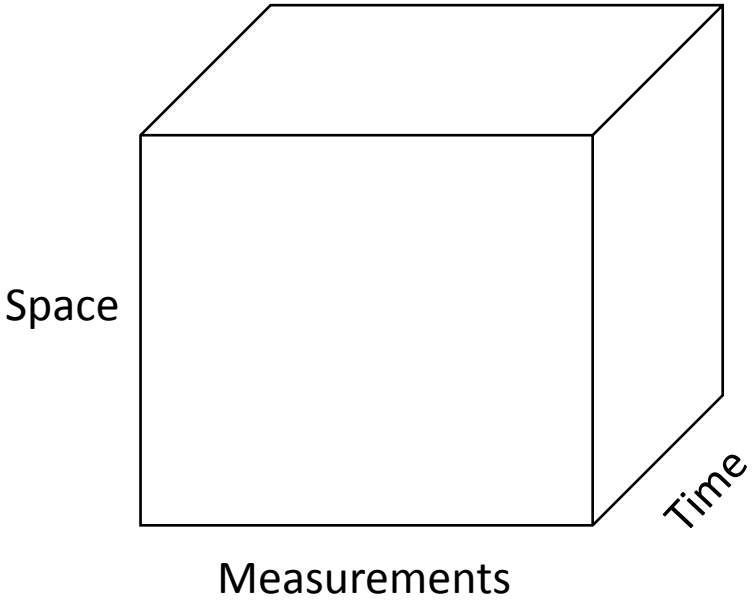
Water Sales in € at **Lisbon** in **2016/04/01**:
10377 € →
X(2,1,1)= 10377

3) Multiway array

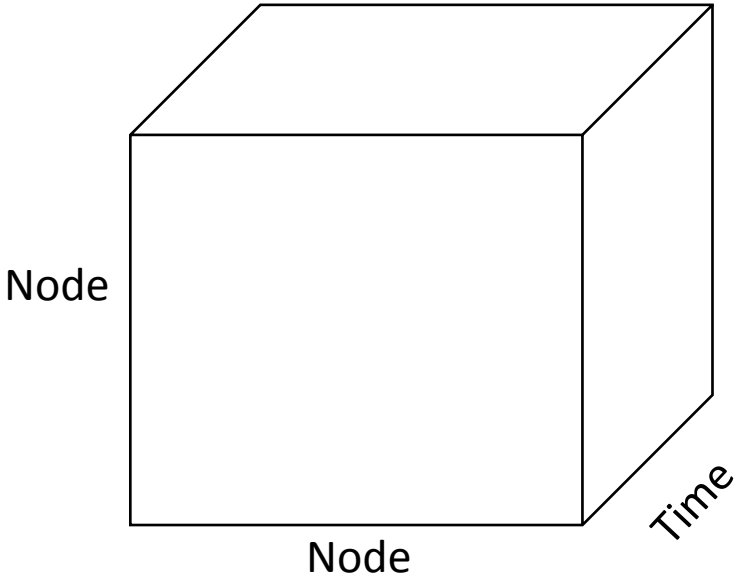


Common tensor datasets

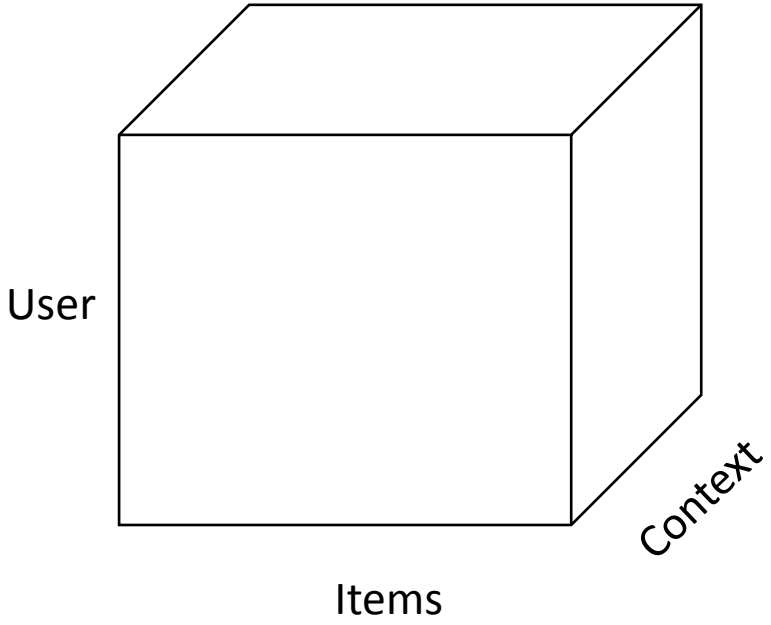
Spatio-temporal Data



Evolving networks



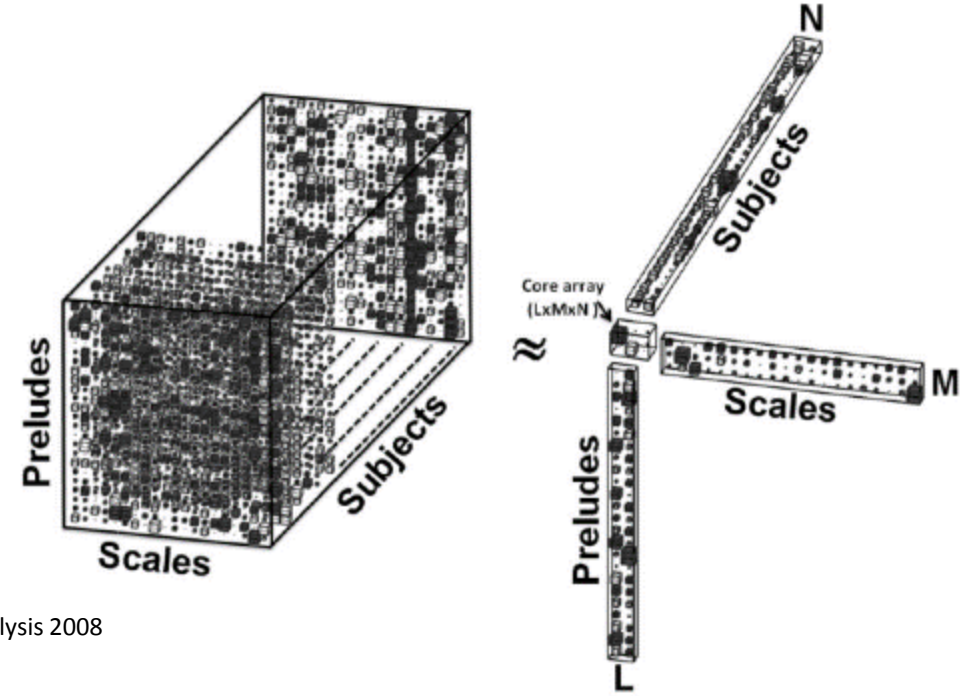
Recommender Systems



Tensor Decomposition

Do you remember PCA?

Tensor Decomposition is extension of PCA for higher order tensors

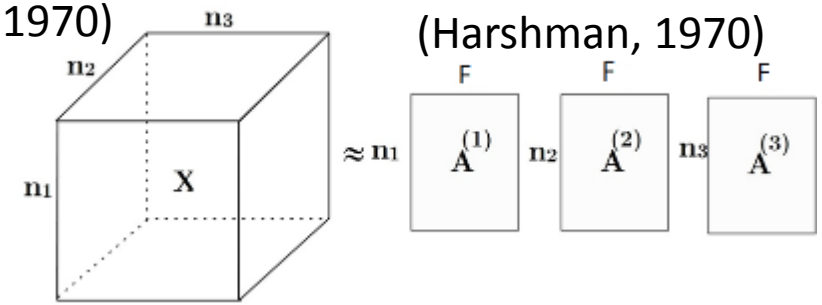


Kroonenberg, Applied Multiway Analysis 2008

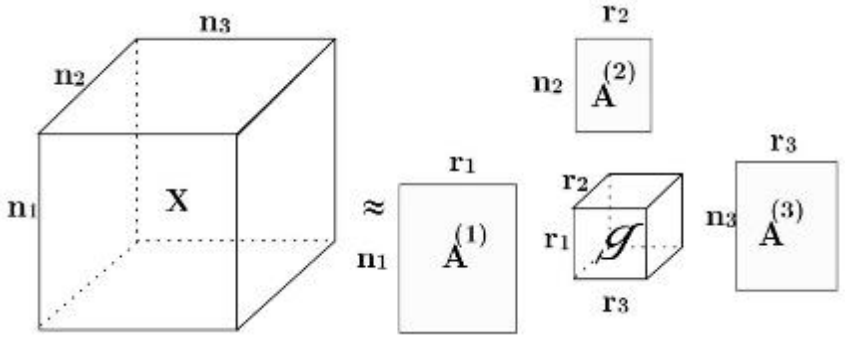
Projection of large data to a summarized subspace → Main goal in data m

Tensor Decomposition Models

(Carrol&Chang, 1970)



(Harshman, 1970)

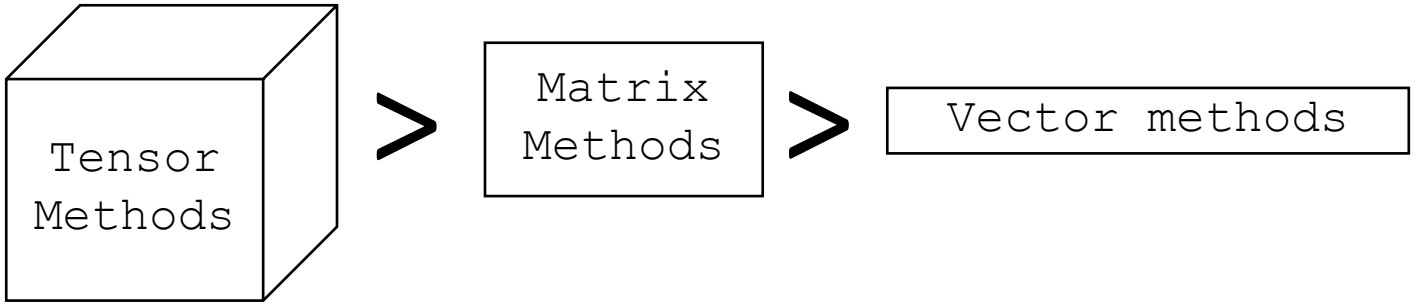


[Tucker, 1966]

- Classical
  - Tucker
  - CP or PARAFAC (Parallel Factor Analysis)
- Modern
  - Block Term Decomposition (BTD)
  - Low multilinear rank approximation (LMRA)
  - Non-Negative Tensor Factorization (NTF) physical interpretation of components
  - Tensor Global-Local Preserving Projections (TGLPP) considers geometric relationships
  - Bayesian Methods use Bayes' theorem to solve tensor decomposition
  - DEDICOM (DEcomposition into DIrectional COMponents) driectional components
  - CTMF (Coupled Tensor-Matrix Factorization) data fusion model
- More : Mørup, Morten. "Applications of tensor (multiway array) factorizations and decompositions in data mining." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (2011): 24-40.

Golden Rule

If data has tensor structure, the performance of data mining task outperforms the non-tensor approach



Benefit for you : you invest your time on something that with high probability improves the existing state-of-arts local method. See some evidences in:

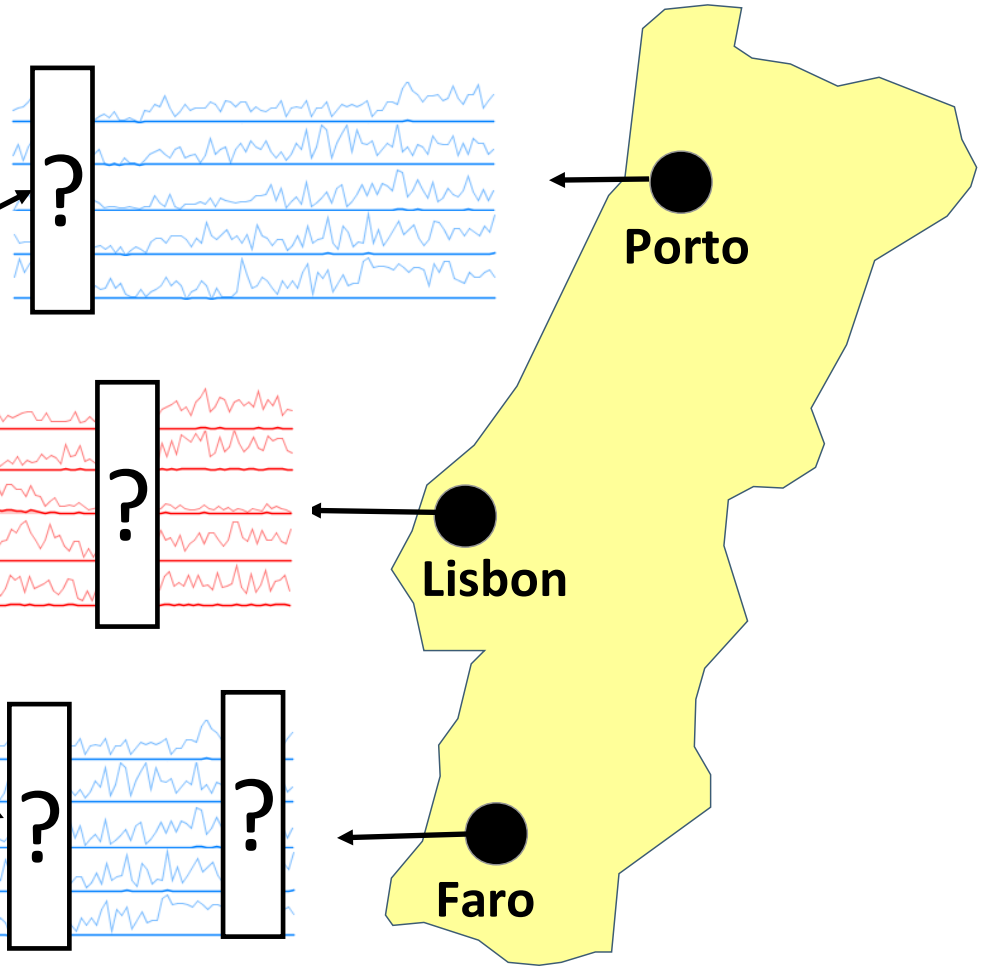
Fanaee-T, Hadi, and Gama, João. "Tensor-based anomaly detection: An interdisciplinary survey." *Knowledge-Based Systems* 98 (2016): 130-147, Elsevier.

Applications

1. Missing value estimation
2. Correspondence analysis
3. Co-Clustering
4. Outlier Detection
5. Event Detection
6. Anomaly Detection
7. Classification
8. Temporal Link Prediction
9. Forecasting
10. Item Recommendation
11. Topic modelling and pattern extraction

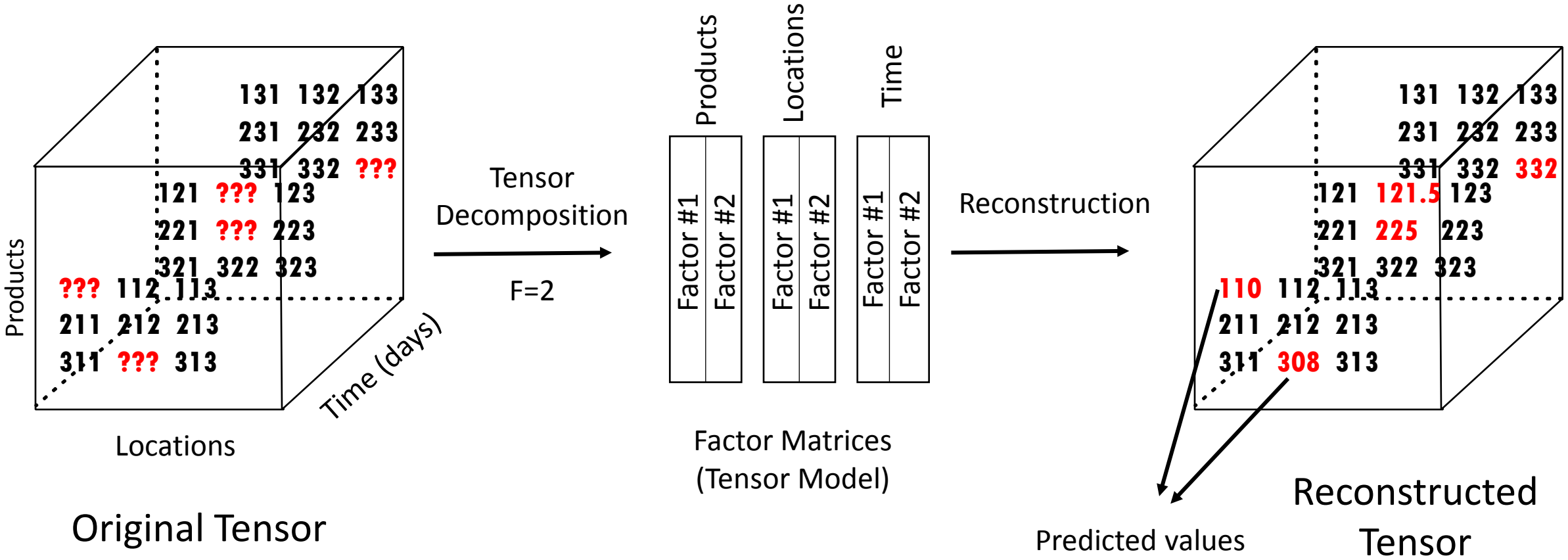
1. Missing value estimation

Sales of multiple products through multiple locations during different weekdays



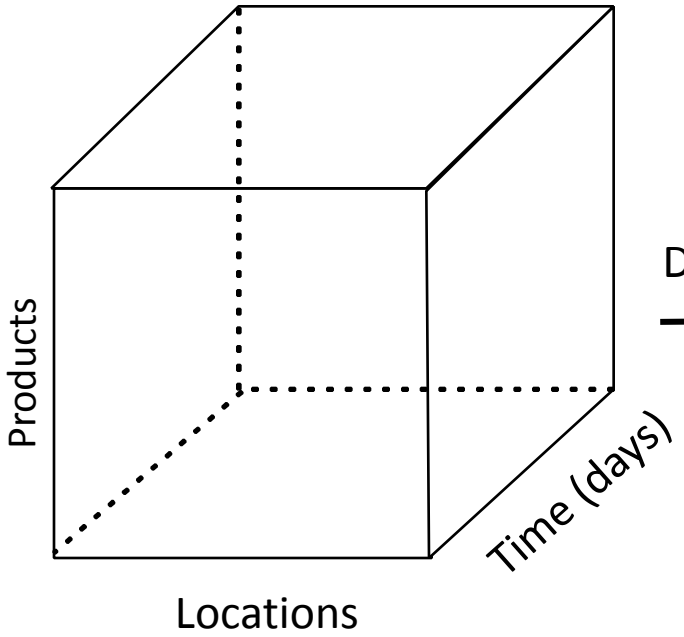
Data is not available due to different reasons: e.g. Errors in informatio n system, power

1. Missing value estimation



Acar, Evrim, et al. "Scalable tensor factorizations for incomplete data." *Chemometrics and Intelligent Laboratory Systems* 106.1 (2011): 41-56.

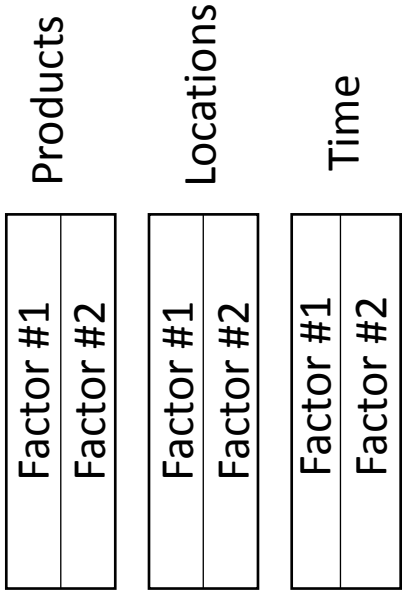
2. Correspondence Analysis



Original Tensor

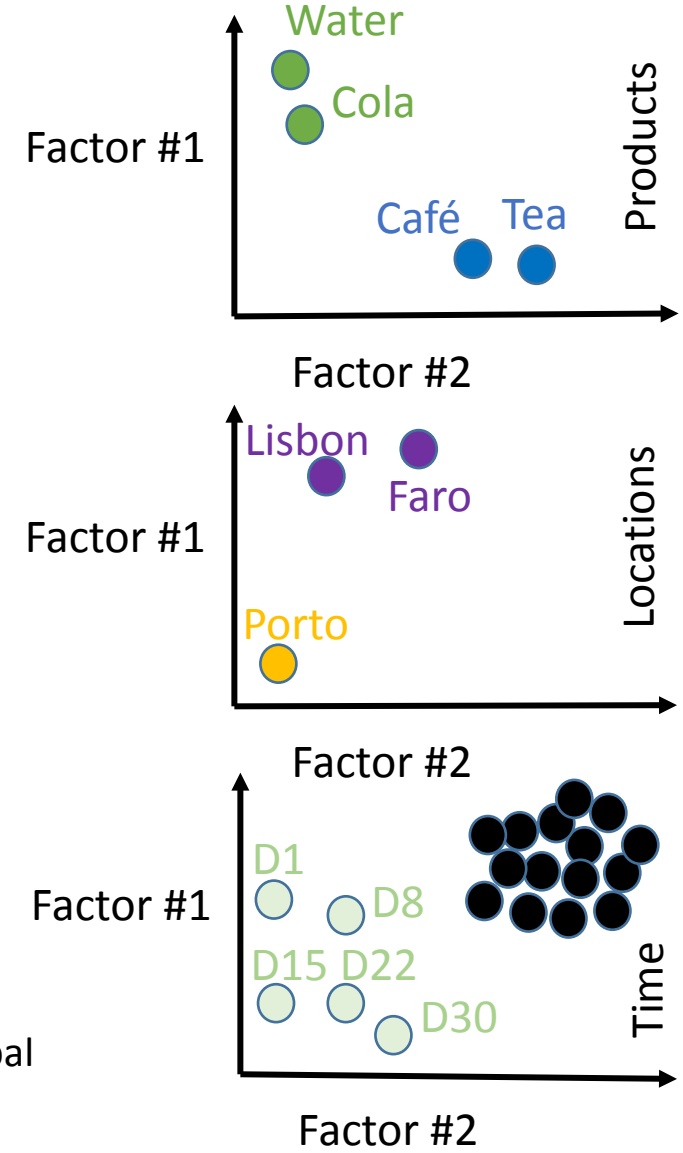
Tensor Decomposition

F=2



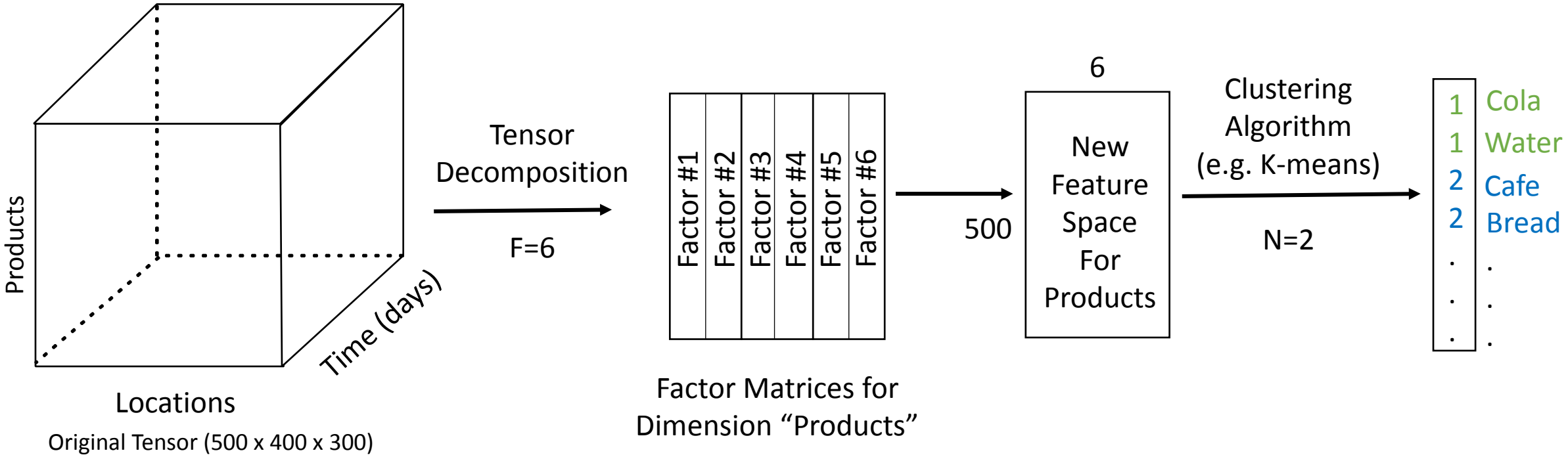
Factor Matrices
(Tensor Model)

Factor Analysis



Andrade, J. M., et al. "3-Way characterization of soils by Procrustes rotation, matrix-augmented principal components analysis and parallel factor analysis." *analytica chimica acta* 603.1 (2007): 20-29.

3. Co-Clustering

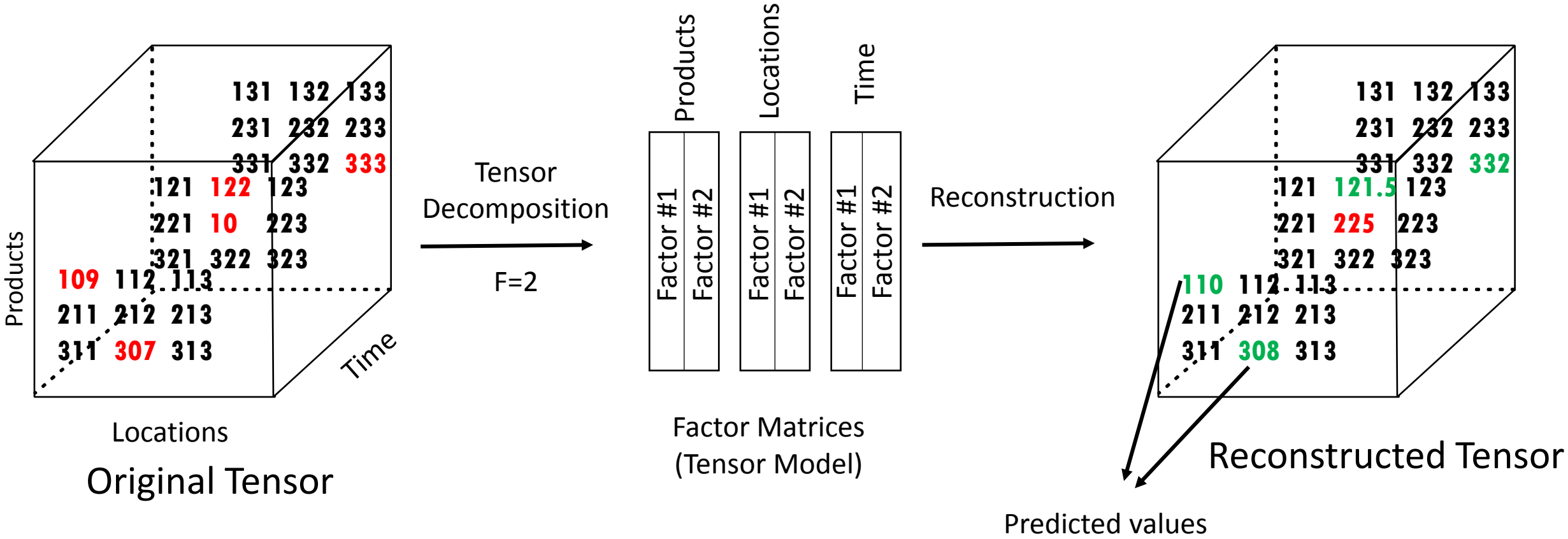


Hadi Fanaee-T, et. al., *Event and Anomaly Detection Using Tucker3 Decomposition*, (ECAI'2012-UDM Workshop), CEUR Workshop Proceedings , Vol 960, PP. 8-12, 2012, DOI:10.13140/RG.2.1.2370.9286

Papalexakis, Evangelos E., Nicholas Sidiropoulos, and Rasmus Bro. "From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors." *Signal Processing, IEEE Transactions on* 61.2 (2013): 493-506.

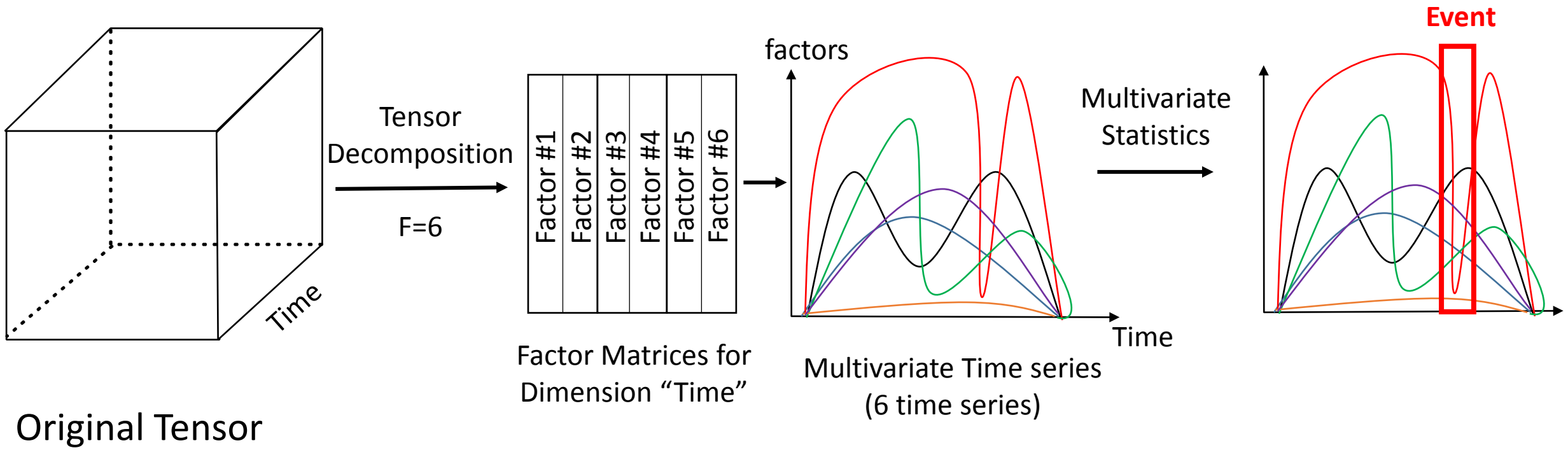
4. Outlier Detection

|225-10|=215 → Outlier
|121.5-122|=0.5 → Regular
|308-307|=1 → Regular



Tan, Huachun, et al. "Traffic volume data outlier recovery via tensor model." *Mathematical Problems in Engineering* 2013 (2013).

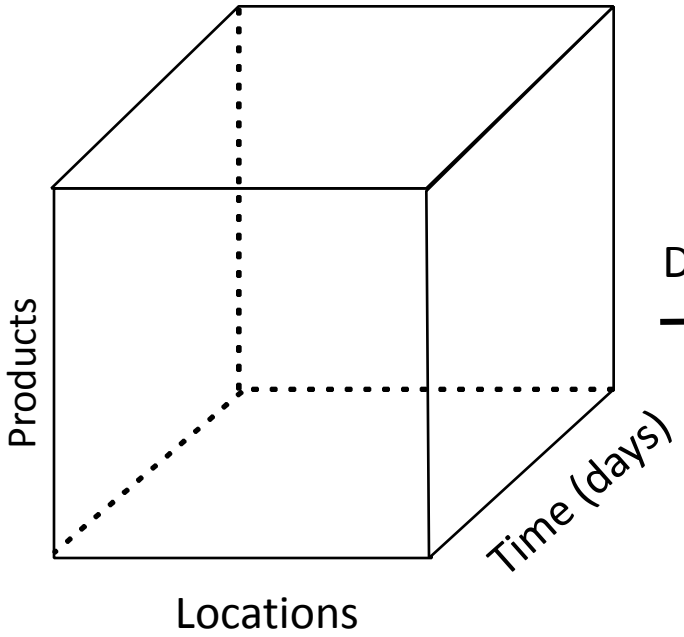
5. Event Detection



Fanaee-T, Hadi, and Gama, João. "Event detection from traffic tensors: A hybrid model." *Neurocomputing* 203 (2016): 22-33, Elsevier.

Meaningful deviation that has a physical meaning

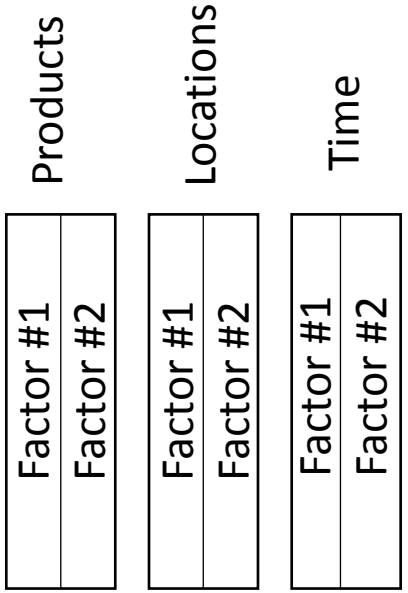
6. Anomaly Detection



Original Tensor

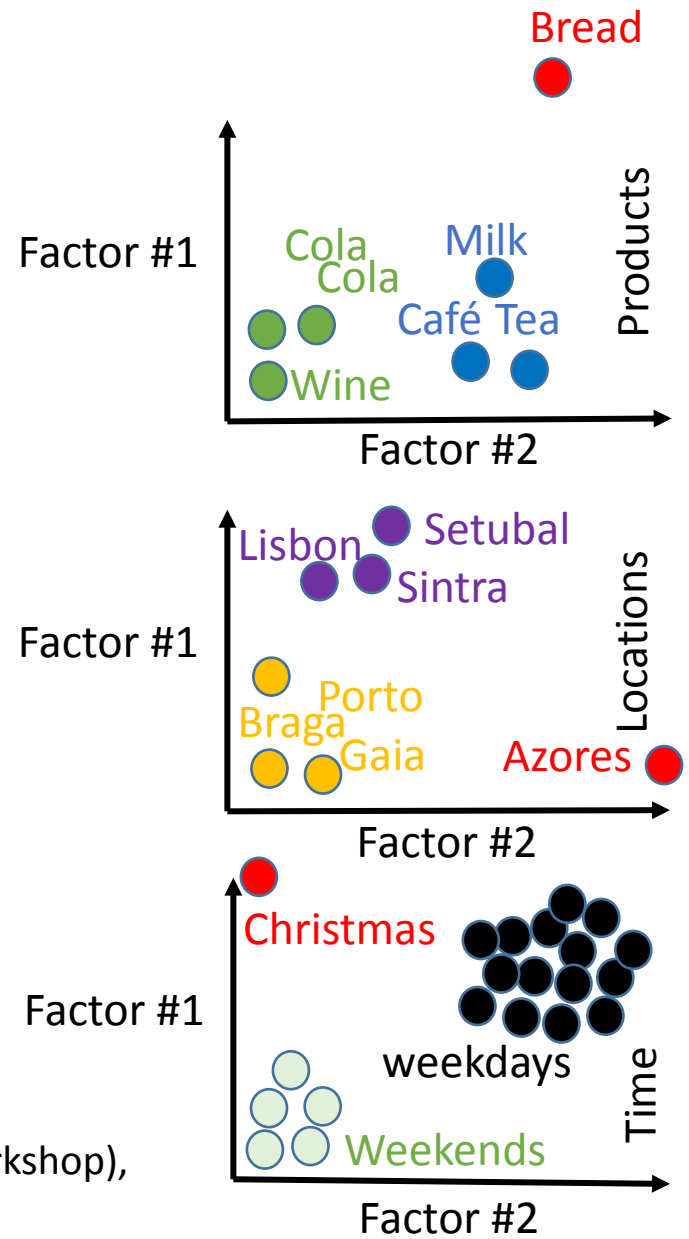
Tensor Decomposition

F=2



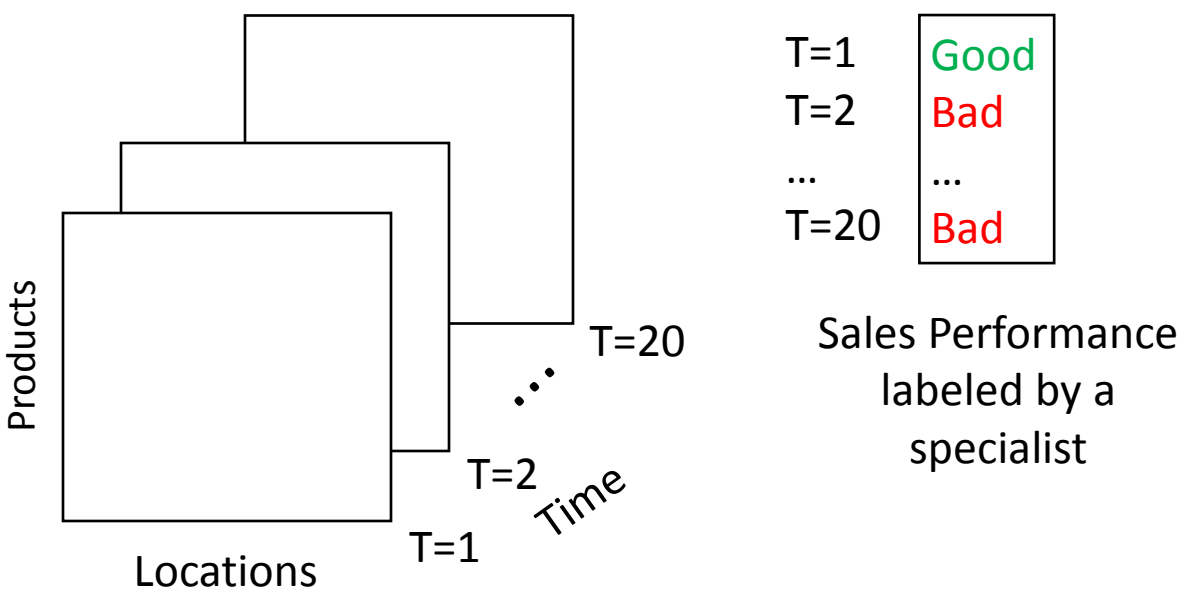
Factor Matrices
(Tensor Model)

Factor Analysis

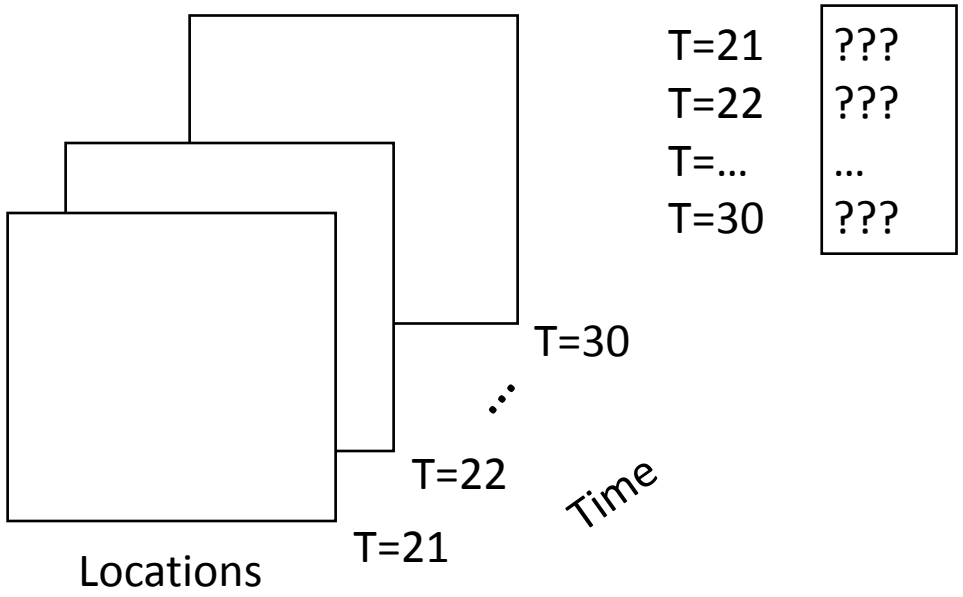


Hadi Fanaee-T, et. al., *Event and Anomaly Detection Using Tucker3 Decomposition*, (ECAI'2012-UDM Workshop), CEUR Workshop Proceedings , Vol 960, PP. 8-12, 2012, DOI:10.13140/RG.2.1.2370.9286

7. Classification



Train Set

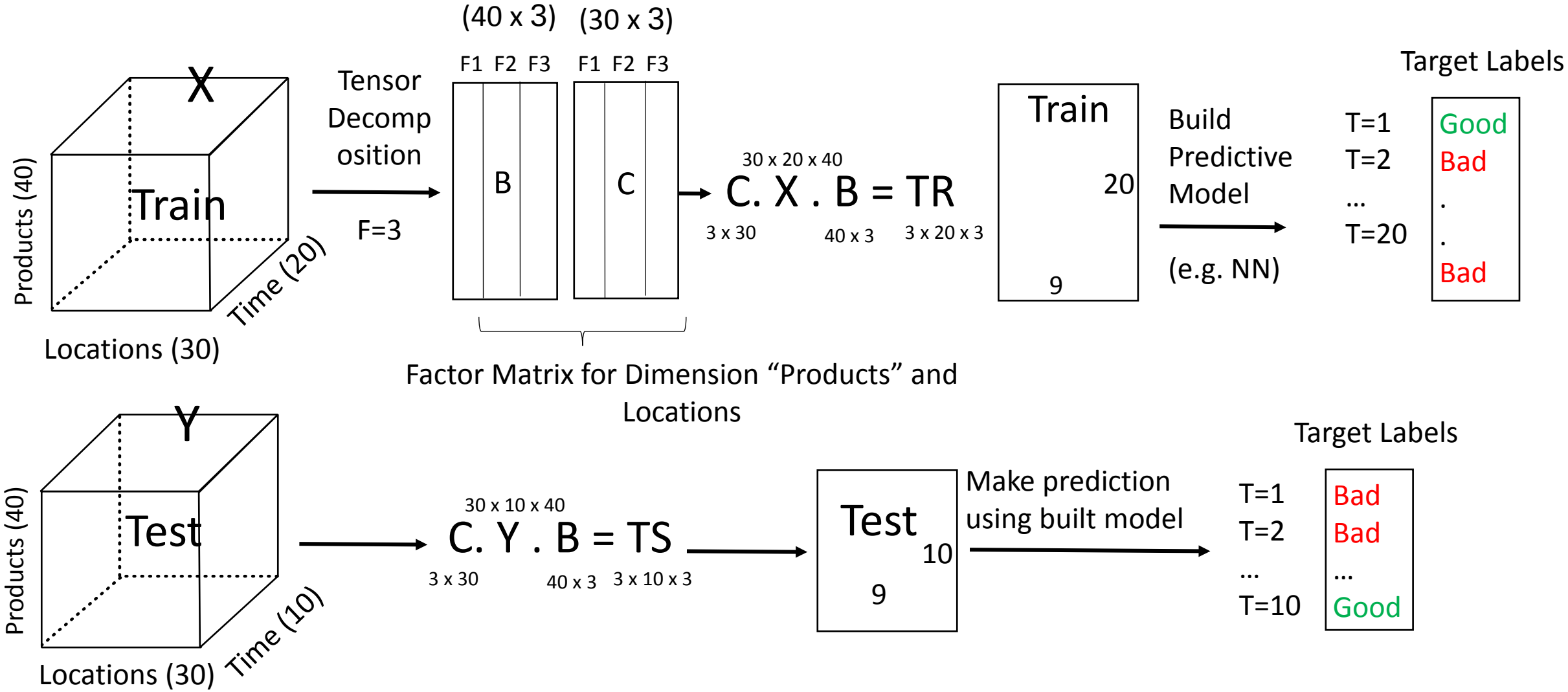


Test Set

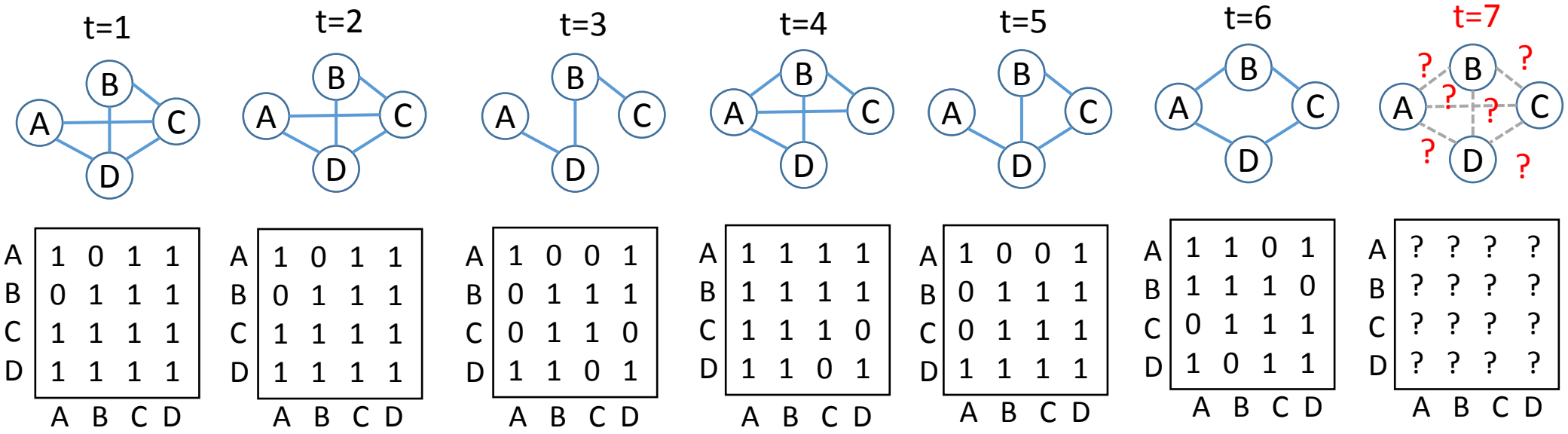
Phan, Anh Huy, and Andrzej Cichocki. "Tensor decompositions for feature extraction and classification of high dimensional datasets." *Nonlinear theory and its applications, IEICE* 1.1 (2010): 37-68

7. Classification

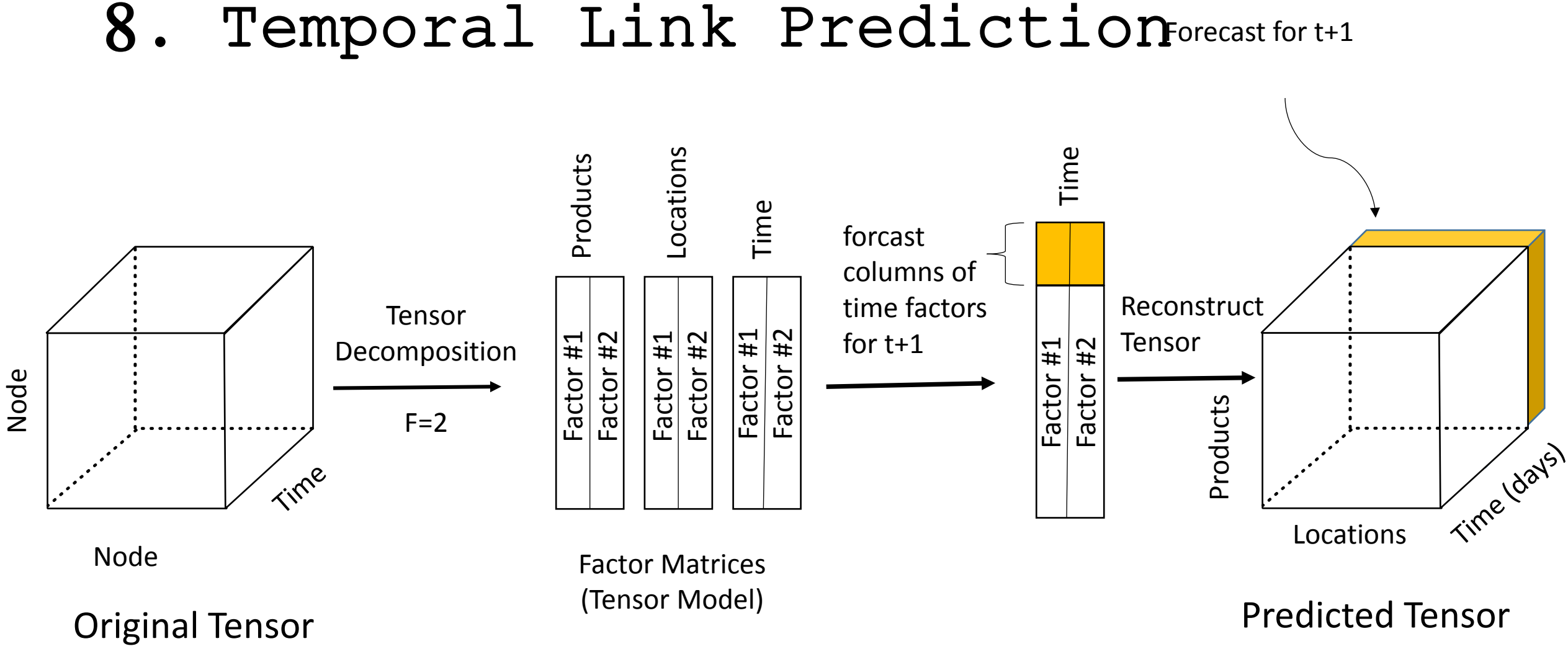
Dimensionality reduction: 40 x 30 = 1200 → 3 x 3 =9



8. Temporal Link Prediction

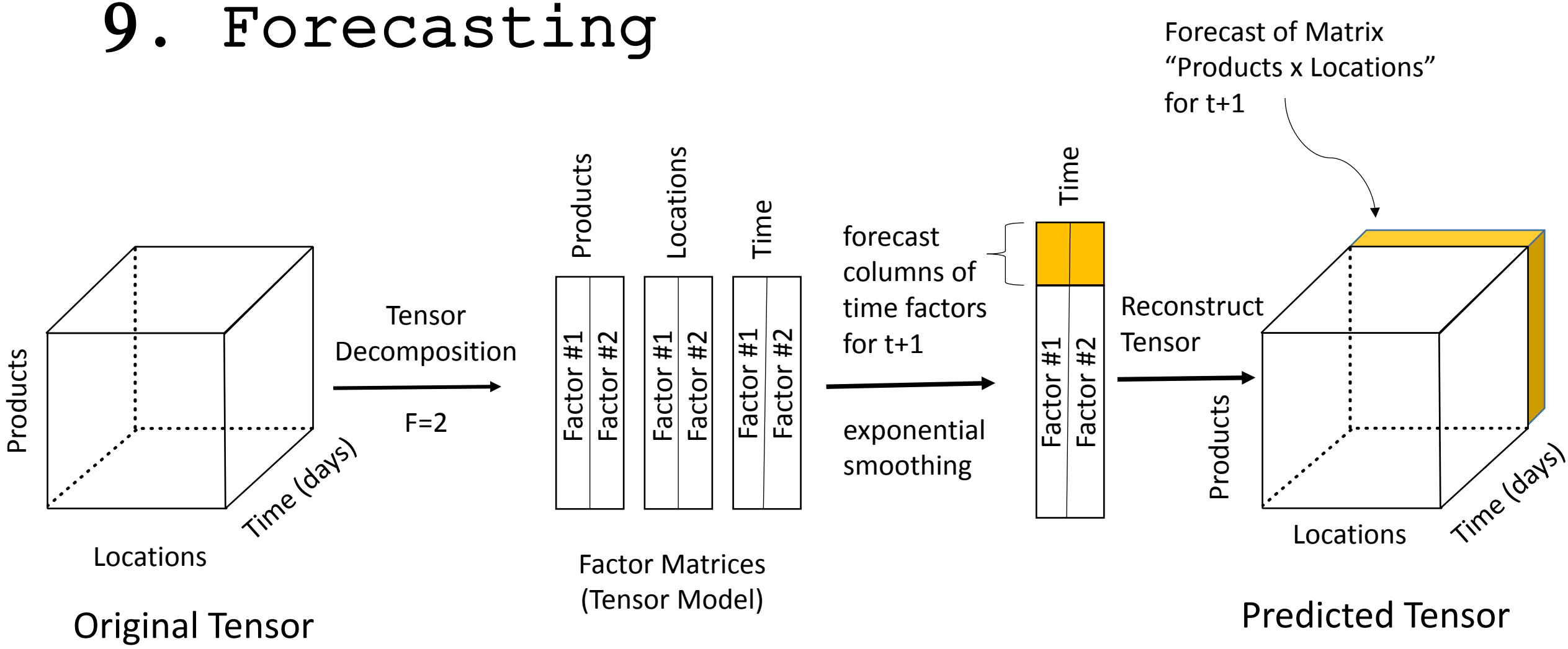


8. Temporal Link Prediction



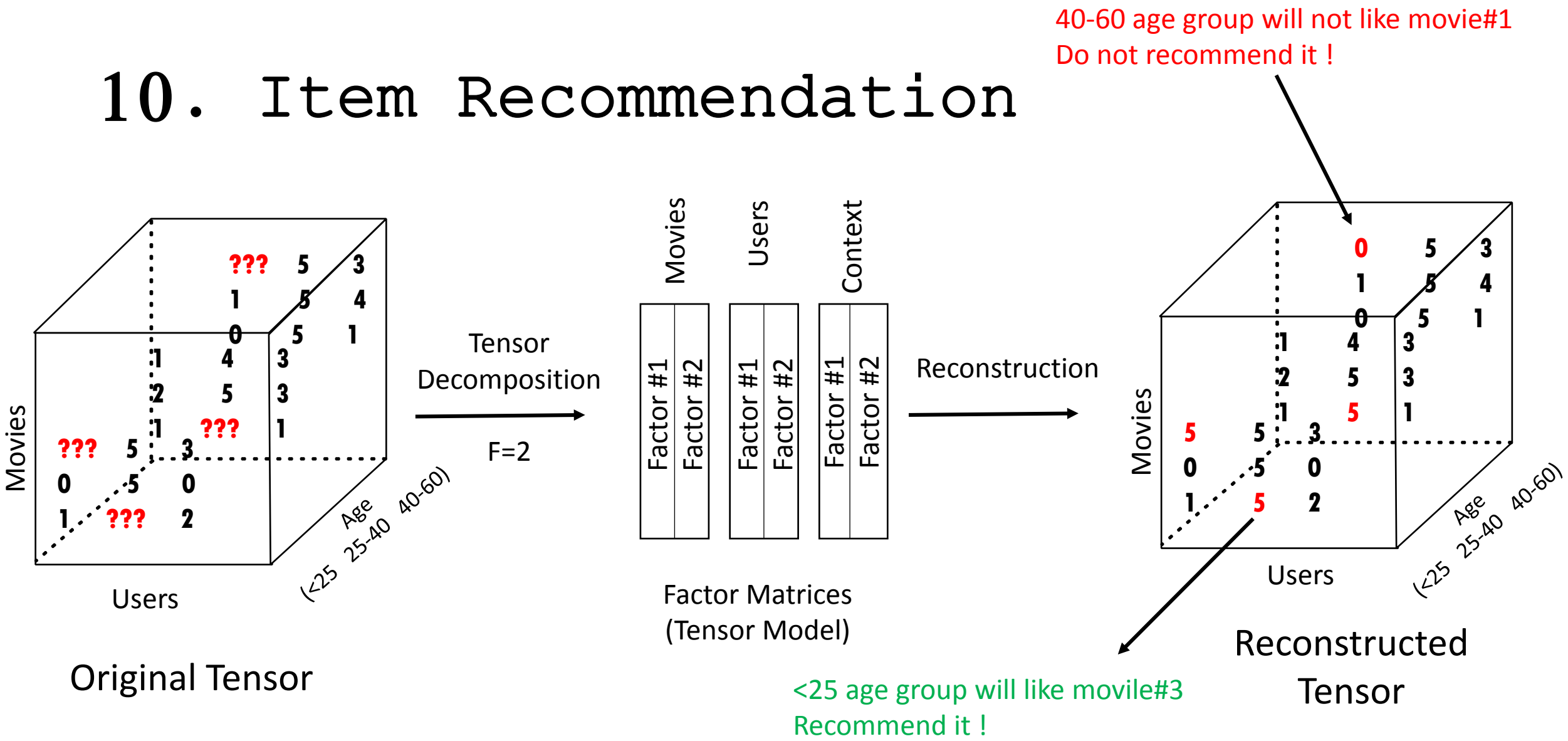
Dunlavy, Daniel M., Tamara G. Kolda, and Evrim Acar. "Temporal link prediction using matrix and tensor factorizations, *TKDD* 5.2 (2011): 10.

9. Forecasting



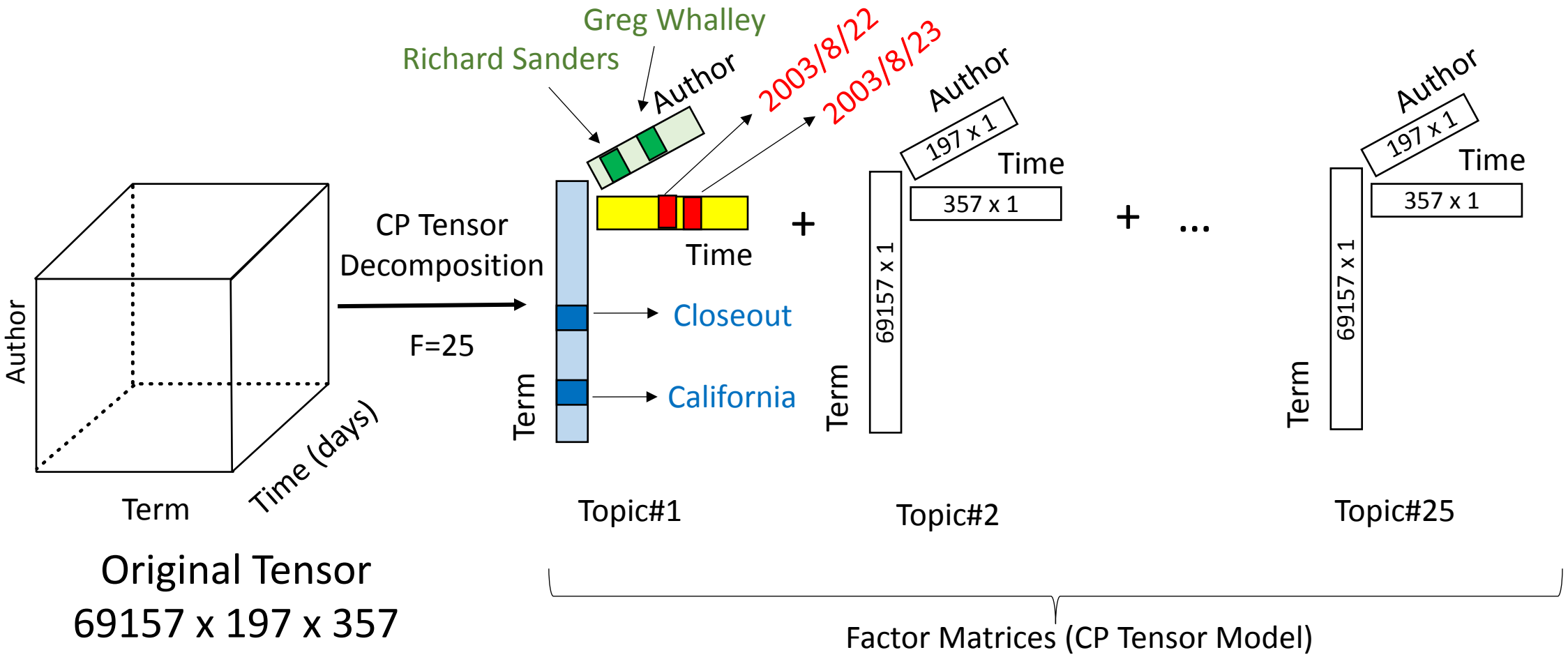
Spiegel, Stephan, et al. "Link prediction on evolving data using tensor factorization." *New Frontiers in Applied Data Mining*. Springer Berlin Heidelberg, 2011. 100-110. (PAKDD 2011)

10. Item Recommendation



Karatzoglou, Alexandros, et al. "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering." *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010.

11. Topic modeling and pattern extraction



Bader, Brett W., Michael W. Berry, and Murray Browne. "Discussion tracking in Enron email using PARAFAC." *Survey of Text Mining II*. Springer London, 2008. 147-163.

Software

- **MATLAB**
  - N-way toolbox (http://www.models.life.ku.dk/nwaytoolbox) July 2013
  - Tensor toolbox (http://www.sandia.gov/~tgkolda/TensorToolbox) Feb 2015
  - TDALAB (http://www.bsp.brain.riken.jp/TDALAB) May 2013
  - TensorBox (http://www.bsp.brain.riken.jp/~phan) Dec 2014
  - **Tensorlab** (http://www.tensorlab.net) March 2016 - ✓**Recommended**
- **R**
  - Threeway and PTAk
- **Python**
  - scikit-tensor
- **Java (Hadoop)**
  - **BigTensor** (http://datalab.snu.ac.kr/bigtensor), 2015 ✓**For Big Data**
    Big data example: freebase dataset: 38M * 532 * 38M (139M nonzeros)
- **Standalone/Multi-platform + Web-based**
  - SimTensor (INESC TEC) ~Nov 2016
  - Tensoraia (INESC TEC) ~ 2017

Quick Start!

- Survey papers
  - Kolda, Tamara G., and Brett W. Bader. "Tensor decompositions and applications." *SIAM review* 51.3 (2009): 455-500.
  - Acar, Evrim, and Bülent Yener. "Unsupervised multiway data analysis: A literature survey." *Knowledge and Data Engineering, IEEE Transactions on* 21.1 (2009): 6-20.
  - Mørup, Morten. "Applications of tensor (multiway array) factorizations and decompositions in data mining." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (2011): 24-40.
  - Fanaee-T, Hadi, and João Gama. "Tensor-based anomaly detection: An interdisciplinary survey." *Knowledge-Based Systems* 98 (2016): 130-147.
- Slides
  - Morten Mørup, Applications of tensor decomposition in data mining and machine learning, 2010
  - Johan Westerhuis, Multiway Data Analysis (www.cac.science.ru.nl/education/Brugcollege/multiway.ppt)
- Books
  - Kroonenberg, Pieter M. Applied multiway data analysis. Vol. 702. John Wiley & Sons, 2008.
  - Cichocki, Andrzej, et al. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

Interdisciplinary showcases

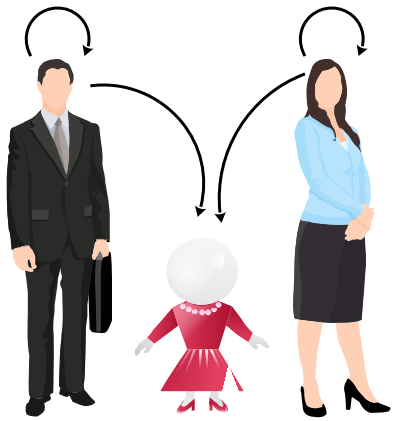
- Psychology
- Process control
- Environmental monitoring
- Video surveillance
- Network security
- Social networks
- Text-based systems
- Neuroscience
- Remote sensing
- Sensors
- Transportation
- Metallurgy Engineering
- Civil structures
- Mechanical systems
- Power systems
- Economy
- Bio-informatics
- Medical diagnosis
- Epidemiology
- Seismology
- Climatology
- Criminology
- Face Recognition
- Image Completion

1-Psychology

Behavior Analysis

Conditions

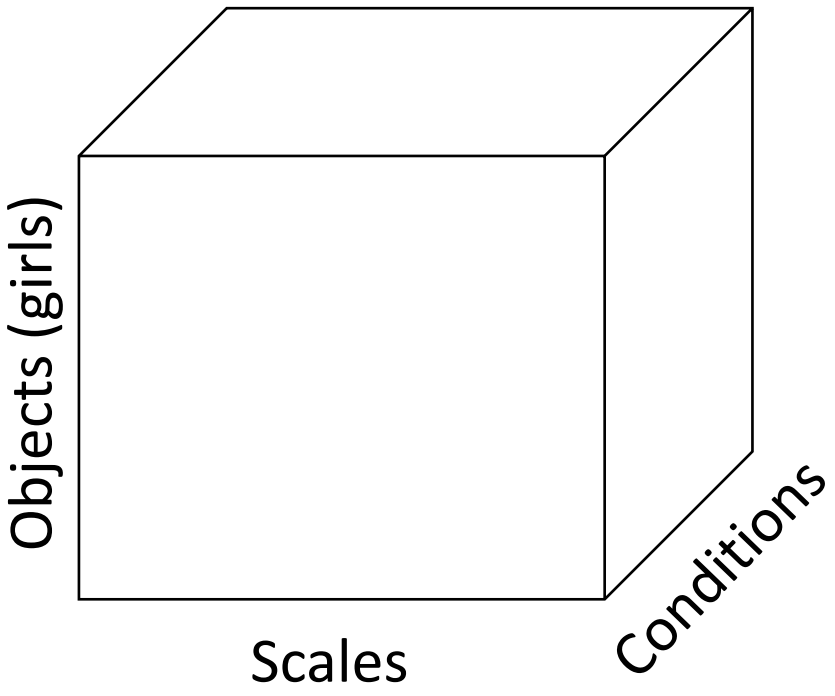
- Father own behavior (F-F)
- Mother-Own behavior (M-M)
- Daughters' judgment of their Father (D-F)
- Daughter ' judgment of their Mother (D-M).



Scales

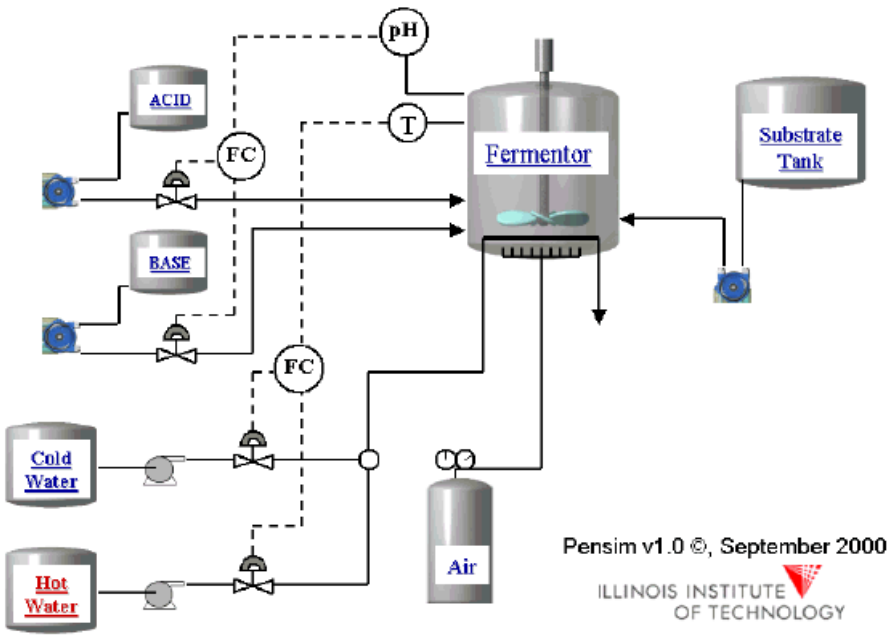
- Parental support or Acceptance
  - Acceptance
  - Acceptance individuation
  - Child centredness
  - Positive involvement
- Behavioural and Psychological control
  - Possessiveness
  - Instilling persistent anxiety
  - Intrusiveness
  - Hostile control
  - Enforcement
  - Control
- Rejection
  - Control through guilt
  - Withdrawal of relations
  - Rejection
  - Hostile detachment
- Lax discipline
  - Inconsistent discipline
  - Lax discipline
  - Extreme autonomy
  - Nonenforcement

$$150 \text{ (girls)} \times 18 \text{ (scales)} \times 4 \text{ (judgement combinations)}$$

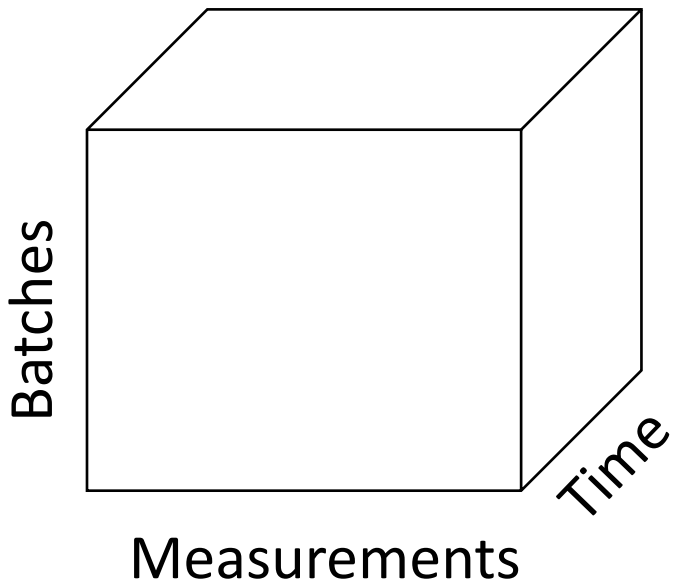


Kroonenberg, Pieter M., Richard A. Harshman, and Takashi Murakami. "Analysing three-way profile data using the Parafac and Tucker3 models illustrated with views on parenting." *Applied Multivariate Research* 13.1 (2009): 5-41.

2.Batch Processes Fault Detection

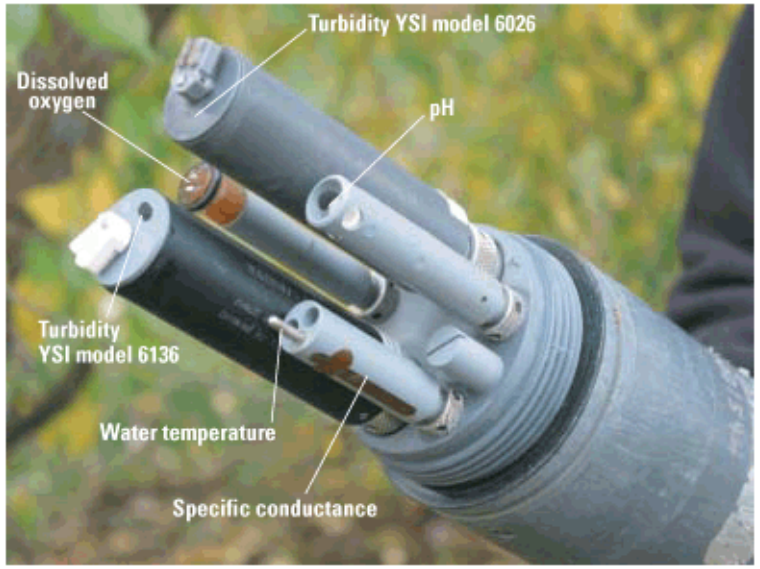


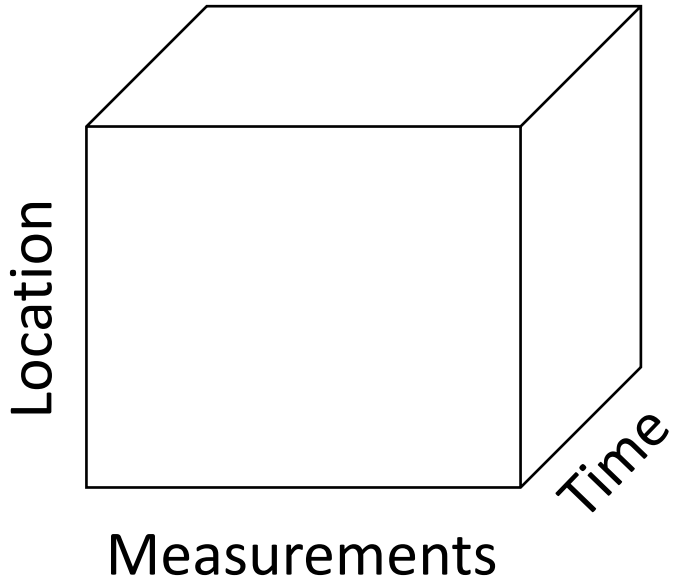
Fed-Batch Penicillin Production



Nomikos, Paul, and John F. MacGregor. "Monitoring batch processes using multiway principal component analysis." *AIChE Journal* 40.8 (1994): 1361-1375.

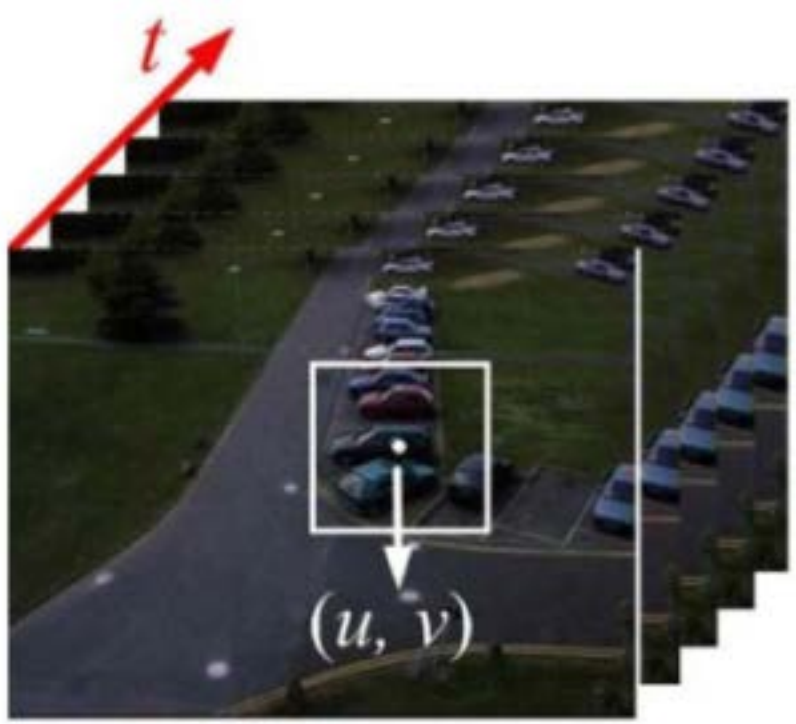
3.Environmental Monitoring Spatio-temporal fluctuations, Quality Monitoring

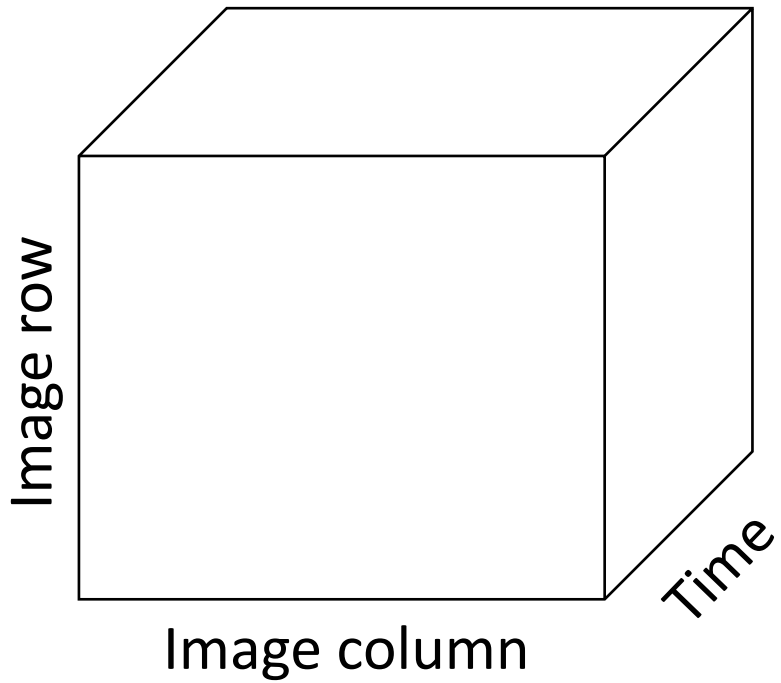




Cid, Fabricio D., et al. "Modelling spatial and temporal variations in the water quality of an artificial water reservoir in the semiarid Midwest of Argentina." *Analytica chimica acta* 705.1 (2011): 243-252.

4.Video Surveillance Anomaly Detection, Object Tracking



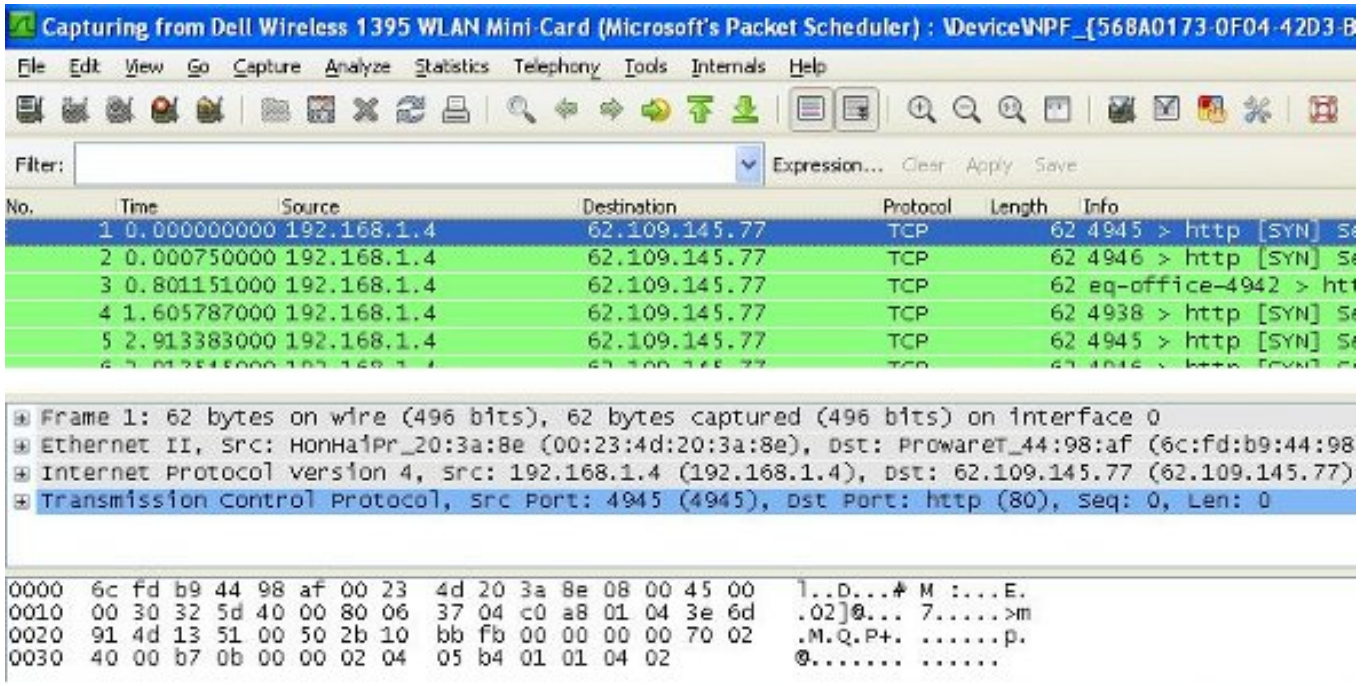


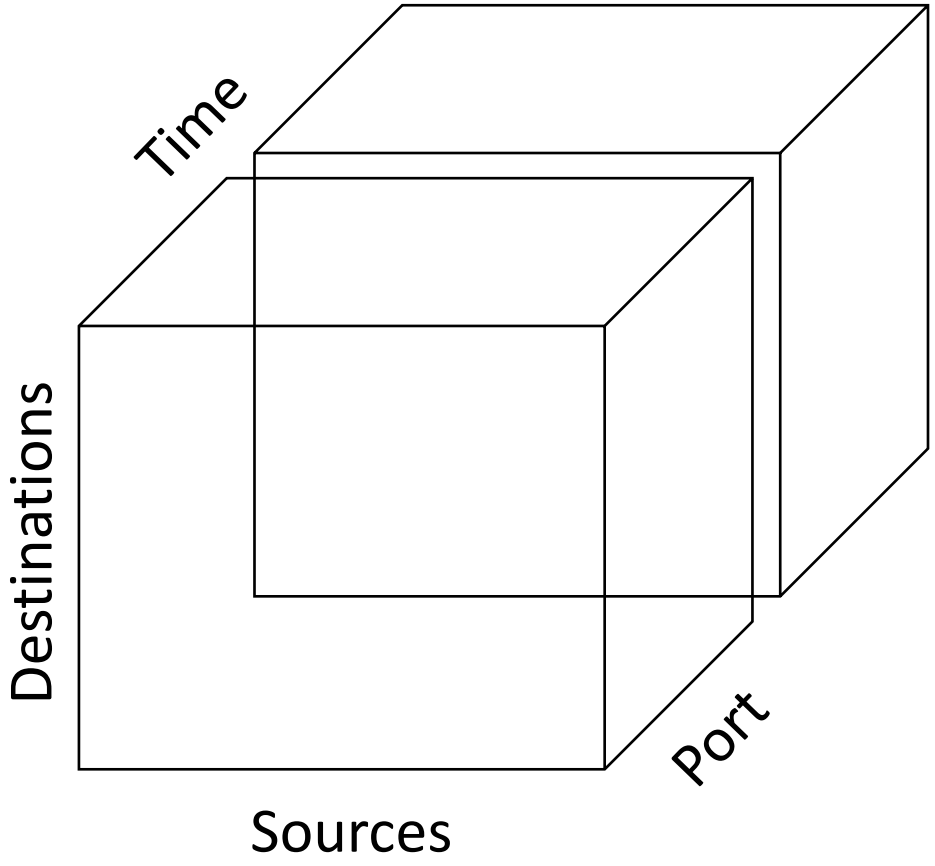
A 2 hours movie with 25 frame/seconds : 180000 images

Hu, Weiming, et al. "Incremental tensor subspace learning and its applications to foreground segmentation and tracking." *International Journal of Computer Vision* 91.3 (2011): 303-327.

5. Network security

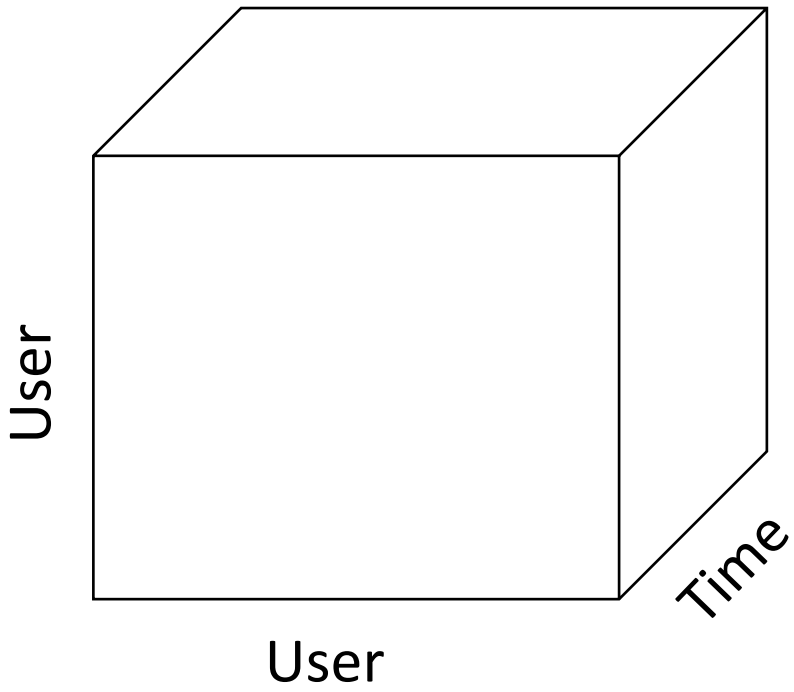
Intrusion Detection, Network Health Monitoring





Sun, Jimeng, Dacheng Tao, and Christos Faloutsos. "Beyond streams and graphs: dynamic tensor analysis." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

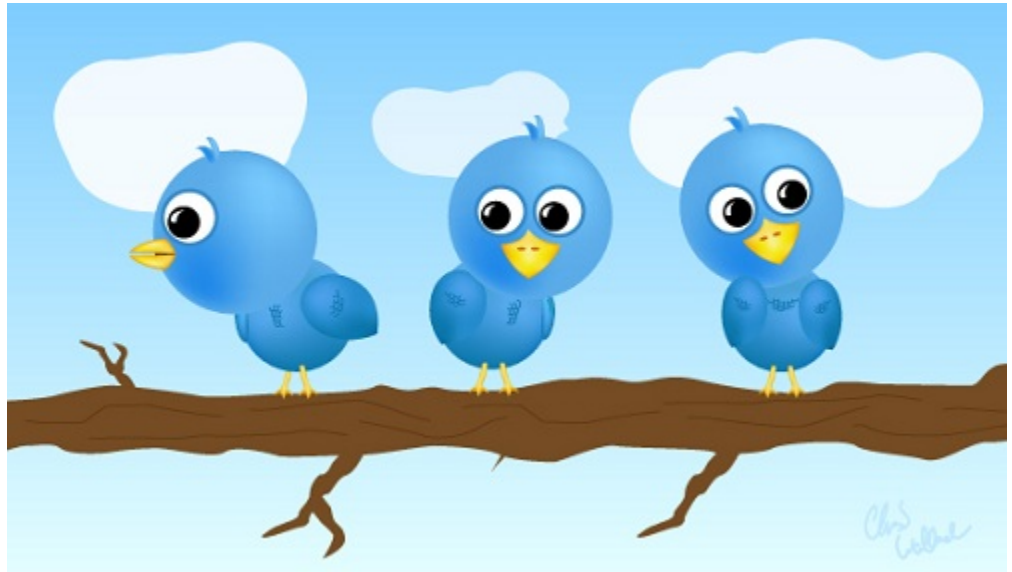
6. Social Networks Event and Anomaly Detection, Community Detection, Link Prediction



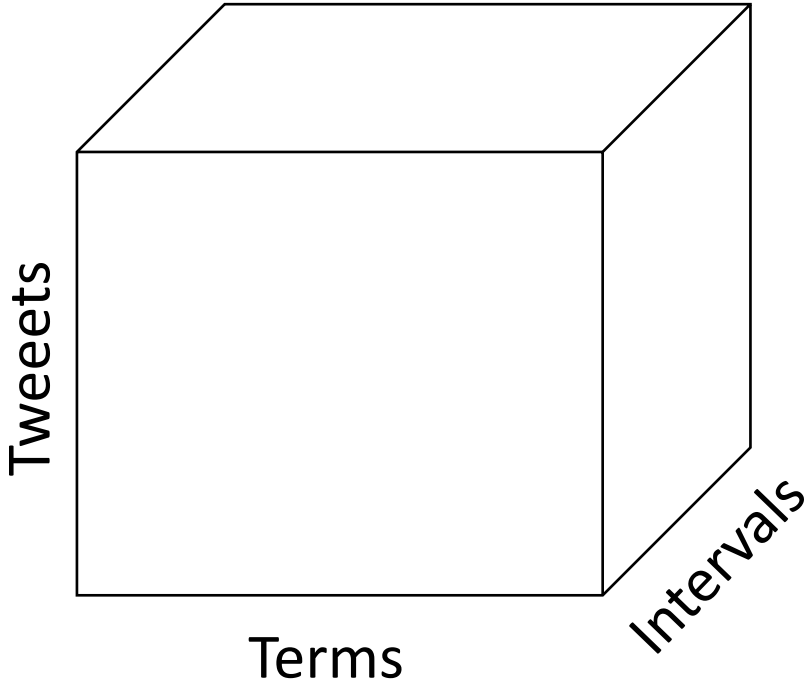
- Koutra, Danai, Evangelos E. Papalexakis, and Christos Faloutsos. "Tensorsplat: Spotting latent anomalies in time." *Informatics (PCI), 2012 16th Panhellenic Conference on*. IEEE, 2012.
- Ermiş, Beyza, Evrim Acar, and A. Taylan Cemgil. "Link prediction in heterogeneous data via generalized coupled tensor factorization." *Data Mining and Knowledge Discovery* 29.1 (2015): 203-236.
- Araujo, Miguel, et al. "Com2: Fast automatic discovery of temporal ('comet') communities." *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, 2014. 271-283.

7. Text-based Systems

Event Detection, Topic Models

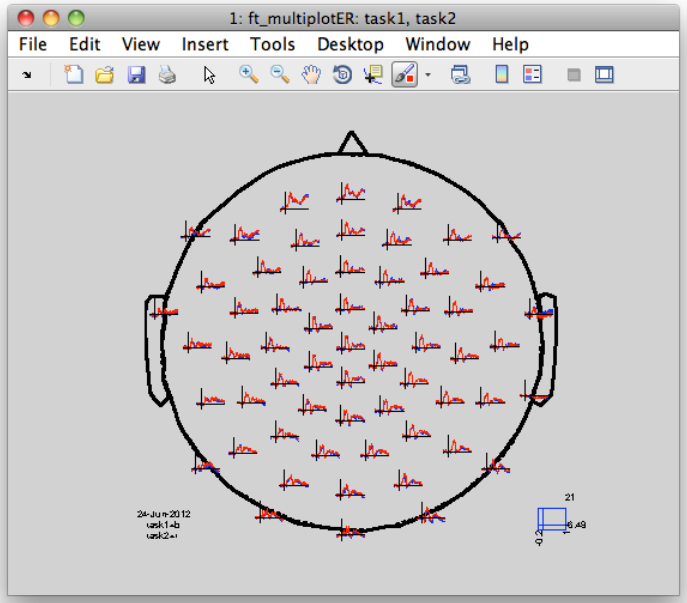


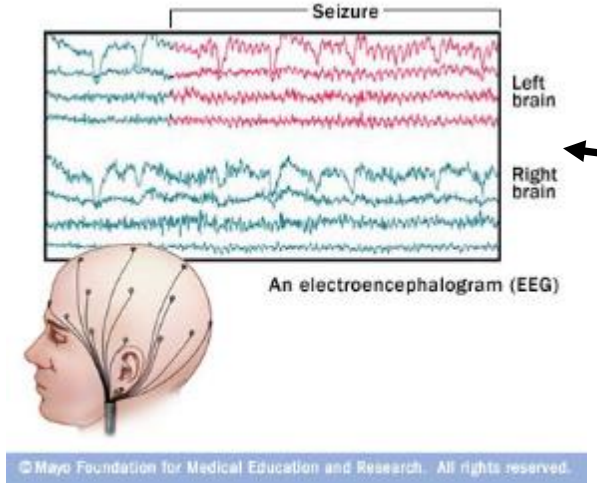
Analysis of Twitter Data

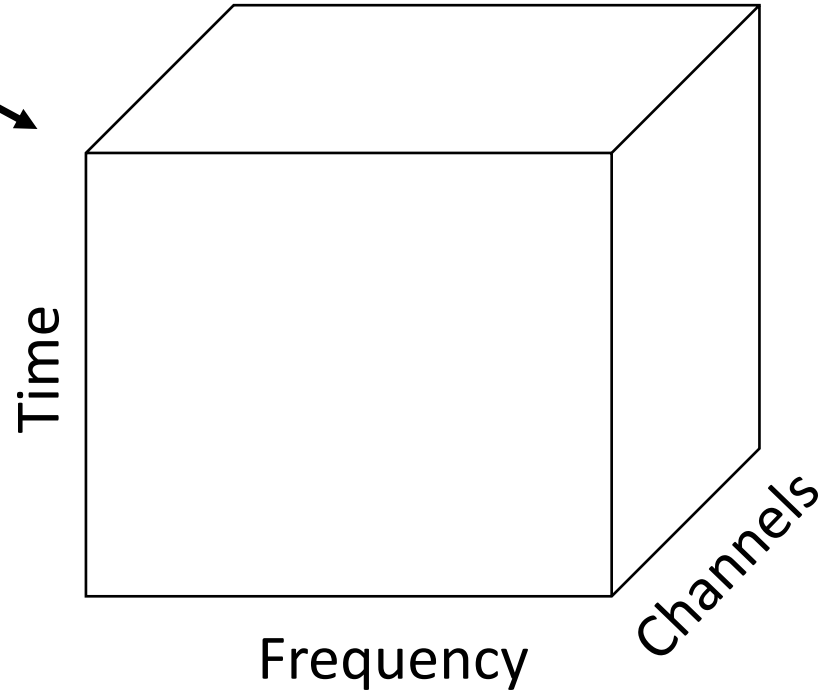


Panisson, André, et al. "Mining Concurrent Topical Activity in Microblog Streams." *arXiv preprint arXiv:1403.1403* (2014).

8.Neuroscience (EEG) Activity Recognition, Disease Detection



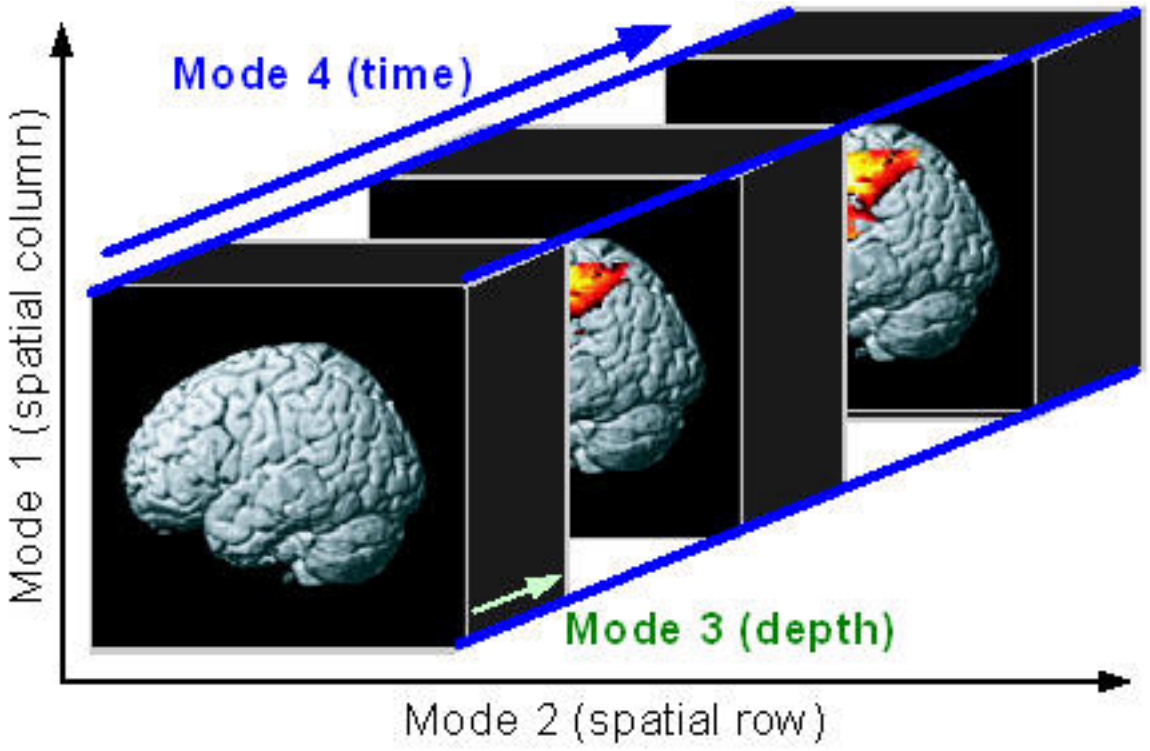


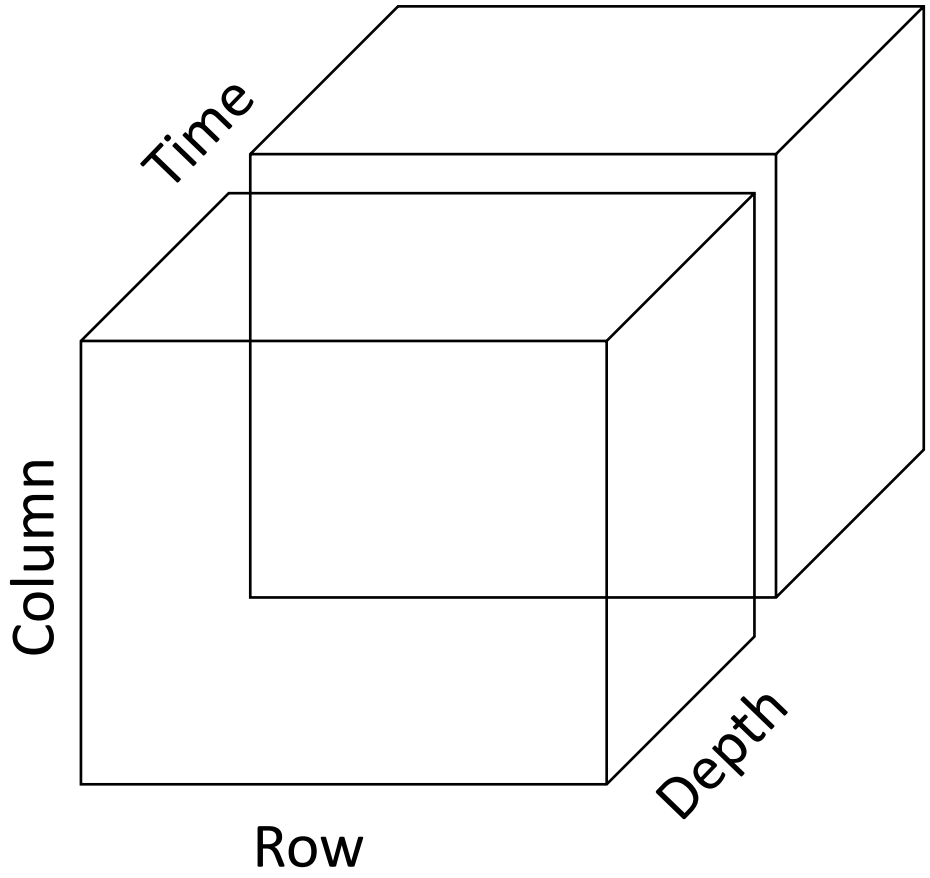


Acar, Evrim, et al. "Multiway analysis of epilepsy tensors." *Bioinformatics* 23.13 (2007): i10-i18.

9. Neuroscience (fMRI)

Activity Recognition, Disease Detection

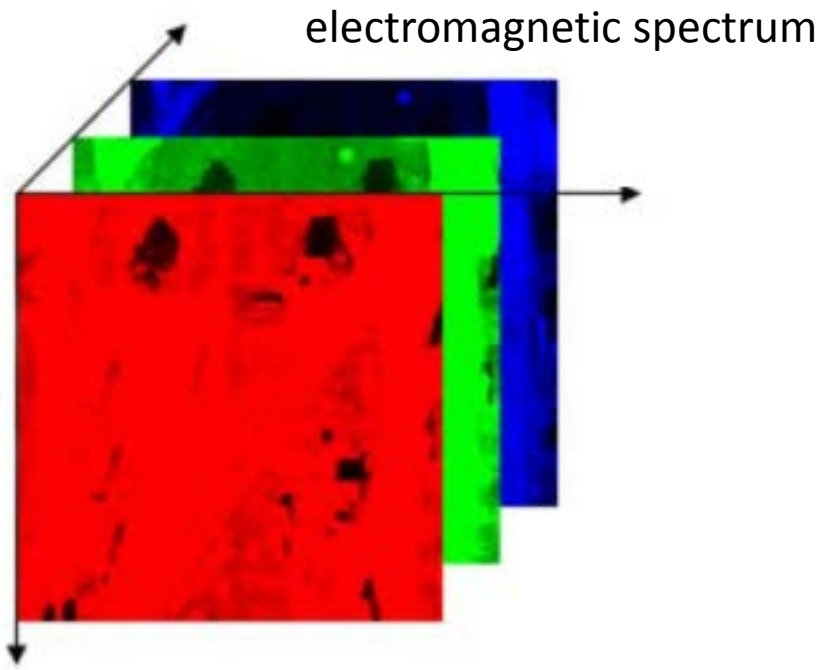




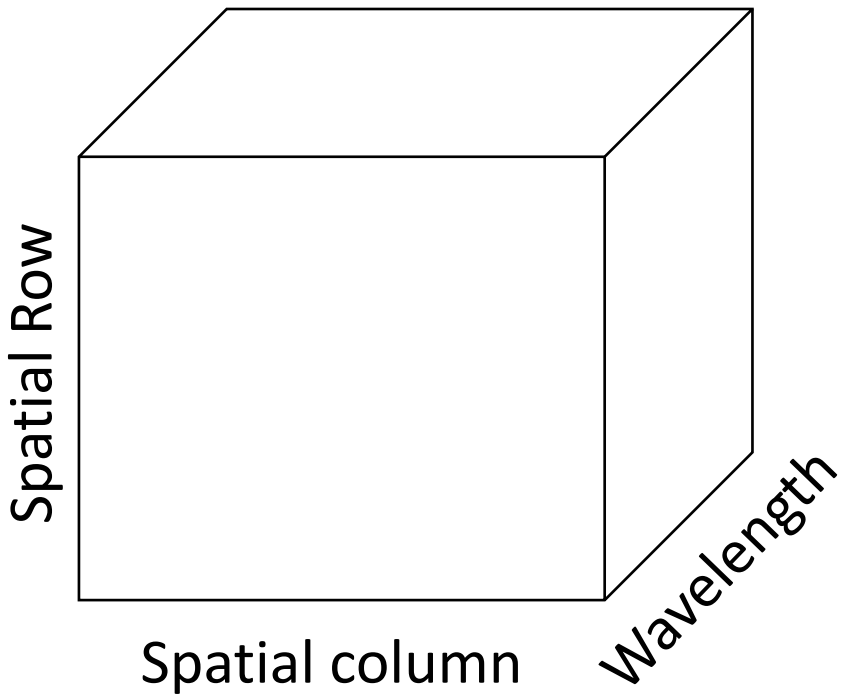
10. Remote sensing

Target Detection, Remote Material Identification

Spectral signatures : Certain objects leave unique 'fingerprints' in the electromagnetic spectrum. Known as For oil spectral signature → find new oil fields

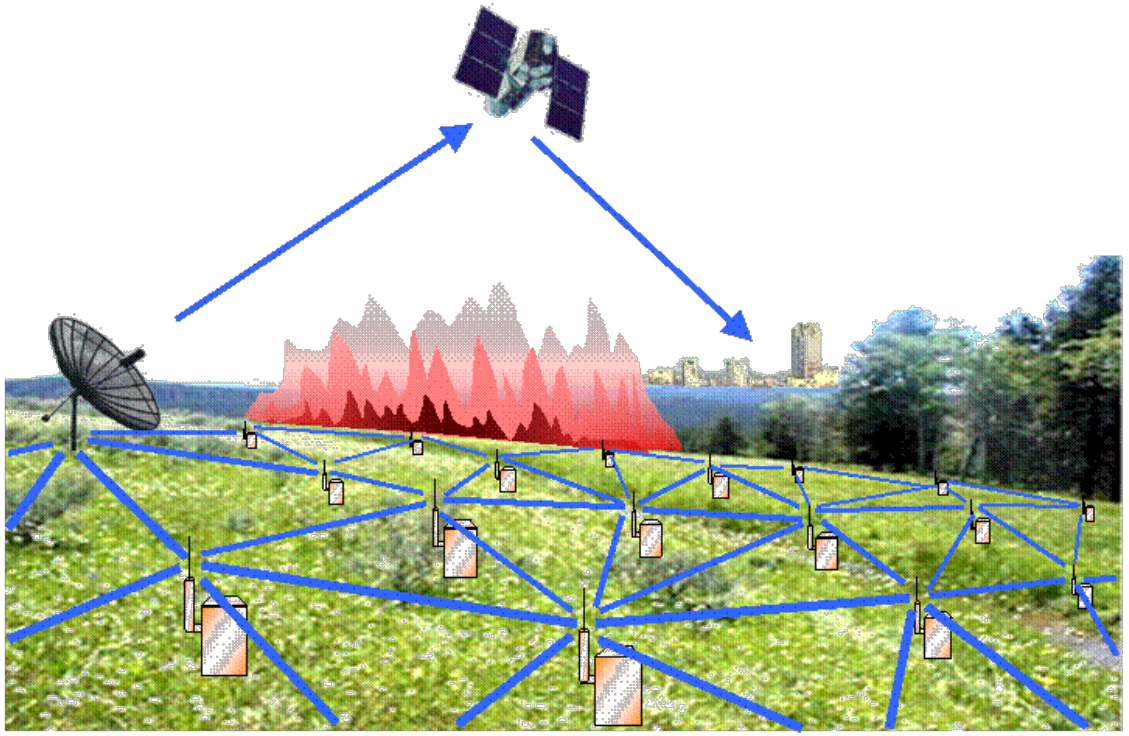


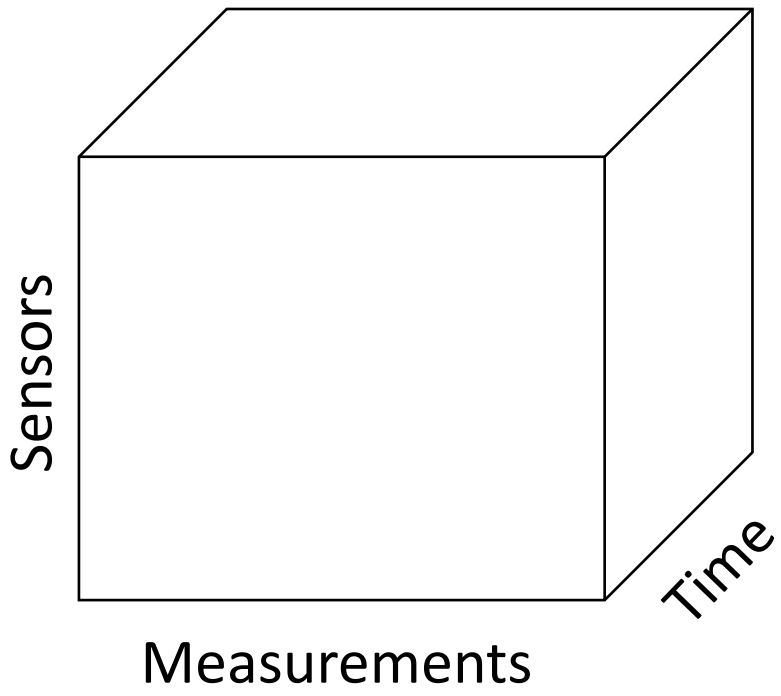
Hyperspectral image



Bourennane, Salah, Caroline Fossati, and Alexis Cailly. "Improvement of classification for hyperspectral images based on tensor modeling." *Geoscience and Remote Sensing Letters, IEEE* 7.4 (2010): 801-805.

11.Sensors Event Detection



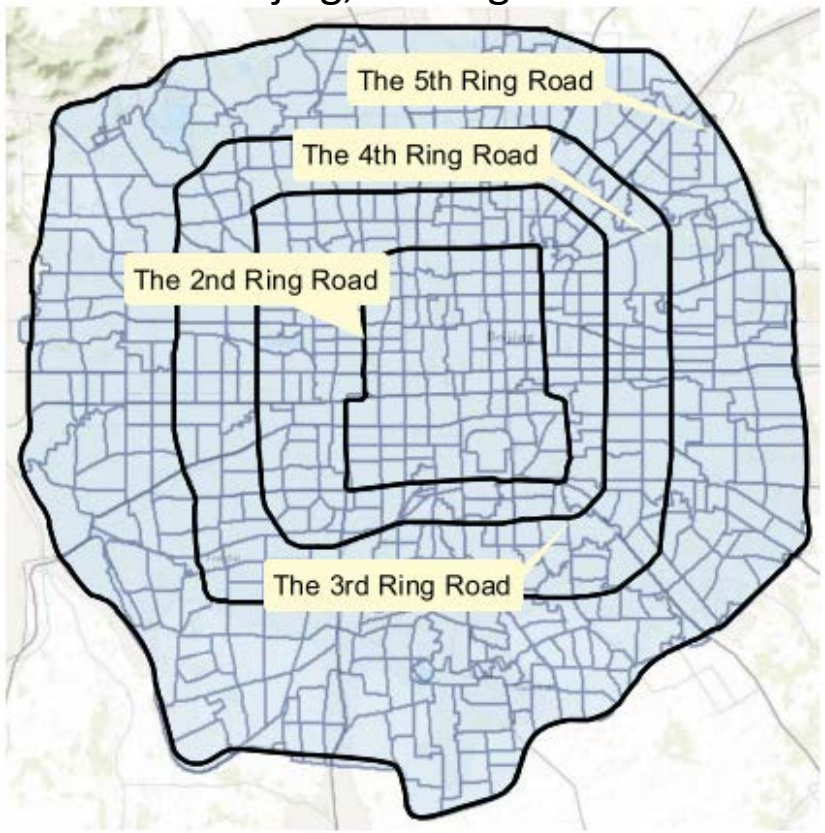


Sun, Jimeng, Spiros Papadimitriou, and S. Yu Philip. "Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams." *ICDM*. Vol. 6. 2006.

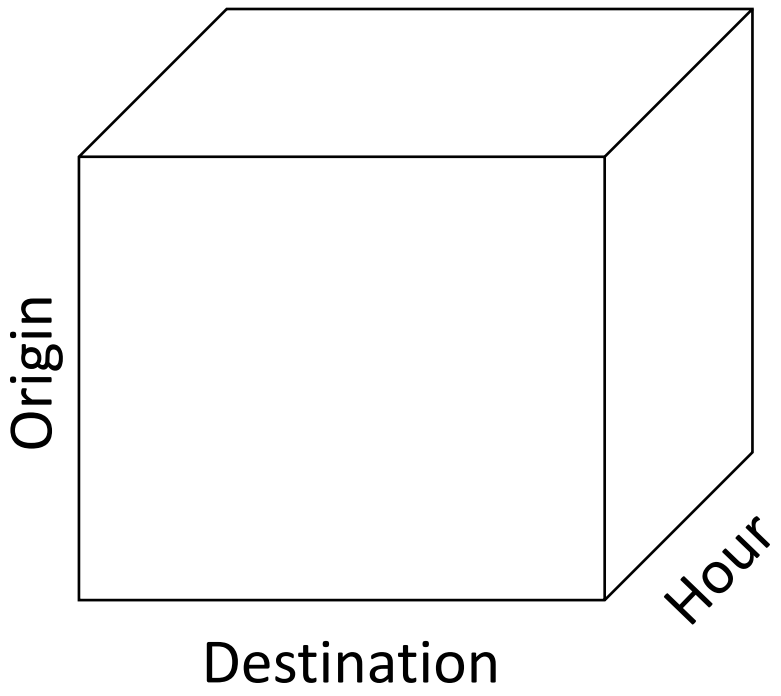
12. Transportation

Discovering Urban Spatio-temporal Structure

Beijing, China grid



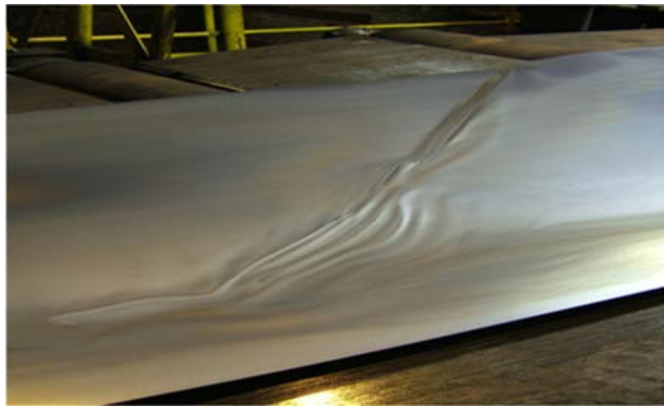
651 Regions x 651 Regions x 24 Hour



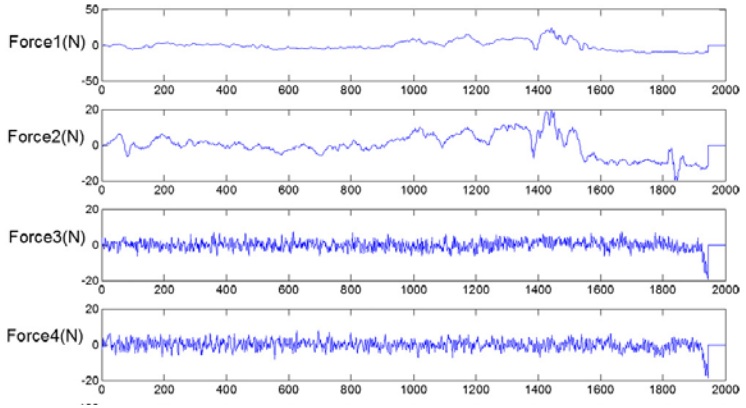
Jingyuan Wang, Fei Gao, Peng Cui, Chao Li, and Zhang Xiong. 2014. Discovering Urban Spatio-temporal Structure from Time-Evolving Traffic Networks. In Web Technologies and Applications. Springer, 93–104.

13. Metallurgy Engineering

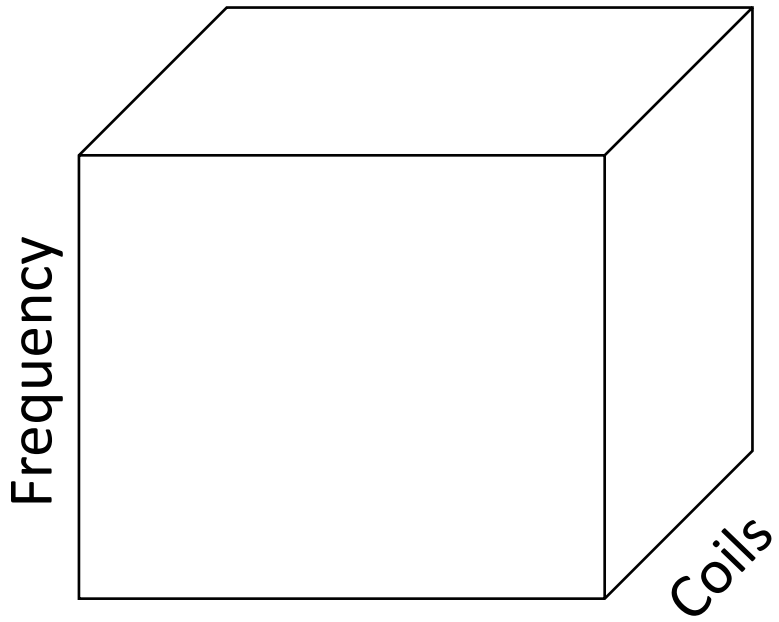
Surface damage detection in hot strip mill



hot strip mill
In steel production



I coils , *J* Signals → Transform via AR model to PSD x Frequency

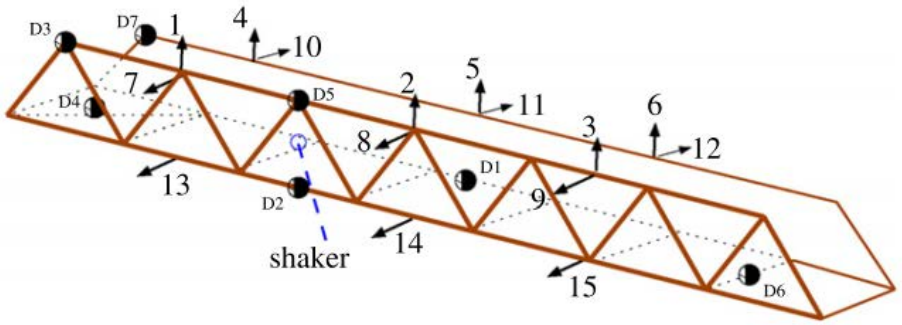


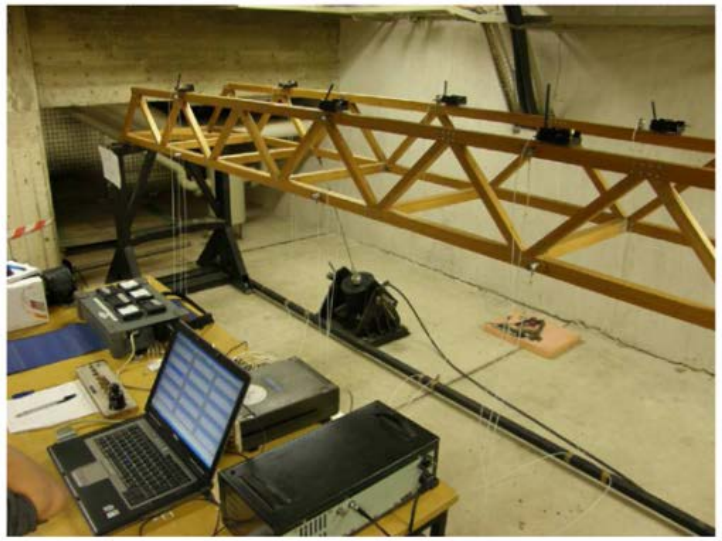
PSD
(Power Spectrum Density)

Chuang, Wei-Li, et al. "Using MPCA of spectra model for fault detection in a hot strip mill." *Journal of Materials Processing Technology* 209.8 (2009): 4162-4168.

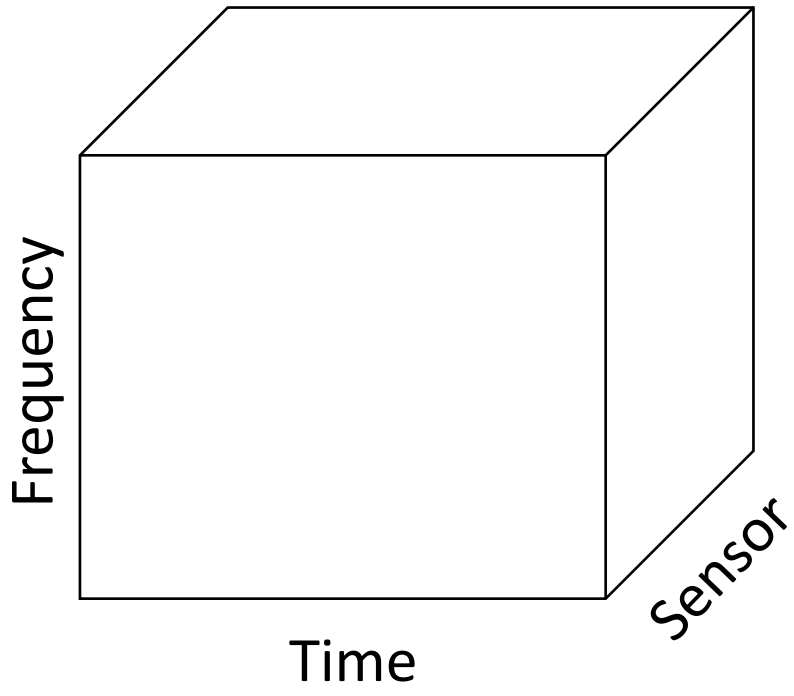
14. Civil Structures

Novelty Detection, Structural health monitoring (vibration monitoring)





Transmissibility magnitudes Signals → Fast Fourier Transforms (FFT) -> Time-Frequency

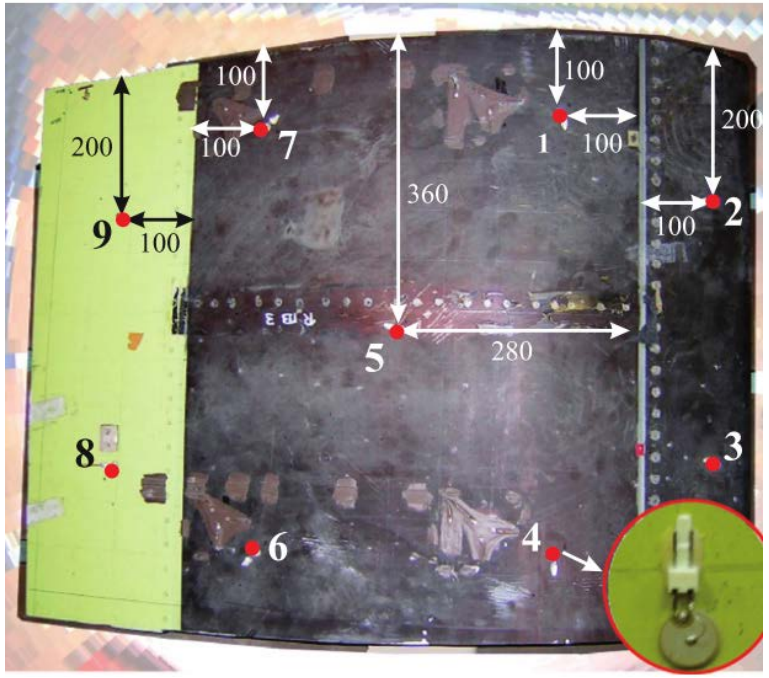


Prada, Miguel A., et al. "Three-way analysis of structural health monitoring data." *Neurocomputing* 80 (2012): 119-128.

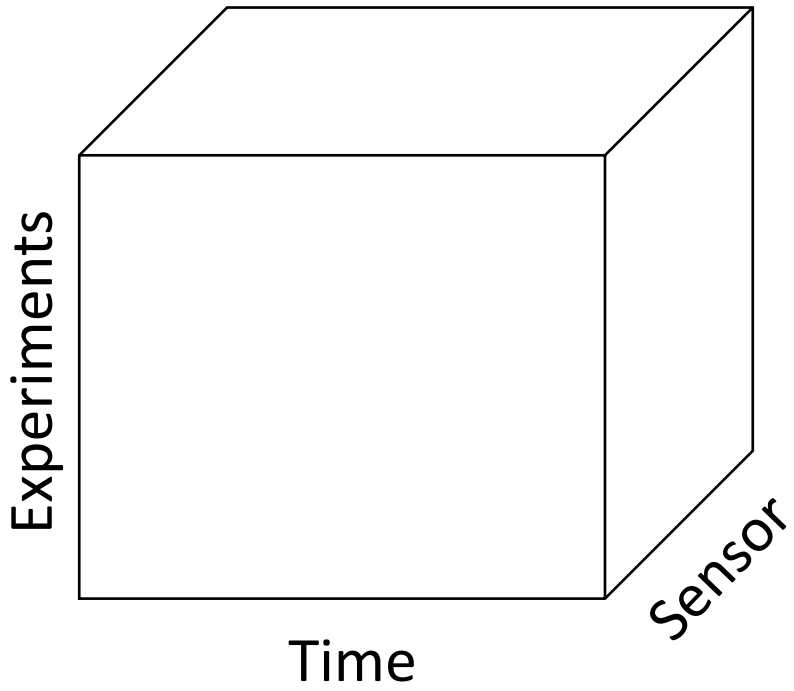
15.Mechanical Systems

Damage detection in an aircraft structure

Aircraft wing flap



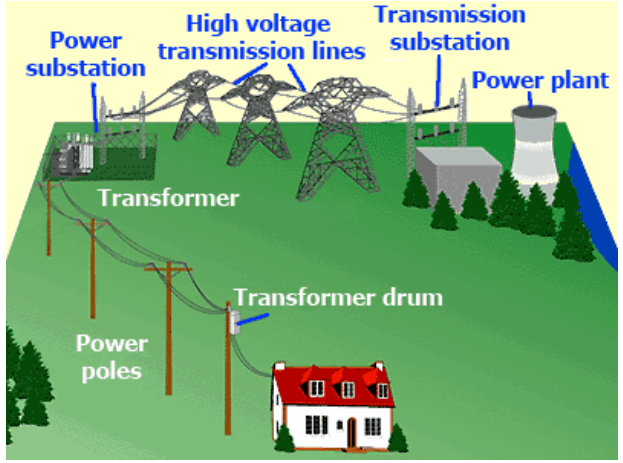
Impact locations

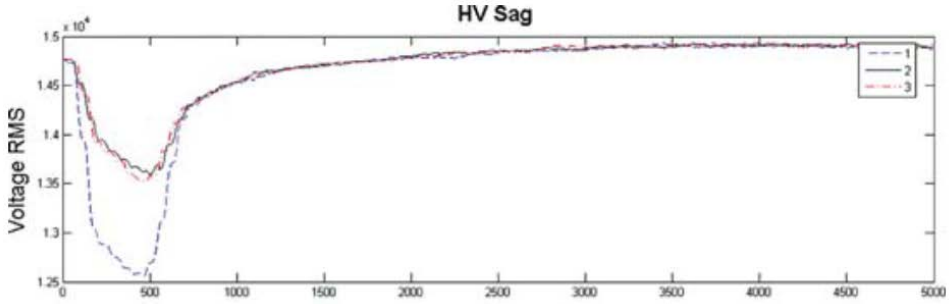


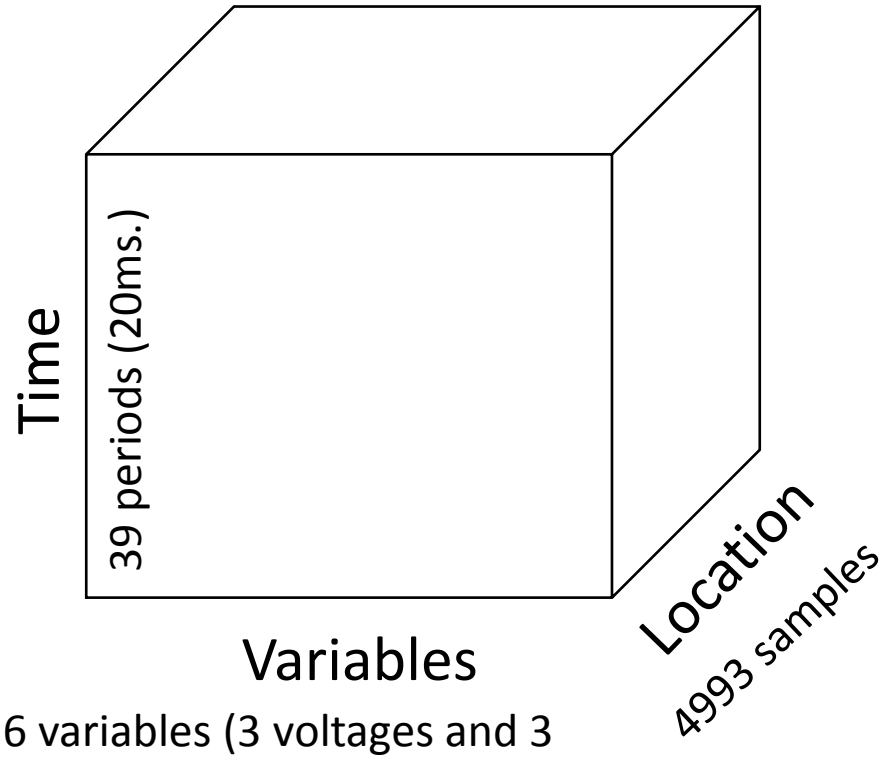
Mujica, Luis E., et al. "Multivariate statistics process control for dimensionality reduction in structural assessment." *Mechanical Systems and Signal Processing* 22.1 (2008): 155-171.

16. Power systems

Voltage Sag detection in power distribution networks





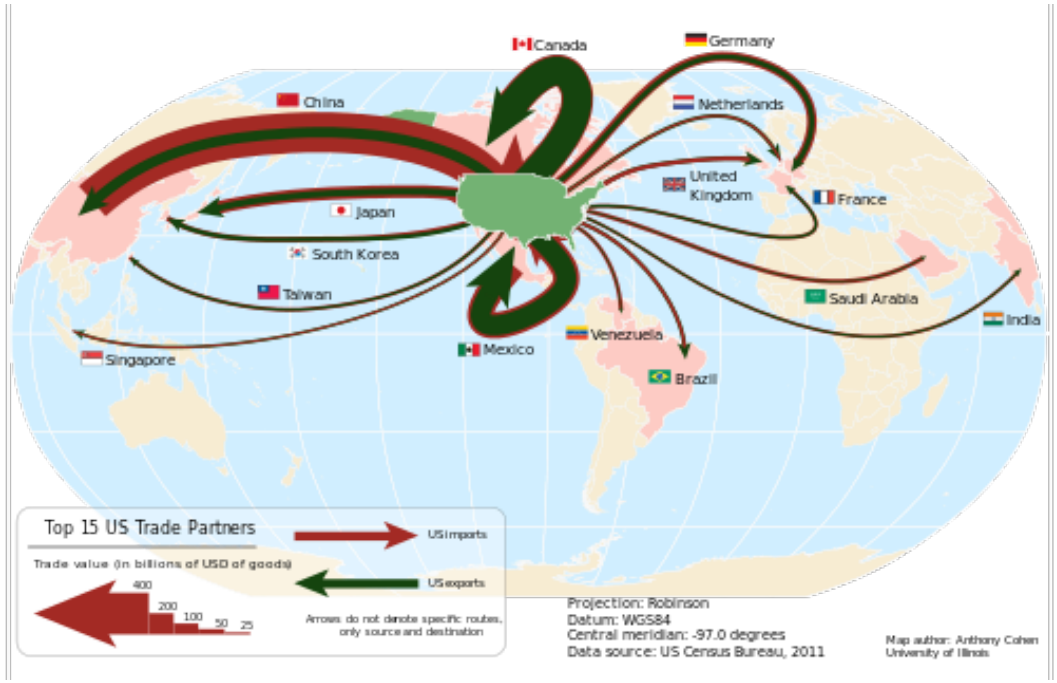


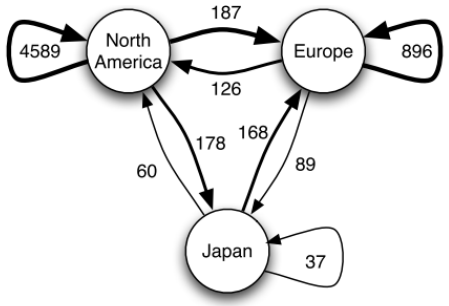
6 variables (3 voltages and 3 currents) sampled at 6400Hz

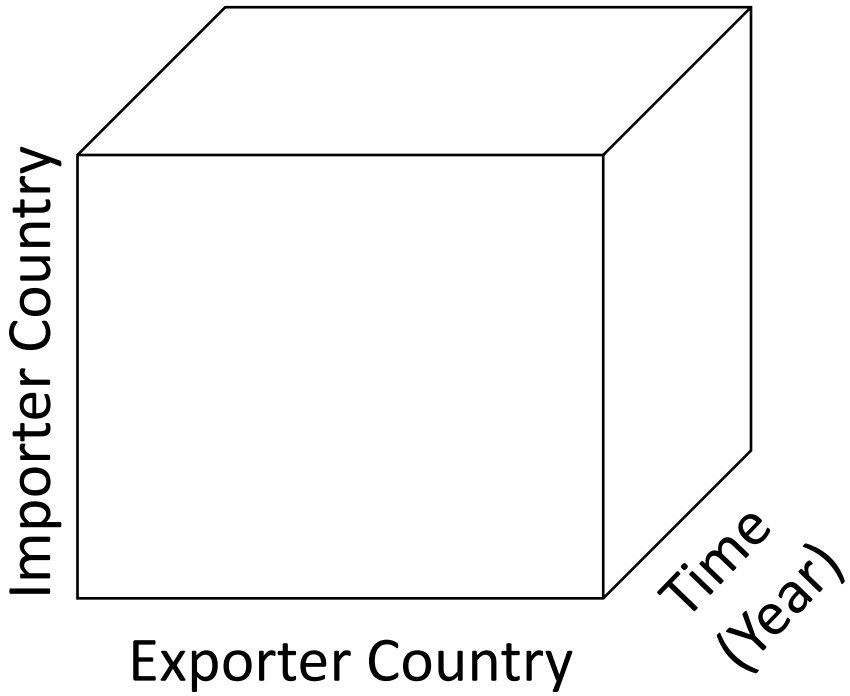
Khosravi, Abbas, et al. "Multiway principal component analysis (MPCA) for upstream/downstream classification of voltage sags gathered in distribution substations." *Advances of Computational Intelligence in Industrial Systems*. Springer Berlin Heidelberg, 2008. 297-312.

17.Economy

Analysis of international trade



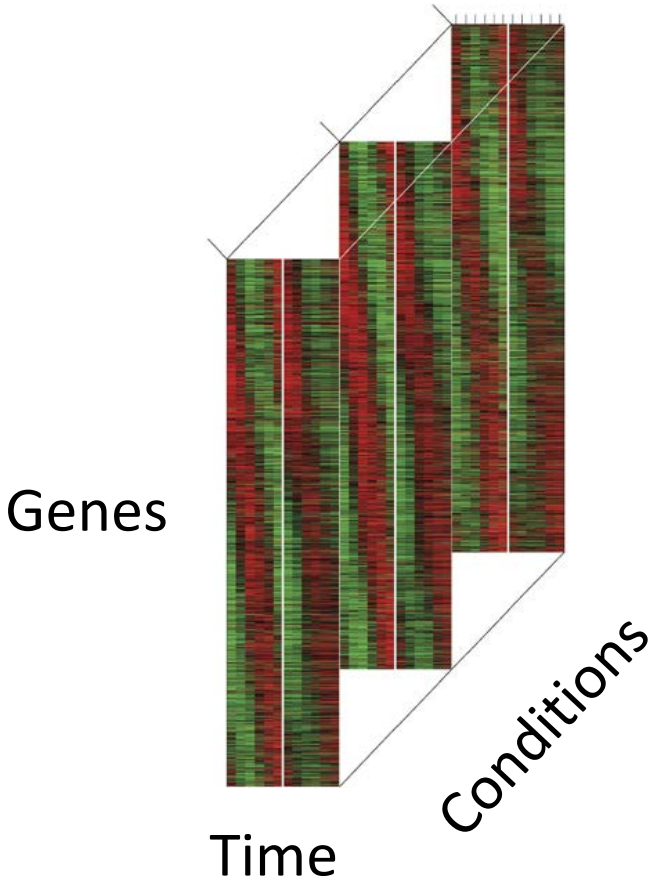


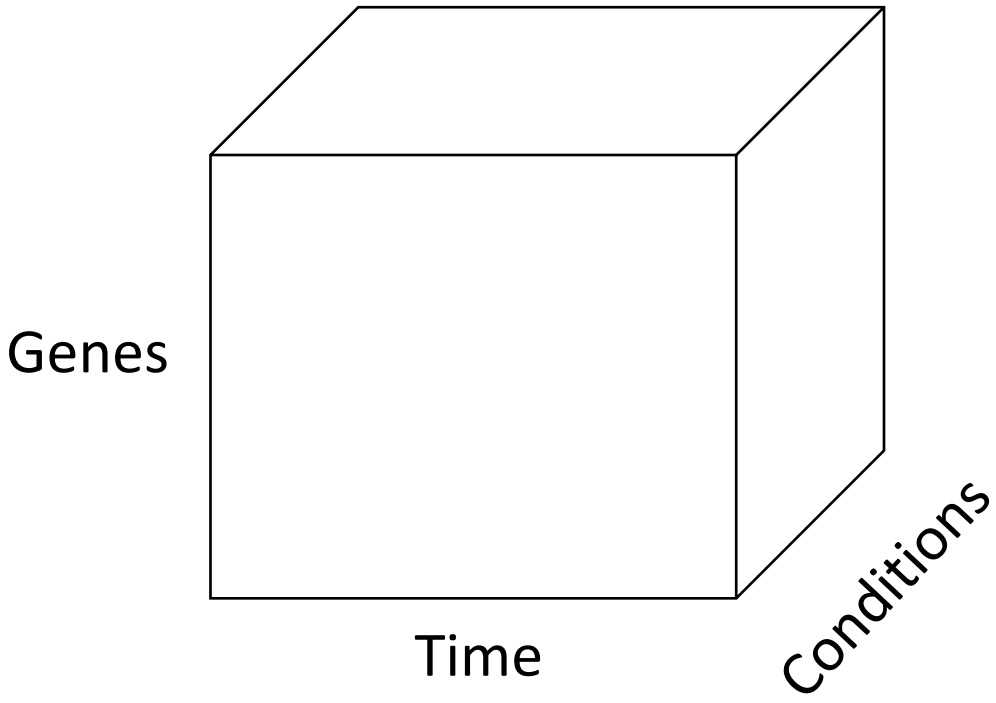


Bader, Brett W., Richard A. Harshman, and Tamara G. Kolda. "Temporal analysis of semantic graphs using ASALSAN." *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007.

18. Bio-informatics

DNA data (micro-array) pre-processing (remove experimental artifacts), Detect related genes



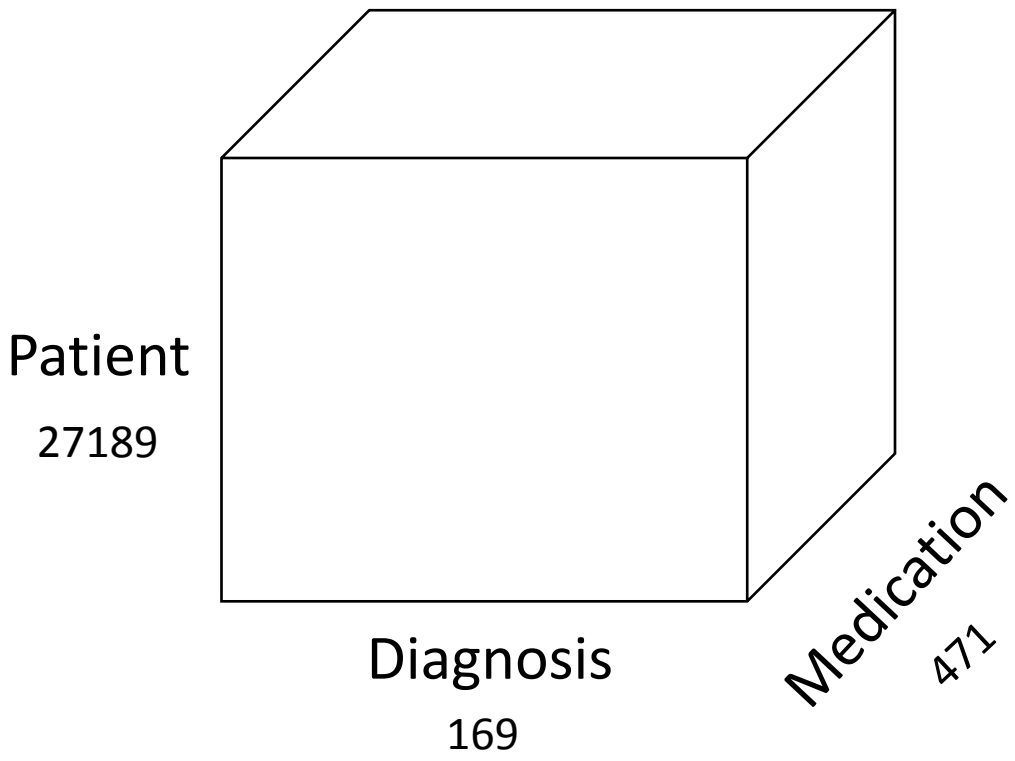


Omberg, Larsson, et al. "Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression." *Molecular systems biology* 5.1 (2009).

19. Medical Diagnosis

Identify multiple candidate phenotypes (clusters) from EHR data
Heart failure predication

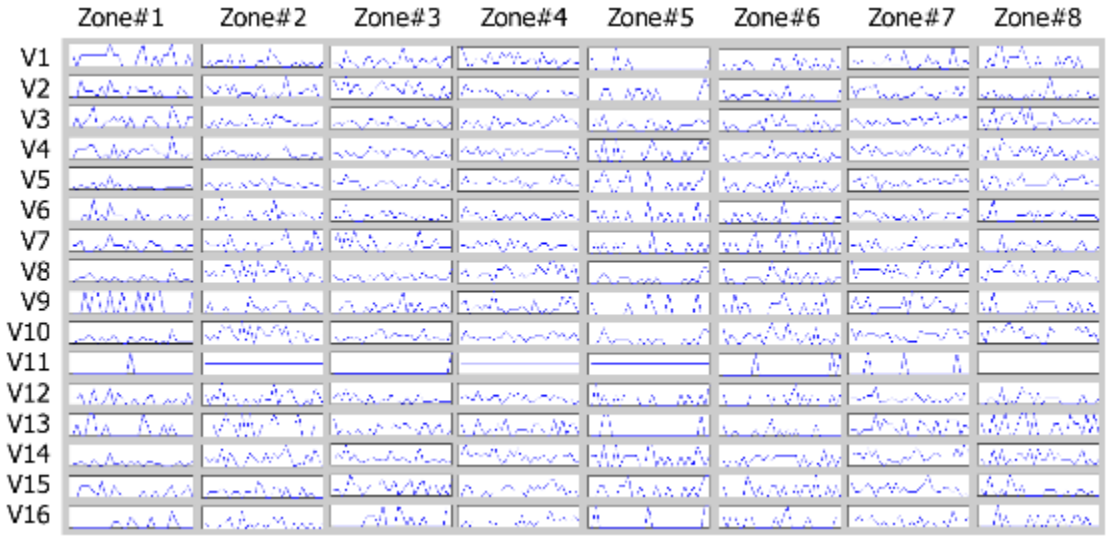
Hyperlipidemia	Moderate Hypertension	Uncomplicated Diabetes	Mild Hypertension	Chronic Respiratory Inflammation/Infection
Phenotype 1 (41.6% of patients)	**Phenotype 2 (31.5% of patients)**	**Phenotype 3 (17.6% of patients)**	**Phenotype 4 (31.1% of patients)**	**Phenotype 5 (36.7% of patients)**
Other Endocrine, Metabolic, and Nutritional Disorders	Hypertension	Diabetes with No or Unspecified Complications	Hypertension	Other Ear, Nose, Throat, and Mouth Disorders
HMG CoA Reductase Inhibitors	Beta Blockers Cardio-Selective	Sulfonylureas	ACE Inhibitors	Viral and Unspecified Pneumonia, Pleurisy
Intestinal Cholesterol Absortpion Inhibitors	Angiotensin II Receptor Antagonists	Biguanides	Thiazides and Thiazide-Like Diuretics	Significant Ear, Nose, and Throat Disorders
Fibric Acid Derivatives	Loop Diuretics	Diagnostic Tests		Cough/Cold/Allergy Combinations
Antihyperlipidemics - Combinations	Potassium	Insulin Sensitizing Agents		Azithromycin
Nicotinic Acid Derivatives	Nitrates	Diabetic Supplies		Fluoroquinolones
Bile Acid Sequestrants	Alpha-Beta Blockers	Meglitinide Analogues		Sympathomimetics
Oil Soluble Vitamins	Vasodilators	Antidiabetic Combinations		Penicillin Combinations
				Antitussives
				Glucocorticosteroids
				Tetracyclines
				Anti-infective Misc. - Combinations
				Clarithromycin
				Cephalosporins - 2nd Generation
				Cephalosporins - 1st Generation
				Expectorants

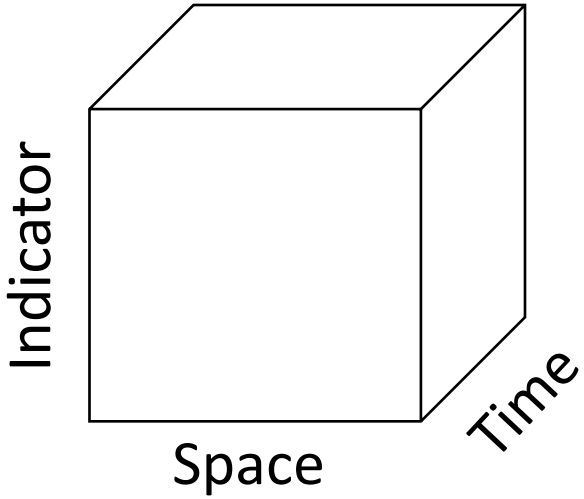


Ho, Joyce C., et al. "Limestone: High-throughput candidate phenotype generation via tensor factorization." *Journal of biomedical informatics* 52 (2014): 199-211.

20. Epidemiology

Syndromic Surveillance : Disease Outbreak Prediction

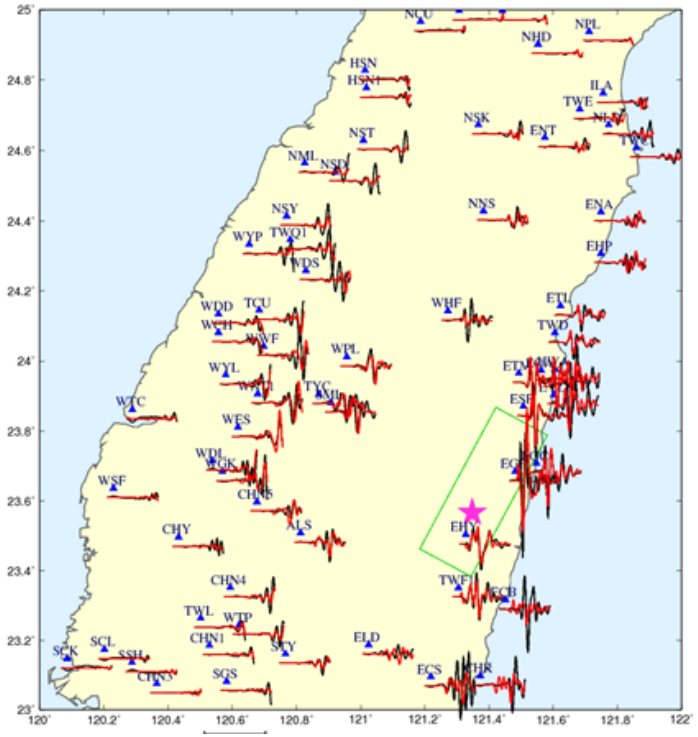




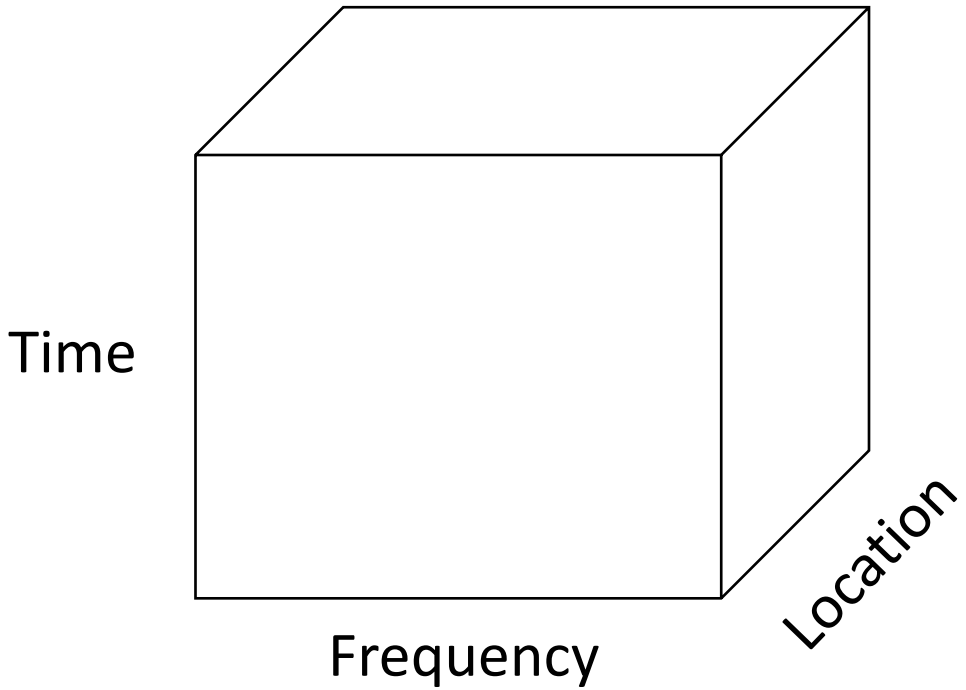
Hadi Fanaee-T, Joao Gama, EigenEvent: An Algorithm for Event Detection from Complex Data streams in Syndromic Surveillance , Intelligent Data Analysis , Volume 19(3), PP. 597–616, June 2015, IOS Press. DOI: 10.3233/IDA-150734

21. Seismology

Predicting Earthquake Ground Motion



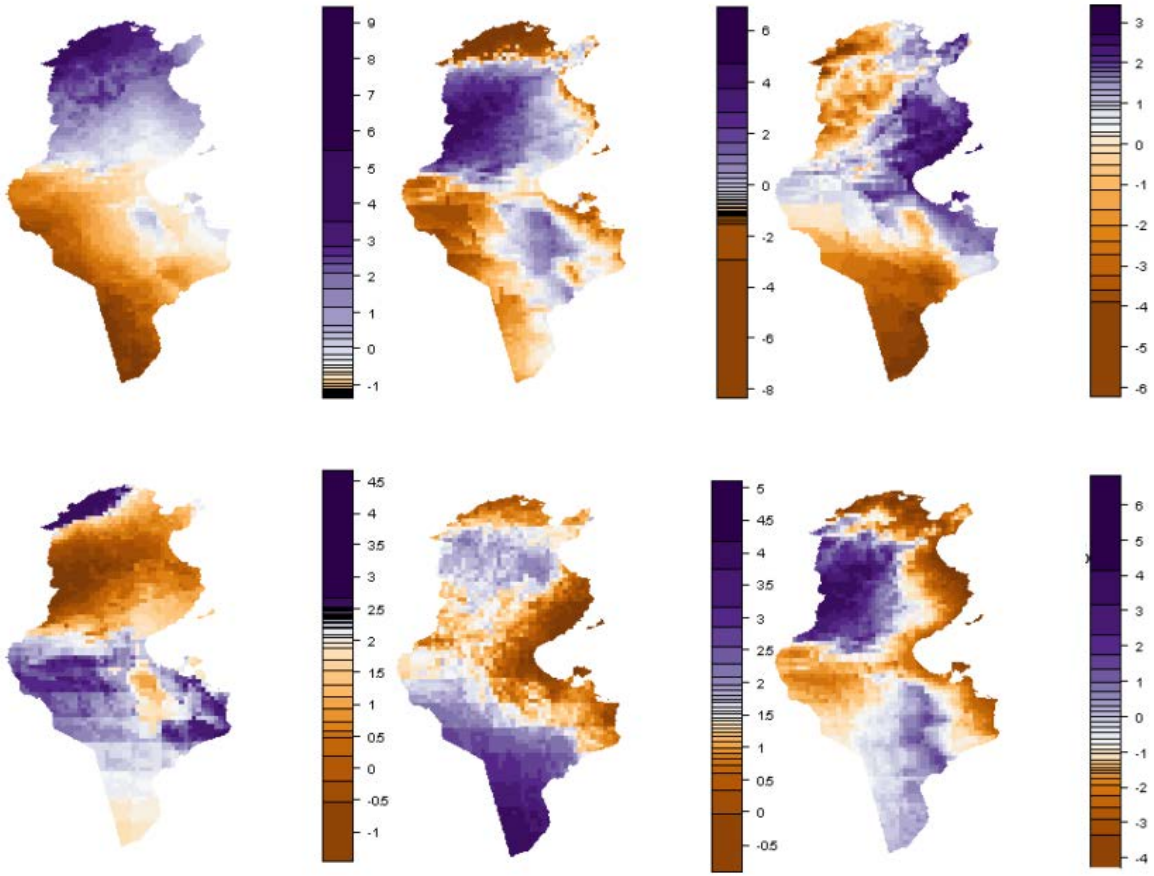
Acceleration records → Wavelet → Time-Frequency

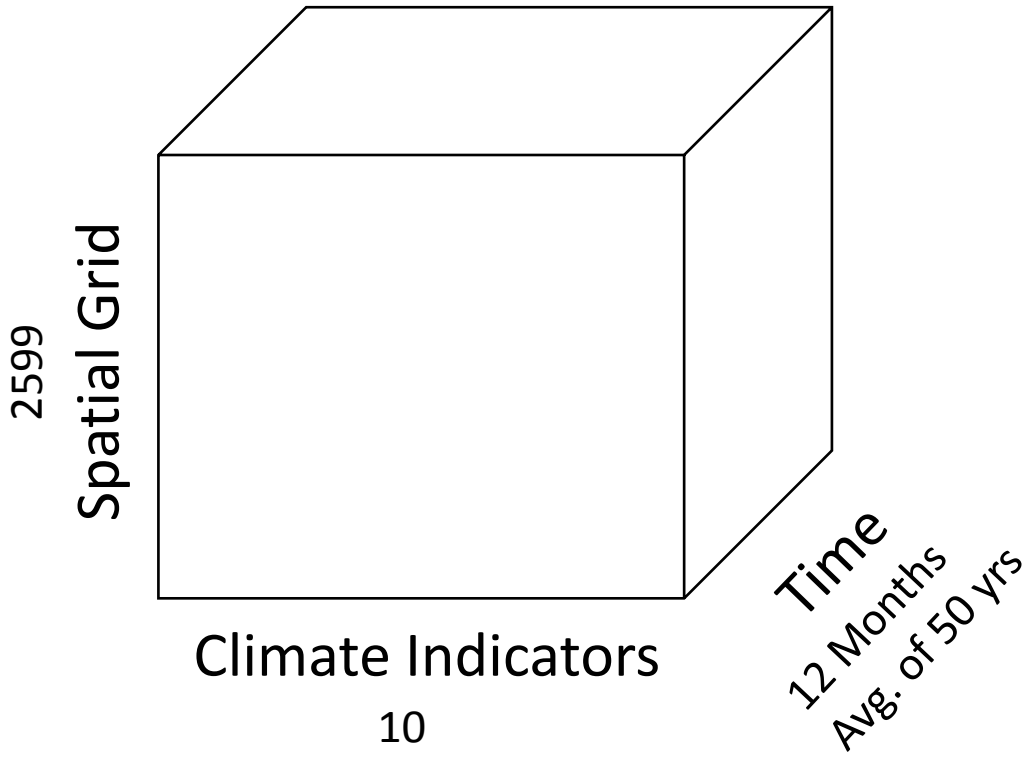


Bai, Yang, et al. "A Multiway Model for Predicting Earthquake Ground Motion." *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2013 14th ACIS International Conference on*. IEEE, 2013.

22. Climatology

Climate Analysis, Climate Change Detection

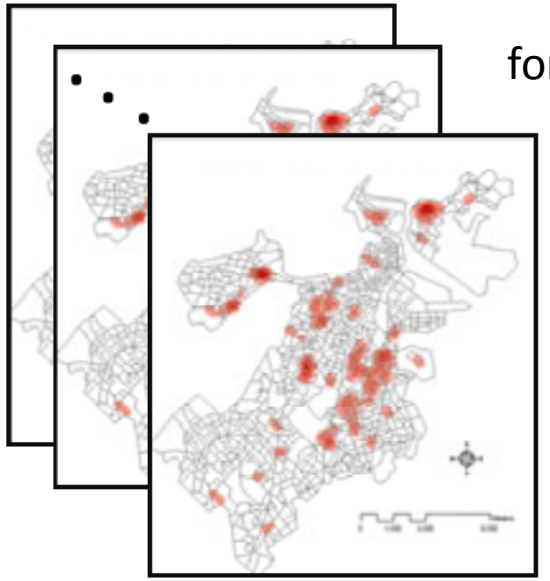




Leibovici, Didier G. "Spatio-temporal multiway decomposition using principal tensor analysis on k-modes: the R package PTAk." *Journal of Statistical Software* 34.10 (2010): 1-34.

23. Criminology

Crime Forecasting



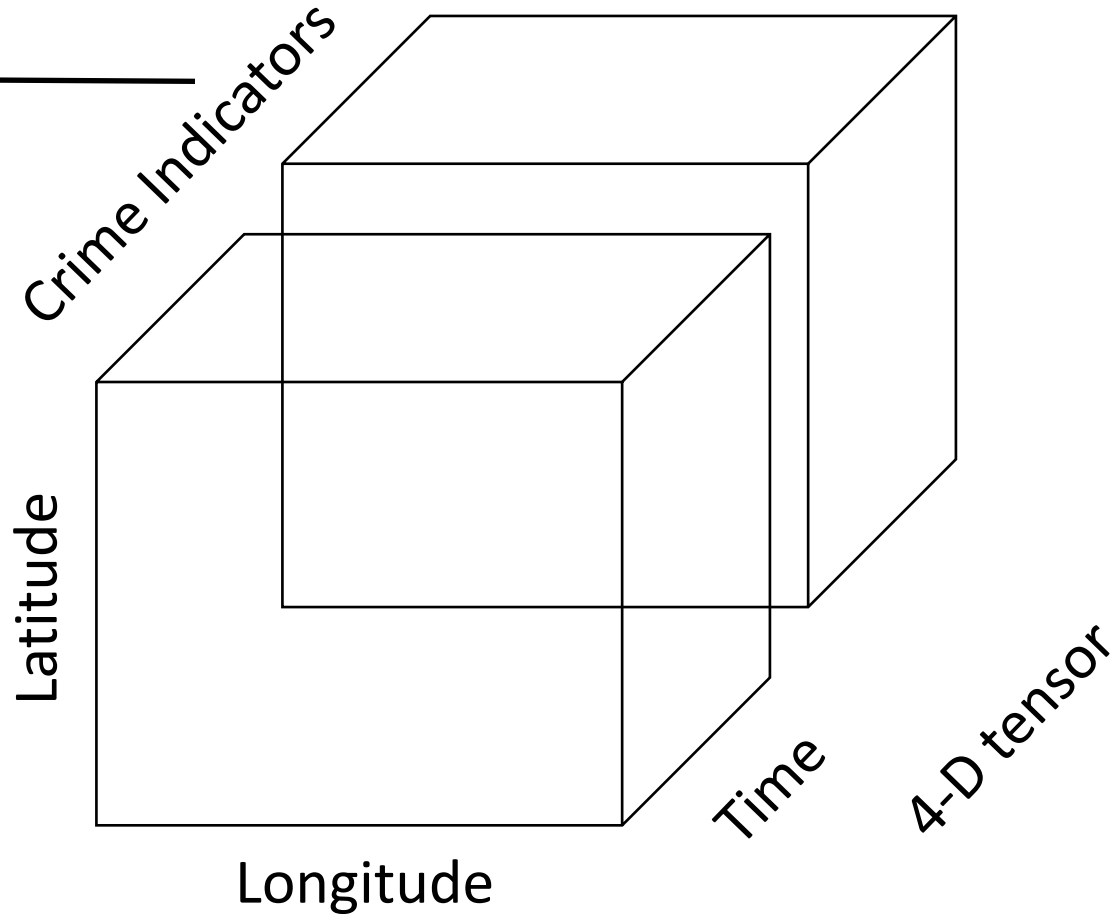
foreclosure

Construction permits

Vehicle larceny

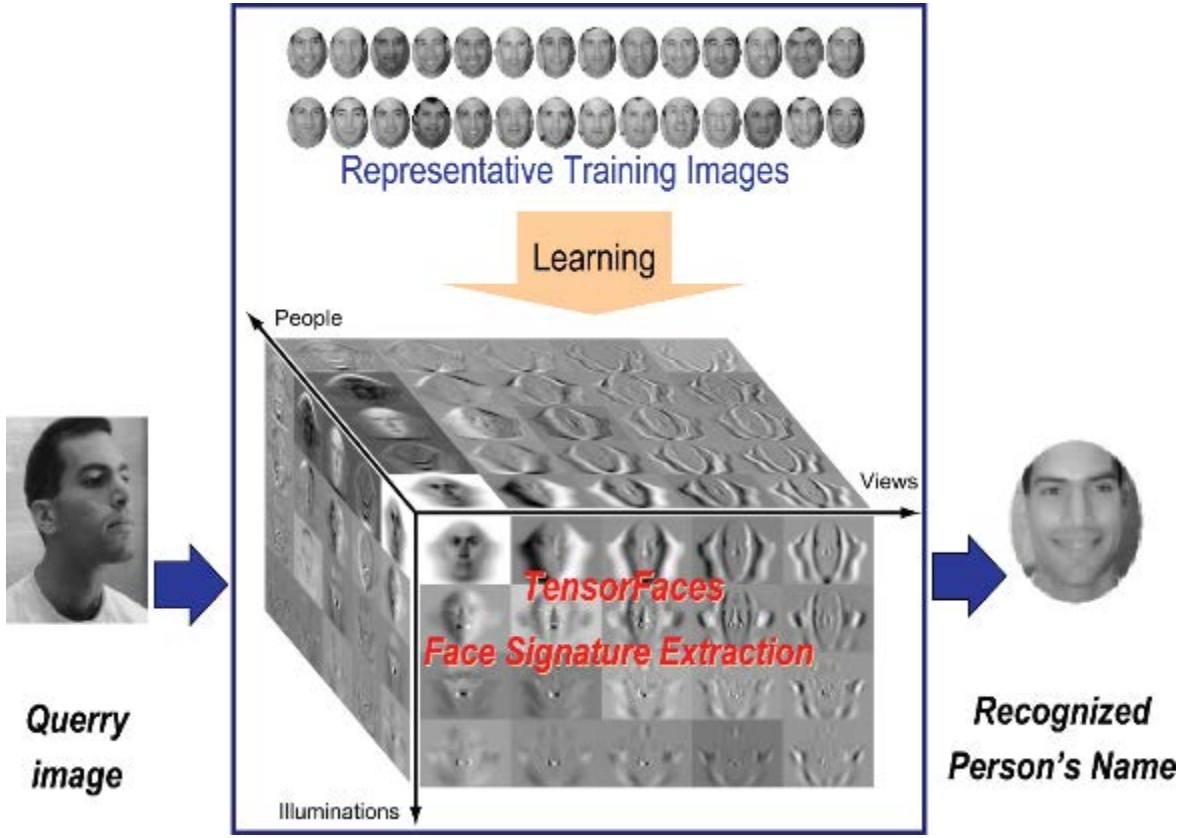
social events

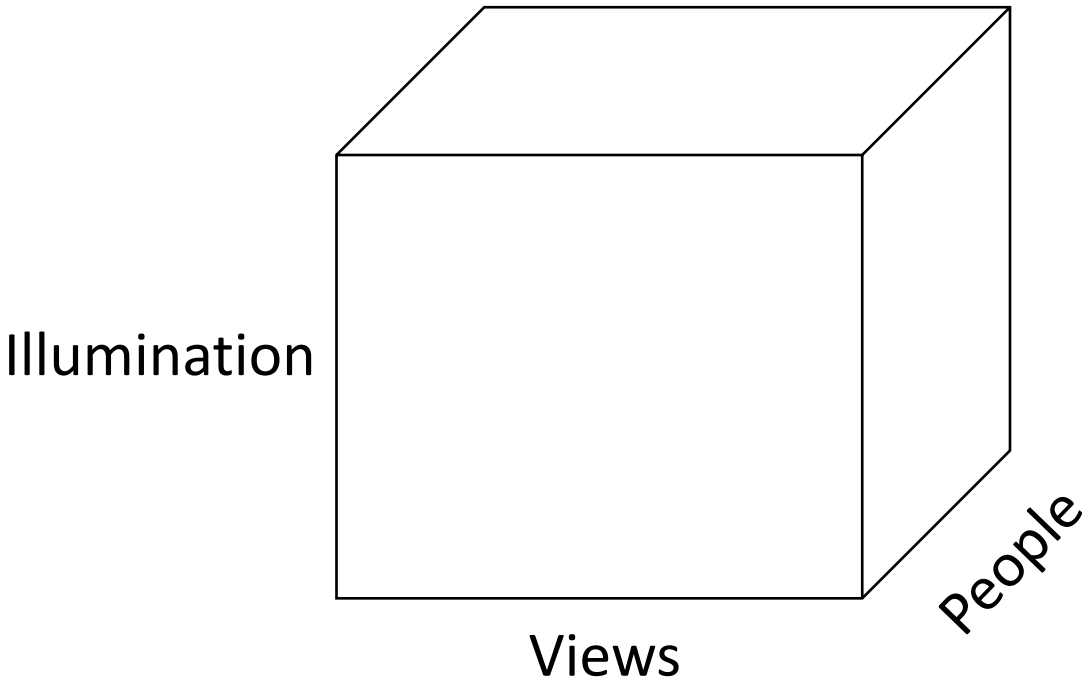
offender



Mu, Yang, et al. "Empirical discriminative tensor analysis for crime forecasting." *Knowledge Science, Engineering and Management*. Springer Berlin Heidelberg, 2011. 293-304.

24. Face Recognition





Vasilescu, M. Alex O., and Demetri Terzopoulos. "Multilinear analysis of image ensembles: Tensorfaces." *Computer Vision—ECCV 2002*. Springer Berlin Heidelberg, 2002. 447-460.

25. Image Completition

Liu, Ji, et al. "Tensor completion for estimating missing values in visual data." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.1 (2013): 208-220.

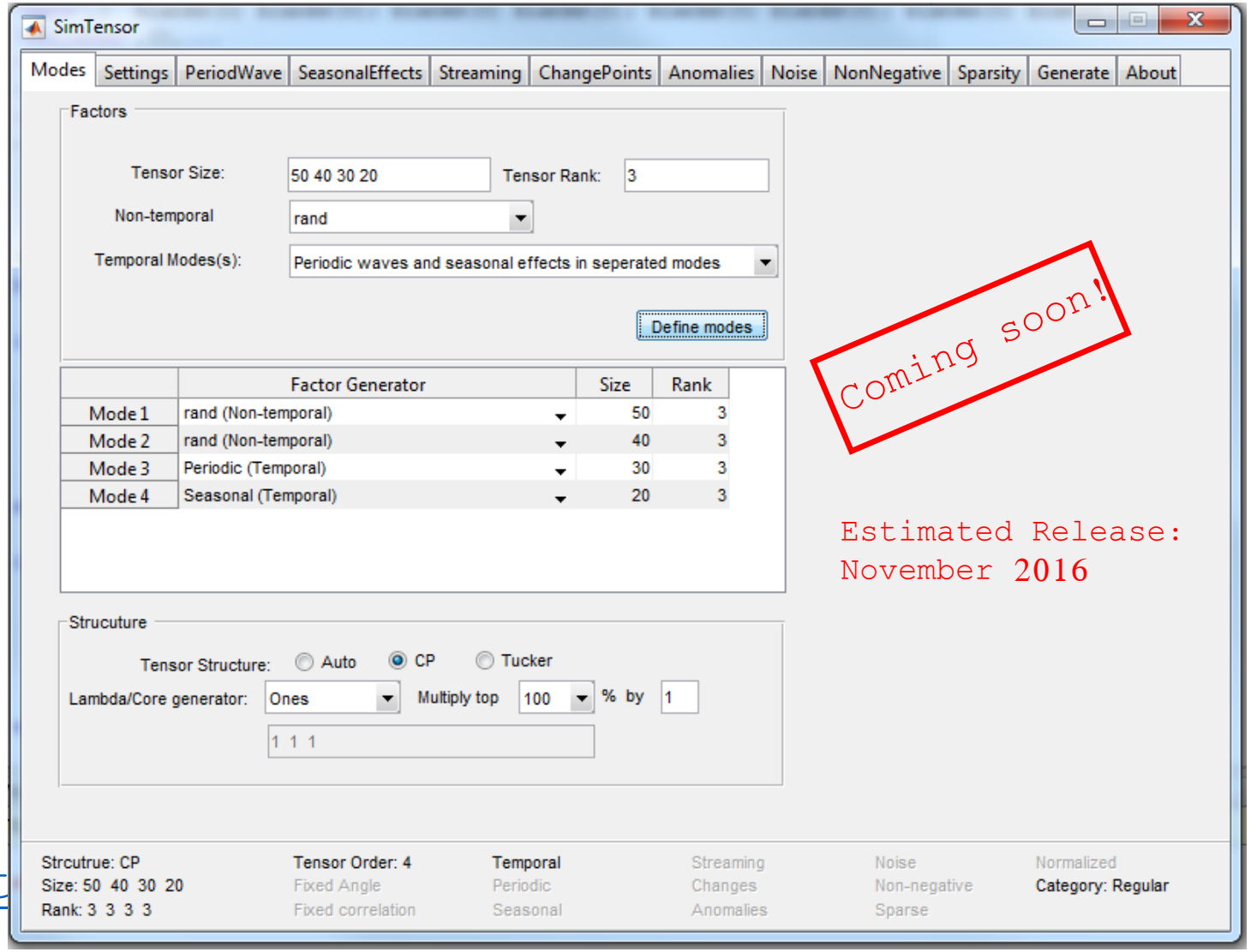
Recent Trends in TD literature

- Bayesian approaches (e.g. BPTF)
- Streaming/Incremental approaches (e.g. DTA, OnlineCP)
- Parallel approaches (e.g. ParCube)
- GPU-based Tensor Decomposition (e.g. GPU-PARAFAC)
- Boolean Tensor Factorization
- TD for unequal-length tensors (e.g. PARAFAC2, GTucker2)
- Algorithms for handling shifts and shape-change (e.g. ShiftCP, ConvCP)
- Efficient methods for tensor completion (e.g. HaLRTC).
- Tensor Regression (e.g. HOPLS)
- Tensor-based data fusion (e.g. CMTF)
- Tensor Projection methods (e.g. TGLPP)

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

LIAAD INESCTEC U.PORTO

Ads by MAESTRA



Questions ?

Hadi Fanaee-T

INESC TEC, Portugal

Email: hadi.fanaee@fe.up.pt