UC RIVERSIDE UNIVERSITY OF CALIFORNIA

Tensor Decompositions for Big Multi-aspect Data Analytics

Evangelos (Vagelis) Papalexakis

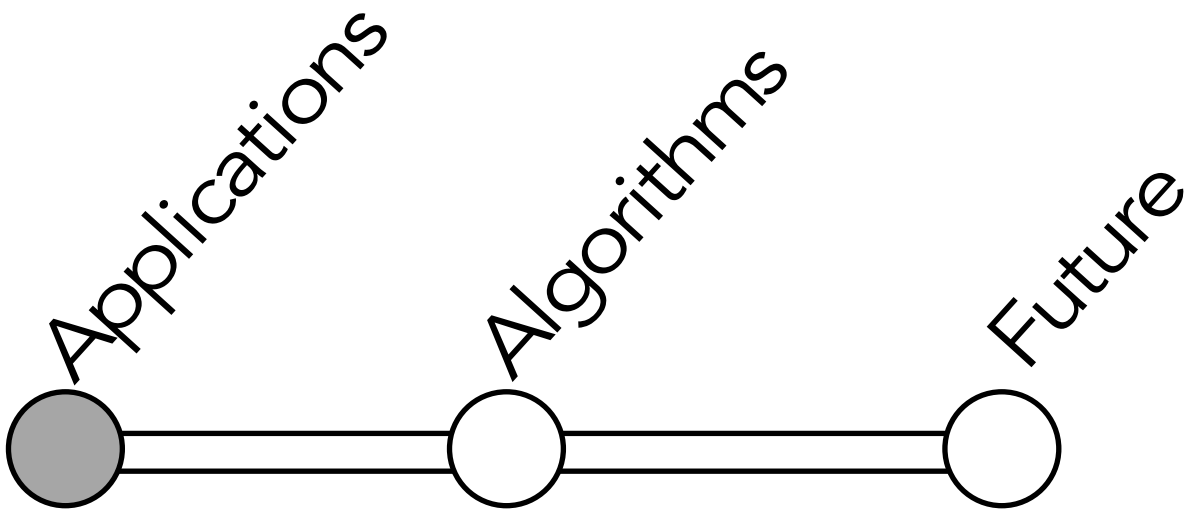
UC Riverside

M A D Lab @ UCR

Multi Aspect Data

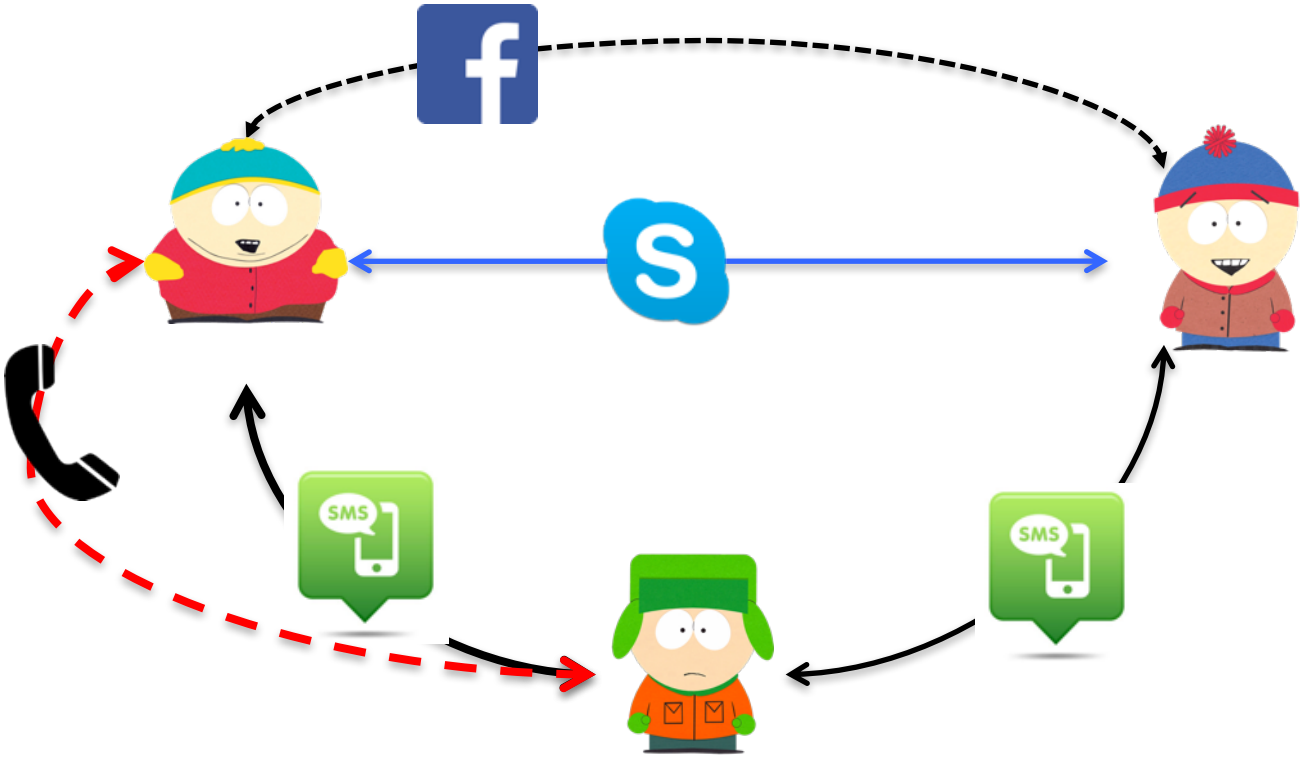
SIAM ALA 2018 – Hong Kong

Roadmap

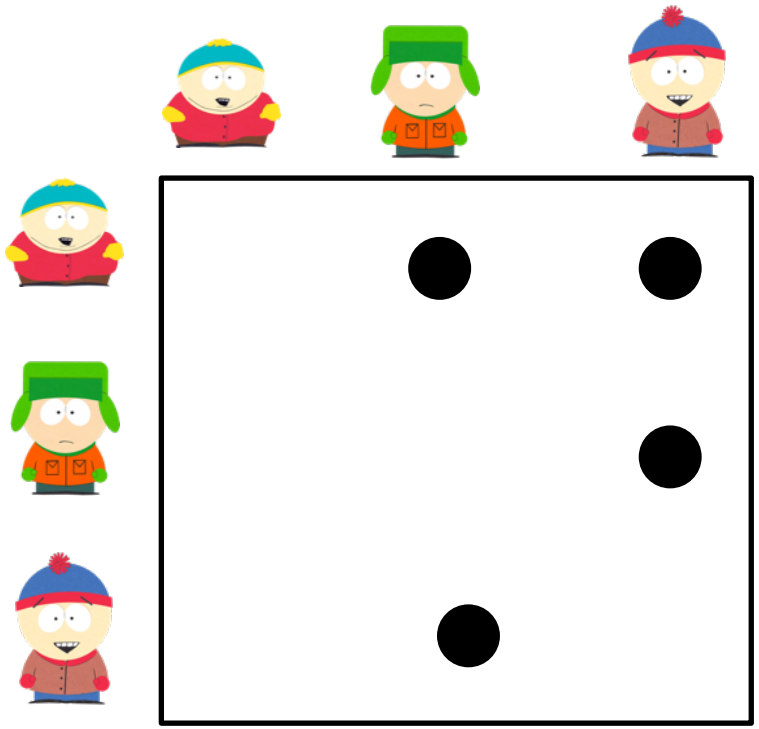


Multi-Aspect Data??

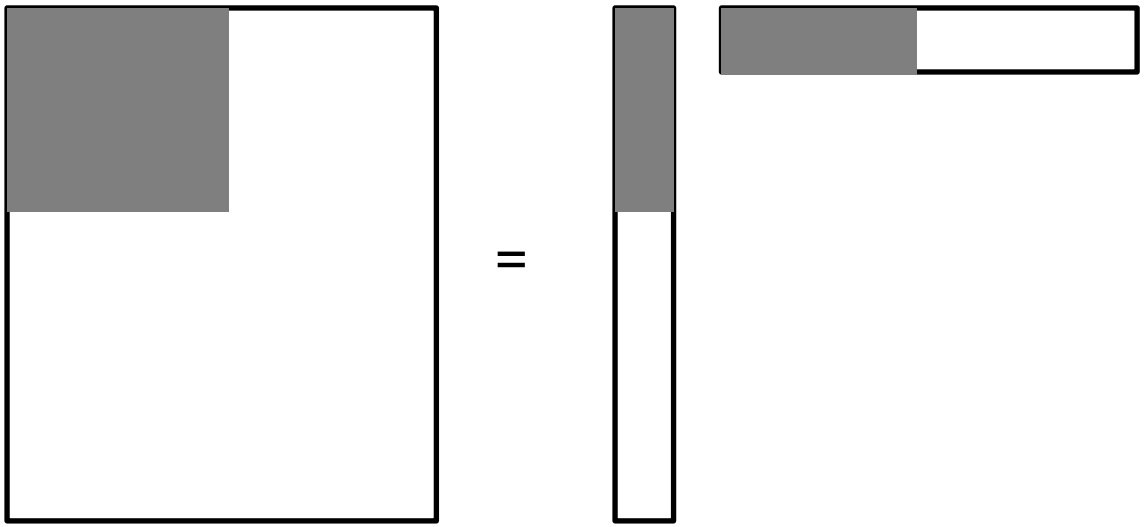
Multi-View Social Networks



Social Network Matrix

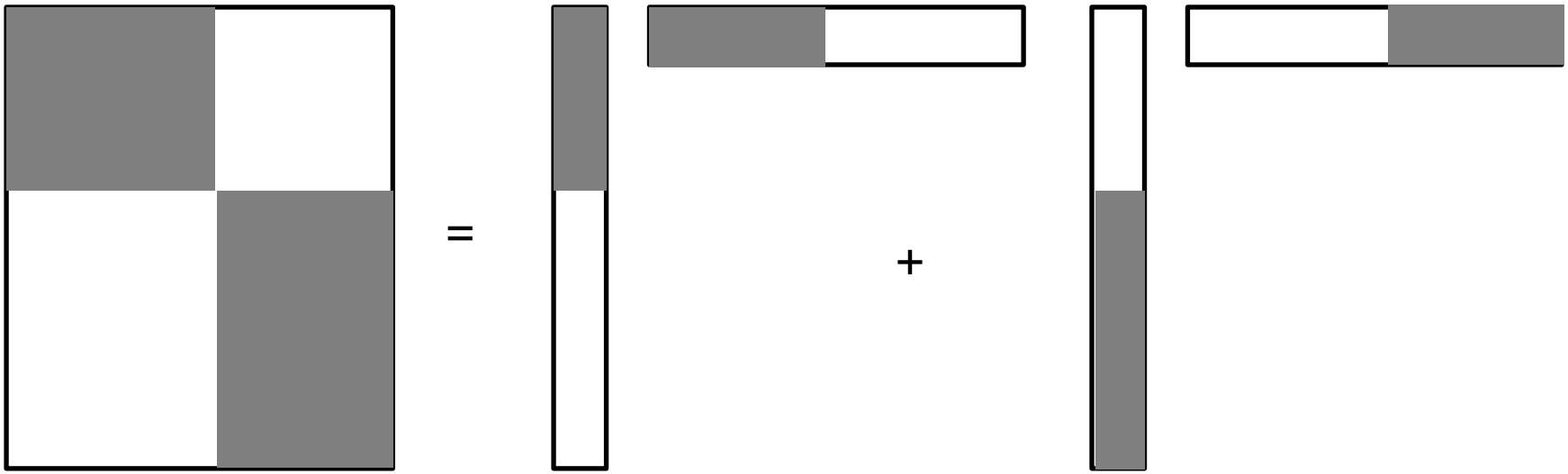


Matrix Factorization



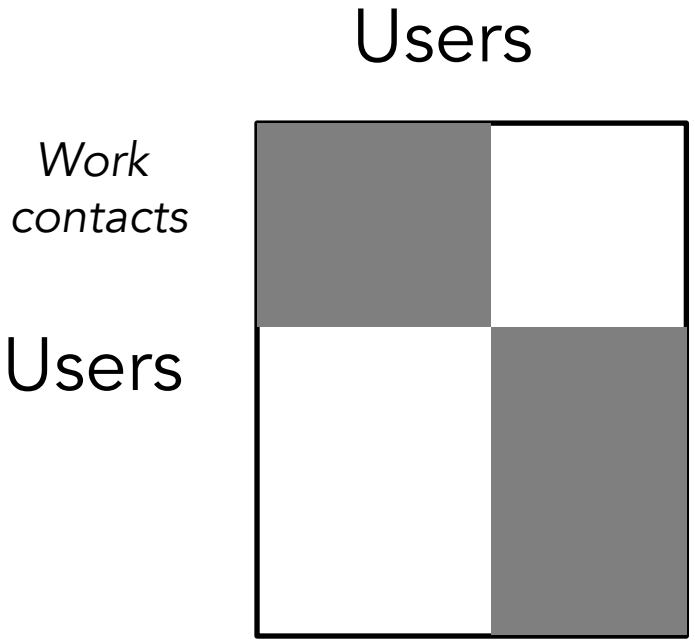
Rank 1

Matrix Factorization



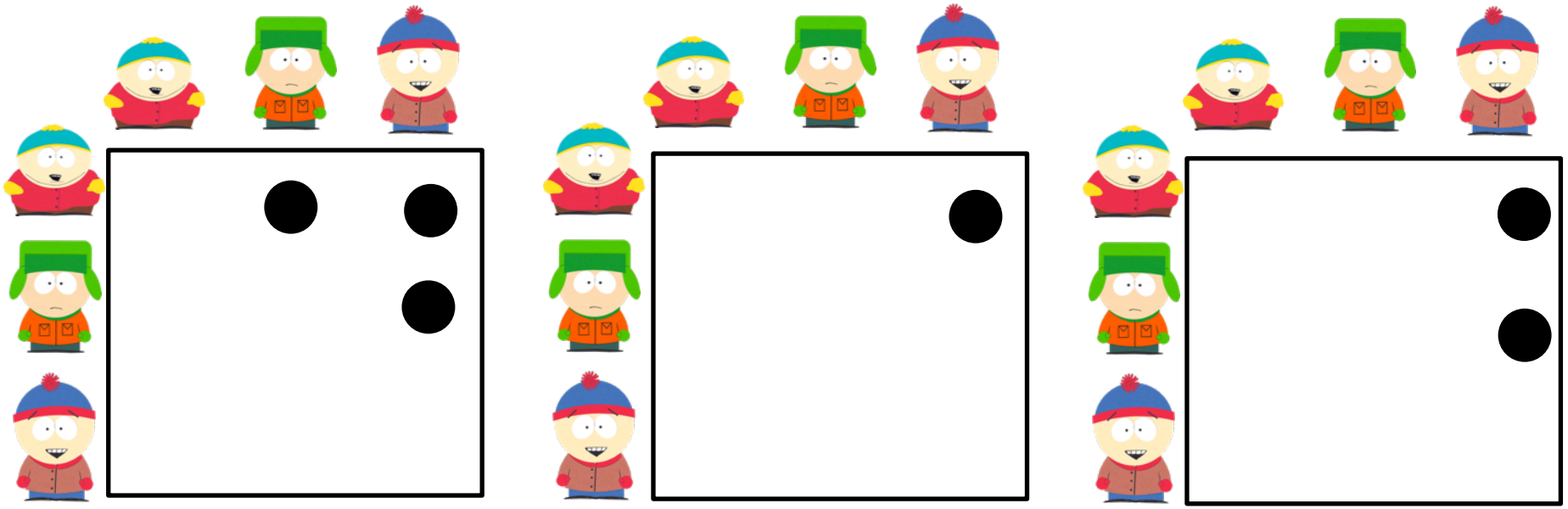
Rank 2

Matrix Factorization



Each block in the data is a latent ("hidden") concept

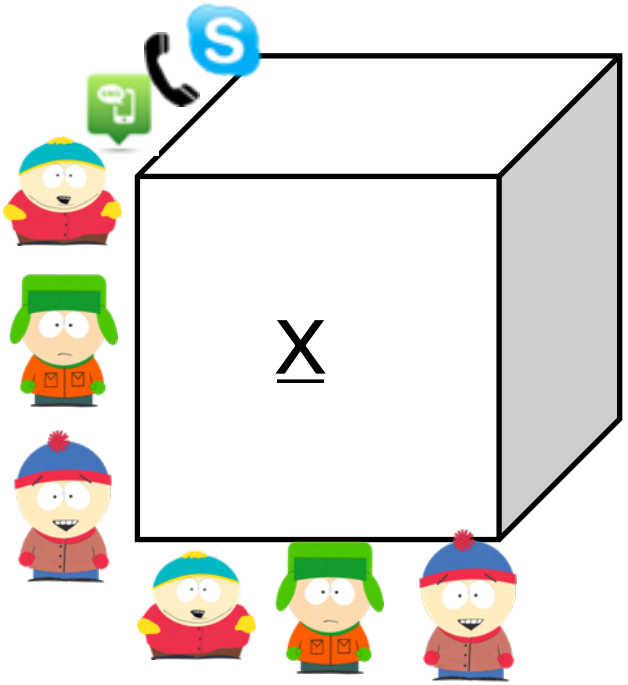
What about the rest of the views??



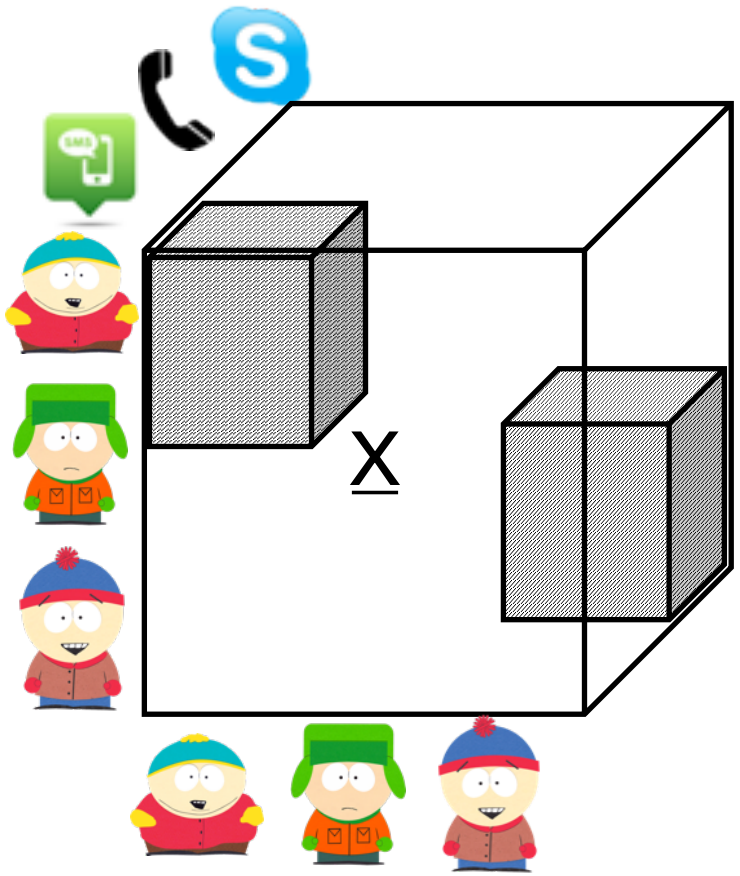
If we aggregate, we ignore important structure!!

Tensors

- Multi-dimensional matrices
- Model multi-aspect datasets
- Long list of applications: Chemometrics, Psychometrics, Signal Processing, Machine Learning, Data Mining

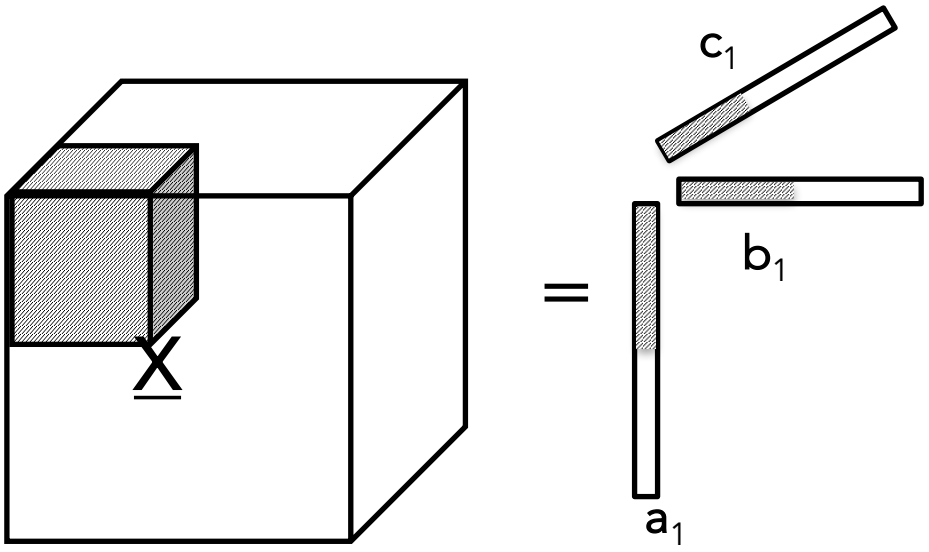


What are we looking for?



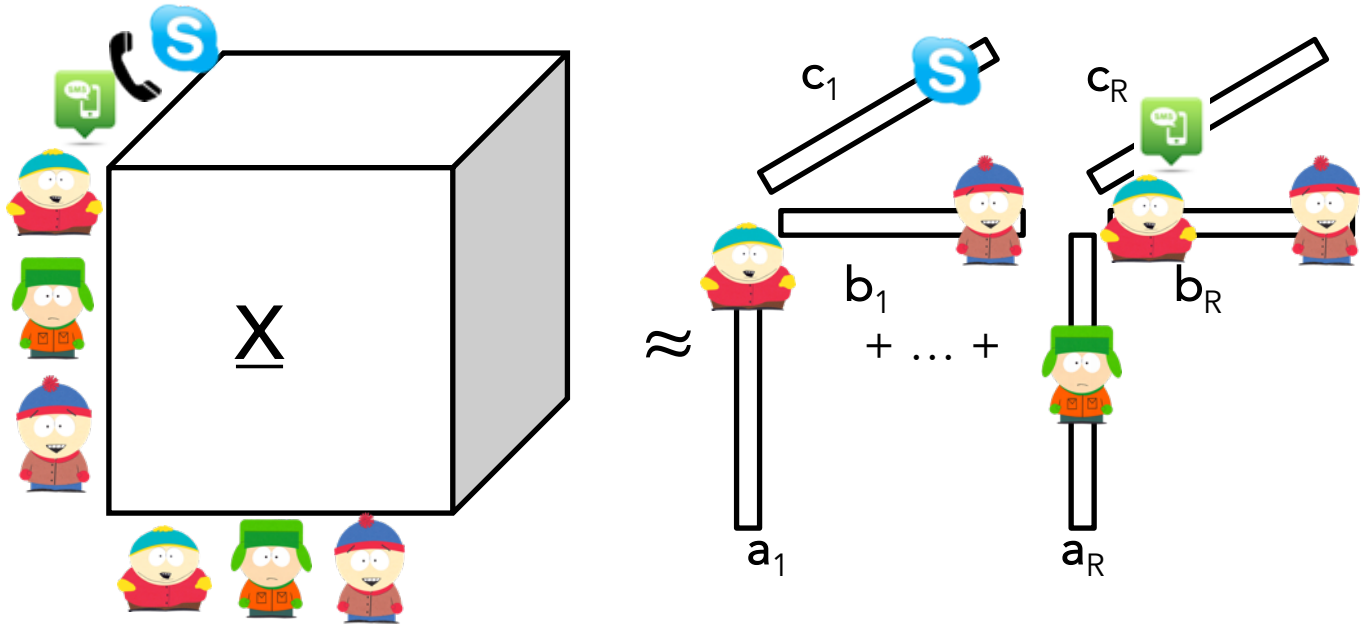
Blocks within the data
Subsets / co-clusters of:
1) Users (“senders”)
2) Users (“receivers”)
3) Means of communication

Blocks are rank-one tensors



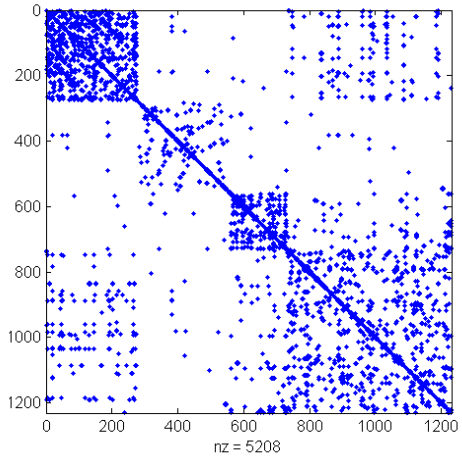
Direct extension of matrix case!

CP/PARAFAC Decomposition

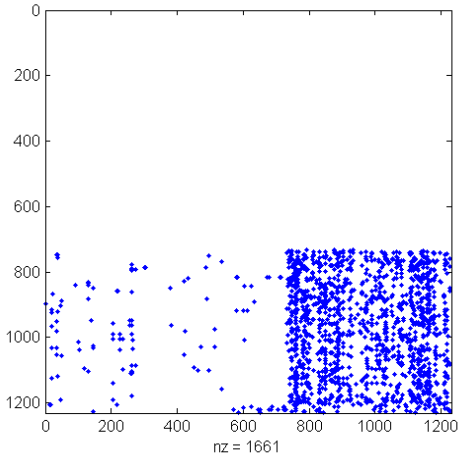


$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\underline{\mathbf{X}} - \sum_{R} \mathbf{a}_R \circ \mathbf{b}_R \circ \mathbf{c}_R\|_F^2$$

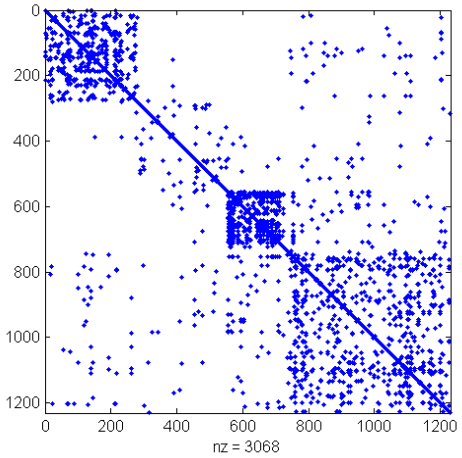
DBLP Multi-View Graph



(a) citation

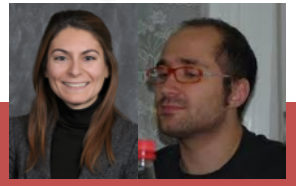


(b) co-auth.



(c) co-term

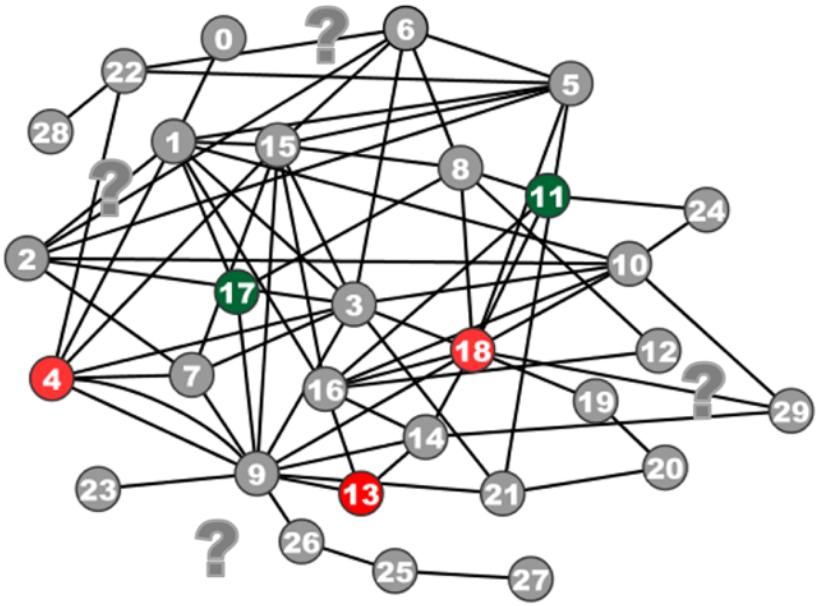
- Assignment of authors to research communities
- Measure NMI (Normalized Mutual Information)
- Baselines
  - Spectral clustering on sum of matrices / views
  - Linked Matrix Factorization [Tang et al. ICDM 2009]
- GRAPHFUSE outperforms "2D" baselines



[Papalexakis, Akoglu, Ienco **Fusion 2013**]

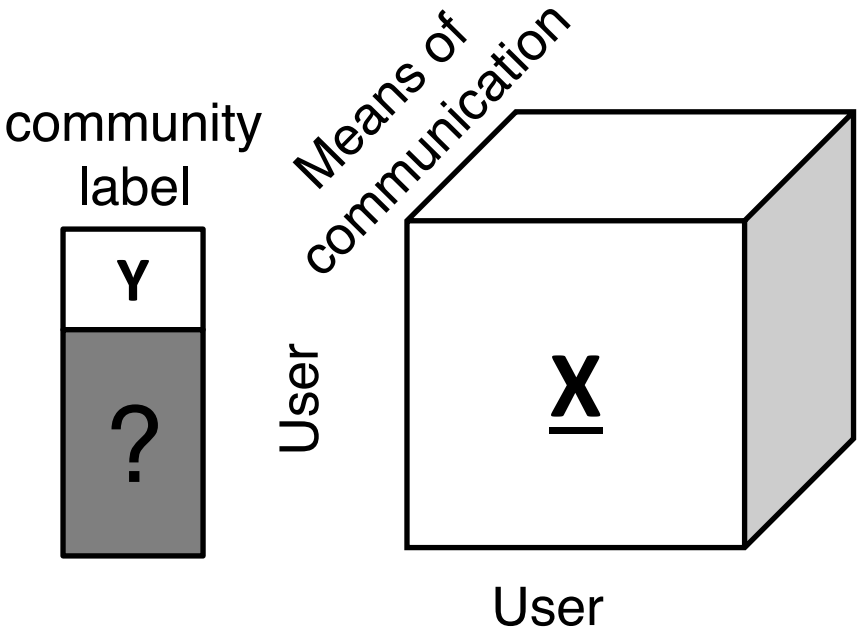
Semi-supervised Community Detection

- What if we have very few community labels?
- Use "Guilt-by-association
  - Also called homophily
- Propagate labels in the graph
- BUT: this ignores multi-view structure!



Fast Belief Propagation [Koutra et al. 2011]

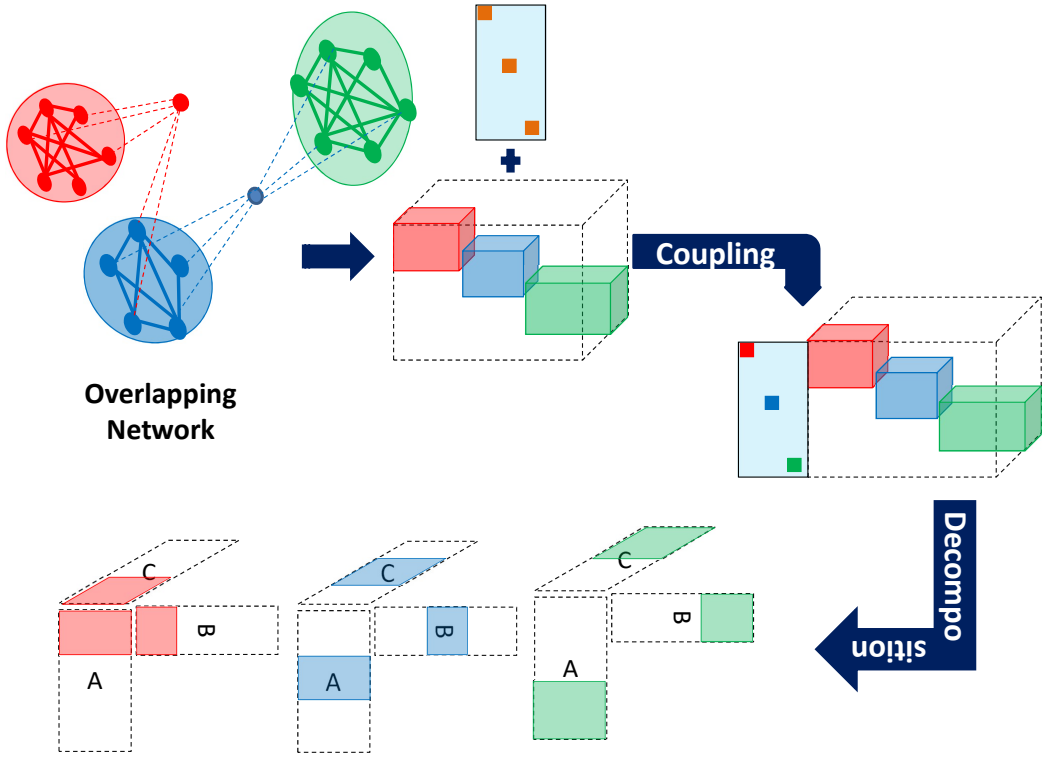
Semi-supervised Community Detection



Coupling as semi-supervision!

SDM 2018 w/ Ekta Gujral

SMACD: Semi-supervised Multi-Aspect Community Detection

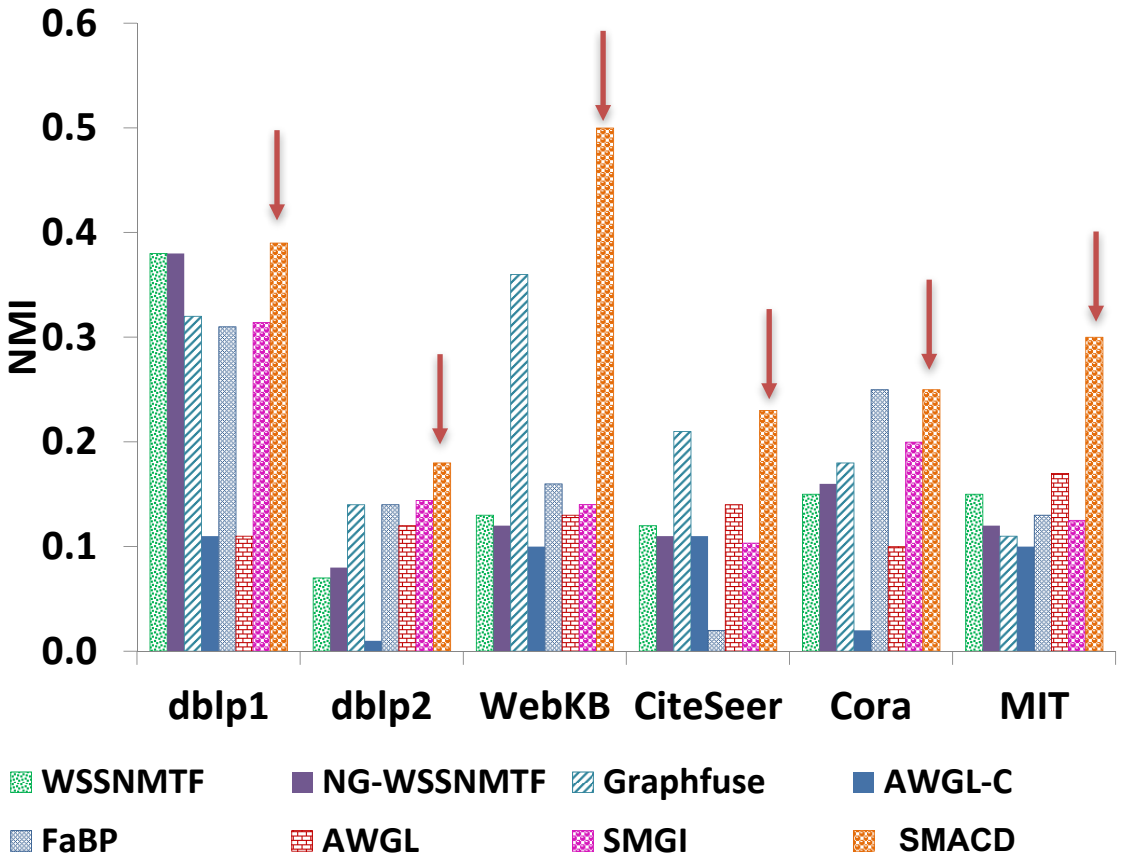


$$\min_{\mathbf{A}\geq\mathbf{0},\mathbf{B}\geq\mathbf{0},\mathbf{C}\geq\mathbf{0},\mathbf{D}\geq\mathbf{0}} \|\underline{\mathbf{X}} - \sum_r \mathbf{A}(:,r)\circ\mathbf{B}(:,r)\circ\mathbf{C}(:,r)\|_F^2 + \|\mathbf{Y} - \mathbf{A}\mathbf{D}^T\|_F^2$$

$$+ \lambda \sum_{i,r} |\mathbf{A}(i,r)| + \lambda \sum_{j,r} |\mathbf{B}(j,r)| + \lambda \sum_{k,r} |\mathbf{C}(k,r)| + \lambda_d \sum_{l,r} |\mathbf{D}(l,r)|$$

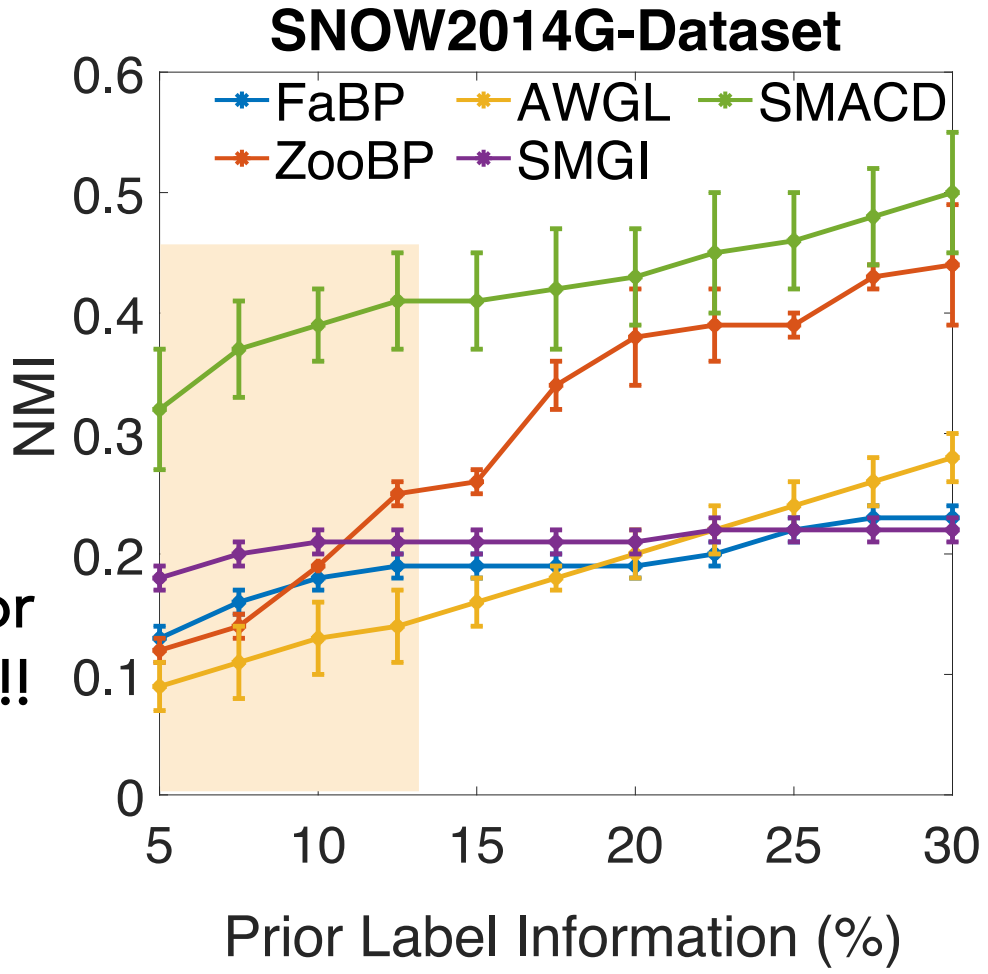
SDM 2018 w/ Ekta Gujral

SMACD: Semi-supervised Multi-Aspect Community Detection



SDM 2018 w/ Ekta Gujral

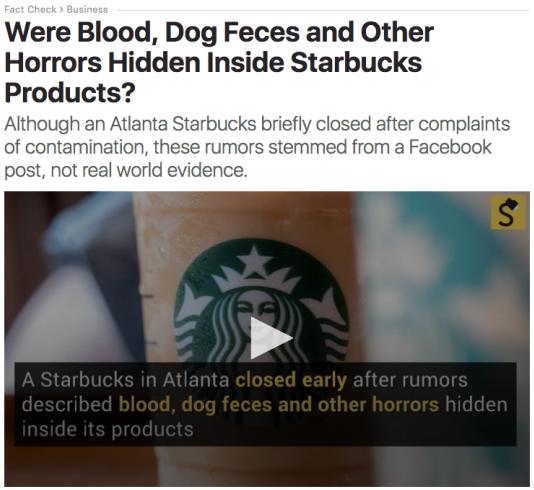
SMACD: Semi-supervised Multi-Aspect Community Detection

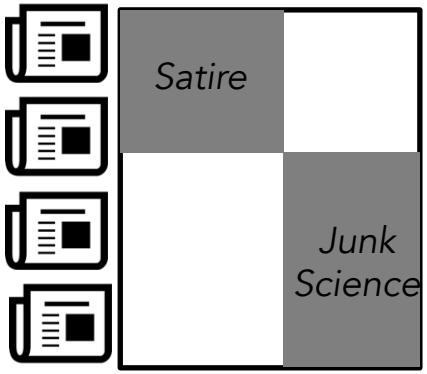


Works well for small #labels!!

SDM 2018 w/ Ekta Gujral

Unsupervised Fake News Identification





- Bag-of-words typically loses context info
- We need to capture context/spatial relations of different (groups of) terms

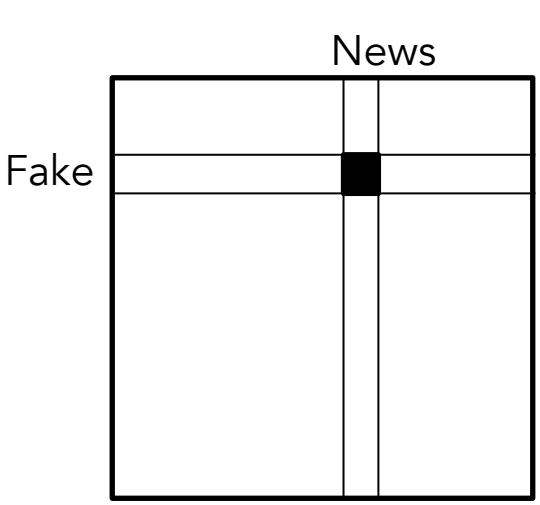
BREAKING BOMBSHELL: NYPD Blows Whistle on New Hillary Emails: Money Laundering, Sex Crimes with Children Child Exploitation, Pay to Play Perjury

Lot of information in title

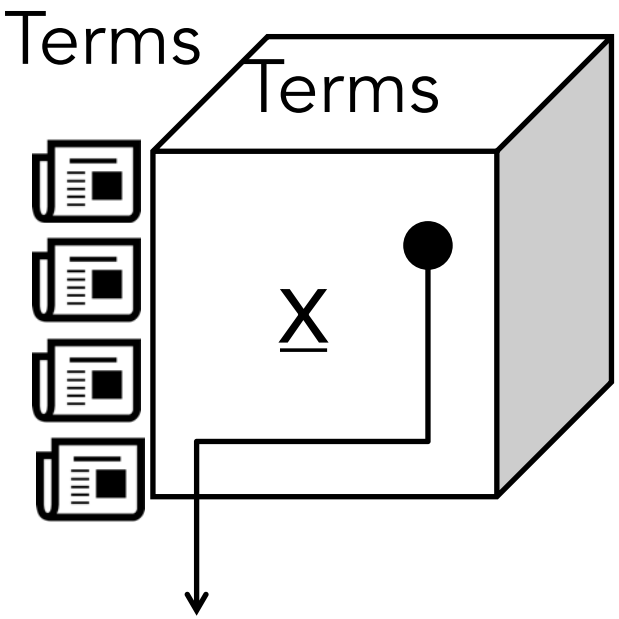
Simple and repetitive content

http://snap.stanford.edu/www2017tutorial/docs/050-hoax.pdf

Unsupervised Fake News Identification



("Fake", "News") co-occur
f times within *k* terms of each other

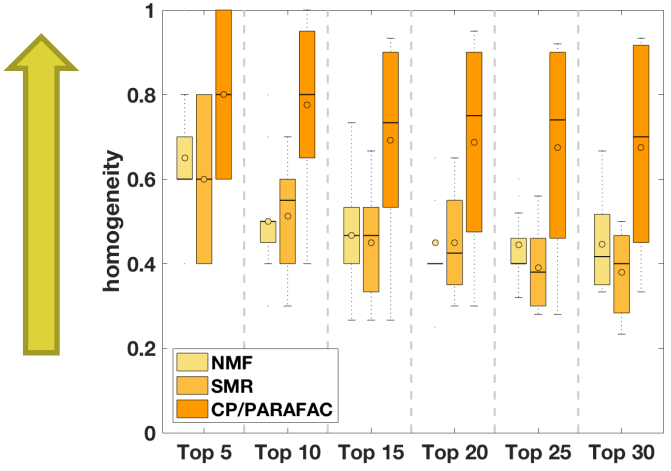


$\underline{X}(i,j,k) = f$: word j and word k co-occur f times in article i

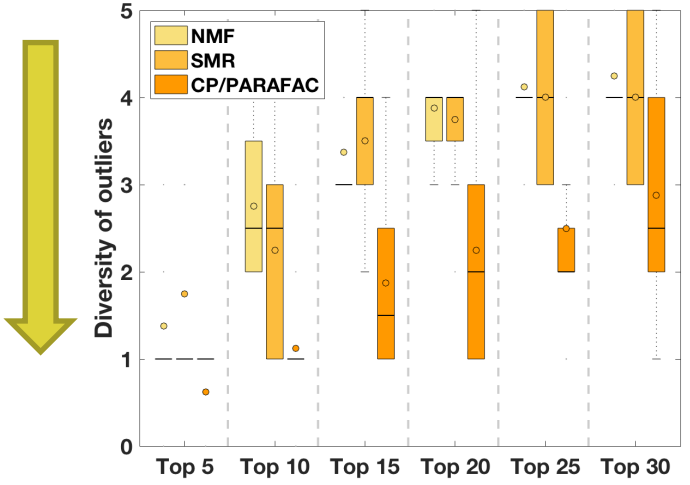
WSDM18 MIS2 Workshop [Best Paper Award] w/ Mehdi Hosseini

Unsupervised Fake News Identification

- **Homogeneity @ K**
  - Sort the values of each latent factor
  - Take the top-K
  - Measure homogeneity of article labels
  - **Higher** is better

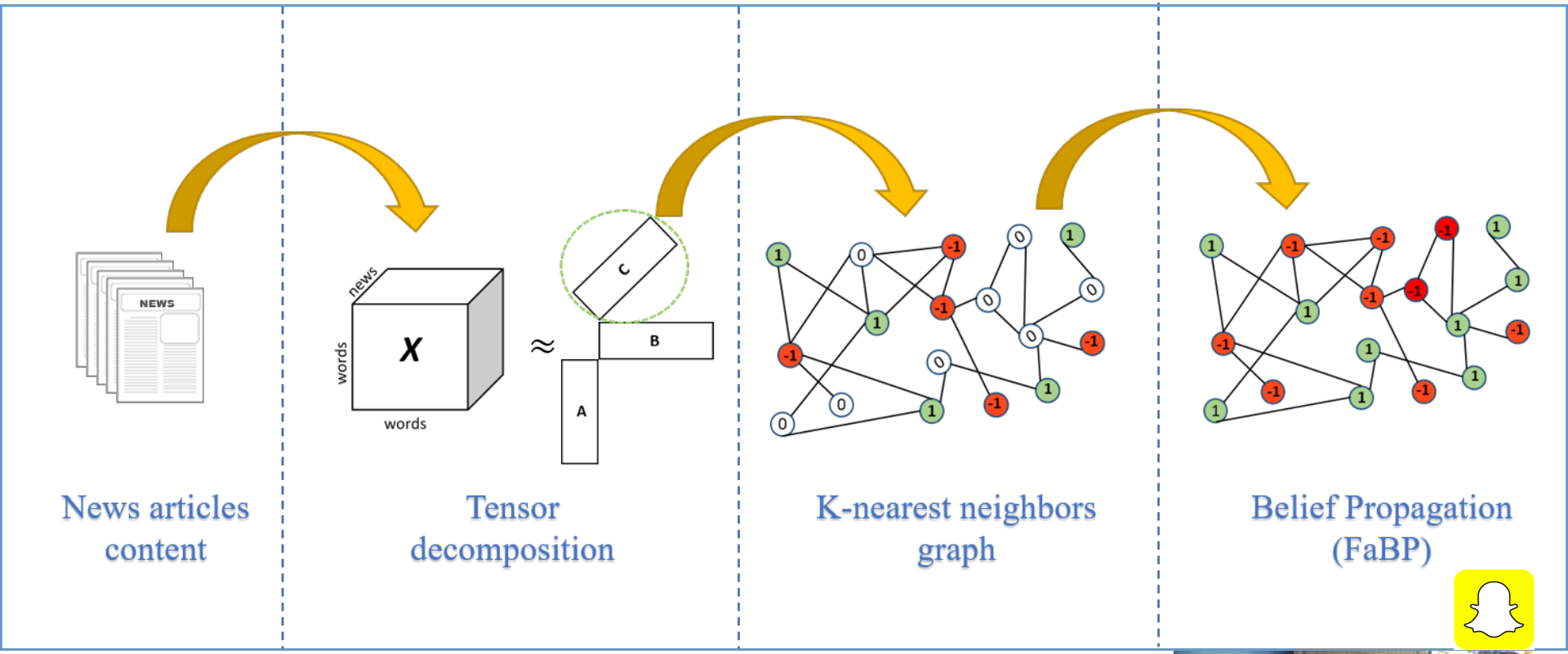


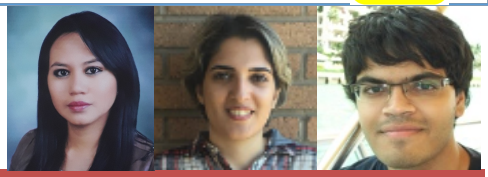
- **Diversity of outliers @ K**
  - Within the top-K find articles with diff. label from the dominant one
  - Count their distinct labels
  - **Lower** is better



WSDM18 MIS2 Workshop [**Best Paper Award**] w/ Mehdi Hosseini

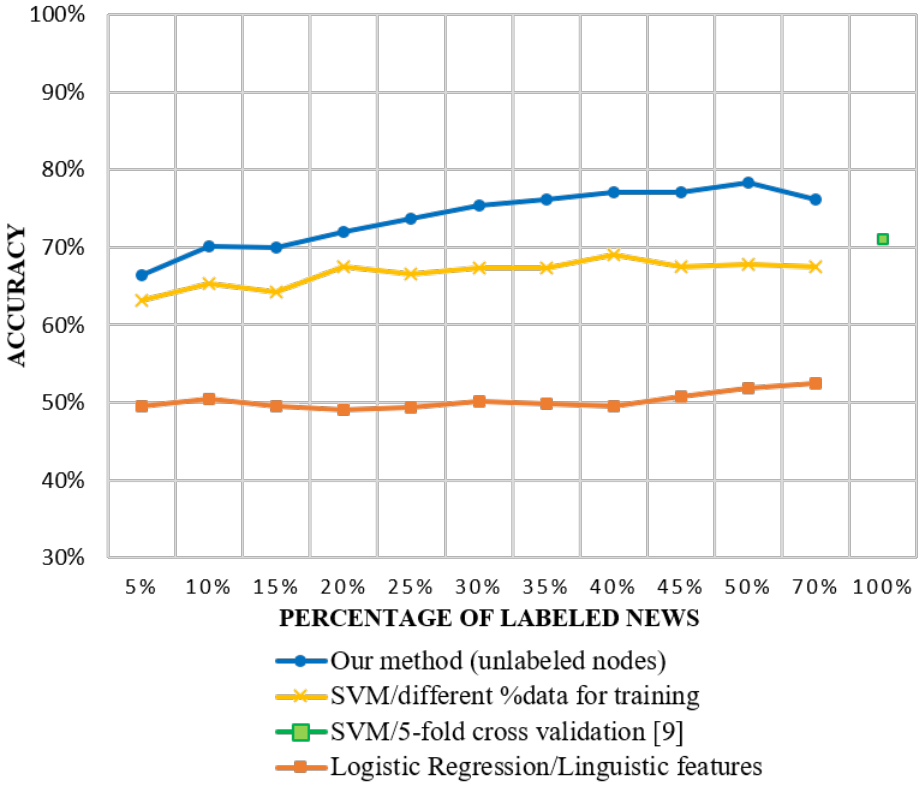
Semi-supervised Fake News Detection

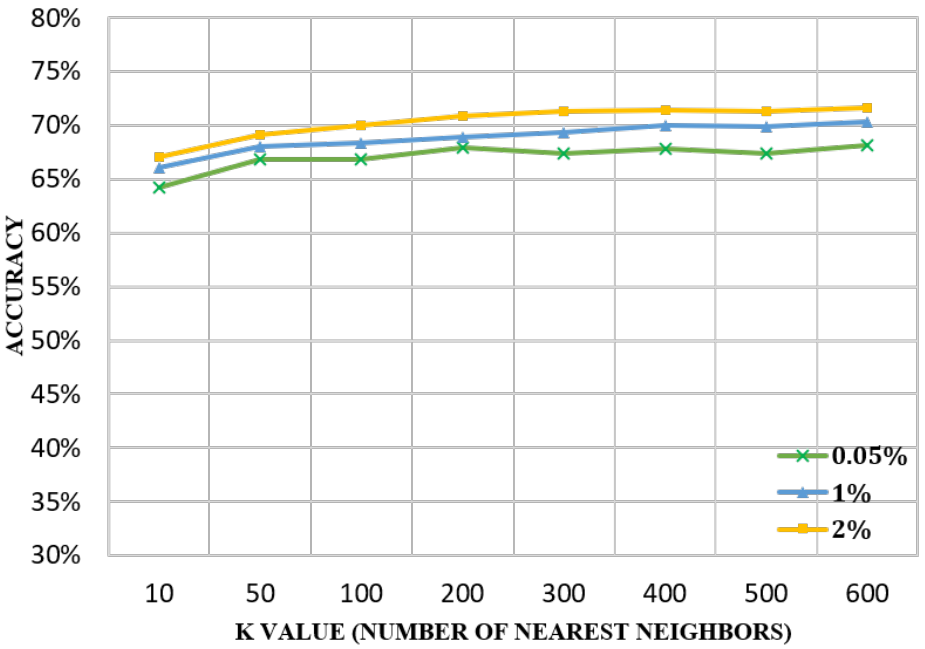




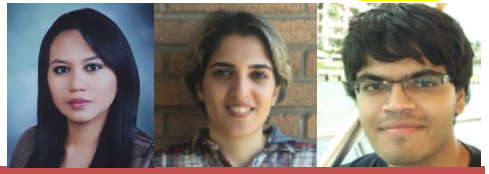
arxiv.org/pdf/1804.09088.pdf w/ Gisel Guacho, Sara Abdali, Neil Shah

Semi-supervised Fake News Detection

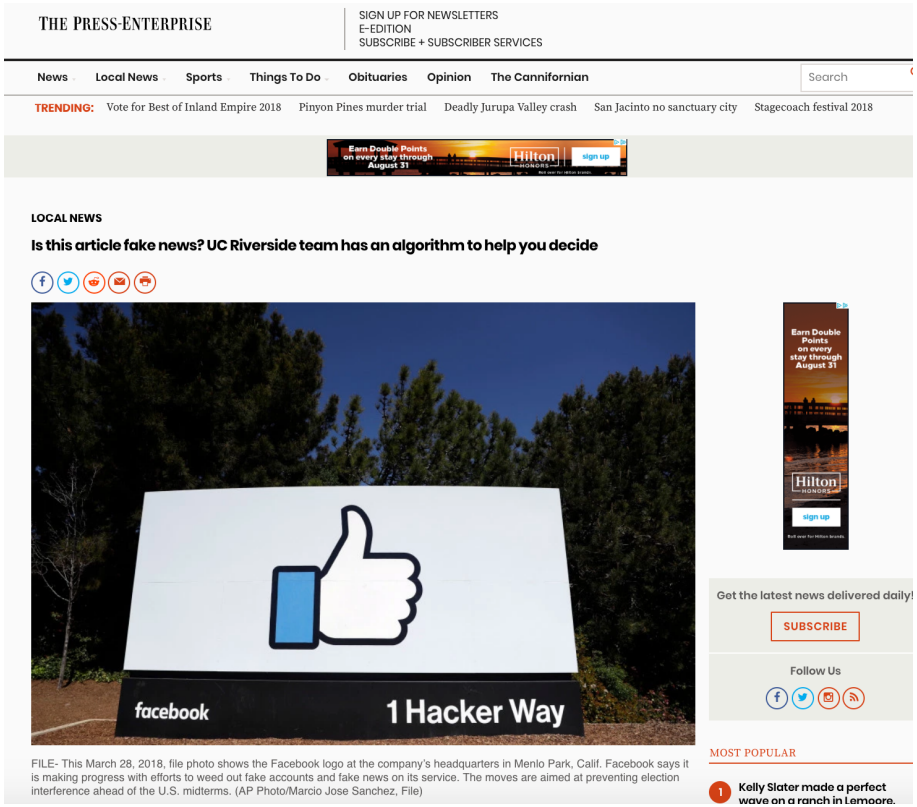




State-of-the-art accuracy with extremely few labels!



arxiv.org/pdf/1804.09088.pdf w/ Gisel Guacho, Sara Abdali, Neil Shah



THE PRESS-ENTERPRISE

SIGN UP FOR NEWSLETTERS
E-EDITION
SUBSCRIBE + SUBSCRIBER SERVICES

News Local News Sports Things To Do Obituaries Opinion The Cannifornian Search

TRENDING: Vote for Best of Inland Empire 2018 Pinyon Pines murder trial Deadly Jurupa Valley crash San Jacinto no sanctuary city Stagecoach festival 2018

LOCAL NEWS

Is this article fake news? UC Riverside team has an algorithm to help you decide

FILE- This March 28, 2018, file photo shows the Facebook logo at the company's headquarters in Menlo Park, Calif. Facebook says it is making progress with efforts to weed out fake accounts and fake news on its service. The moves are aimed at preventing election interference ahead of the U.S. midterms. (AP Photo/Marcio Jose Sanchez, File)

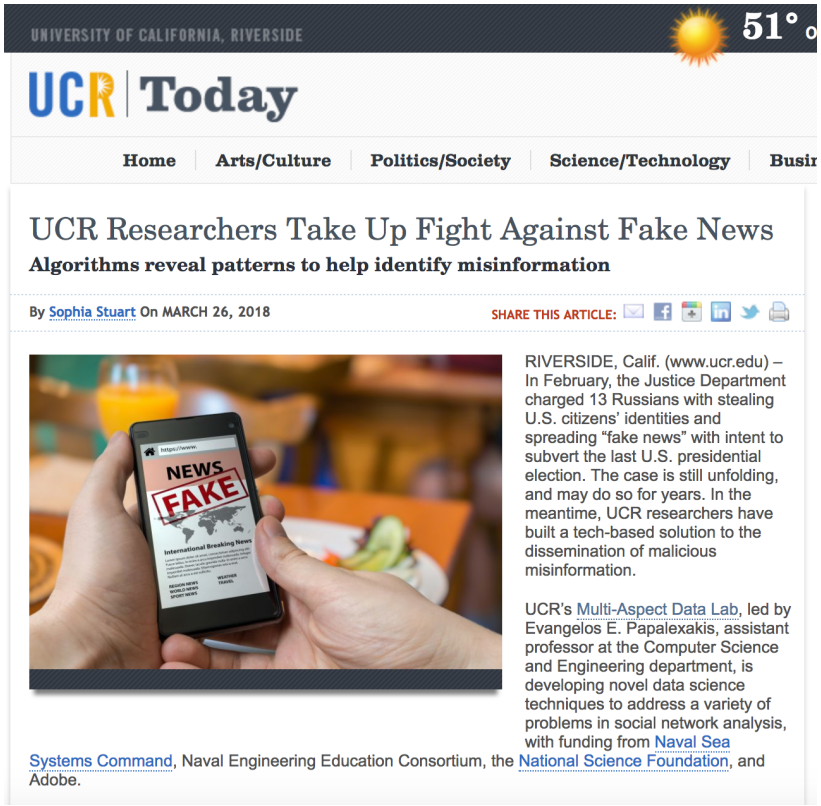
Get the latest news delivered daily!

SUBSCRIBE

Follow Us

MOST POPULAR

1 Kelly Slater made a perfect wave on a ranch in Lemoore.



UNIVERSITY OF CALIFORNIA, RIVERSIDE 51°

UCR | Today

Home Arts/Culture Politics/Society Science/Technology Busir

UCR Researchers Take Up Fight Against Fake News

Algorithms reveal patterns to help identify misinformation

By Sophia Stuart On MARCH 26, 2018 SHARE THIS ARTICLE:

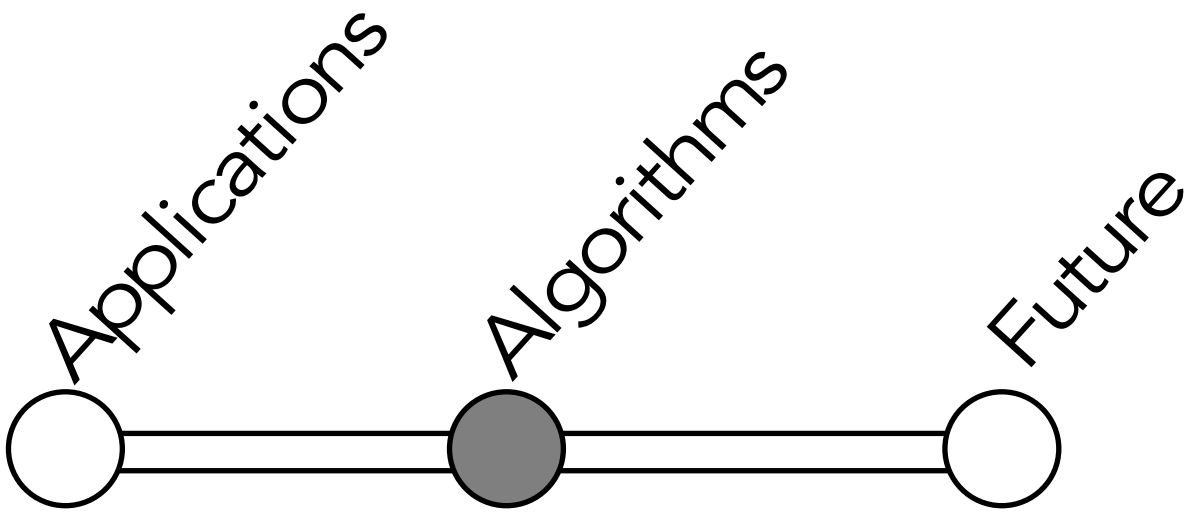
RIVERSIDE, Calif. (www.ucr.edu) – In February, the Justice Department charged 13 Russians with stealing U.S. citizens' identities and spreading "fake news" with intent to subvert the last U.S. presidential election. The case is still unfolding, and may do so for years. In the meantime, UCR researchers have built a tech-based solution to the dissemination of malicious misinformation.

UCR's Multi-Aspect Data Lab, led by Evangelos E. Papalexakis, assistant professor at the Computer Science and Engineering department, is developing novel data science techniques to address a variety of problems in social network analysis, with funding from Naval Sea Systems Command, Naval Engineering Education Consortium, the National Science Foundation, and Adobe.

https://www.pe.com/2018/04/13/is-this-article-fake-news-uc-riverside-team-has-an-algorithm-to-help-you-decide/

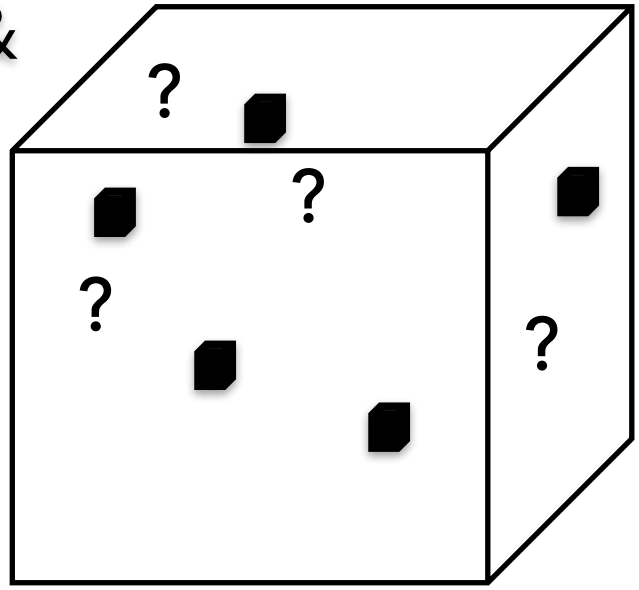
https://ucrtoday.ucr.edu/52434

Roadmap



Tensors in Data Science

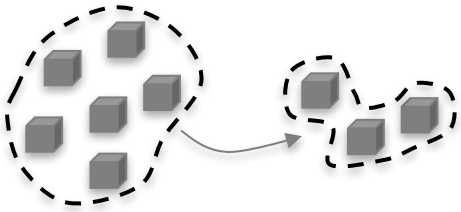
- Naturally model multi-aspect data
- Very powerful modeling tools
- **Big Challenges**
  - **C1**: Data Size & Scalability
  - **C2**: Model Selection, Quality & Interpretability



Fast and Scalable Tensor Decompositions

- Exploiting Sparsity
  - Tensor Toolbox for Matlab [Kolda et al.]
  - GigaTensor [Kang et al. 2012]
  - FlexiFaCT [Beutel et al. 2014]
  - DFacto [Choi et al. 2014]
  - SPLATT [Smith et al. 2015]
  - …
- All above methods are exact
  - Most of them focus on the "MTTKRP" operation
- Can we do something by **approximating**?

Approximate "Sketching" Methods



Sampling

Tensor CUR [Mahoney et al. 2008]

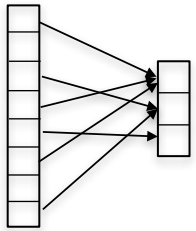
MACH [Tsourakakis 2010]

ParCube [Papalexakis et al. 2012]

Walk'n'Merge [Erdos et al 2013]

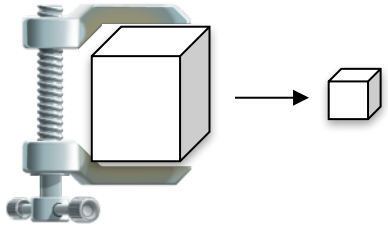
SPALS [Cheng et al 2016]

CPRAND[Battaglino et al 2017]



Hashing

[Wang et al. 2015]

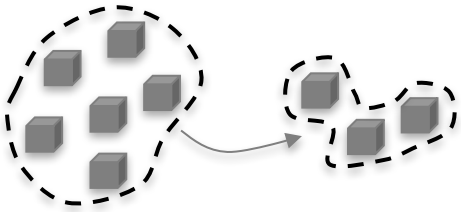


Compression

Tucker Compression
[Bro et al. 1998]

PARACOMP [Sidiropoulos et al. 2014]

Approximate "Sketching" Methods



Sampling

Tensor CUR [Mahoney et al. 2008]

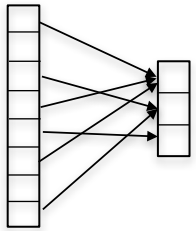
MACH [Tsourakakis 2010]

ParCube [Papalexakis et al. 2012]

Walk'n'Merge [Erdos et al 2013]

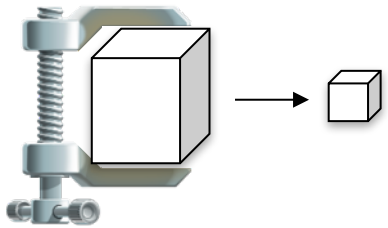
SPALS [Cheng et al 2016]

CPRAND[Battaglino et al 2017]



Hashing

[Wang et al. 2015]

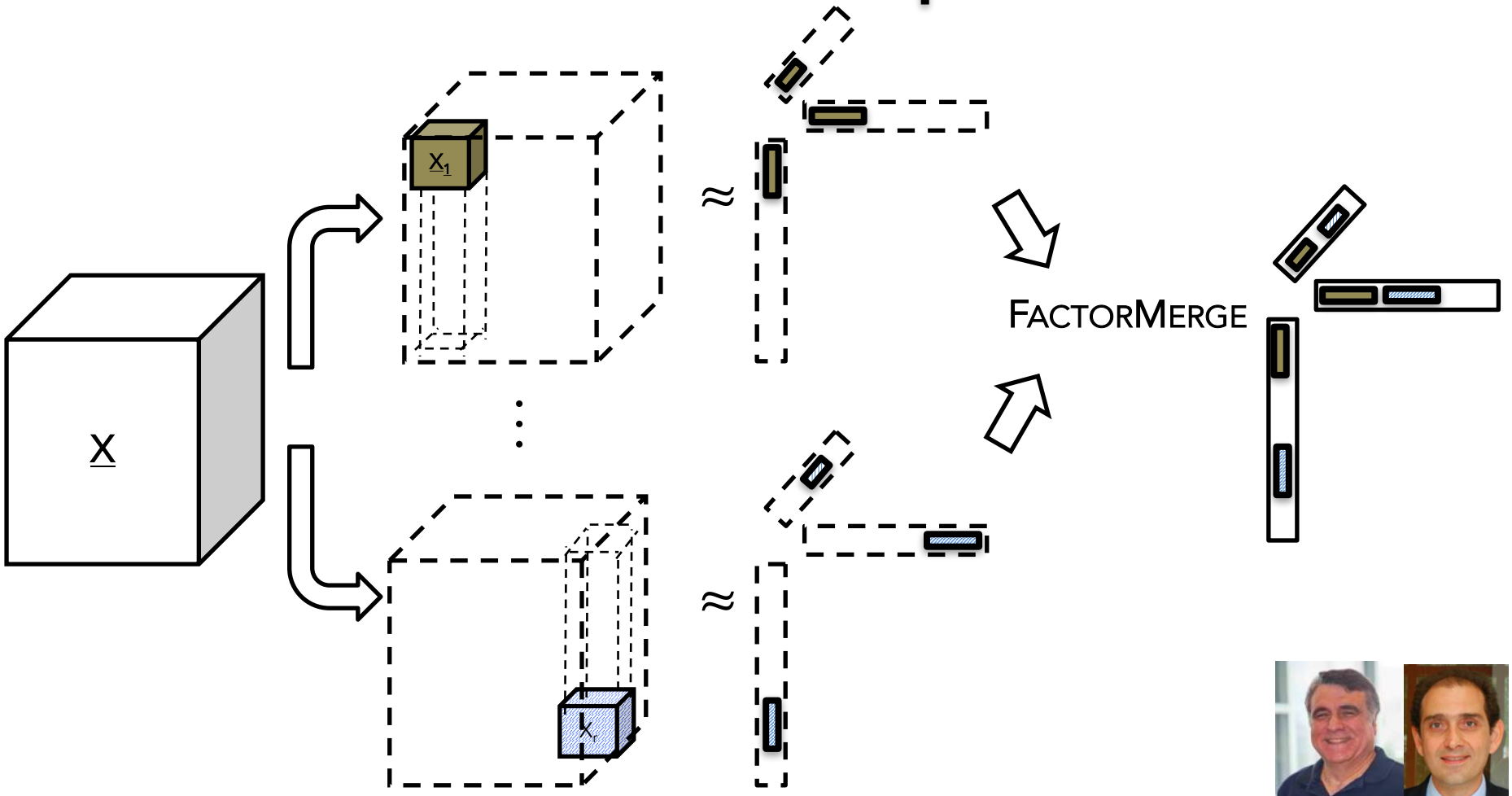


Compression

Tucker Compression
[Bro et al. 1998]

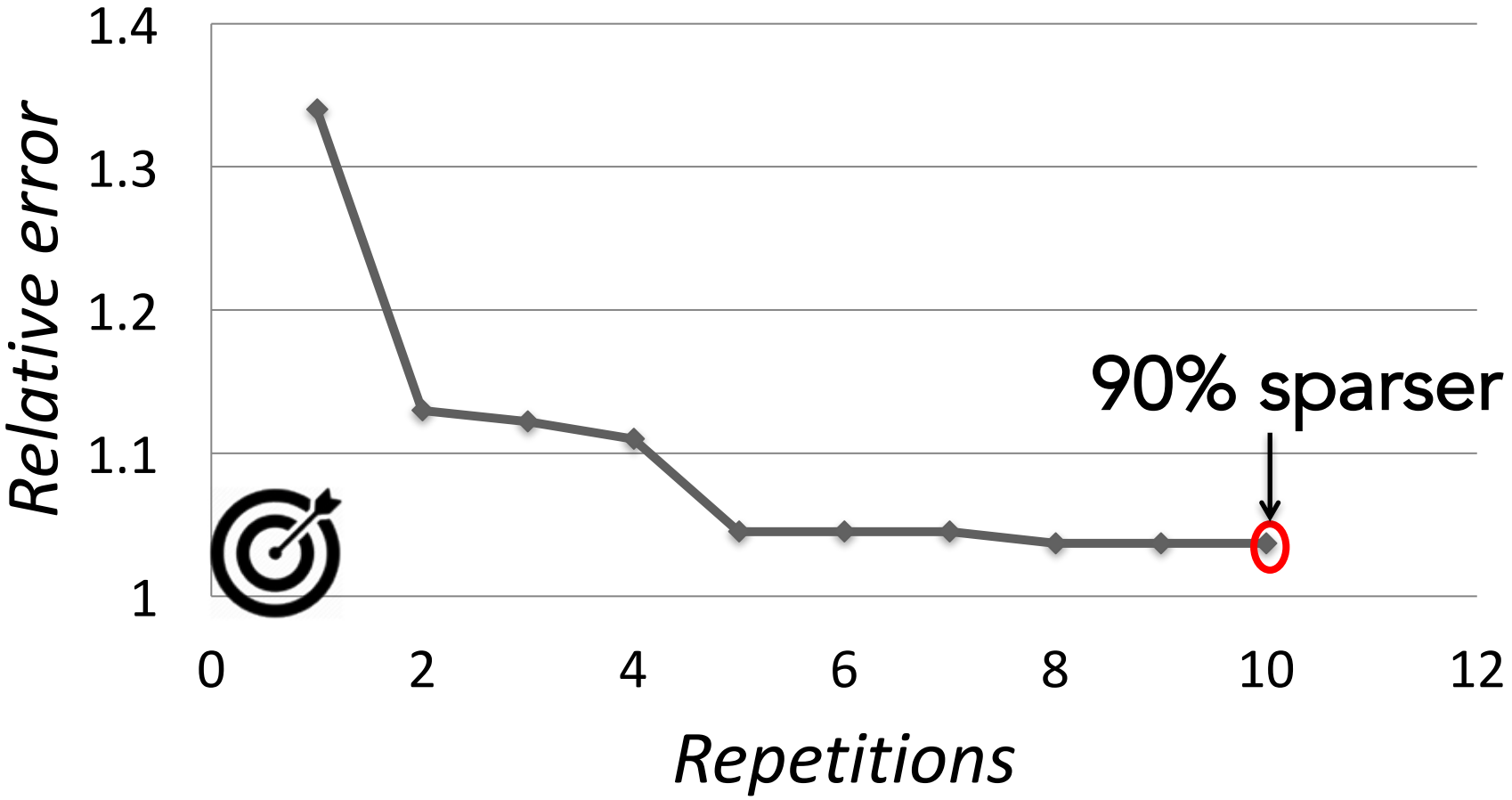
PARACOMP [Sidiropoulos et al. 2014]

ParCube: Sampling-based Parallel Tensor Decomposition



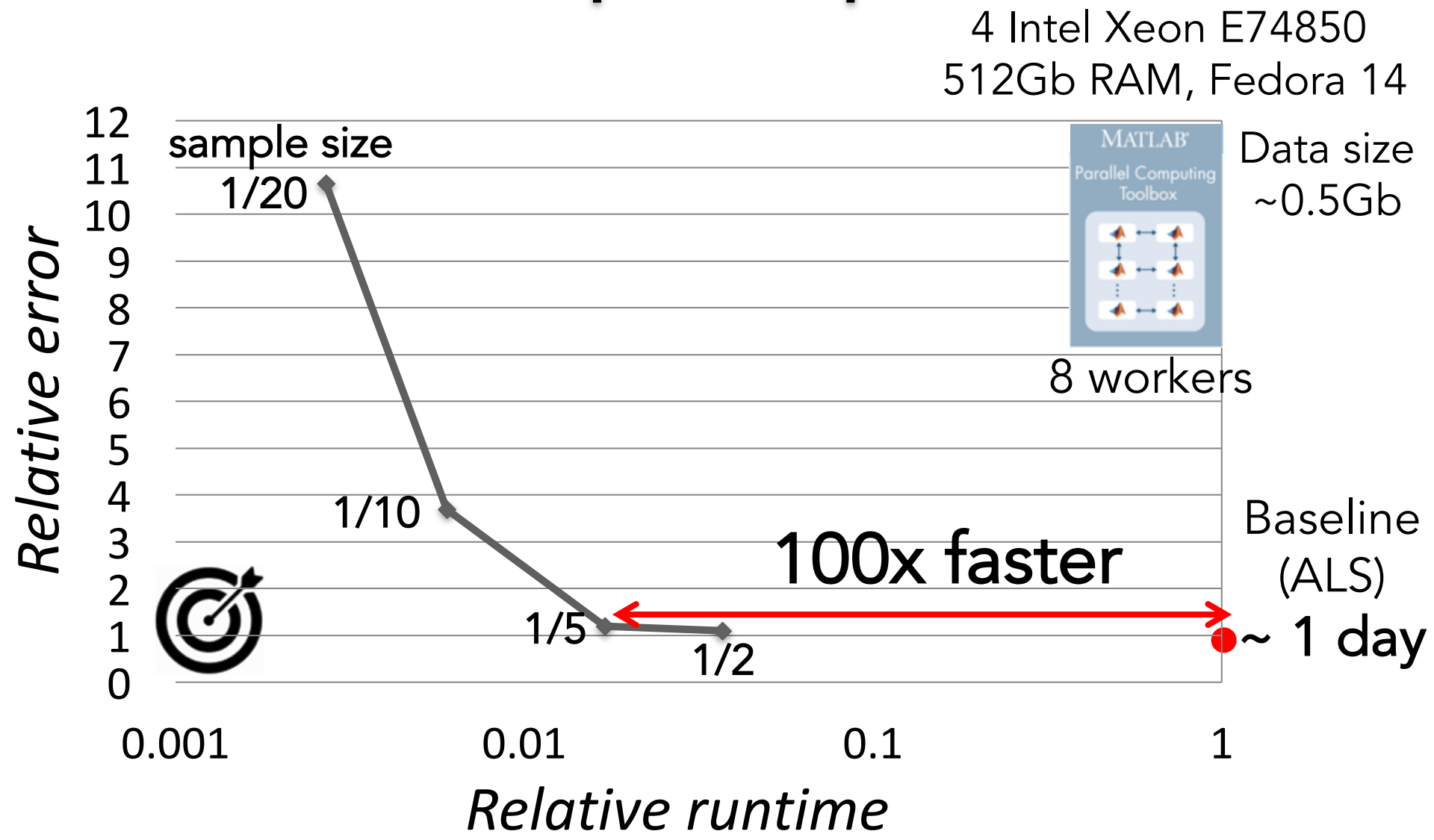
Papalexakis et al. ECML-PKDD 2012 w/ Christos Faloutsos, Nikos Sidiropoulos

Does it work?

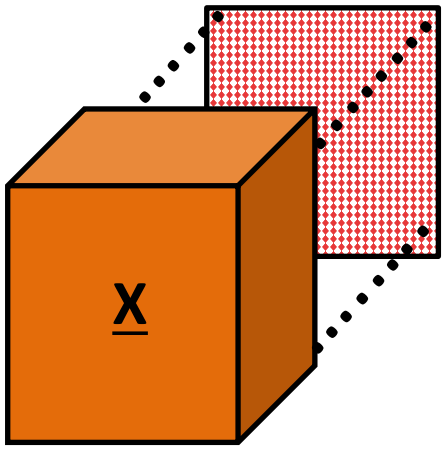


Achieves comparable accuracy to exact algorithm

Speedup


4 Intel Xeon E74850
512Gb RAM, Fedora 14
MATLAB
Parallel Computing Toolbox
8 workers
Data size
~0.5Gb
sample size
1/20
1/10
1/5
1/2
100x faster
Baseline
(ALS)
~ 1 day
Relative error
Relative runtime
0
1
2
3
4
5
6
7
8
9
10
11
12
0.001
0.01
0.1
1

Incremental Decomposition

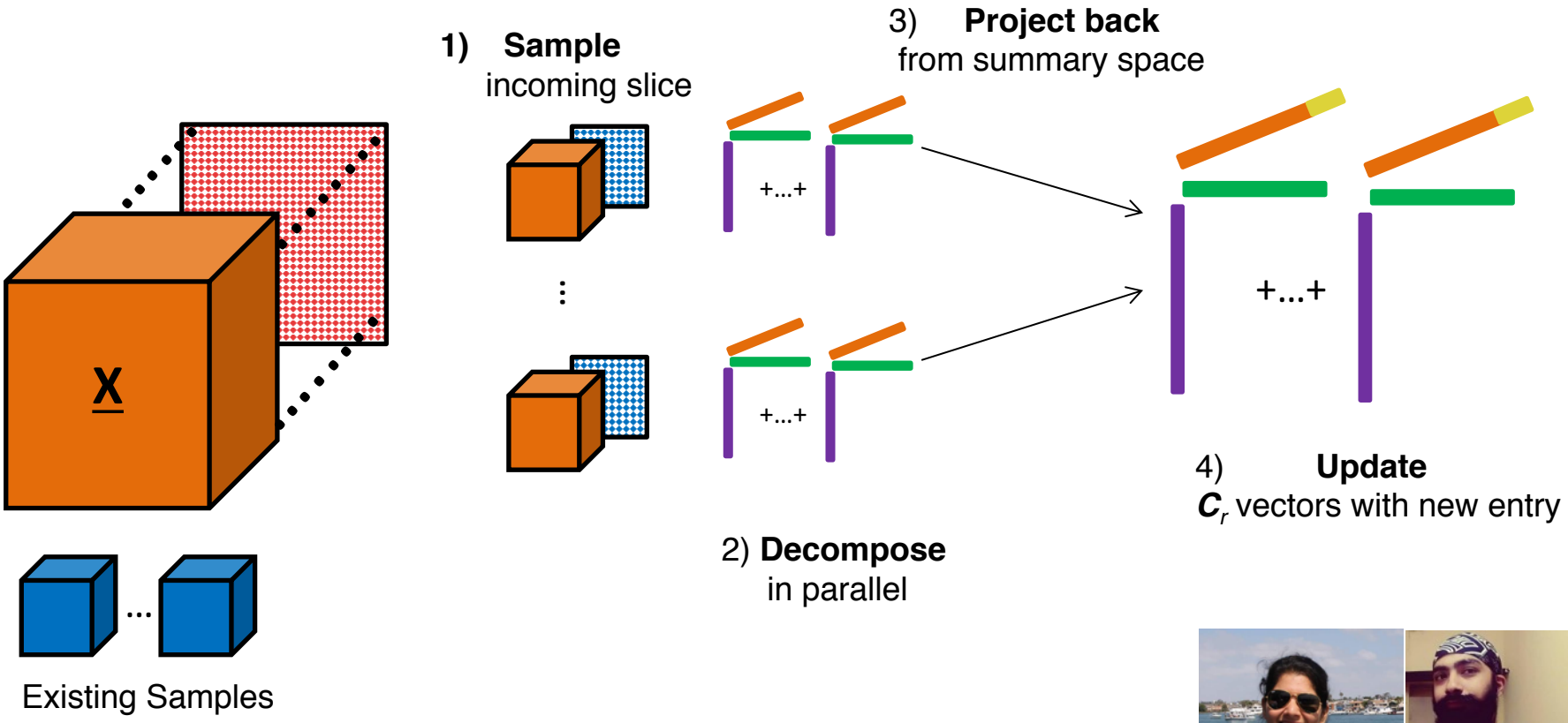


- Tensor is updated in a streaming fashion
- New slices arrive
  - New snapshots on a temporal graph
  - New article
  - …

How can we ***incrementally*** update the decomposition?

SDM 2018 w/ Ekta Gujral & Ravdeep Pasricha

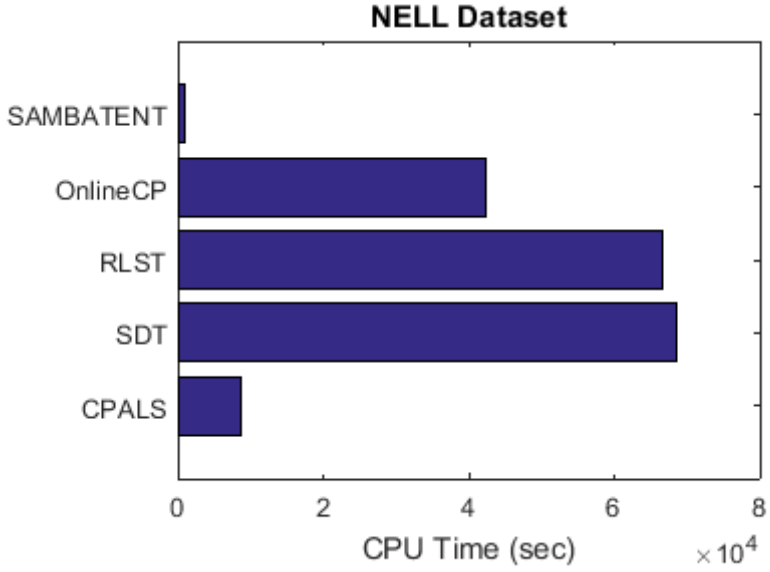
SamBaTen: Sampling-based Batch Incremental Tensor Decomposition

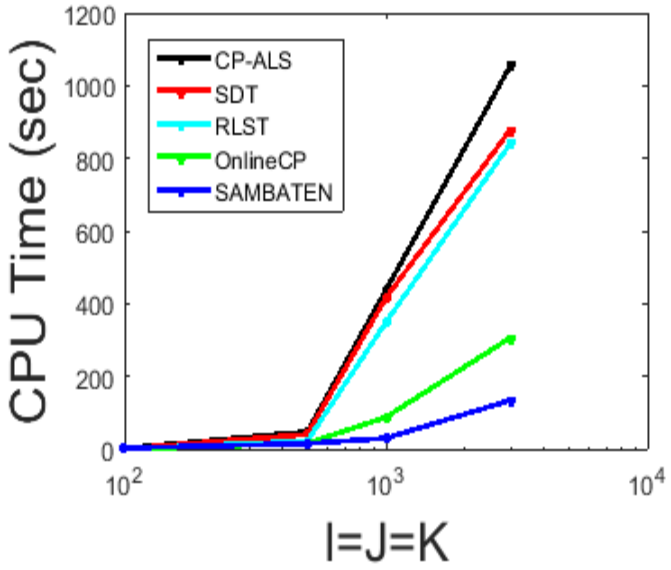


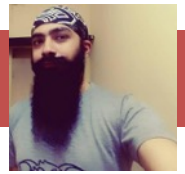
SDM 2018 w/ Ekta Gujral & Ravdeep Pasricha

SamBaTen: Sampling-based Batch Incremental Tensor Decomposition

Dataset	CPU Time (sec)					Fitness SAMBATEN w.r.t			
	CP_{ALS}	OnlineCP	SDT	RSLT	SAMBATEN	CP_{ALS}	OnlineCP	SDT	RSLT
NIPS	177.46	372.03	1608.23	1596.07	**43.98**	0.96	0.98	**0.78**	0.82
NELL	8783.27	42325.22	65325.22	63485.98	**983.83**	0.95	0.81	**0.76**	0.81
Facebook-wall	3041.98	N/A	N/A	N/A	**736.07**	0.97	N/A	N/A	N/A
Facebook-links	2689.69	N/A	N/A	N/A	**343.32**	0.96	N/A	N/A	N/A
Amazon	N/A	N/A	N/A	N/A	**4892.07**	N/A	N/A	N/A	N/A
Patent	N/A	N/A	N/A	N/A	**8068.27**	N/A	N/A	N/A	N/A



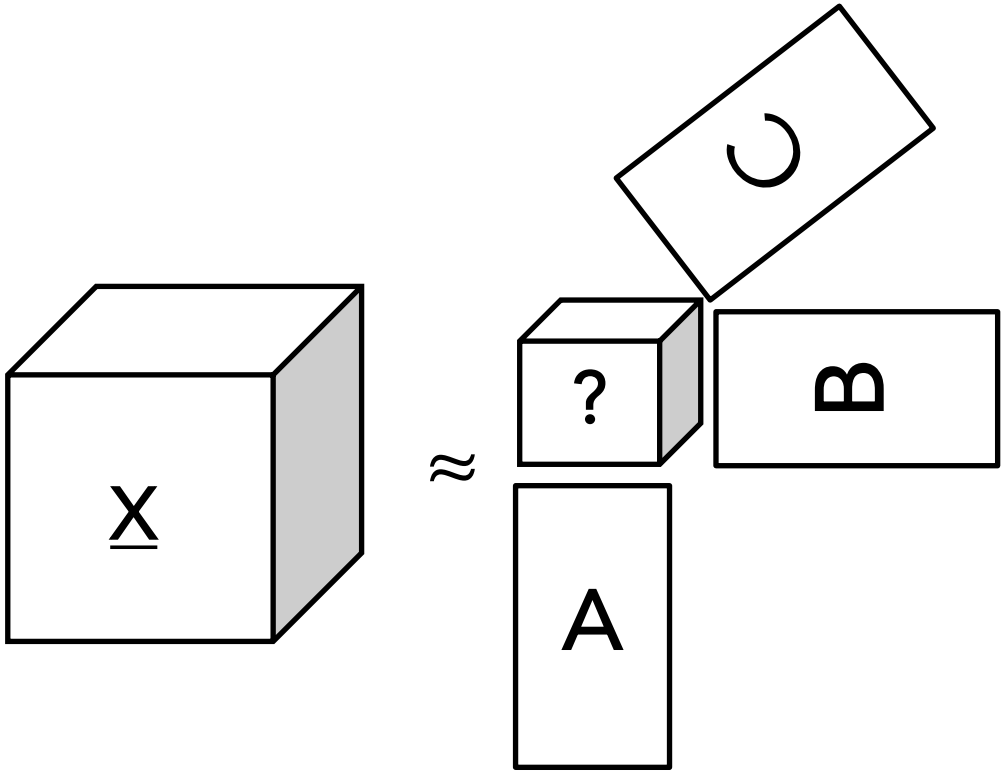




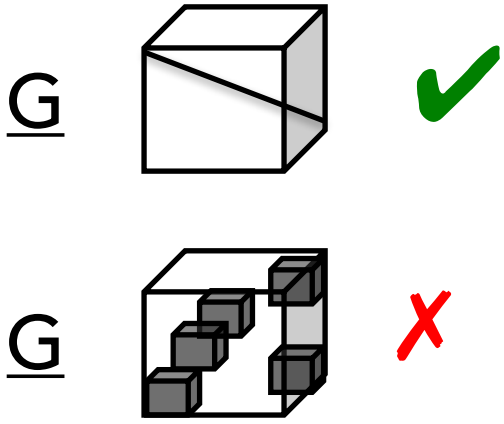
Model Selection & Quality

- **Rank Estimation**
  - Given a model (e.g. PARAFAC), choose the right number of components
  - Do this without any ground truth

Core Consistency Diagnostic 101



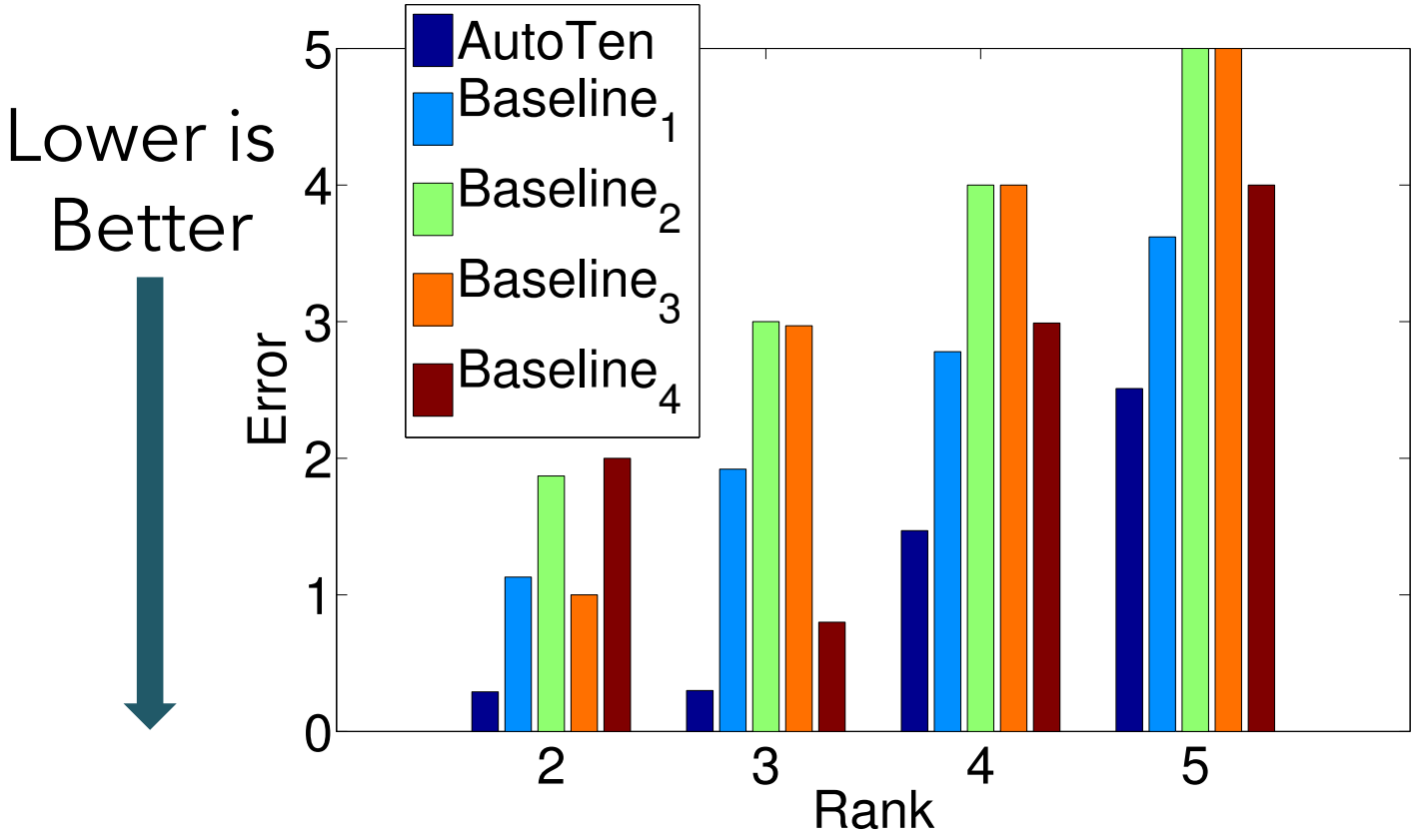
Core has important modeling quality info



$$\min_{\underline{\mathbf{G}}} \| vec\left(\underline{\mathbf{X}}\right) - \left(\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}\right) vec\left(\underline{\mathbf{G}}\right) \|_F^2$$

[Bro, Kiers Journal of Chemometrics 2003]

Rank Estimation for CP/PARAFAC

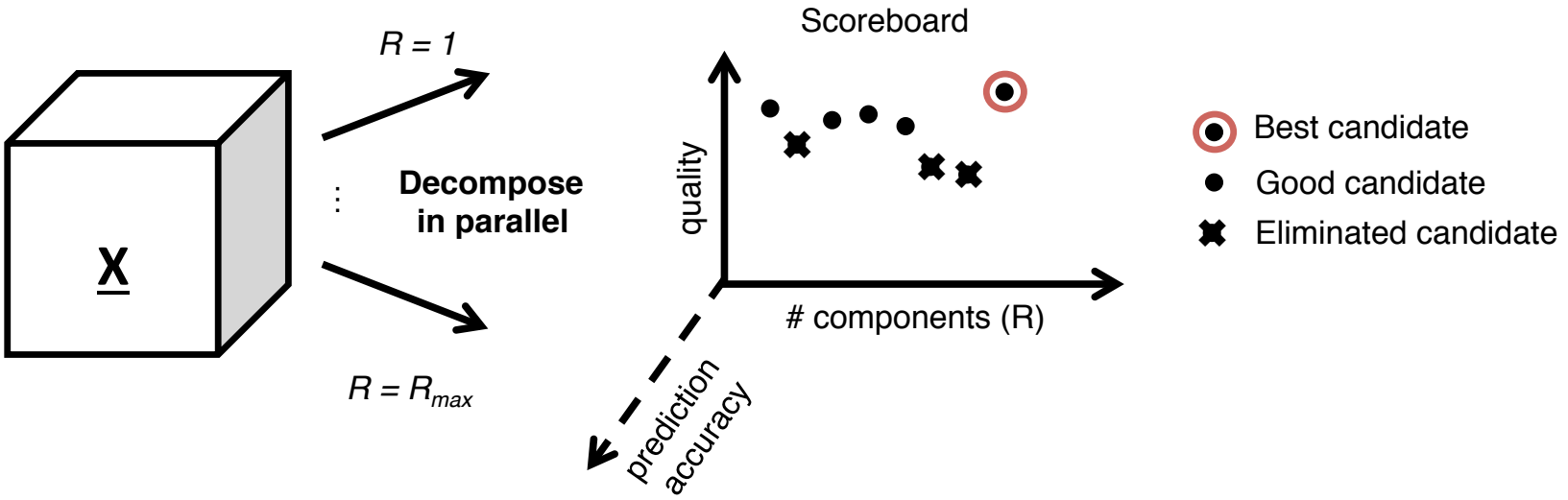


- Maximize both #components and "quality" of decomposition
- Quality is defined through Core Consistency [Bro et al. 2003]

Papalexakis SDM'16 [Best Student Paper Award]

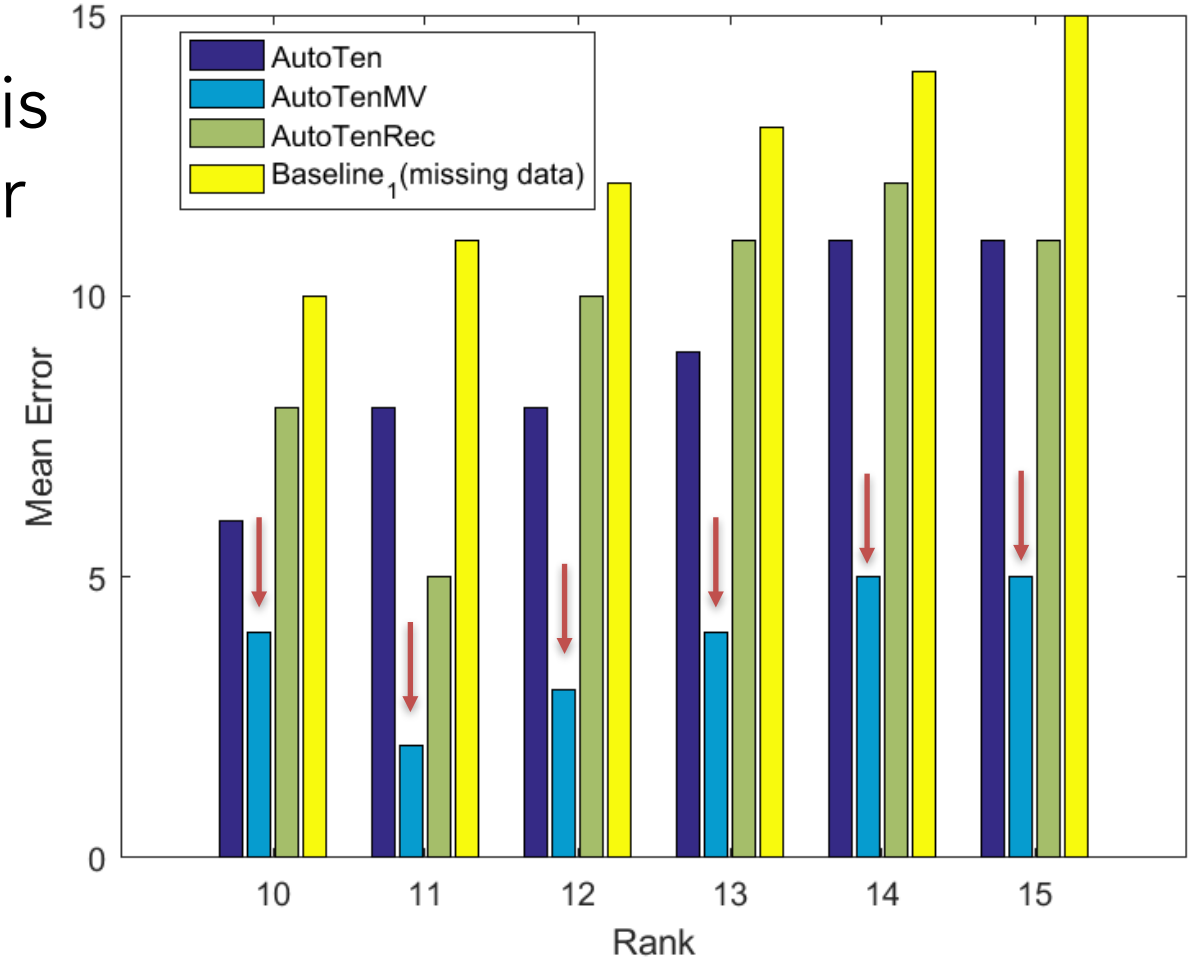
Balancing Interpretability and Predictive Quality

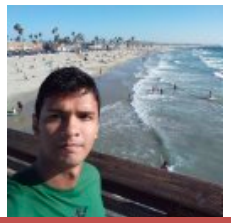
- CP/PARAFAC has been successful in Collaborative Filtering [Xiong et al 2010 SDM]
- Cross-validation on held-out has been used by the N-way Toolbox for CP/PARAFAC.
- What about scoring a balance between prediction and interpretability?



Balancing Interpretability and Predictive Quality

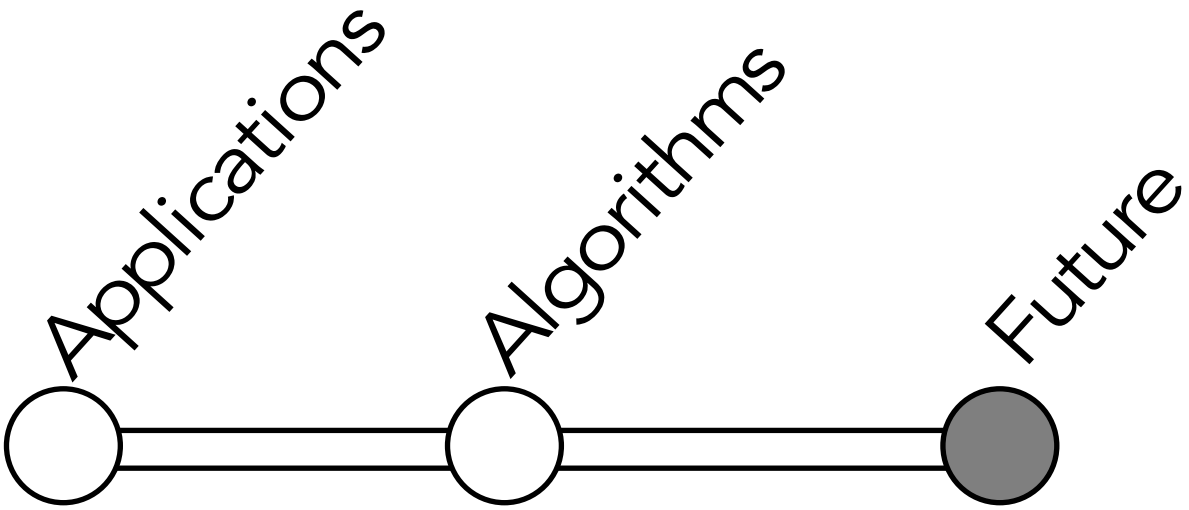
Lower is Better





Work in progress – ASILOMAR 2017 w/ Ishmam Zabir

Roadmap



Future Directions

- Fake News & User Personality Mining
- Mental State Estimation Using Online Behavior & Smartphone Usage
- Sports Analytics using Tensors
- Interplay of Tensor Methods and Deep Learning

Thank you! Questions?

- How to reach me: http://www.cs.ucr.edu/~epapalex/

I

Tensors

Supported by:

NAVSEA NAVAL SEA SYSTEMS COMMAND

NSF

Adobe

UCR Bourns College of Engineering

M A D Lab @ UCR
Multi Aspect Data