Tackling Big Data with Tensor Methods

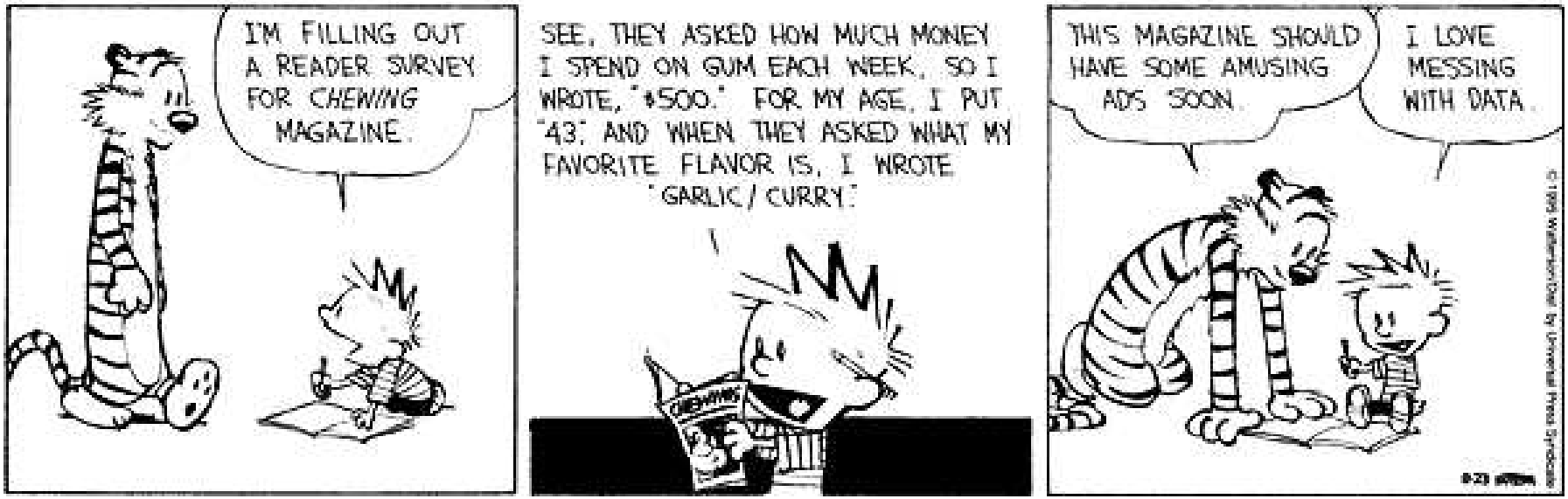
Anima Anandkumar

U.C. Irvine

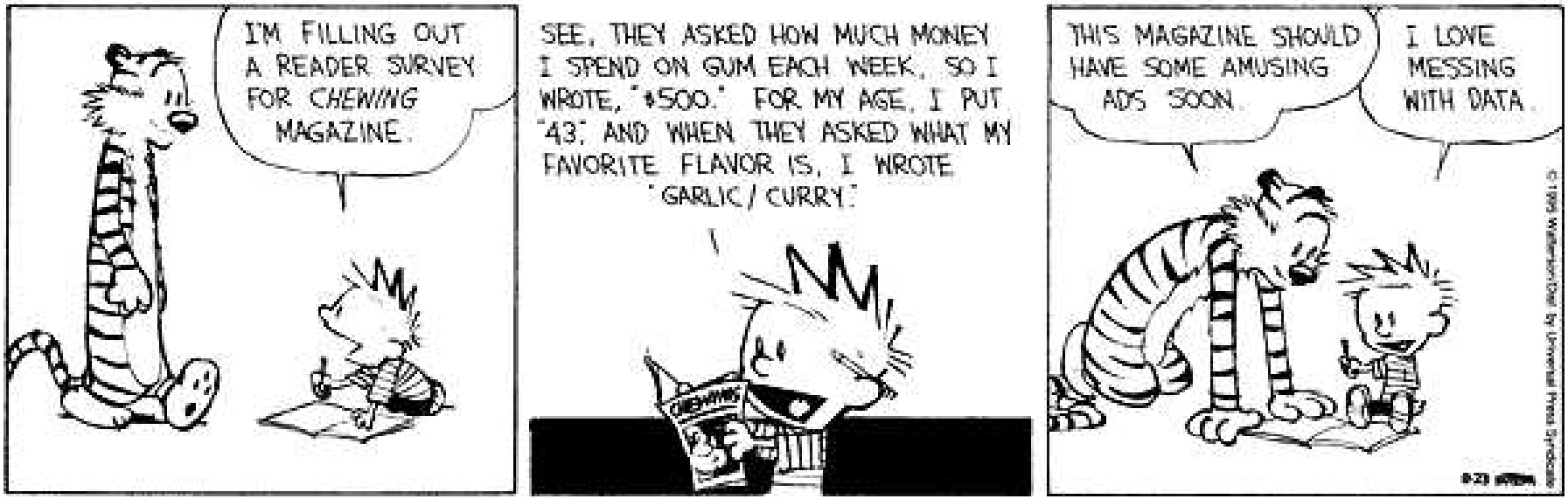
Learning with Big Data

Data vs. Information

Data vs. Information

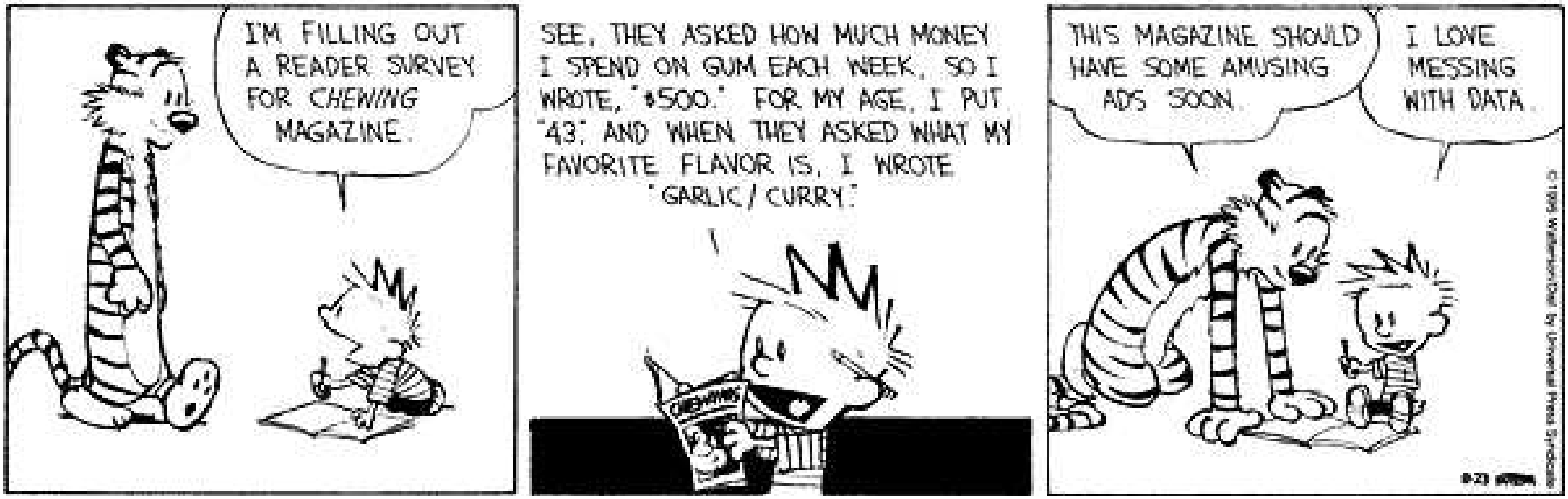


Data vs. Information



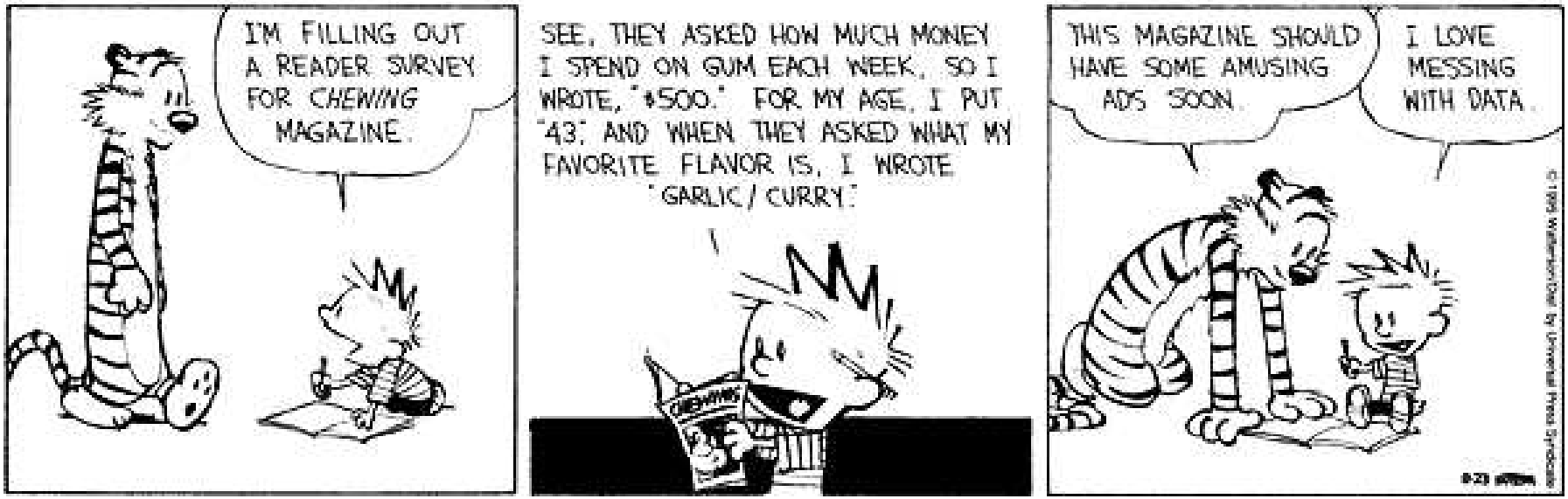
- Missing observations, gross corruptions, outliers.

Data vs. Information



- Missing observations, gross corruptions, outliers.
- High dimensional regime: as data grows, more variables !

Data vs. Information



- Missing observations, gross corruptions, outliers.
- High dimensional regime: as data grows, more variables !

Data deluge also a data desert!

Learning in High Dimensional Regime

- Useful information: low-dimensional structures.
- Learning with big data: ill-posed problem.

Learning in High Dimensional Regime

- Useful information: low-dimensional structures.
- Learning with big data: ill-posed problem.

Learning is finding needle in a haystack

Learning in High Dimensional Regime

- Useful information: low-dimensional structures.
- Learning with big data: ill-posed problem.

Learning is finding needle in a haystack

- Learning with big data: computationally challenging!

Principled approaches for finding low dimensional structures?

How to model information structures?

Latent variable models

- Incorporate hidden or latent variables.
- Information structures: Relationships between latent variables and observed data.

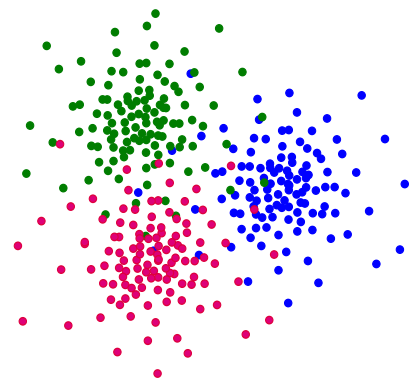
How to model information structures?

Latent variable models

- Incorporate hidden or latent variables.
- Information structures: Relationships between latent variables and observed data.

Basic Approach: mixtures/clusters

- Hidden variable is categorical.



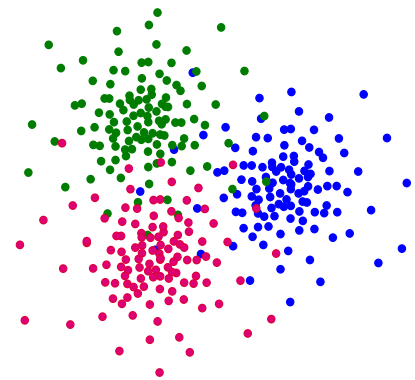
How to model information structures?

Latent variable models

- Incorporate hidden or latent variables.
- Information structures: Relationships between latent variables and observed data.

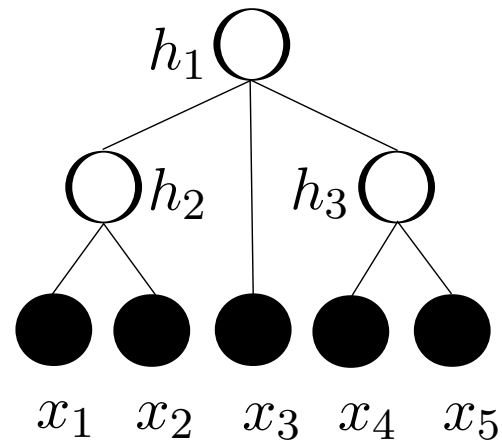
Basic Approach: mixtures/clusters

- Hidden variable is categorical.



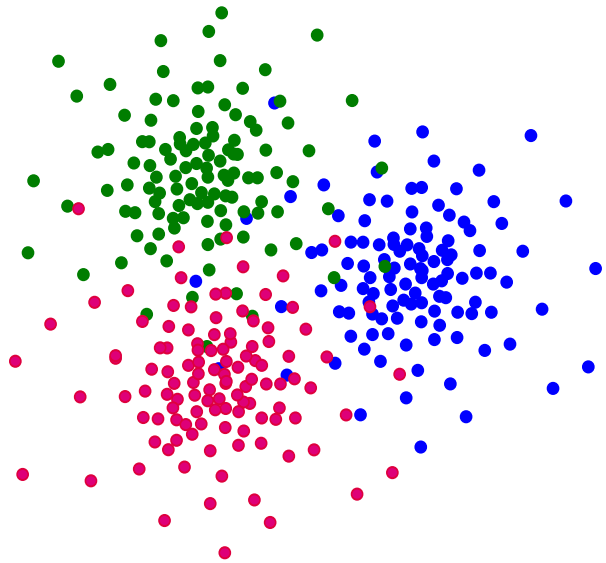
Advanced: Probabilistic models

- Hidden variables have more general distributions.
- Can model mixed membership/hierarchical groups.



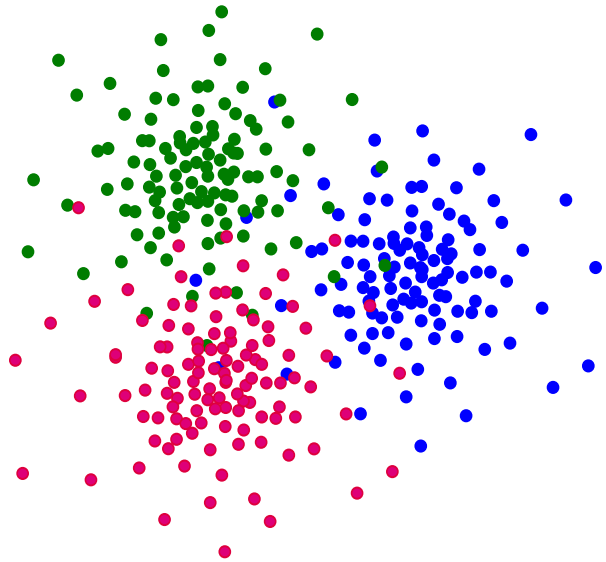
Application 1: Clustering

- Basic operation of grouping data points.
- Hypothesis: each data point belongs to an unknown group.



Application 1: Clustering

- Basic operation of grouping data points.
- Hypothesis: each data point belongs to an unknown group.



Probabilistic/latent variable viewpoint

- The groups represent different distributions. (e.g. Gaussian).
- Each data point is drawn from one of the given distributions. (e.g. Gaussian mixtures).

Application 2: Topic Modeling

Document modeling

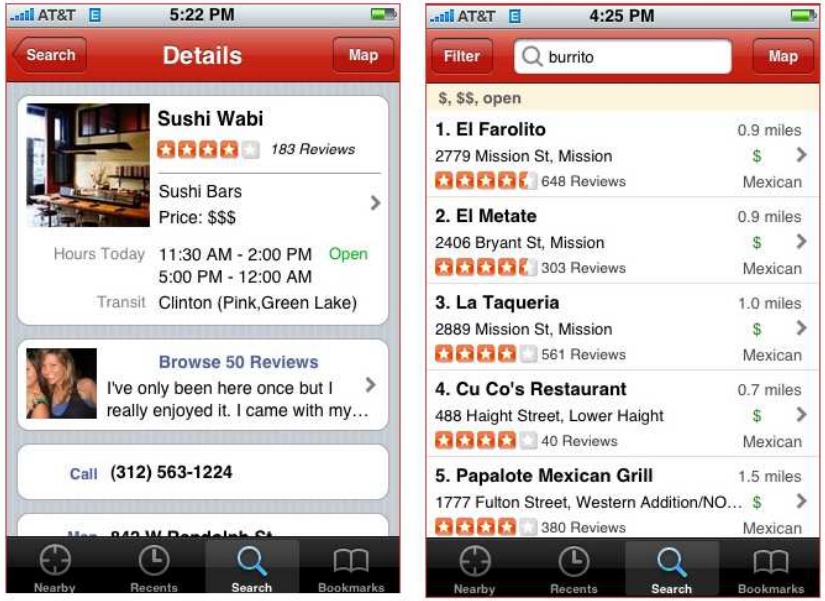
- Observed: words in document corpus.
- Hidden: topics.
- Goal: carry out document summarization.

Application 3: Understanding Human Communities

Social Networks

- Observed: network of social ties, e.g. friendships, co-authorships
- Hidden: groups/communities of actors.

Application 4: Recommender Systems



Recommender System

- Observed: Ratings of users for various products, e.g. yelp reviews.
- Goal: Predict new recommendations.
- Modeling: Find groups/communities of users and products.

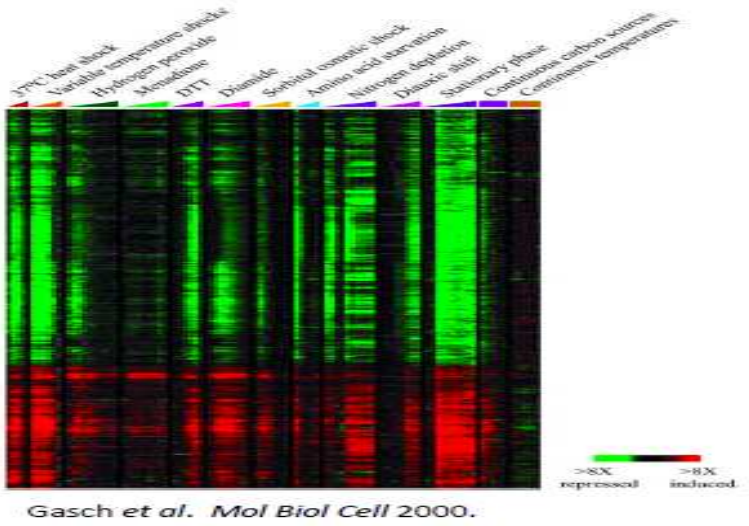
Application 5: Feature Learning

Label	Features			
0	2.1	5.2	0	0
1	0	0	2	1
1	1.1	0	0	0
0	0	0	7	0
⋮	⋮	⋮	⋮	⋮

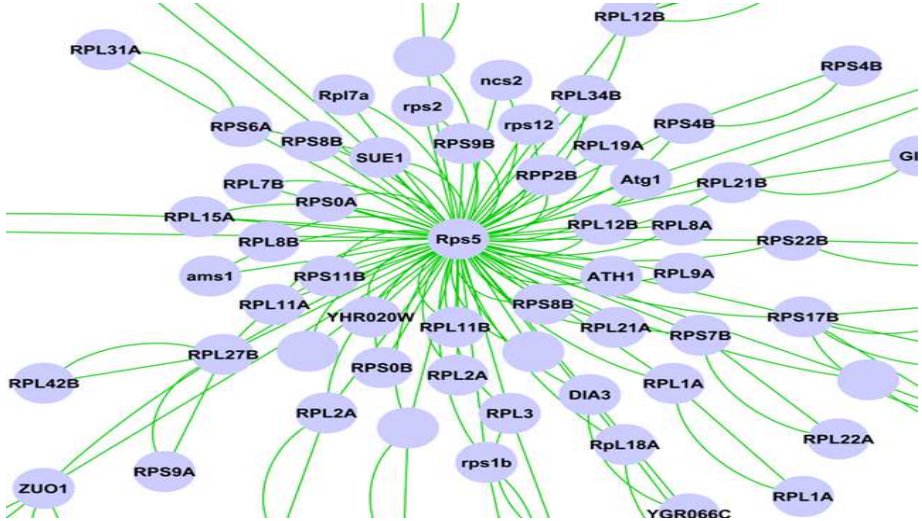
Feature Engineering

- Learn good features/representations for classification tasks, e.g. image and speech recognition.
- Sparse representations, low dimensional hidden structures.

Application 6: Computational Biology

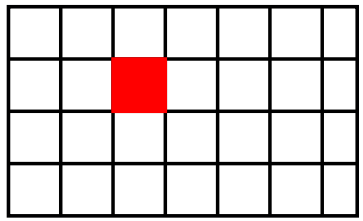


Gasch *et al. Mol Biol Cell* 2000.

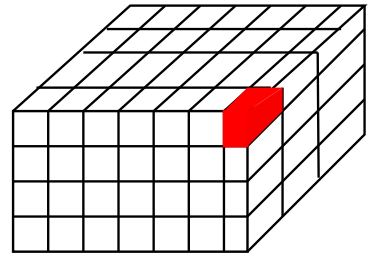


- Observed: gene expression levels
- Goal: discover gene groups
- Hidden variables: regulators controlling gene groups

Learning Algorithms through Tensor Factorization



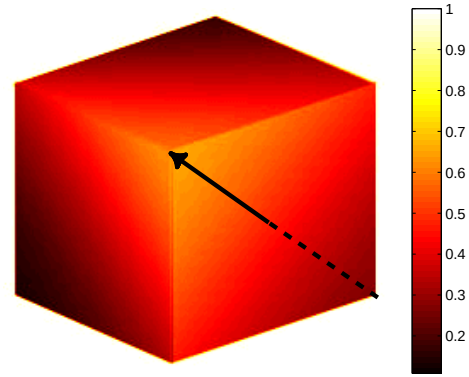
vs.



- Co-occurrence of three-words in a document, e.g. [apple, orange, banana].

Tensor Eigenvectors

- Can learn the hidden topics by finding tensor eigenvectors.
- Common friends (neighbors) of triplets of nodes in a social networks.



Experimental Results on Yelp

Lowest error business categories & largest weight businesses

Rank	Category	Business	Stars	Review Counts
1	Latin American	Salvadoreno Restaurant	4.0	36
2	Gluten Free	P.F. Chang's China Bistro	3.5	55
3	Hobby Shops	Make Meaning	4.5	14
4	Mass Media	KJZZ 91.5FM	4.0	13
5	Yoga	Sutra Midtown	4.5	31

Experimental Results on Yelp

Lowest error business categories & largest weight businesses

Rank	Category	Business	Stars	Review Counts
1	Latin American	Salvadoreno Restaurant	4.0	36
2	Gluten Free	P.F. Chang's China Bistro	3.5	55
3	Hobby Shops	Make Meaning	4.5	14
4	Mass Media	KJZZ 91.5FM	4.0	13
5	Yoga	Sutra Midtown	4.5	31

Bridgeness: Distance from vector $[1/\hat{k}, \ldots, 1/\hat{k}]^\top$

Top-5 bridging nodes (businesses)

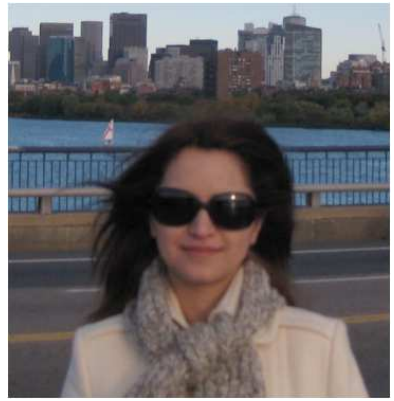
Business	Categories
Four Peaks Brewing	Restaurants, Bars, American, Nightlife, Food, Pubs, Tempe
Pizzeria Bianco	Restaurants, Pizza, Phoenix
FEZ	Restaurants, Bars, American, Nightlife, Mediterranean, Lounges, Phoenix
Matt's Big Breakfast	Restaurants, Phoenix, Breakfast& Brunch
Cornish Pasty Co	Restaurants, Bars, Nightlife, Pubs, Tempe

My Research Group and Resources

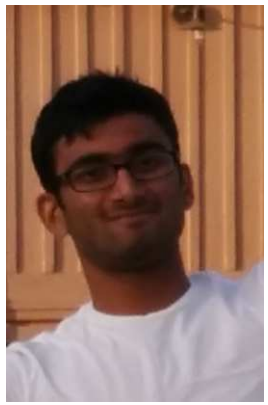
Furong Huang

Majid Janzamin

Hanie Sedghi



Niranjan UN



Forough Arabshahi

- ML summer school lectures available at `http://newport.eecs.uci.edu/anandkumar/MLSS.html`