Discovery of Latent Factors in High-dimensional Data via Spectral Methods

Furong Huang

University of Maryland

Presented by Xuchen You

2018 SIAM Annual Meeting (AN18)

Machine Learning - Excitements

Success of Supervised Learning

Image classification

Speech recognition

Text processing

Key to Success

- Deep composition of nonlinear units
- Enormous labeled data
- Computation power growth

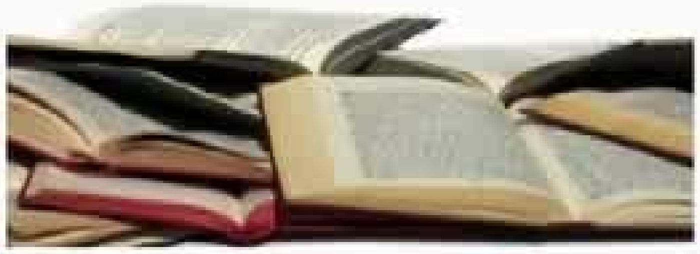
Machine Learning - Modern Challenges

Automated discovery of features and categories?

Real AI requires **Unsupervised Learning**

Filter bank learning

Feature extraction



Embeddings, Topics

- Summarize key features in data
  - State-of-the-art: Humans are better than machines
  - Goal: Intelligent machines that summarize key features in data
- Interpretable modeling and learning of the data
  - Theoretically guaranteed learning
  - Extracted features are interpretable

Unsupervised Learning with Big Data

Information Extraction

- High dimension **observation** vs Low dimension **representation**

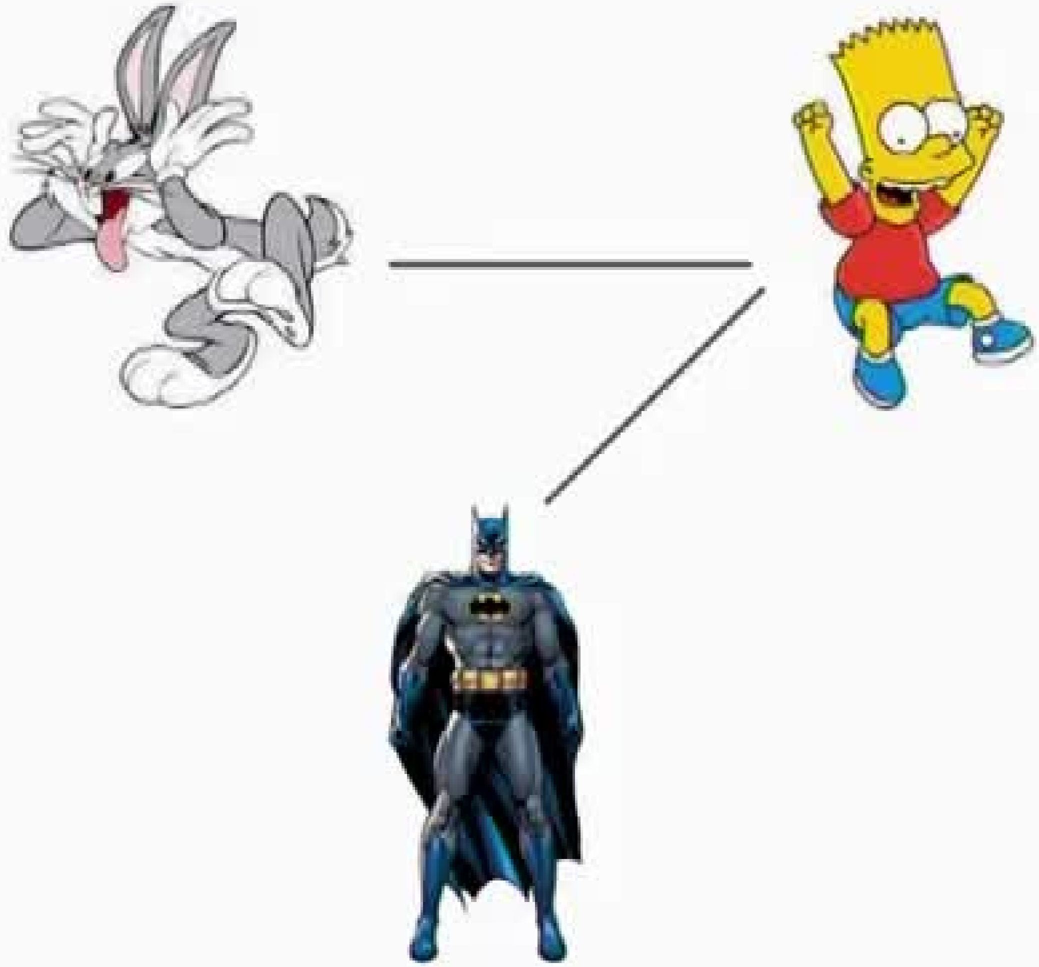
My Solution: A Unified Tensor Decomposition Framework

App 1: Automated Categorization of Documents

Document modeling

- Observed: words in document corpus: search logs, emails etc
- Hidden: (mixed) topics: personal interests, professional area etc

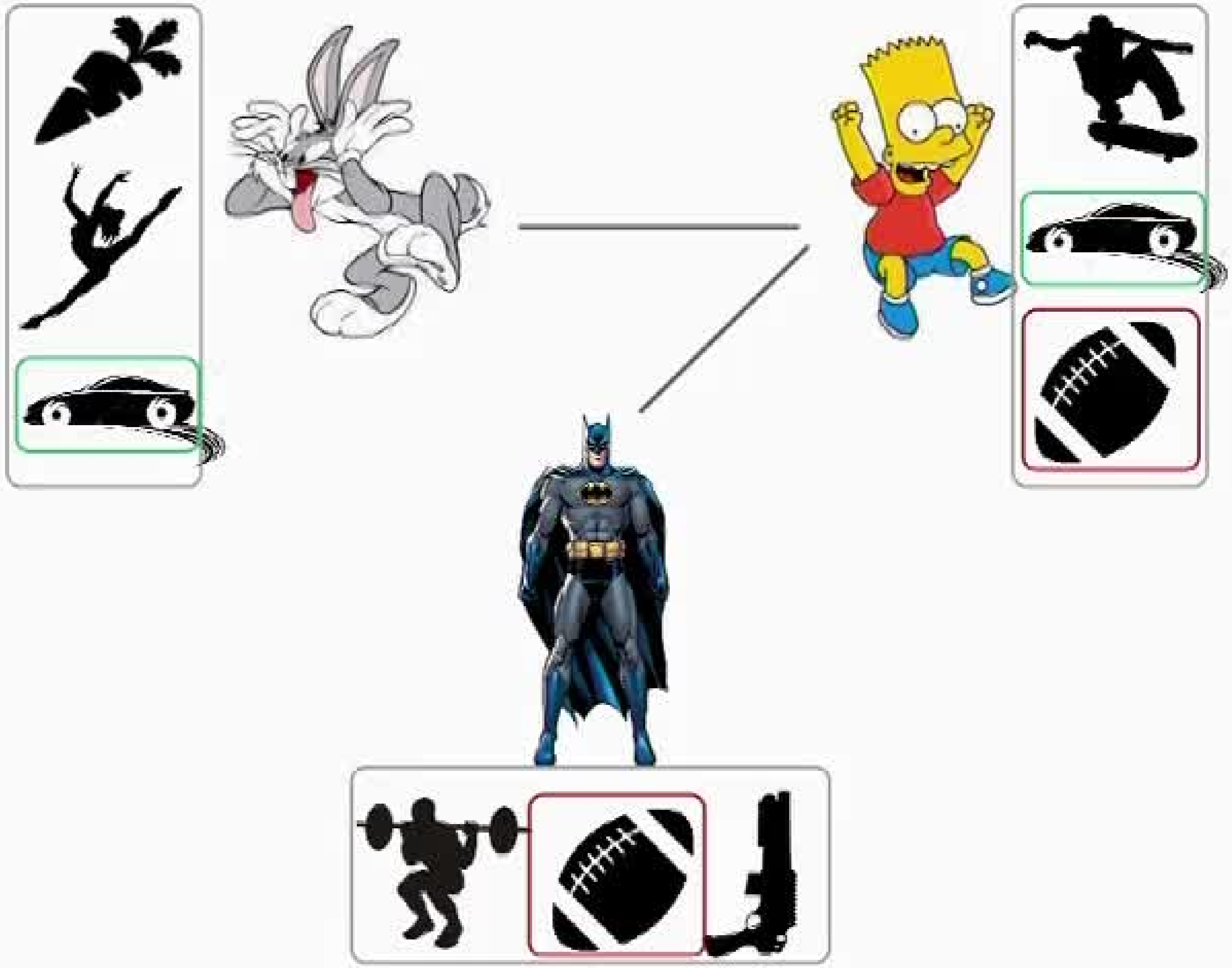
App 2: Community Extraction From Connectivity Graph



Social Networks

- Observed: network of social ties: friendships, transactions etc
- Hidden: (mixed) groups/communities of social actors

App 2: Community Extraction From Connectivity Graph

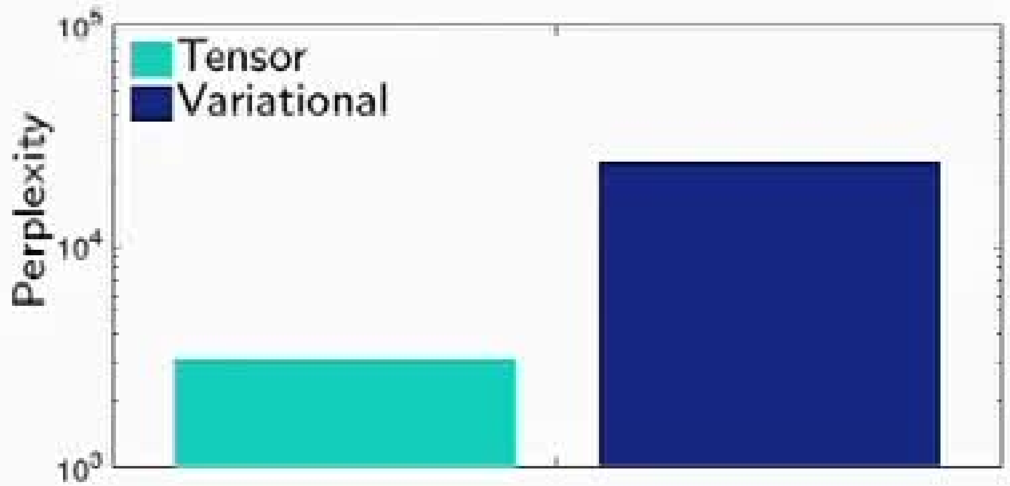


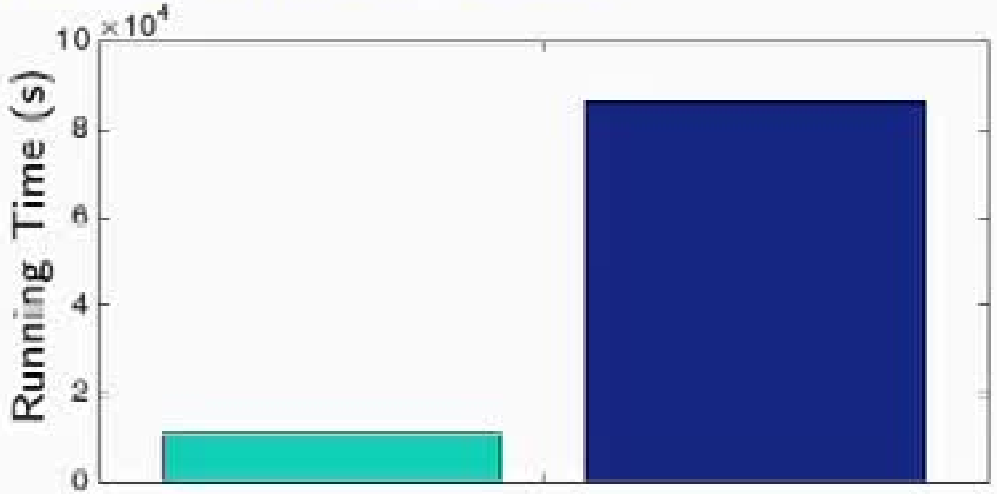
Social Networks

- Observed: network of social ties: friendships, transactions etc
- Hidden: (mixed) groups/communities of social actors

Tensor Methods Compared with Variational Inference

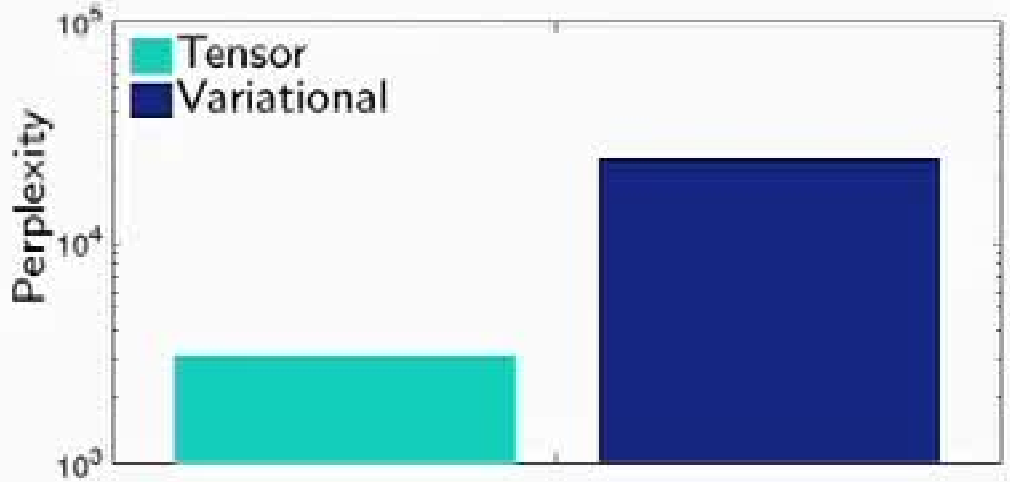
Learning Topics from PubMed on Spark: 8 million docs

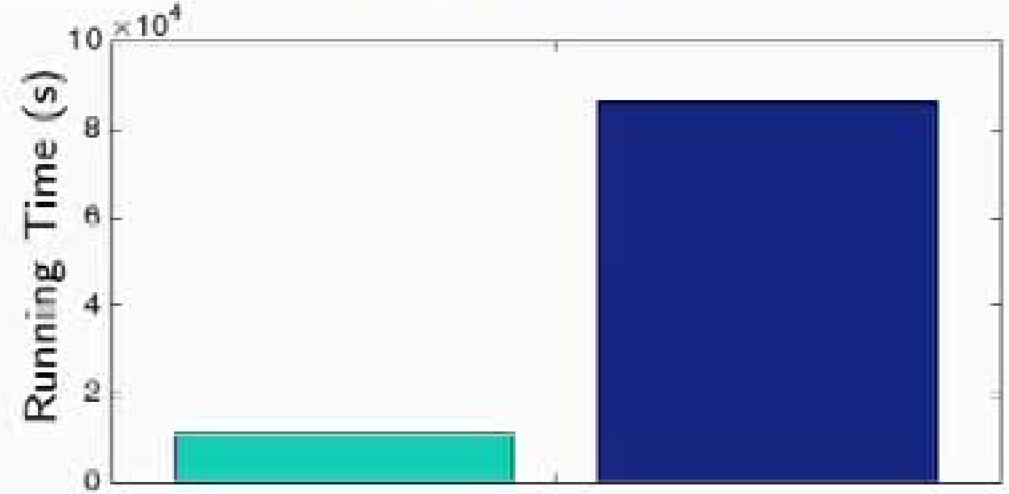




Tensor Methods Compared with Variational Inference

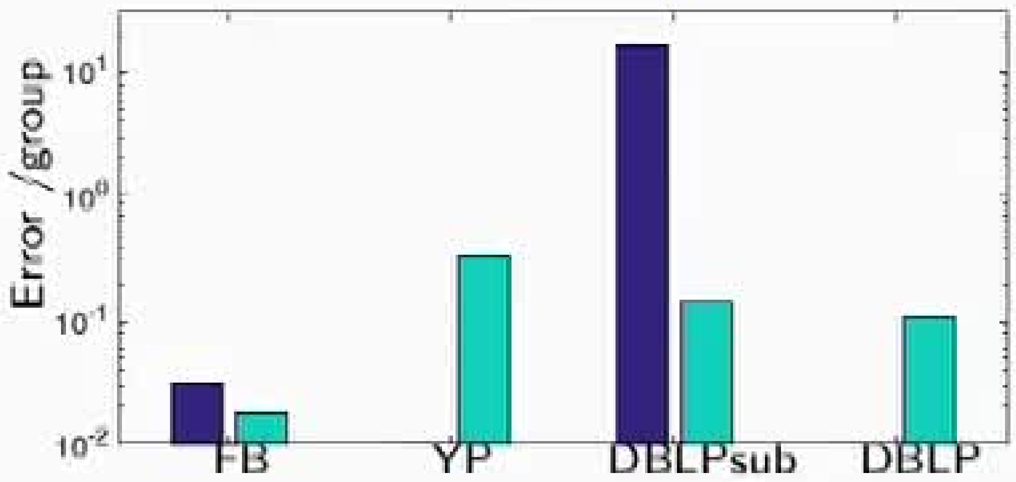
Learning Topics from PubMed on Spark: 8 million docs

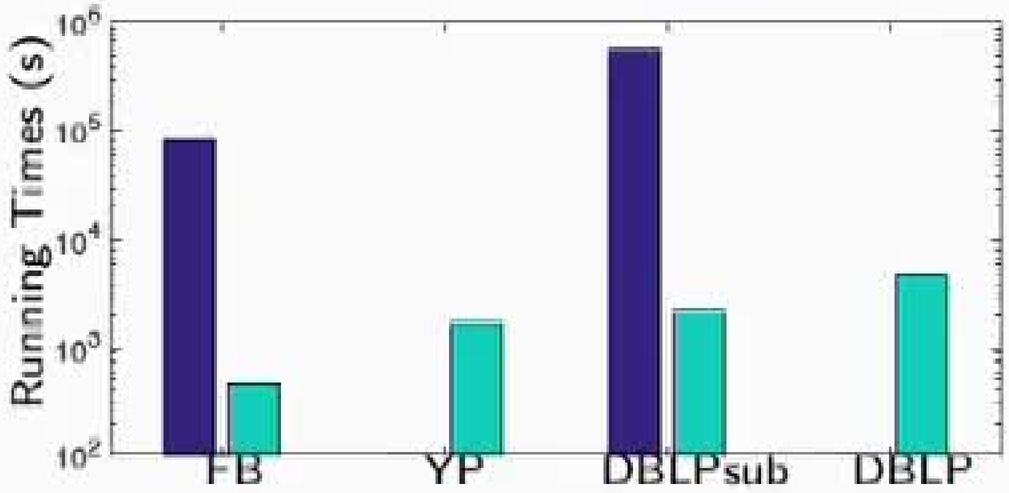




Learning Communities from Graph Connectivity

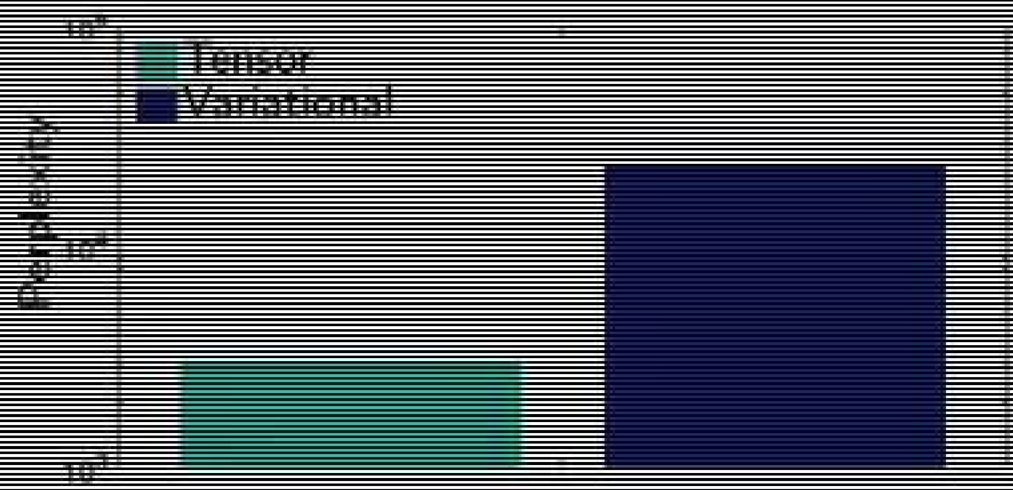
Facebook: $n \sim 20k$ Yelp: $n \sim 40k$ DBLPsub: $n \sim 0.1m$ DBLP: $n \sim 1m$

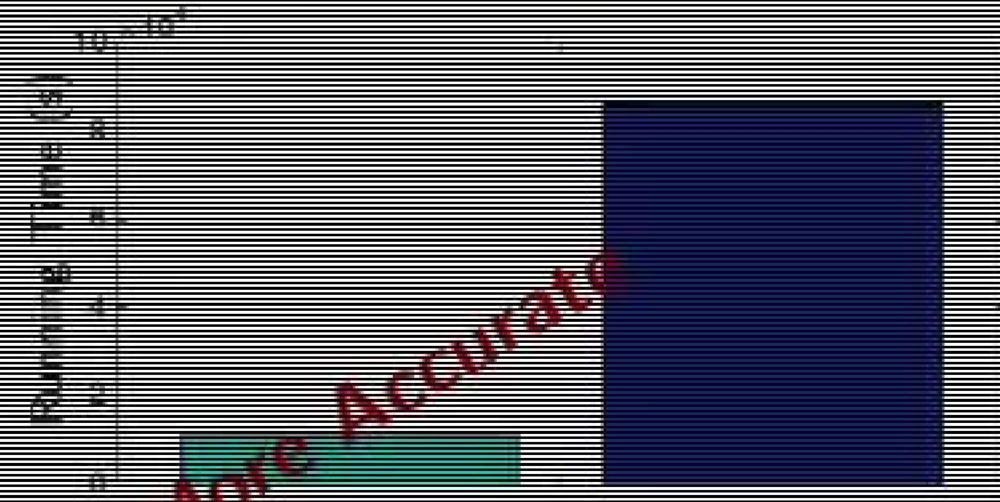




Tensor Methods Compared with Variational Inference

Learning Topics from PubMed on Spark: 8 million docs



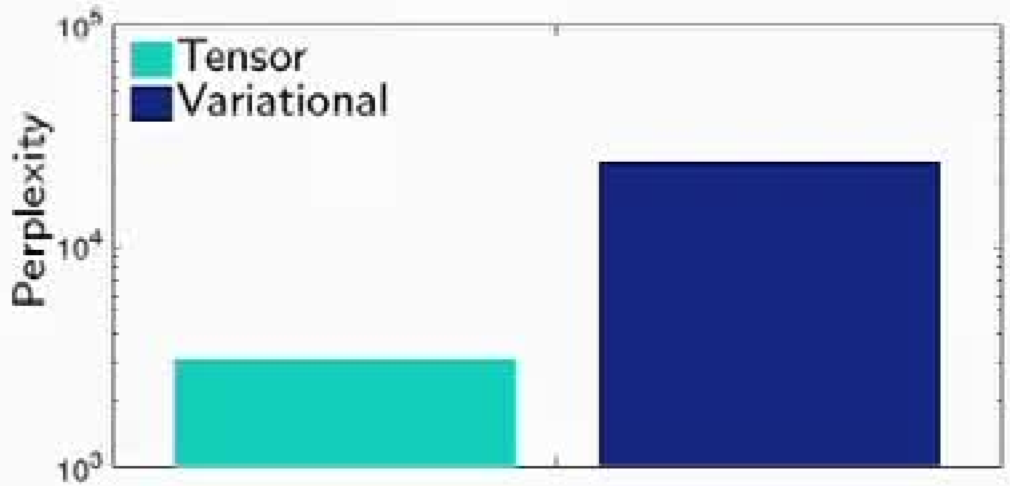


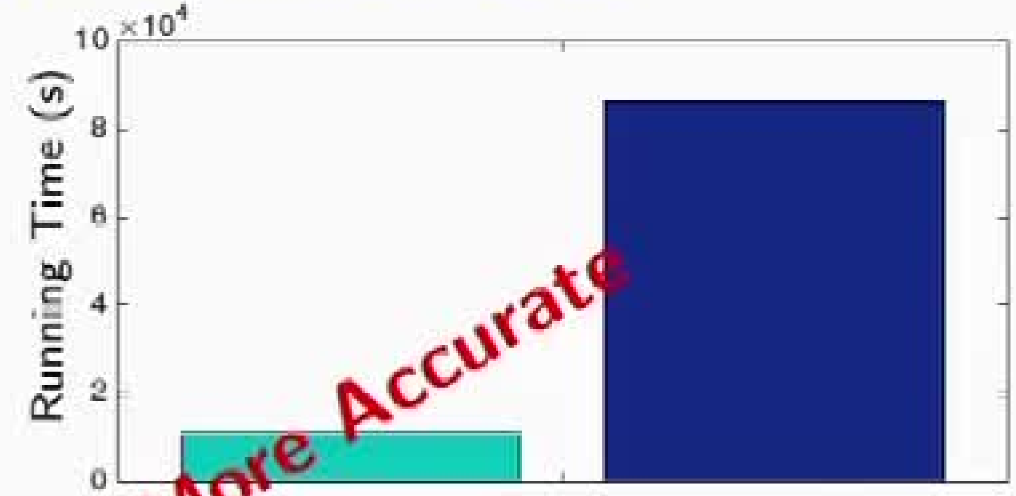
Learning Communities from Graph Connectivity

ub: $n \sim 0.1m$ DBLP: $n \sim 1m$

Tensor Methods Compared with Variational Inference

Learning Topics from PubMed on Spark: 8 million docs

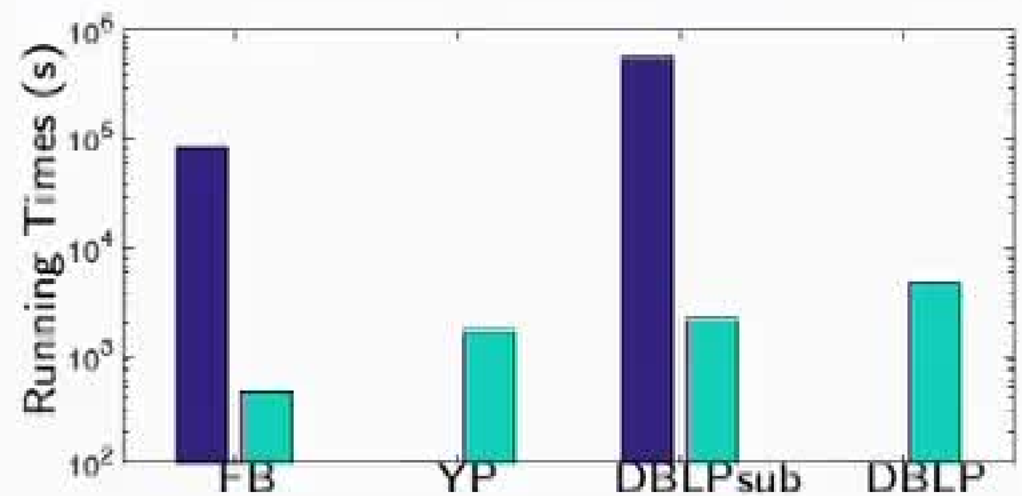




Orders of Magnitude Faster & More Accurate

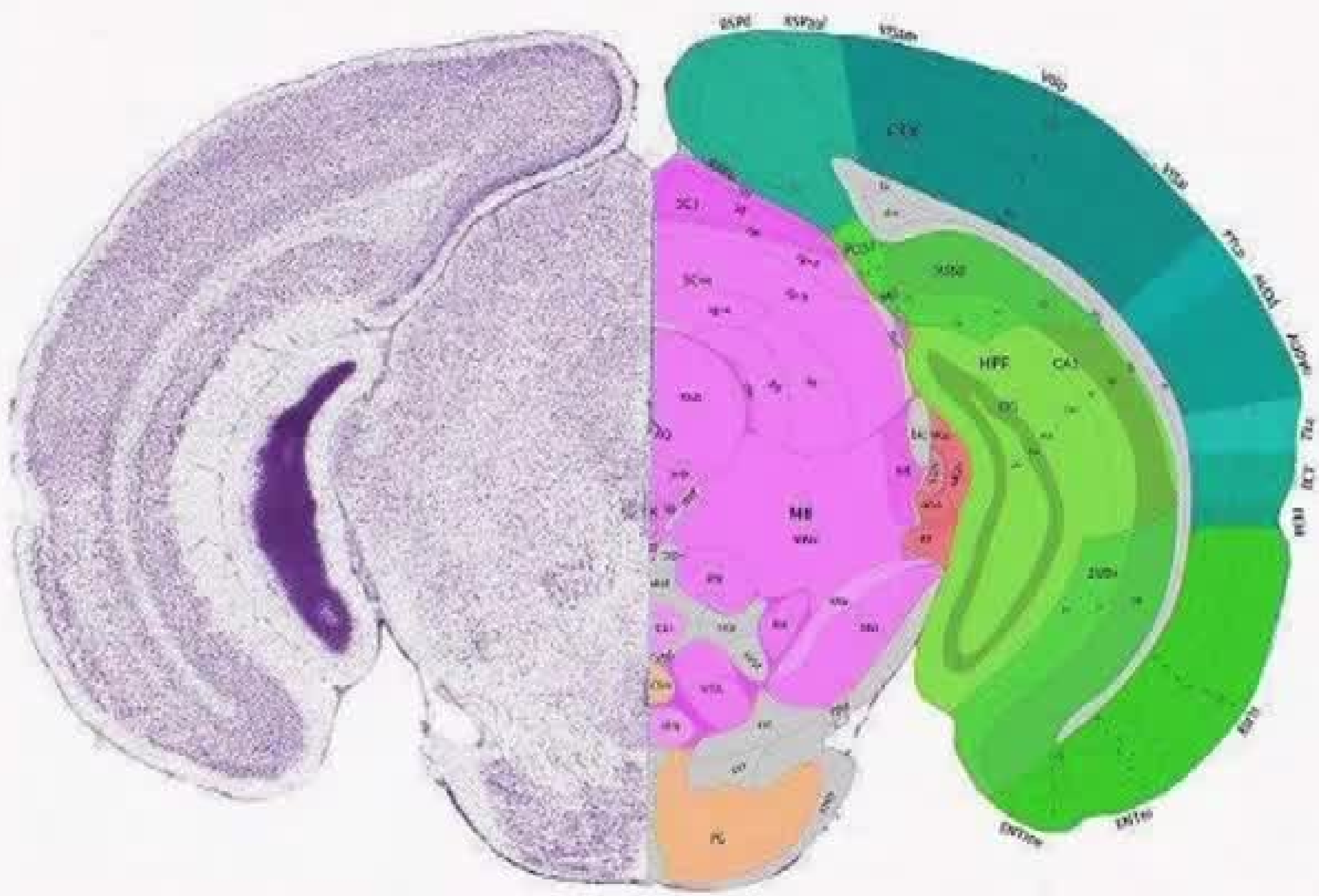
Learning Communities from Graph Connectivity

Facebook: $n \sim 20k$ Yelp: $n \sim 40k$ DBLPsub: $n \sim 0.1m$ DBLP: $n \sim 1m$



"Online Tensor Methods for Learning Latent Variable Models", F. Huang, U. Niranjan, M. Hakeem, A. Anandkumar, JMLR14.
"Tensor Methods on Apache Spark", F. Huang, A. Anandkumar, Oct. 2015.

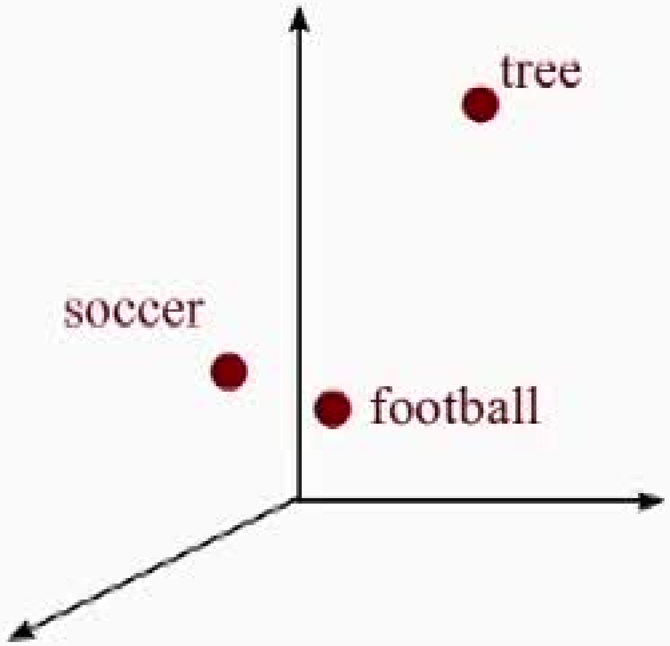
App 3: Cataloging Neuronal Cell Types In the Brain



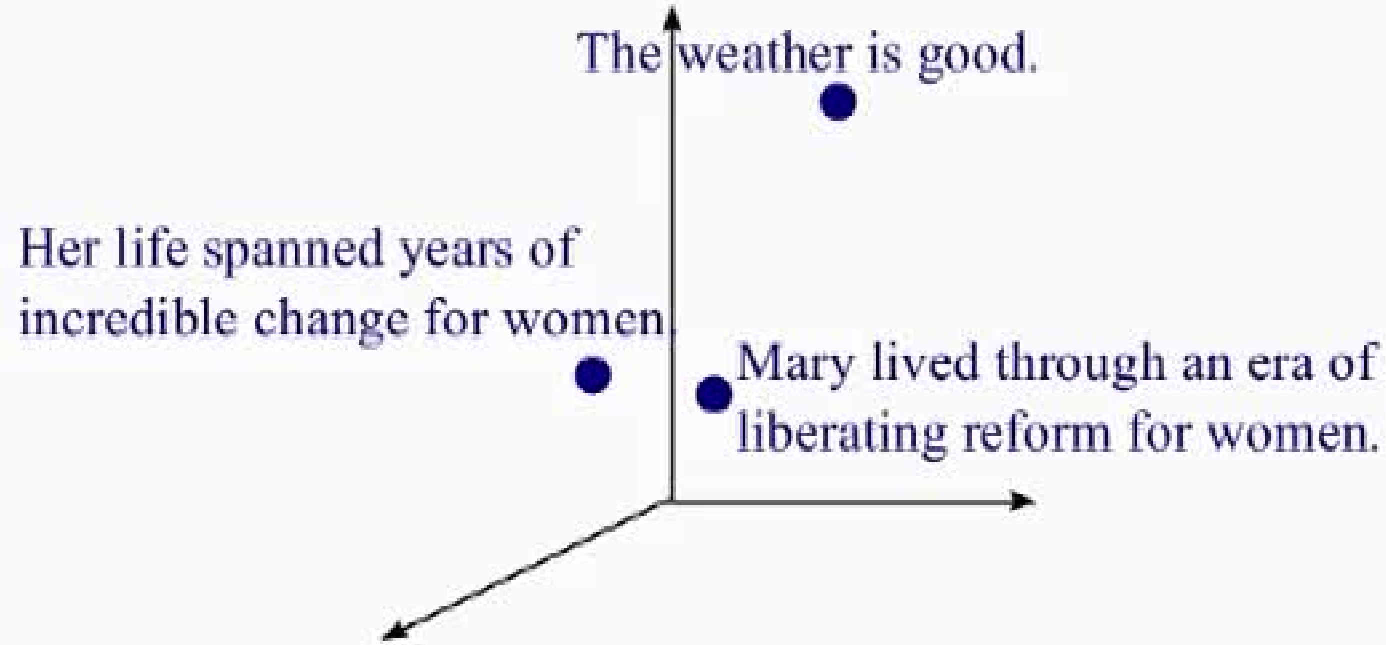
Neuroscience

- Observed: cellular-resolution brain slices
- Hidden: neuronal cell types

App 4: Word Sequence Embedding Extraction



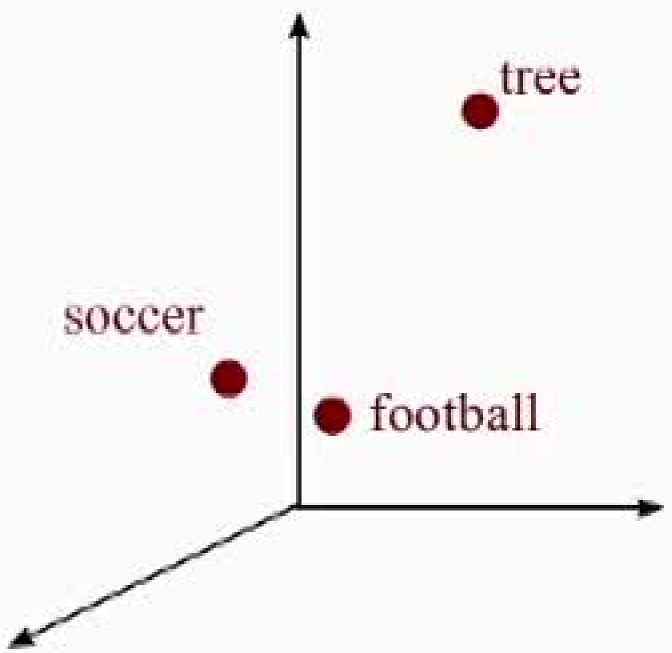
Word Embedding



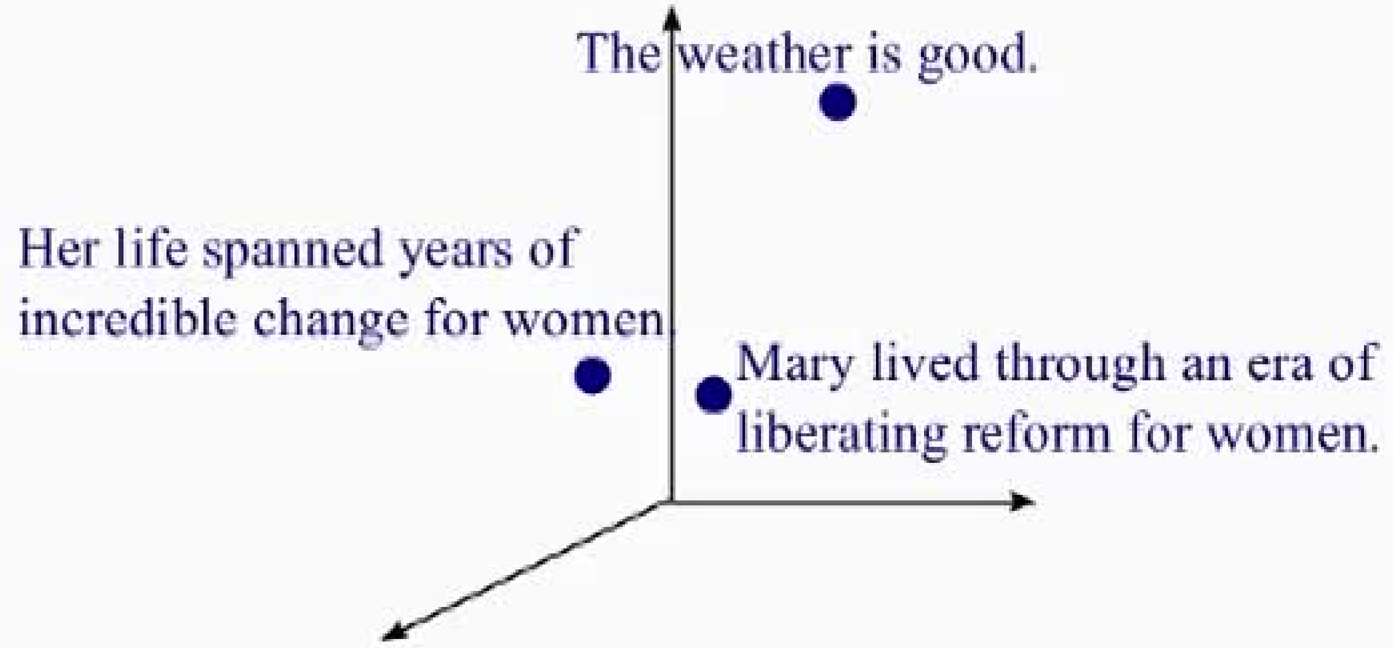
Word Sequence Embedding

"Convolutional Dictionary Learning through Tensor Factorization", by F. Huang, A. Anandkumar, In Proceedings of JMLR 2015.

App 4: Word Sequence Embedding Extraction



Word Embedding

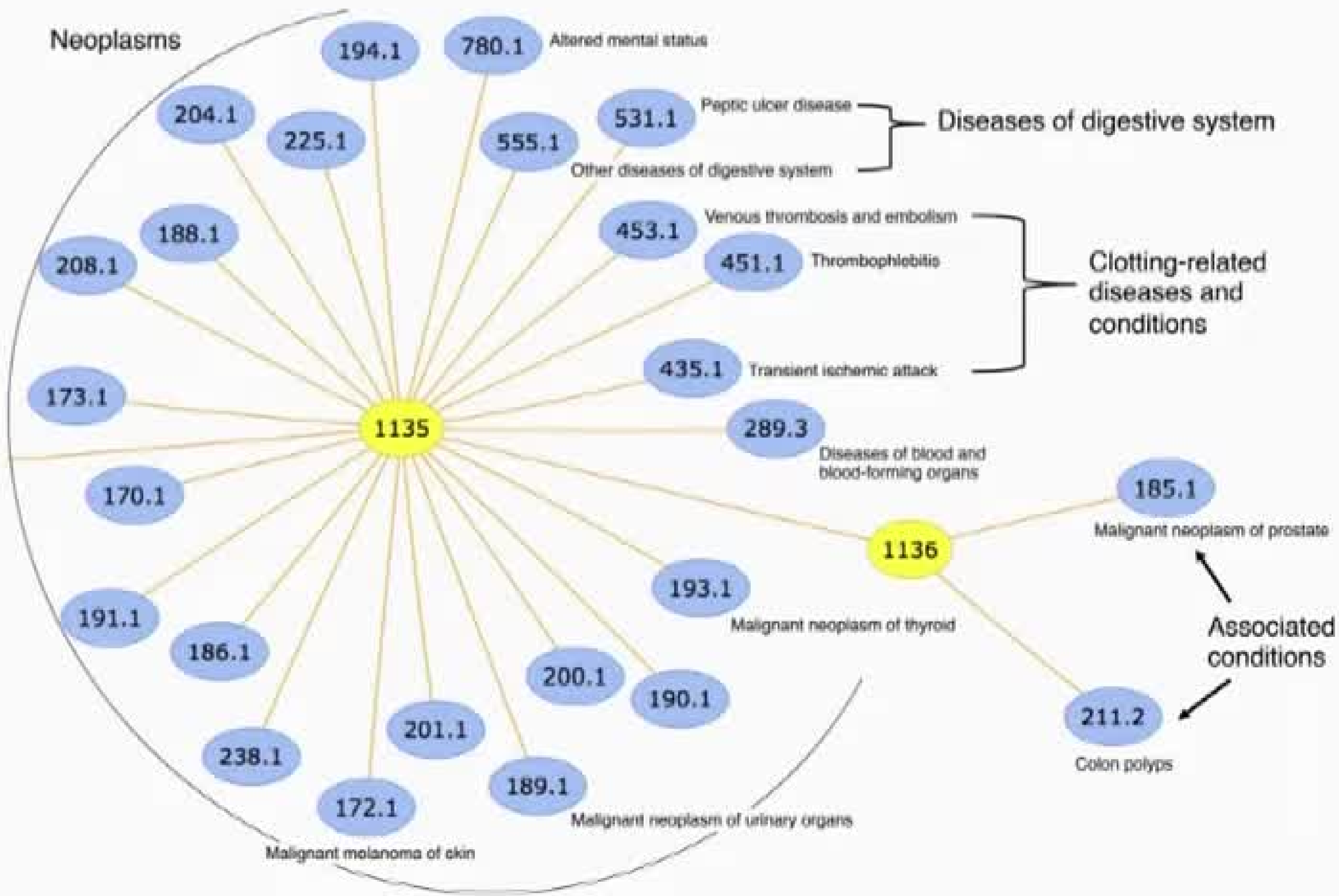


Word Sequence Embedding

"Convolutional Dictionary Learning through Tensor Factorization", by F. Huang, A. Anandkumar, In Proceedings of JMLR 2015.

App 5: Human Disease Hierarchy Discovery

CMS: 1.6 million patients, 168 million diagnostic events, 11 k diseases.



- **Observed:** co-occurrence of diseases on patients
- **Hidden:** disease similarity/hierarchy

" Scalable Latent TreeModel and its Application to Health Analytics " by F. Huang, N. U.Niranjan, I. Perros, R. Chen, J. Sun, A. Anandkumar, NIPS 2015 MLHC workshop.

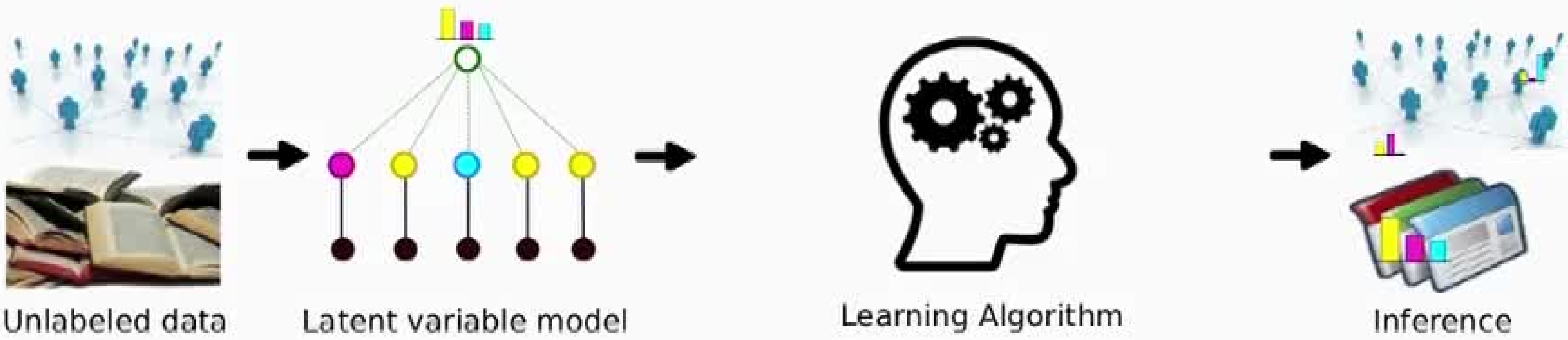
Involve discovering the hidden and compact structure

that is embedded in the high-dimensional complex observed data

Challenges in Learning

Basic goal in all mentioned applications

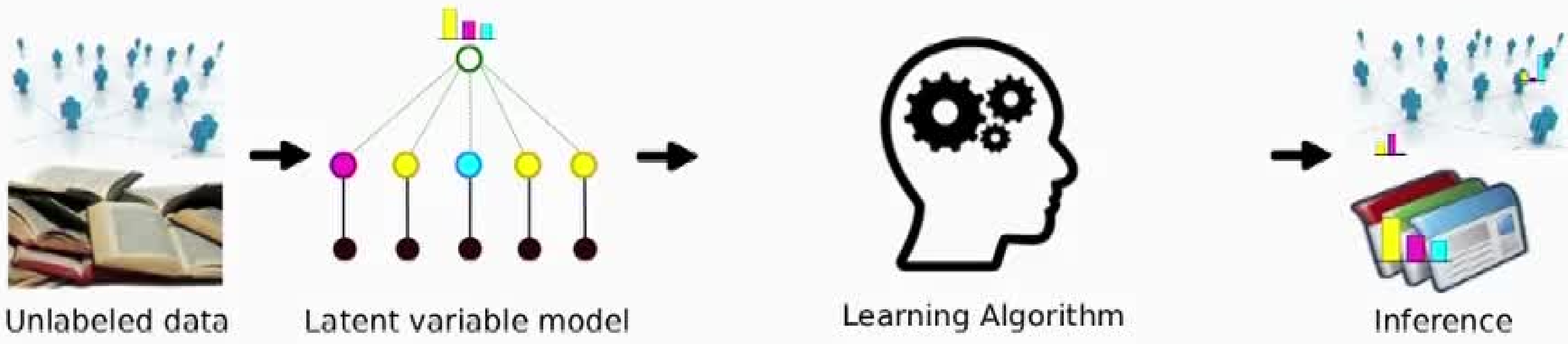
Discover hidden structure in data: unsupervised learning.



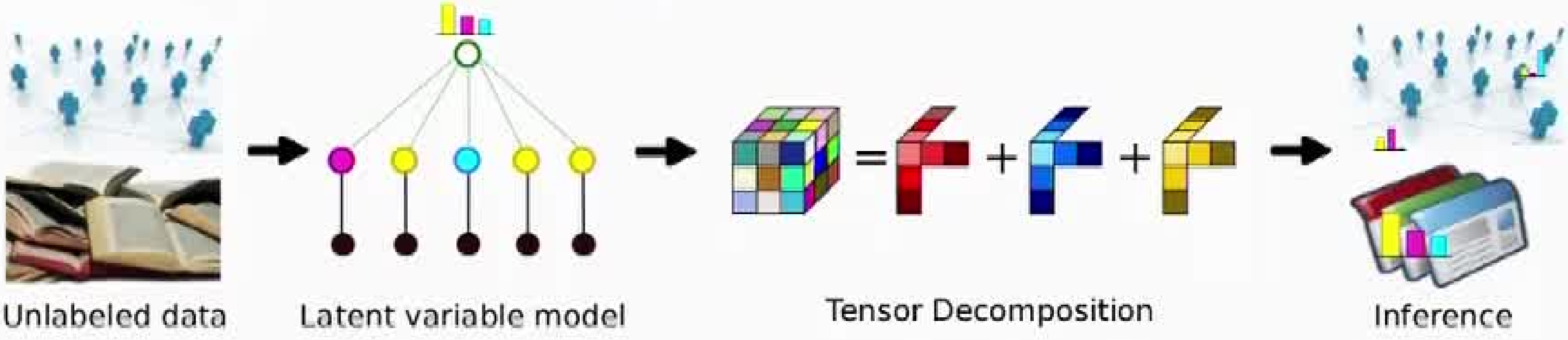
Challenges in Learning

Basic goal in all mentioned applications

Discover hidden structure in data: unsupervised learning.



Challenges in Learning – find hidden structure in data



Challenge: Conditions for Identifiability

- Whether can model be identified given infinite computation and data?
- Are there tractable algorithms under identifiability?

Challenge: Efficient Learning of Latent Variable Models

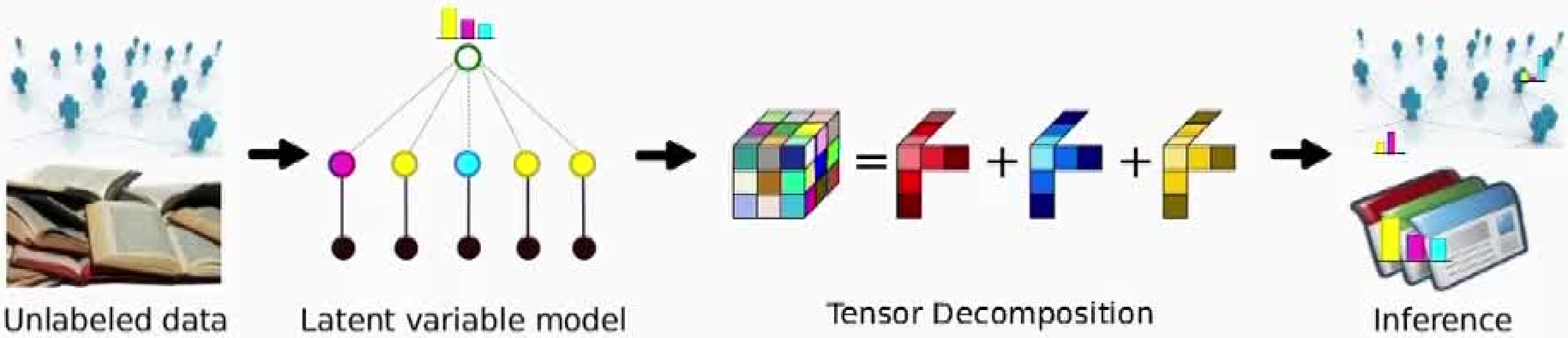
- MCMC: random sampling, slow

  Exponential mixing time
- Likelihood: non-convex, not scalable

  Exponential critical points
- Efficient computational and sample complexities?

Guaranteed and efficient learning through spectral methods

Unsupervised Learning via Probabilistic Models



tensor decomposition → correct model

Contributions

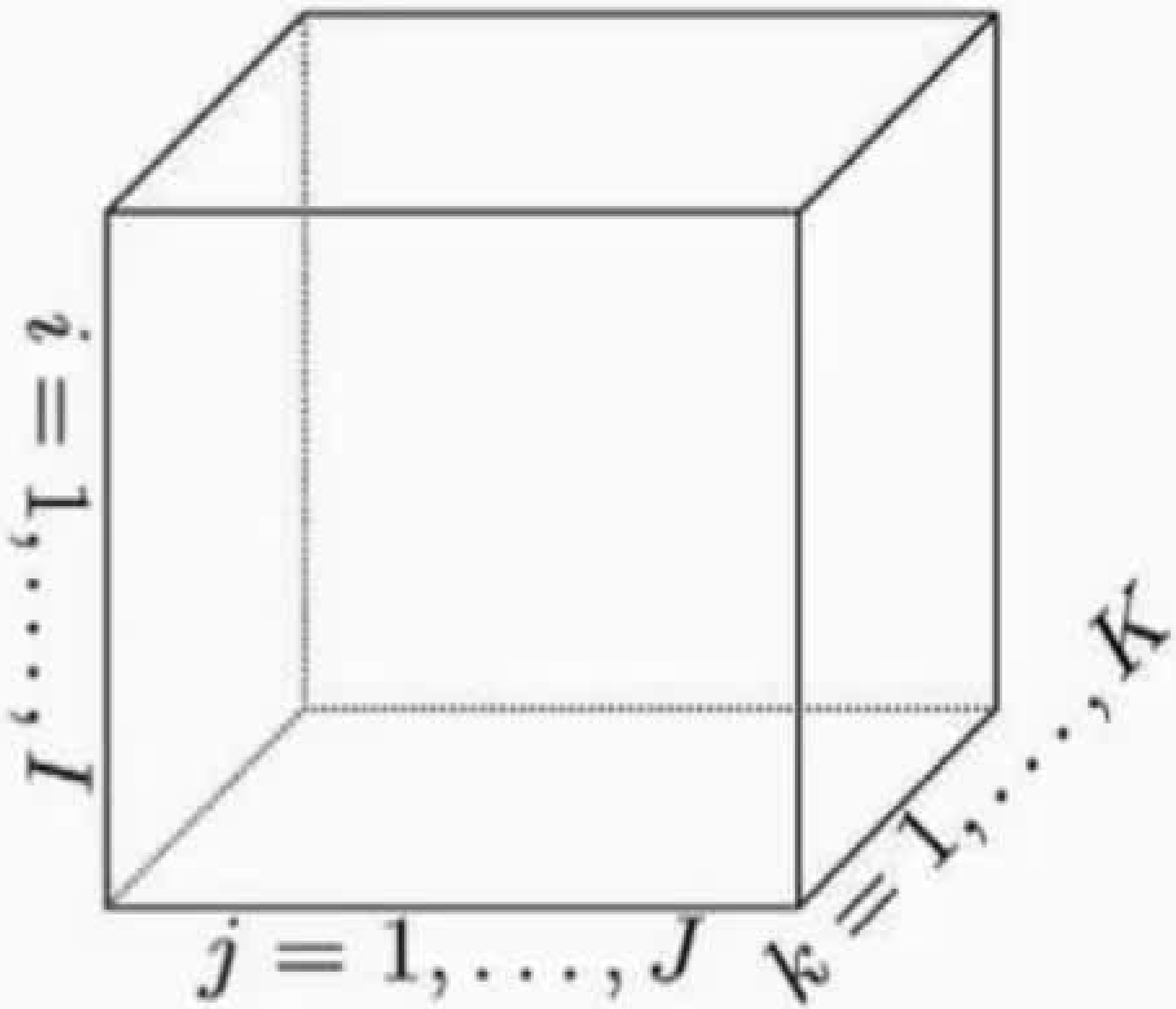
- Guaranteed online algorithm with global convergence guarantee
- Highly scalable, highly parallel, dimensionality reduction
- Tensor library on CPU/GPU/Spark
- Interdisciplinary applications
- Extension to model with group invariance

Outline

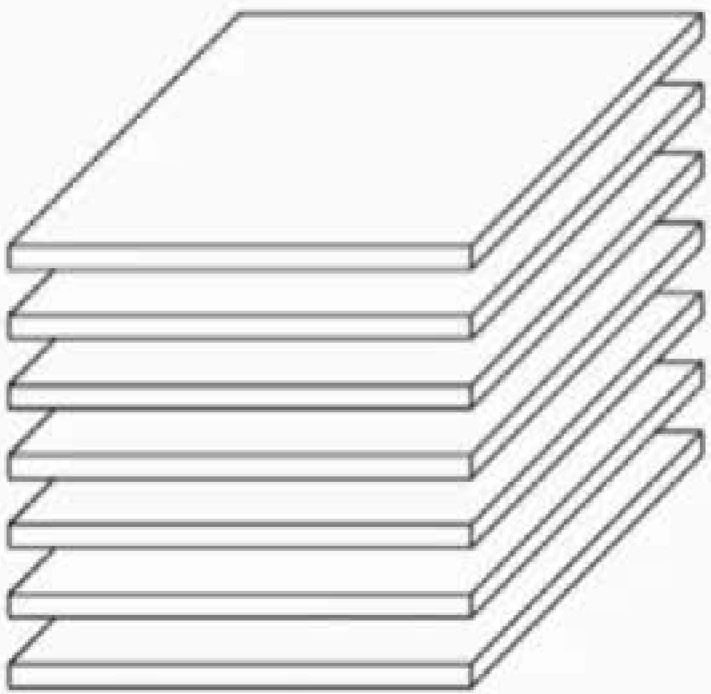
What is a tensor?

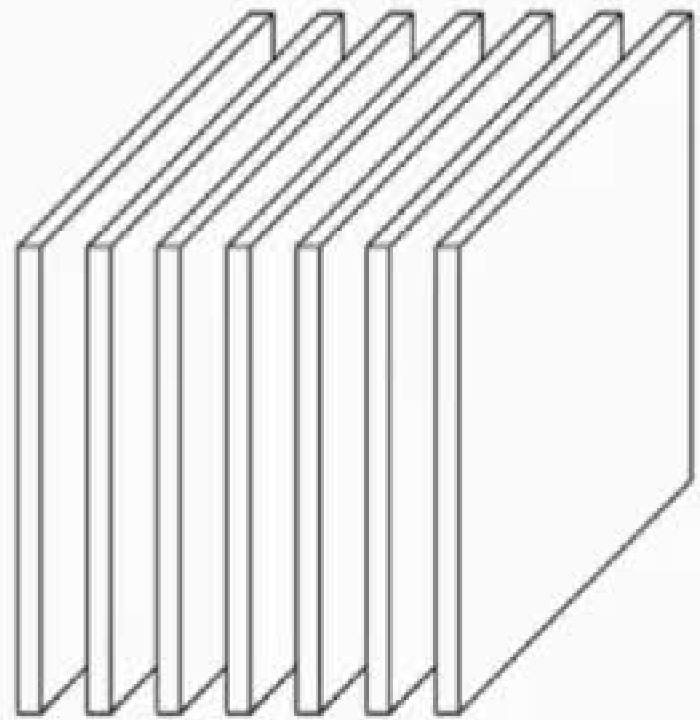
Multi-dimensional Array

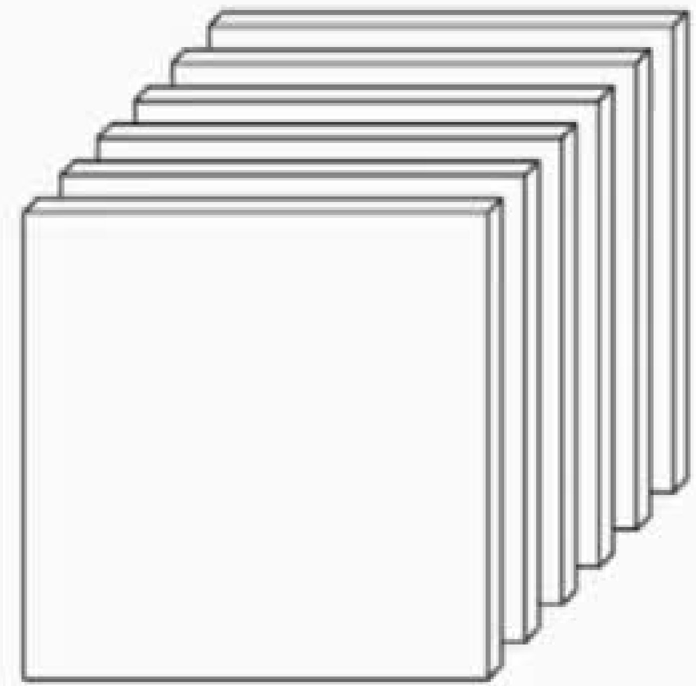
- Tensor - Higher order matrix
- The number of dimensions is called tensor order.



Slices

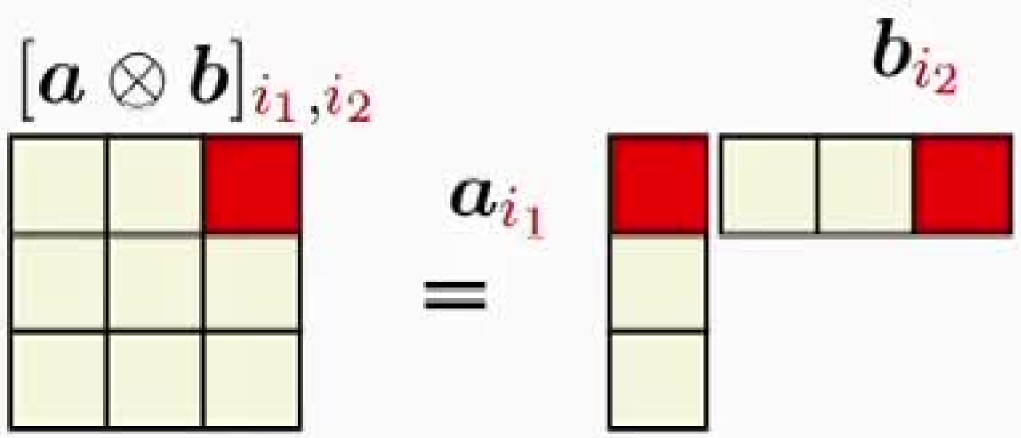


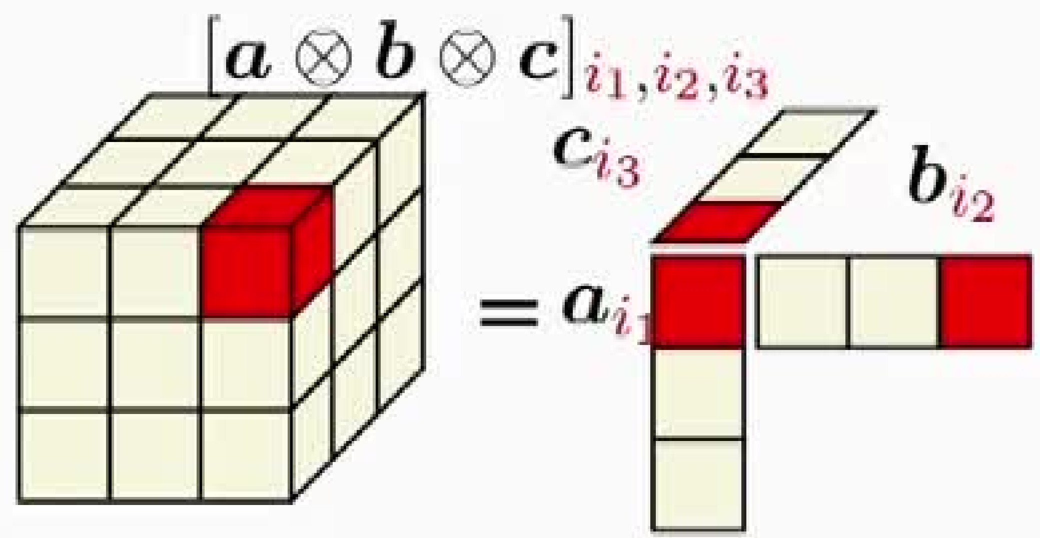




- Horizontal slices
- Lateral slices
- Frontal slices

Tensor Product

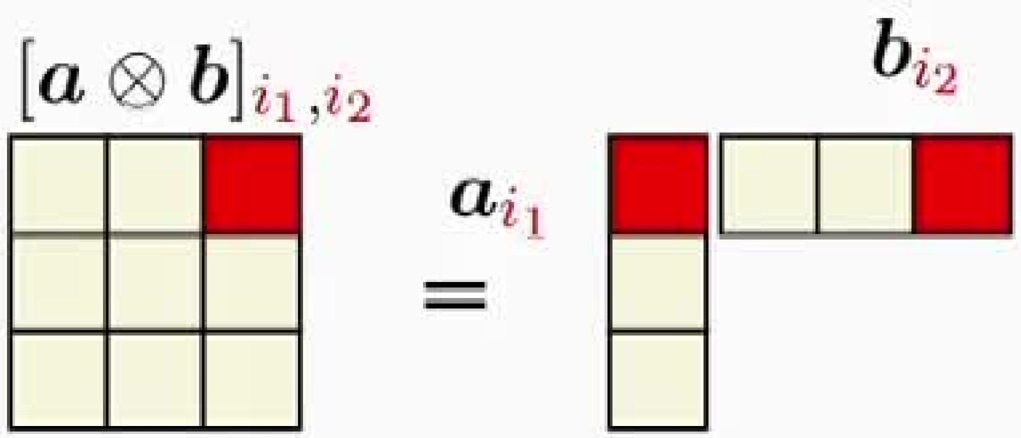


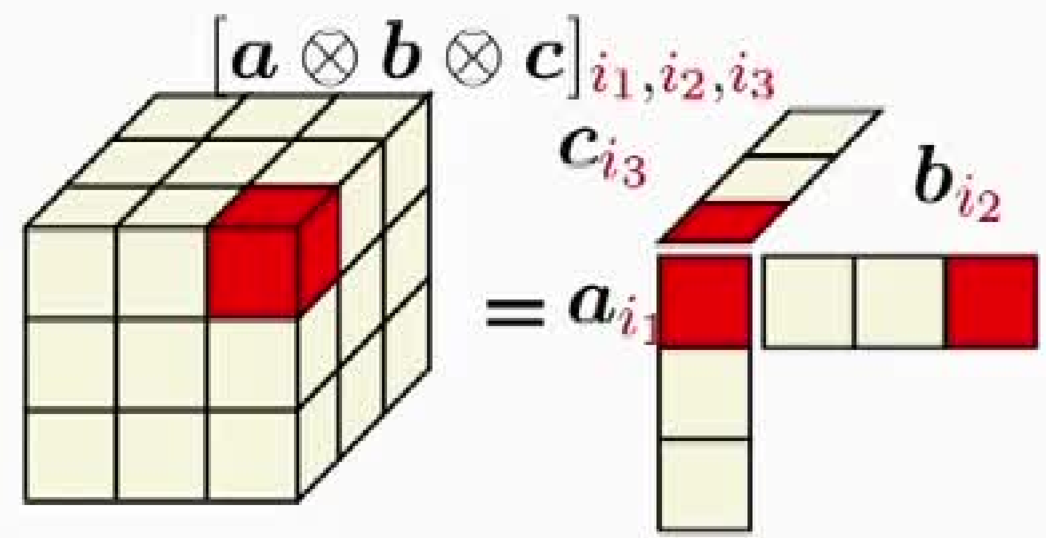


- $[\boldsymbol{a} \otimes \boldsymbol{b}]_{i_1,i_2} = \boldsymbol{a}_{i_1} \boldsymbol{b}_{i_2}$
- Rank-1 matrix

- $[\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}]_{i_1,i_2,i_3} = \boldsymbol{a}_{i_1} \boldsymbol{b}_{i_2} \boldsymbol{c}_{i_3}$
- Rank-1 tensor

Tensor Product





- $[\boldsymbol{a} \otimes \boldsymbol{b}]_{i_1,i_2} = \boldsymbol{a}_{i_1} \boldsymbol{b}_{i_2}$
- Rank-1 matrix

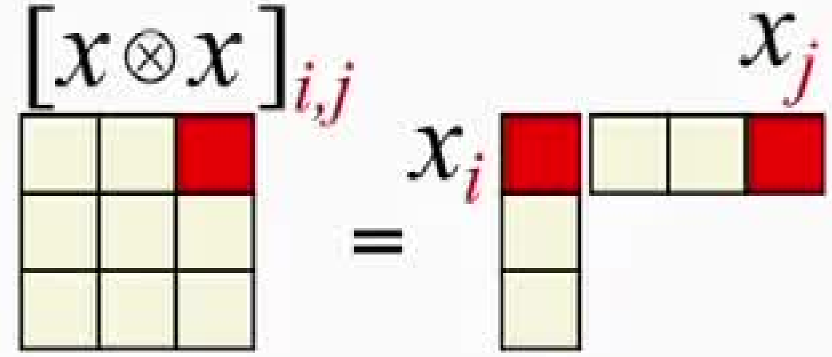
- $[\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}]_{i_1,i_2,i_3} = \boldsymbol{a}_{i_1} \boldsymbol{b}_{i_2} \boldsymbol{c}_{i_3}$
- Rank-1 tensor

Tensors in Method of Moments

Matrix: Pair-wise relationship

- Signal or data observed $\boldsymbol{x} \in \mathbb{R}^d$
- Rank 1 matrix: $[\boldsymbol{x} \otimes \boldsymbol{x}]_{i,j} = \boldsymbol{x}_i \boldsymbol{x}_j$
- Aggregated pair-wise relationship

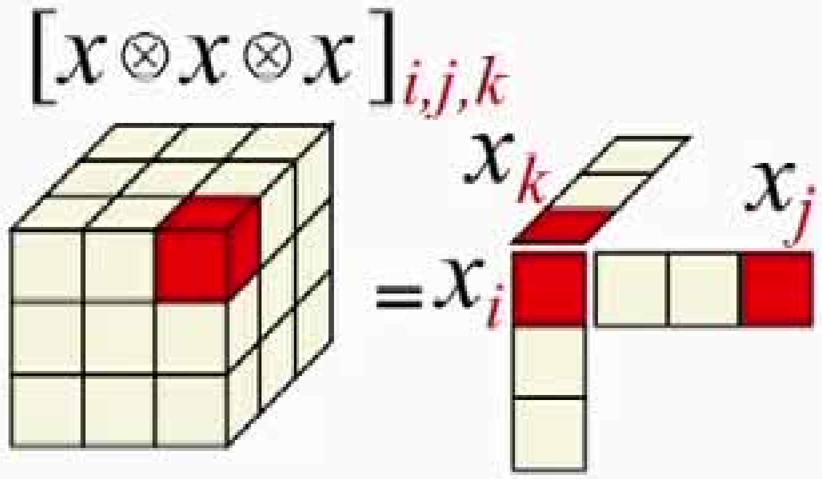
$$\boldsymbol{M}_2 = \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}]$$



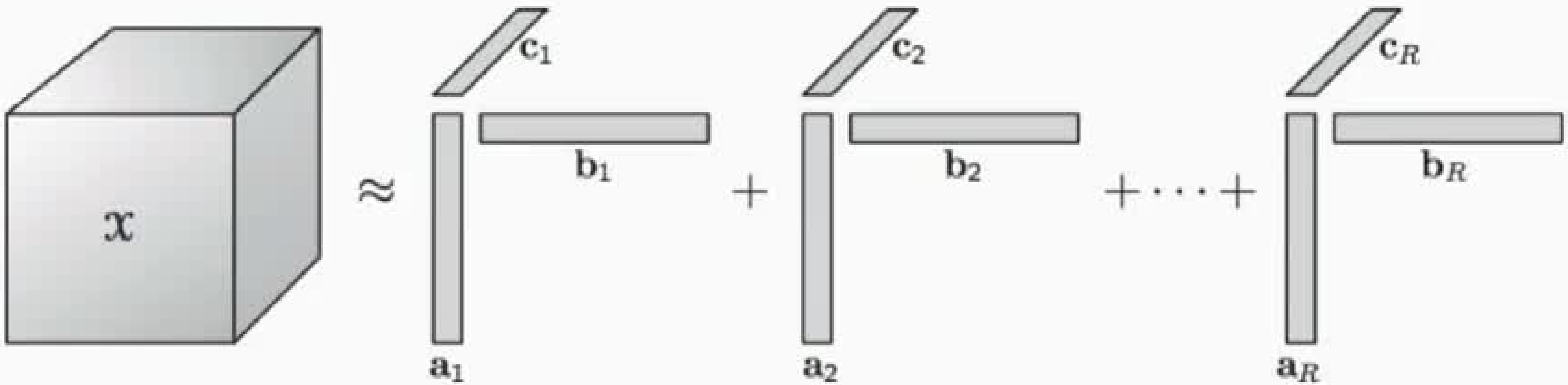
Tensor: Triple-wise relationship or higher

- Signal or data observed $\boldsymbol{x} \in \mathbb{R}^d$
- Rank 1 tensor:
  $[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}]_{i,j,k} = \boldsymbol{x}_i \boldsymbol{x}_j \boldsymbol{x}_k$
- Aggregated triple-wise relationship

$$\mathcal{M}_3 = \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] = \mathbb{E}[\boldsymbol{x} \otimes^3]$$



CP decomposition

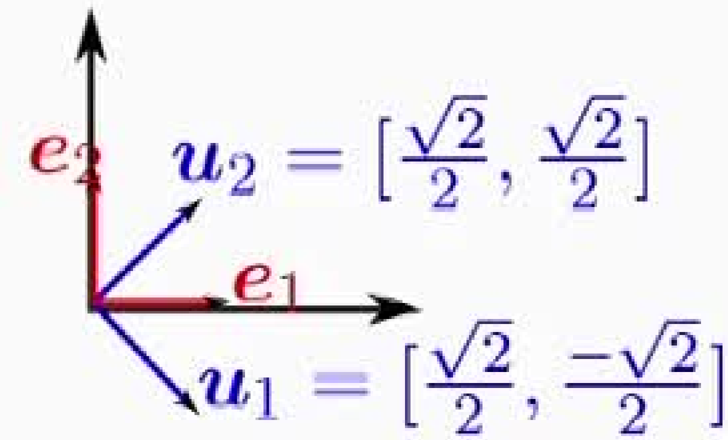


- $\boldsymbol{\mathcal{X}} = \sum_{h=1}^{R} \boldsymbol{a}_h \otimes \boldsymbol{b}_h \otimes \boldsymbol{c}_h$
- Summation of rank-1 tensors

Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap
  $$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$$
- Unique with eigenvalue gap



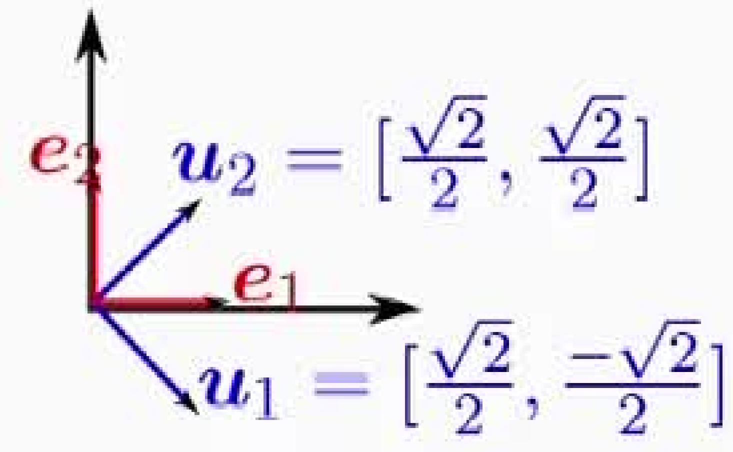
Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap

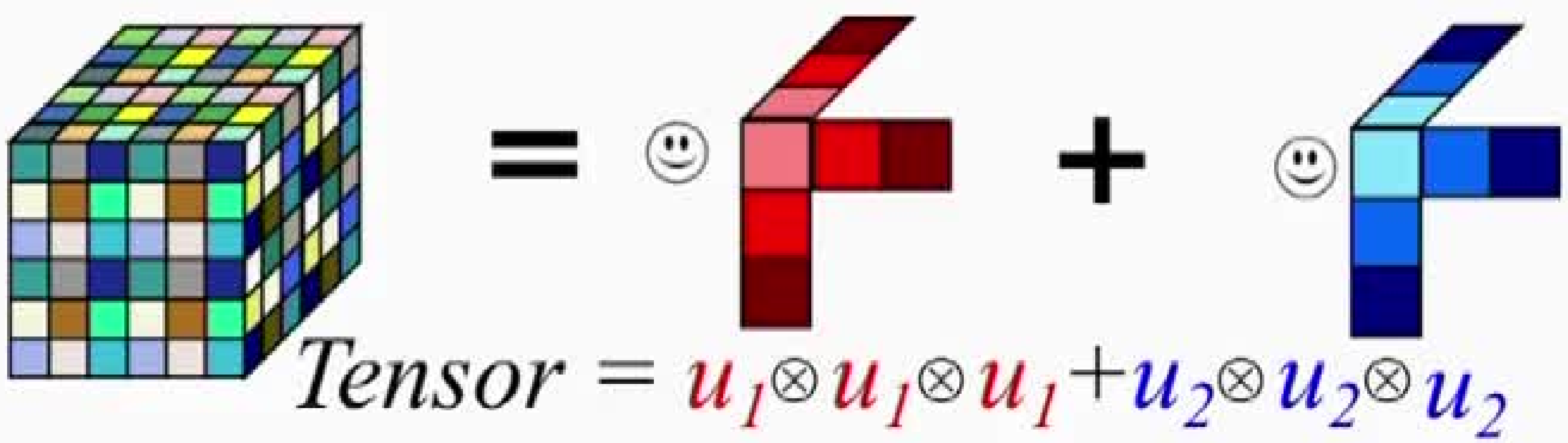
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = e_1 e_1^\top + e_2 e_2^\top = u_1 u_1^\top + u_2 u_2^\top$$

- Unique with eigenvalue gap



Tensor Orthogonal Decomposition (Harshman, 1970)

- Unique: eigenvalue gap not needed



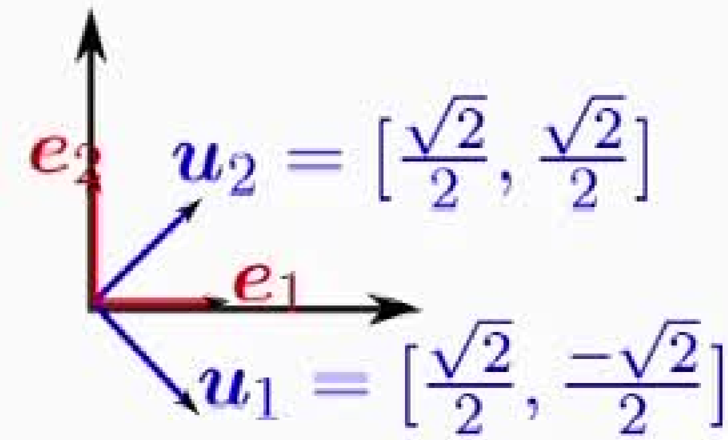
Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$$

- Unique with eigenvalue gap



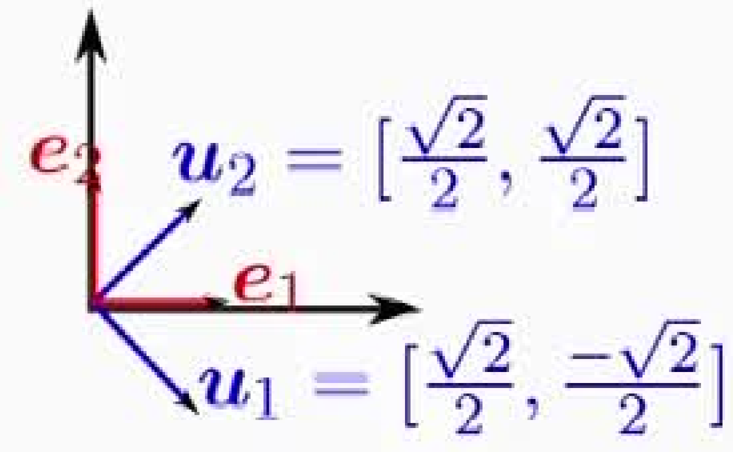
Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap

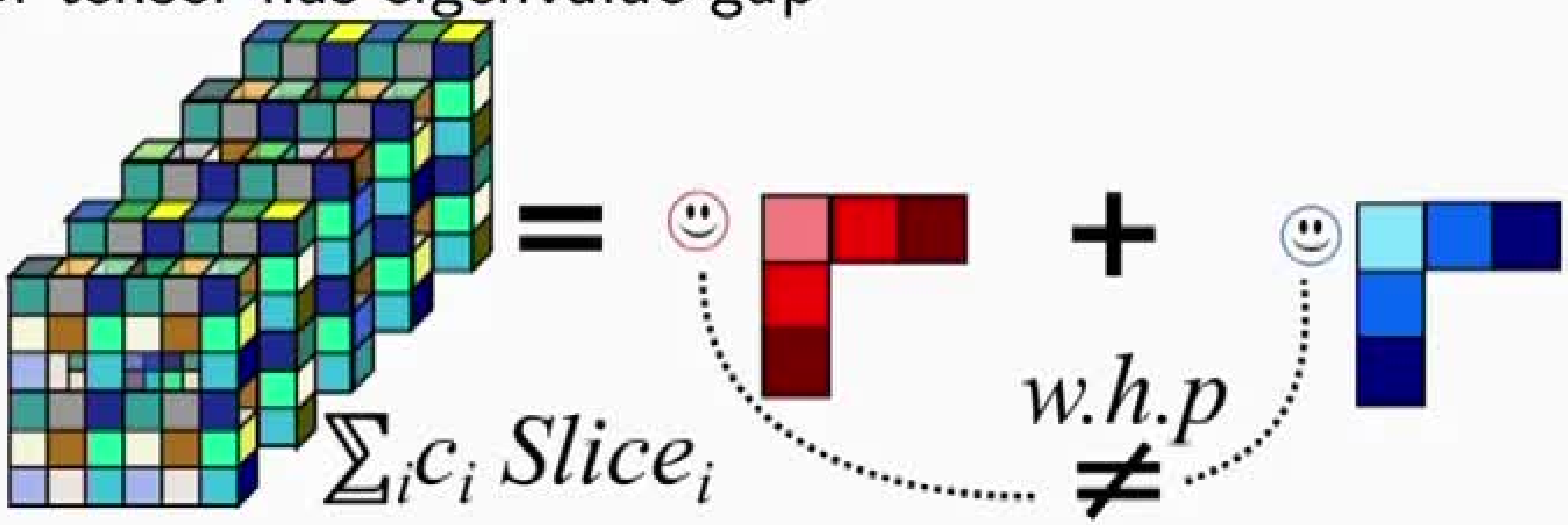
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$$

- Unique with eigenvalue gap



Tensor Orthogonal Decomposition (Harshman, 1970)

- Unique: eigenvalue gap not needed
- Slice of tensor has eigenvalue gap



Outline

1. Introduction
2. Introduction of Method of Moments and Tensor Notations
3. LDA and Community Models
   - From Data Aggregates to Model Parameters
   - Guaranteed Online Algorithm
4. Conclusion

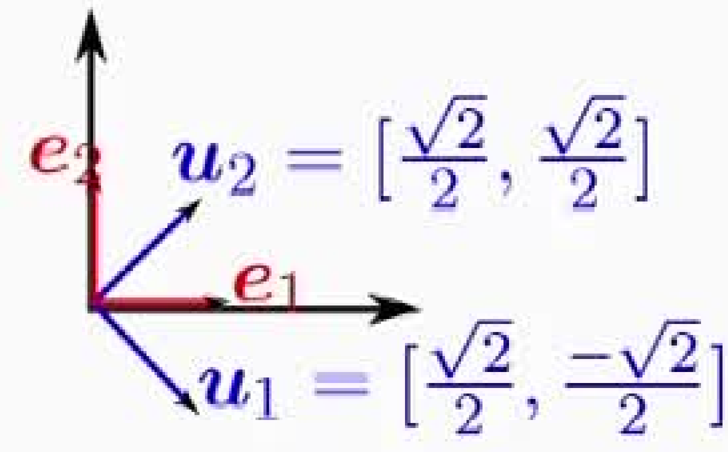
Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap

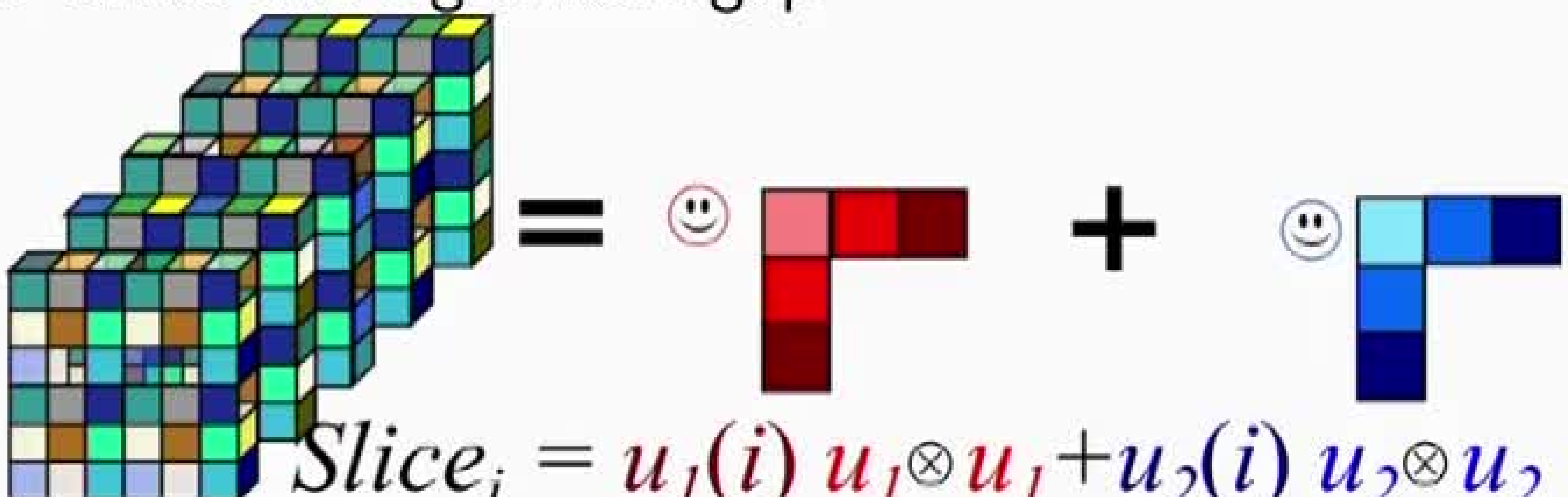
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$$

- Unique with eigenvalue gap



Tensor Orthogonal Decomposition (Harshman, 1970)

- Unique: eigenvalue gap not needed
- Slice of tensor has eigenvalue gap

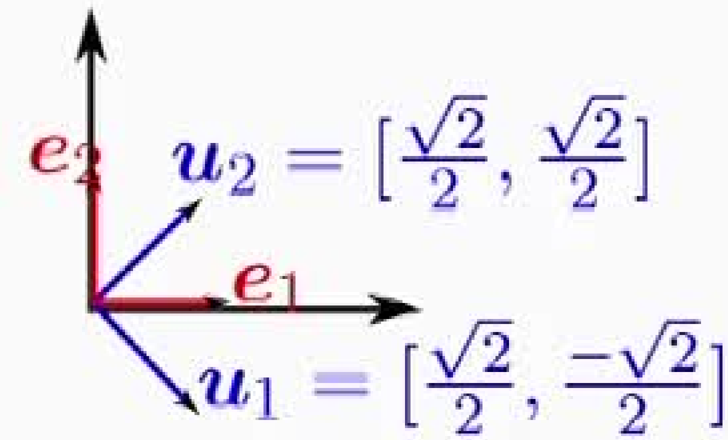


$$Slice_i = u_1(i)\, u_1 \otimes u_1 + u_2(i)\, u_2 \otimes u_2$$

Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap
  $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$
- Unique with eigenvalue gap



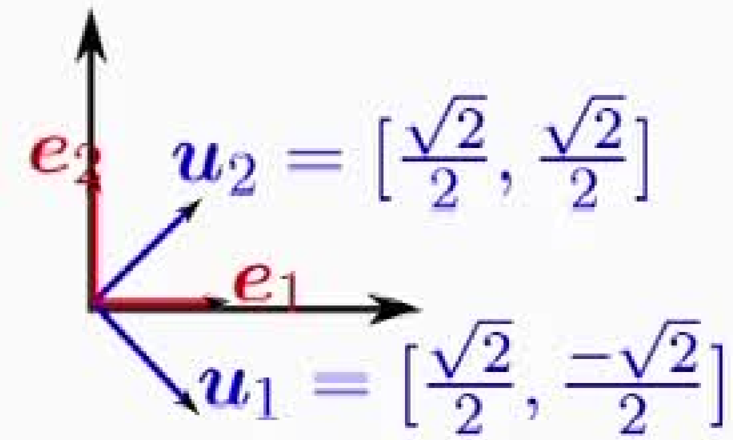
Why are tensors powerful?

Matrix Orthogonal Decomposition

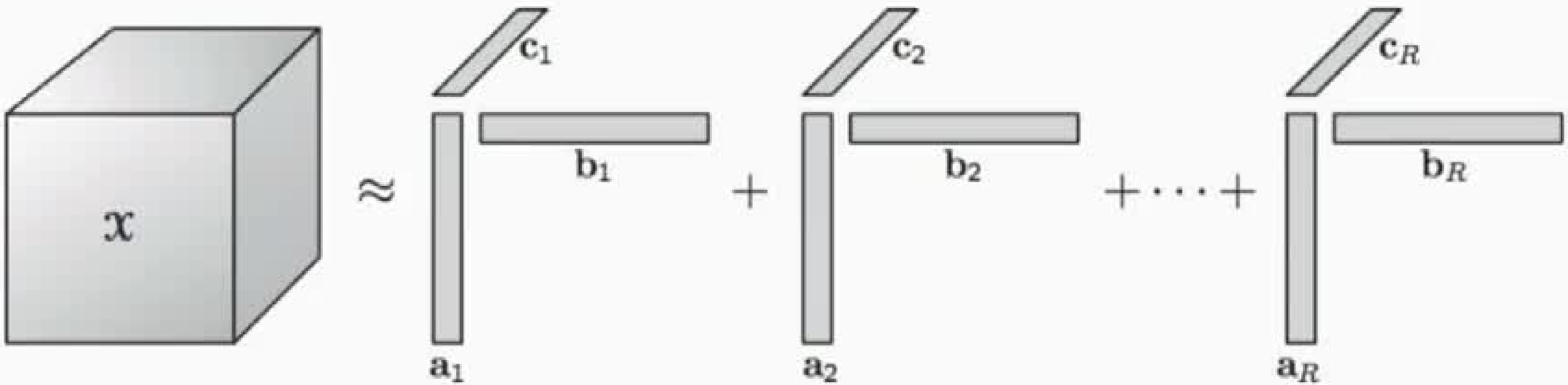
- Not unique without eigenvalue gap

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = e_1 e_1^\top + e_2 e_2^\top = u_1 u_1^\top + u_2 u_2^\top$$

- Unique with eigenvalue gap



CP decomposition



- $\mathcal{X} = \sum_{h=1}^{R} \boldsymbol{a}_h \otimes \boldsymbol{b}_h \otimes \boldsymbol{c}_h$
- Summation of rank-1 tensors

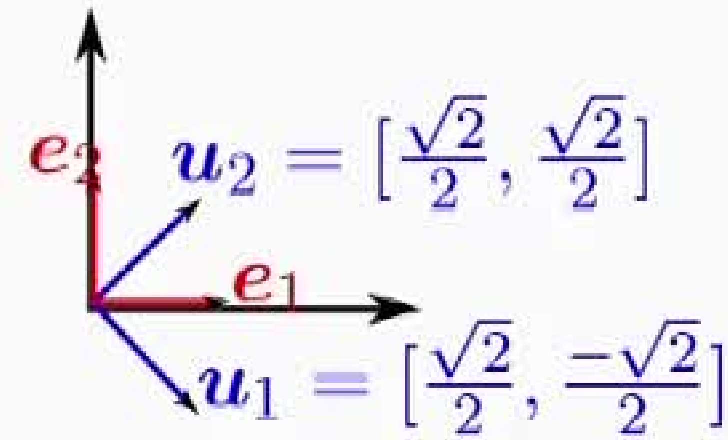
Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap

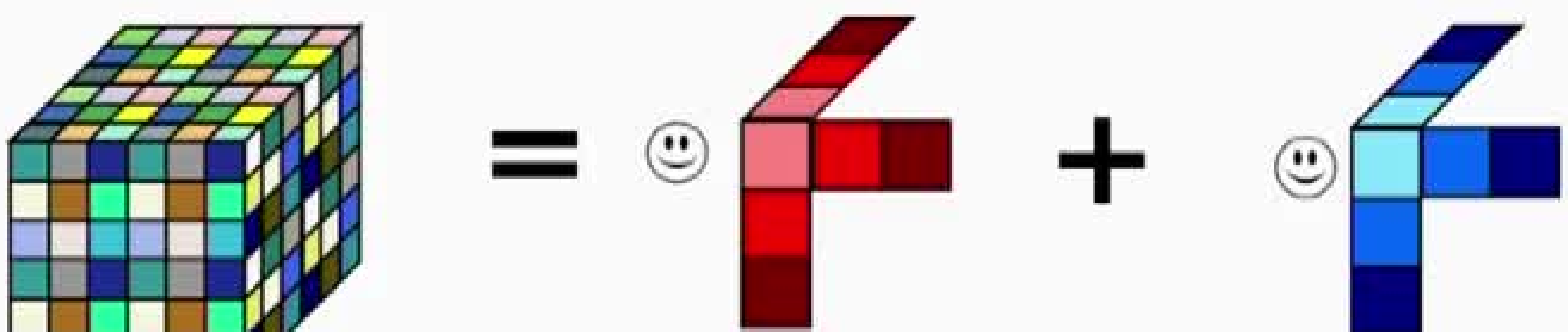
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$$

- Unique with eigenvalue gap



Tensor Orthogonal Decomposition (Harshman, 1970)

- Unique: eigenvalue gap not needed



$$Tensor = u_1 \otimes u_1 \otimes u_1 + u_2 \otimes u_2 \otimes u_2$$

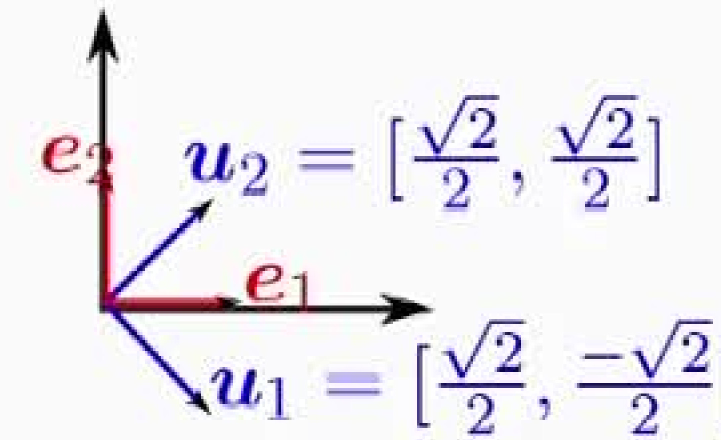
Why are tensors powerful?

Matrix Orthogonal Decomposition

- Not unique without eigenvalue gap

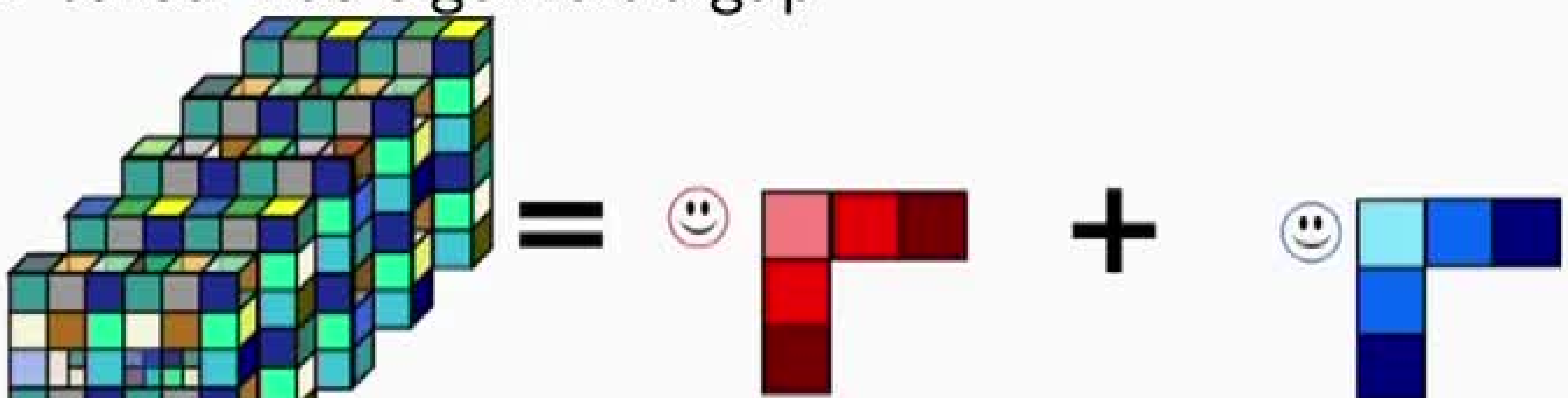
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_2\boldsymbol{e}_2^\top = \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top$$

- Unique with eigenvalue gap



Tensor Orthogonal Decomposition (Harshman, 1970)

- Unique: eigenvalue gap not needed
- Slice of tensor has eigenvalue gap



$$Slice_i = u_1(i)\, u_1 \otimes u_1 + u_2(i)\, u_2 \otimes u_2$$

Outline

1. Introduction
2. Introduction of Method of Moments and Tensor Notations
3. LDA and Community Models
   - From Data Aggregates to Model Parameters
   - Guaranteed Online Algorithm
4. Conclusion

Probabilistic Topic Models - LDA

Bag of words

- Infer topics of documents
- Learn hidden process drives the obs.

Generative model

- Topic proportion $\sim \mathrm{Dir}(\alpha)$ for a doc
- Draw a topic, then a word for a token

Goal

- Topic-word matrix $\boxed{\mathbb{P}[\text{word} = e_i | \text{topic} = j]}$

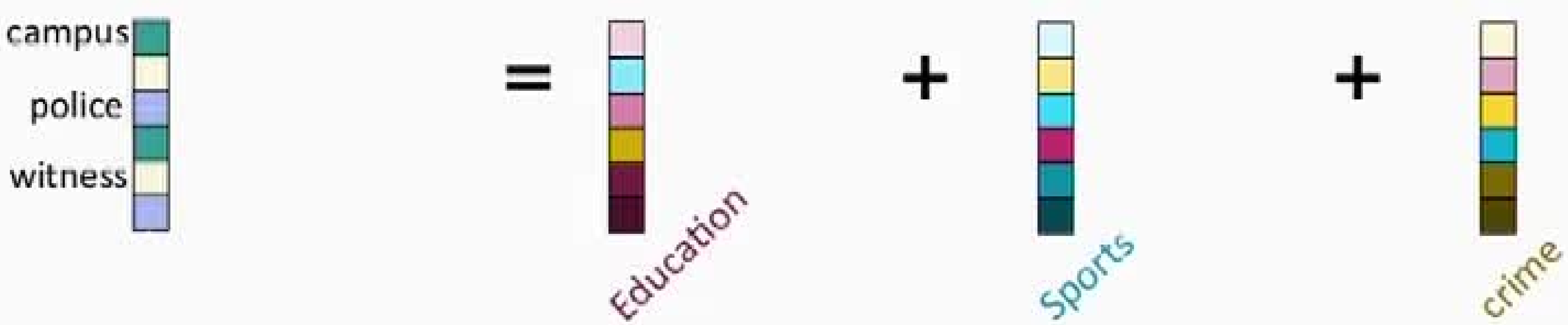
Moments Matching

Goal: Linearly independent topic-word table

$$\mathbb{E}[\text{word}|\text{topic} = j] = \sum_i \mathbb{P}[\text{word} = e_i|\text{topic} = j]e_i = \text{column } j$$

M_1: Occurrence Frequency of Words

$$\mathbb{E}[\text{word}] = \sum_j \mathbb{E}[\text{word}|\text{topic} = j]\mathbb{P}[\text{topic} = j]$$



No unique decomposition of vectors

Moments Matching

Goal: Linearly independent topic-word table

$$\mathbb{E}[\text{word}|\text{topic} = j] = \sum_i \mathbb{P}[\text{word} = e_i|\text{topic} = j]e_i = \text{column } j$$

M_2: Modified Co-occurrence Frequency of Word Pairs

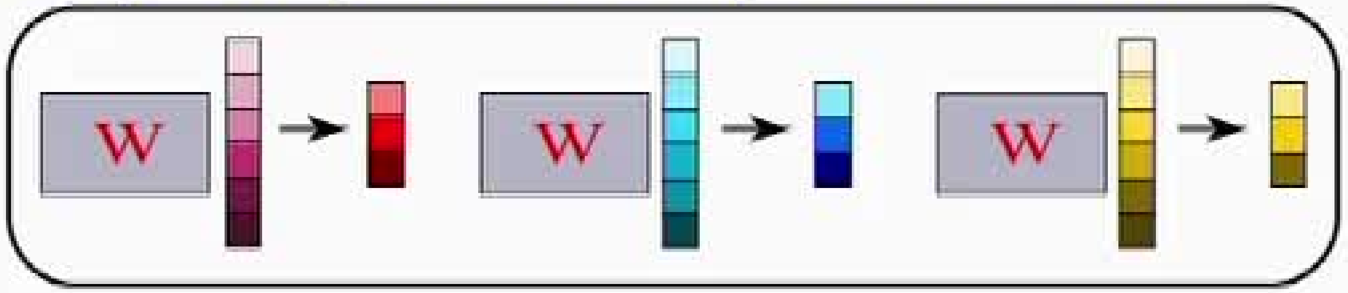
$$\mathbb{E}[\text{word}_1 \otimes \text{word}_2] = \sum_{j,j'} \mathbb{E}[\text{word}_1|\text{topic}_1 = j] \otimes \mathbb{E}[\text{word}_2|\text{topic}_2 = j']\mathbb{P}[\text{topic}_1 = j, \text{topic}_2 = j']$$

Matrix decomposition recovers subspace, not actual model

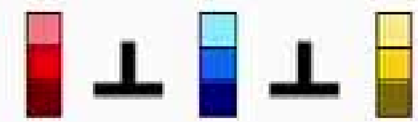
Moments Matching

Goal: Linearly independent topic-word table

Find a W



such that



M_2: Modified Co-occurrence Frequency of Word Pairs

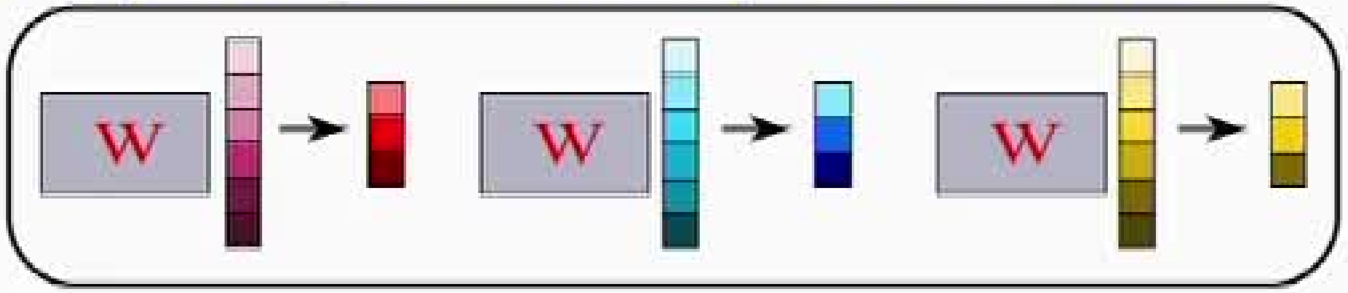
$$\mathbb{E}[\text{word}_1 \otimes \text{word}_2] = \sum_{j,j'} \mathbb{E}[\text{word}_1|\text{topic}_1 = j] \otimes \mathbb{E}[\text{word}_2|\text{topic}_2 = j']\,\mathbb{P}[\text{topic}_1 = j, \text{topic}_2 = j']$$

Many such W's, find one, project data with W

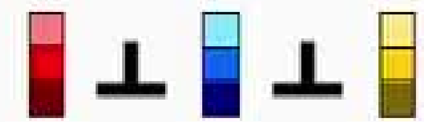
Moments Matching

Goal: Linearly independent topic-word table

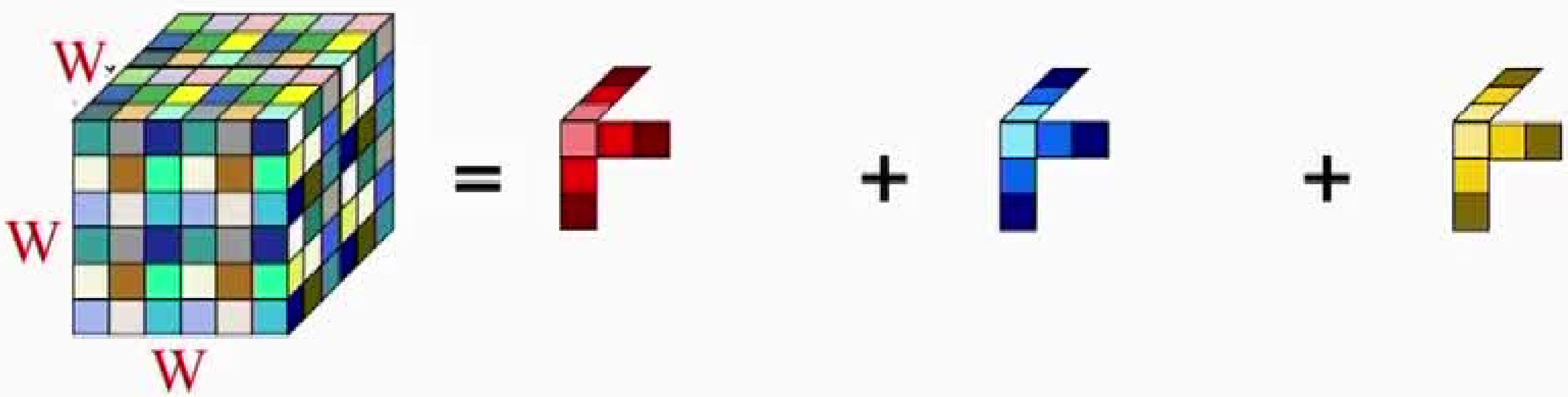
Know a W



such that



M_3: Modified Co-occurrence Frequency of Word Triplets

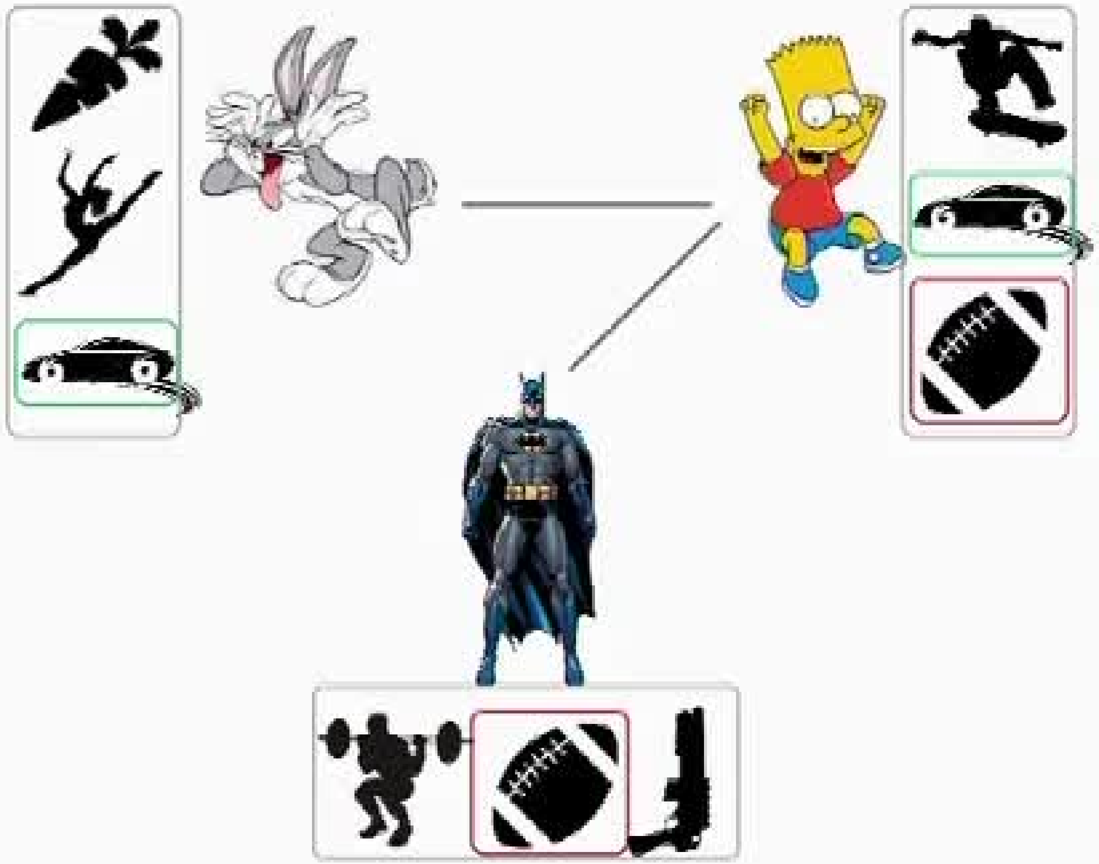


Tensor decomposition uniquely discovers the correct model

Learning Topic Models through Matrix/Tensor Decomposition

Mixed Membership Community Models

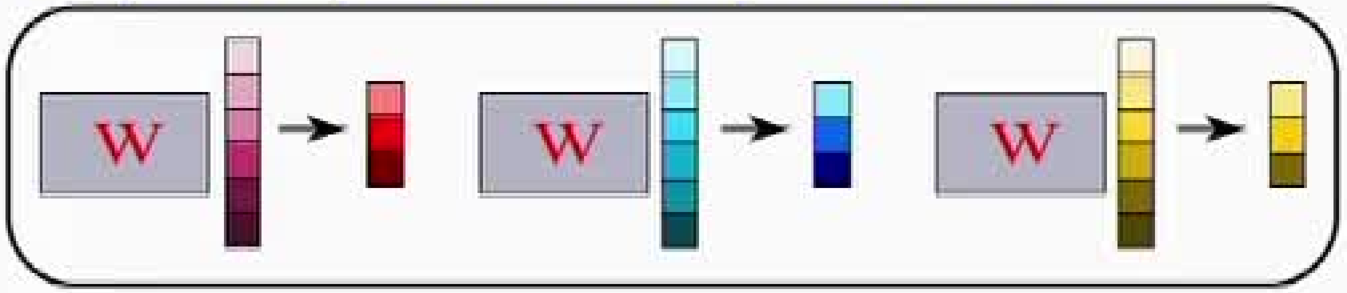
Mixed memberships



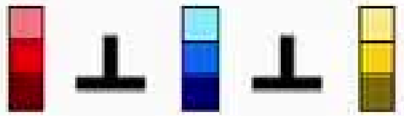
Moments Matching

Goal: Linearly independent topic-word table

Know a W



such that

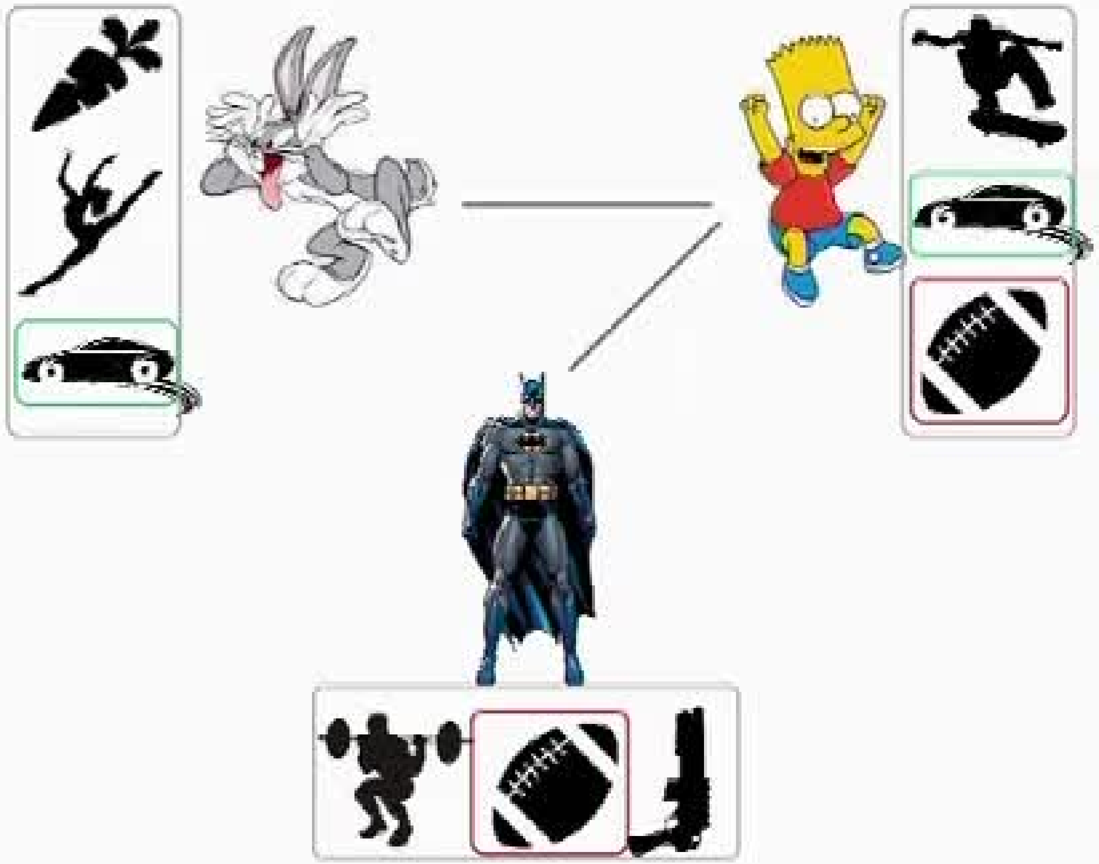


M_3: Modified Co-occurrence Frequency of Word Triplets

Tensor decomposition uniquely discovers the correct model

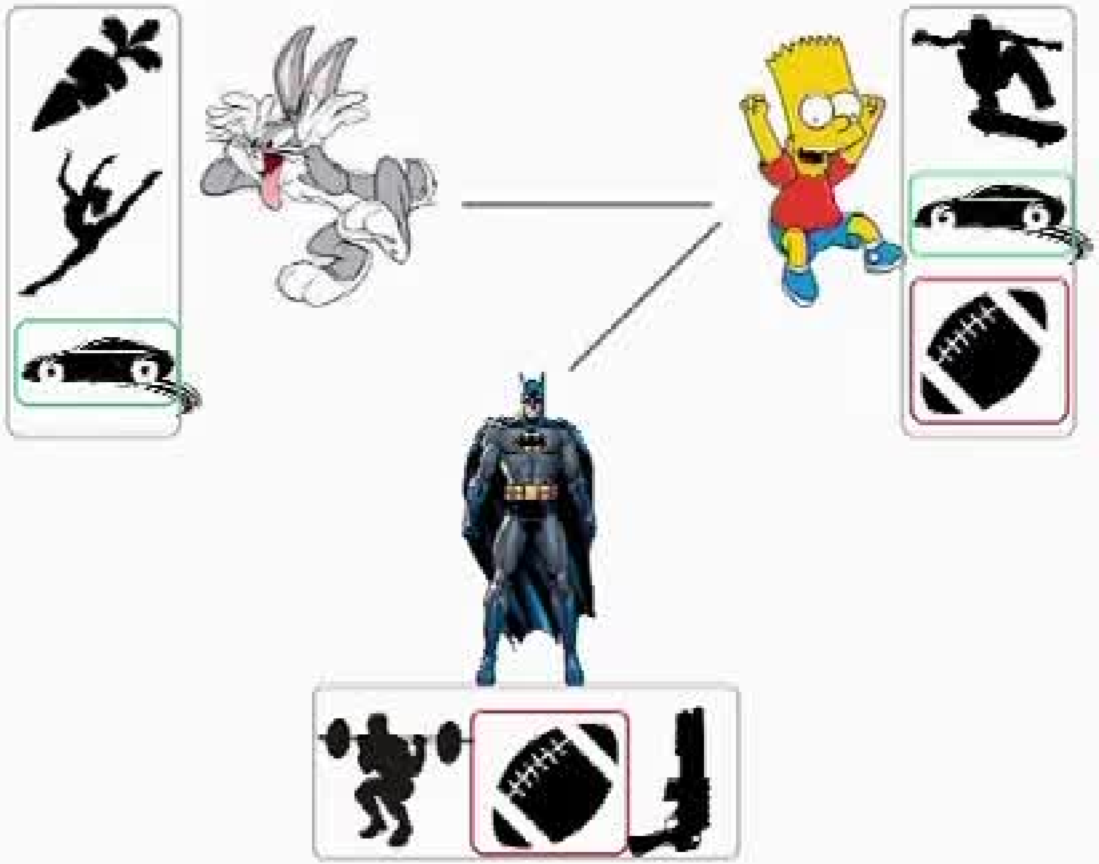
Mixed Membership Community Models

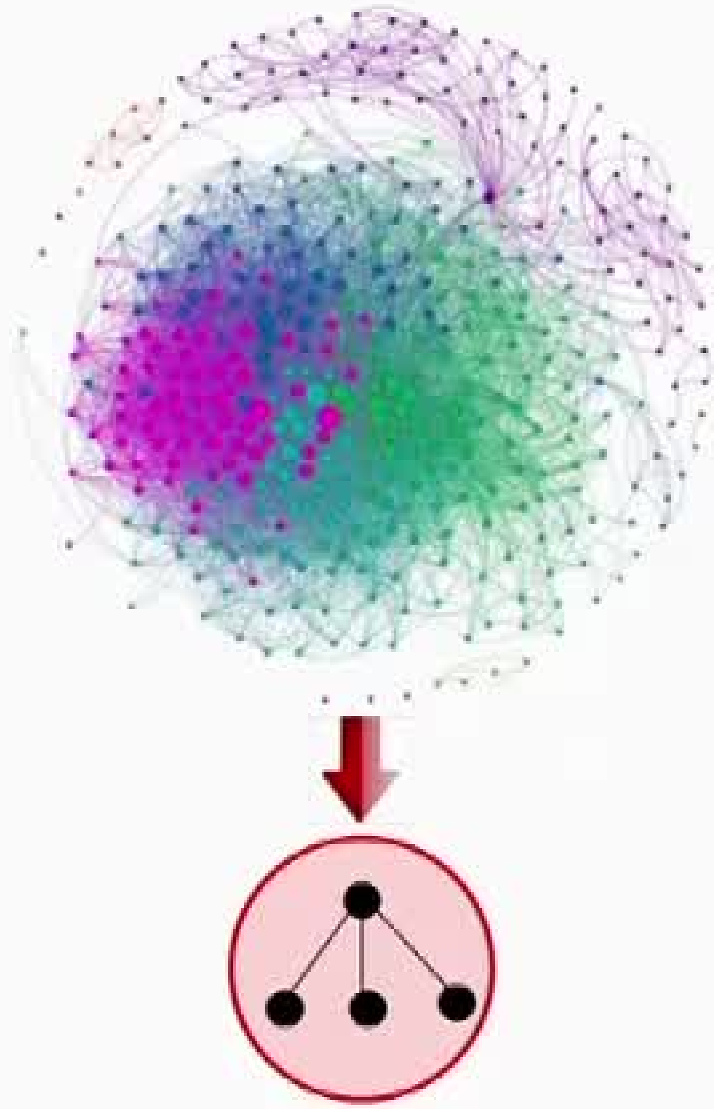
Mixed memberships



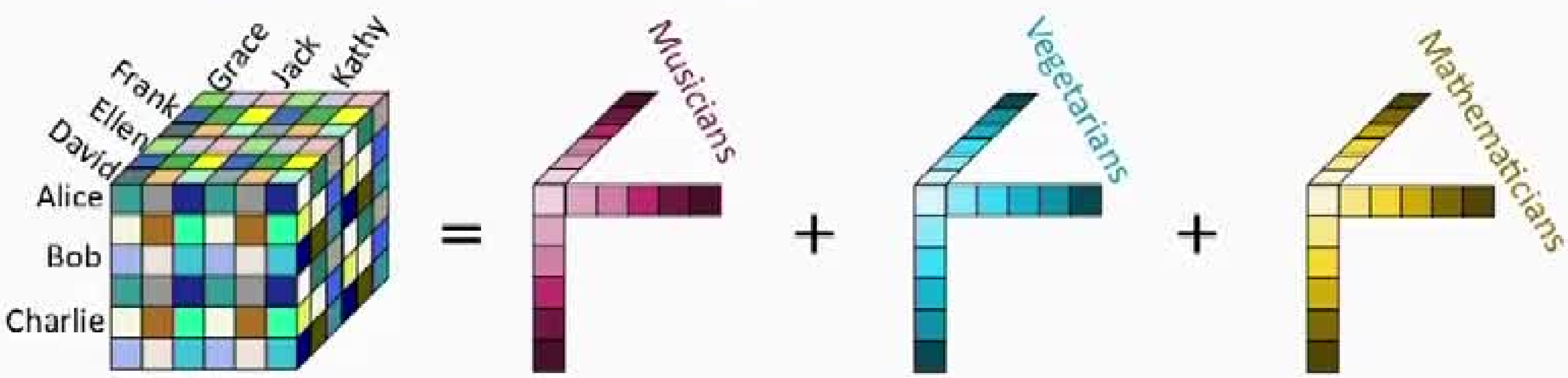
Mixed Membership Community Models

Mixed memberships





What ensures guaranteed learning?



Guaranteed Online Tensor Decomposition

Model is uniquely identifiable! How to identify?

Online Tensor Decomposition

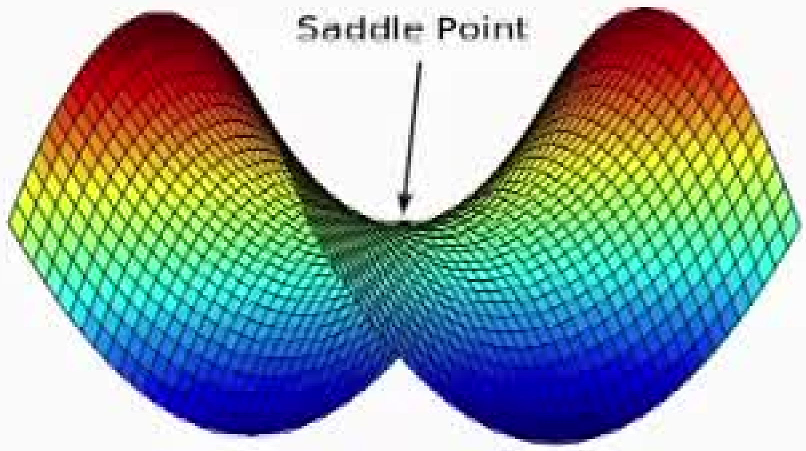
- Tensor $T = \sum_i a_i \otimes a_i \otimes a_i \otimes a_i$, where $\|a_i\| = 1, a_i^\top a_j = 0$

Objective?

- Objective $\min_{\forall i, \|u_i\|^2=1} \quad \sum_{i \neq j} T(u_i, u_i, u_j, u_j)$ **Non-convex!**

Theorem: The proposed objective function has equivalent local optima.

Will SGD work?

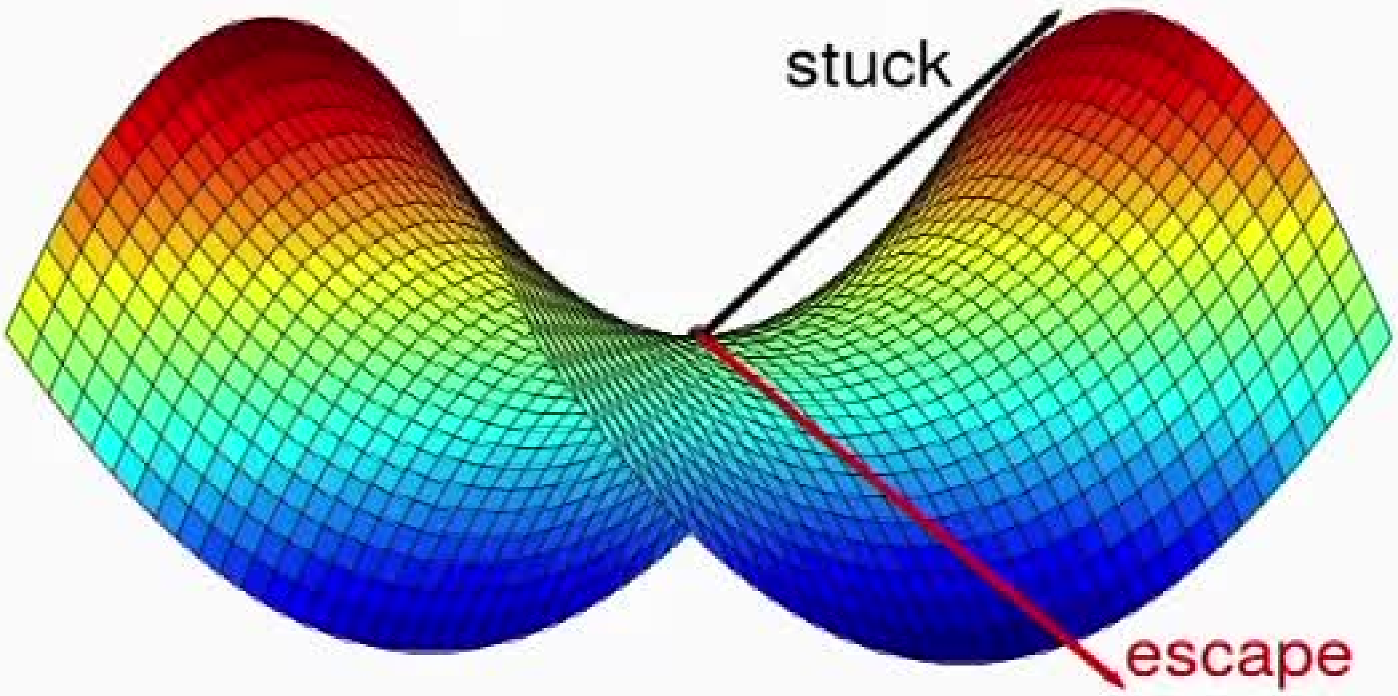


Theorem: For smooth, twice-diff fn. with non-degenerate saddle points, noisy SGD converges to a local optimum in polynomial steps.

Global Convergence Guarantee For Online Tensor Decomposition

Why could we escape from saddle points?

Stochastic Gradient Descent with Noise



- Saddle point has 0 gradient
- Non-degenerate saddle: Hessian has $\pm$ eigenvalue
- Negative eigenvalue: direction of escape

Noise could help!

"Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition",by R. Ge, F. Huang, C. Jin, Y. Yuan, COLT 2015.

Outline

1. Introduction
2. Introduction of Method of Moments and Tensor Notations
3. LDA and Community Models
   - From Data Aggregates to Model Parameters
   - Guaranteed Online Algorithm
4. Conclusion

Summary

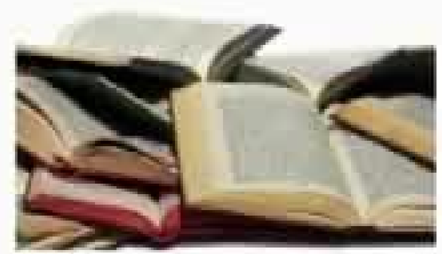
Spectral methods reveal hidden structure

- Text/Image processing
- Social networks
- Neuroscience, healthcare ...

Summary

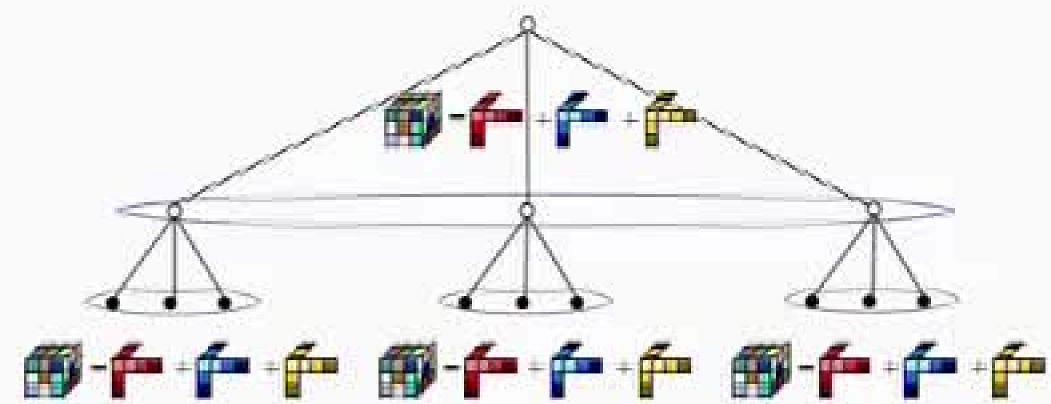
Spectral methods reveal hidden structure

- Text/Image processing
- Social networks
- Neuroscience, healthcare ...



Versatile for latent variable models

- Flat model → hierarchical model
- Sparse coding → convolutional model
- Efficient, convergence guarantee



A list of papers related to this talk

Topic Model, Community Detection, Feature Learning

- "Online Tensor Methods for Learning Latent Variable Models", by F. Huang, U.N. Niranjan, M.U. Hakeem, A. Anandkumar, JMLR 2014.
- "Convolutional Dictionary Learning through Tensor Factorization", by F. Huang and A. Anandkumar, conference and workshop proceeding of JMLR, vol.44, Dec 2015.
- "Tensor Methods on Apache Spark", by F. Huang, A. Anandkumar, Oct. 2015.

Guaranteed tensor decomposition

- "Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition", by R. Ge, F. Huang, C. Jin, Y. Yuan, COLT 2015.

Application in Health Analytics and Neuroscience

- "Scalable Latent TreeModel and its Application to Health Analytics", by F. Huang, U.N. Niranjan, I. Perros, R. Chen, J. Sun, A. Anandkumar, NIPS 2015 MLHC workshop.
- "Discovering Neuronal Cell Types and Their Gene Expression Profiles Using a Spatial Point Process Mixture Model", by F. Huang, A. Anandkumar, C. Borgs, J. Chayes, E. Fraenkel, M. Hawrylycz, E. Lein, A. Ingrosso, S. Turaga, NIPS 2015 BigNeuro workshop.

Collaborators

Anima Anandkumar
Caltech

Rong Ge
Duke University

Srini Turaga
Janelia Research

Chi Jin
UC Berkeley

Jennifer Chayes
MSR

Christian Borgs
MSR

Ernest Fraenkel
MIT

Yang Yuan
Cornell U

UN Niranjan
UC Irvine