Data Exploration in Data Science

Annie Qu

University of Illinois at Urbana-Champaign
anniequ@illinois.edu

Spring, 2018

What are Data?

- Collection of data objects and their attributes
- Data could have a variety of structures:
  - Structured data: Vector, matrix, tensor
  - Unstructured data (images, texts, videos, etc.)
- Purpose of analyzing data:
  - Model building, model validation
  - Estimation and hypothesis testing for interested model parameters
  - Find associations among interested variables
  - Make predictions for future events
  - Statistical inference

Data Examples

- A few concrete examples to illustrate the importance of data analysis
  - Traditional data: Vector data
  - Longitudinal data: HIV data
  - Survey data: Presidential election data
  - Matrix data: MovieLens data
  - Tensor data: IRi marketing data
  - Network data: ADHD data (Attention Deficit Hyperactivity Disorder)
  - Text data: TripAdvisor data, Chinese corpus data
  - Imaging data: Lung cancer, ADNI data (Alzheimer Disease Neuroimaging Initiative), ABIDE (Autism Disorder)
  - Other data

Most Common Data Structure: Vector

- Each row is a sample subject: measurements of multiple variables
- Each column is a variable

ID	Refund	Marital Status	Taxable Income	$\cdots$
1	Yes	Single	125K	$\cdots$
2	No	Married	100K	$\cdots$
3	No	Single	70K	$\cdots$
4	Yes	Married	120K	$\cdots$
5	No	Divorced	95K	$\cdots$
$\vdots$				

Longitudinal Data (Panel Data)

- **Multiple measurements** from the same subject (usually over time)
  - **Longitudinal Data** (biomedical): e.g., clinical trial
  - **Panel Data** (economics): e.g., company's revenues, profits over years
  - **Time series** is a special case of panel data (one dimension)

A Simple Example: Balanced and Unbalanced Data

- Balanced data: same number of observations for all individuals
- Unbalanced data: different number of observations for different individuals

Balanced Data

person	year	income	age	sex
1	2001	1300	27	1
1	2002	1600	28	1
1	2003	2000	29	1
2	2001	2000	38	2
2	2002	2300	39	2
2	2003	2400	40	2

Unbalanced Data

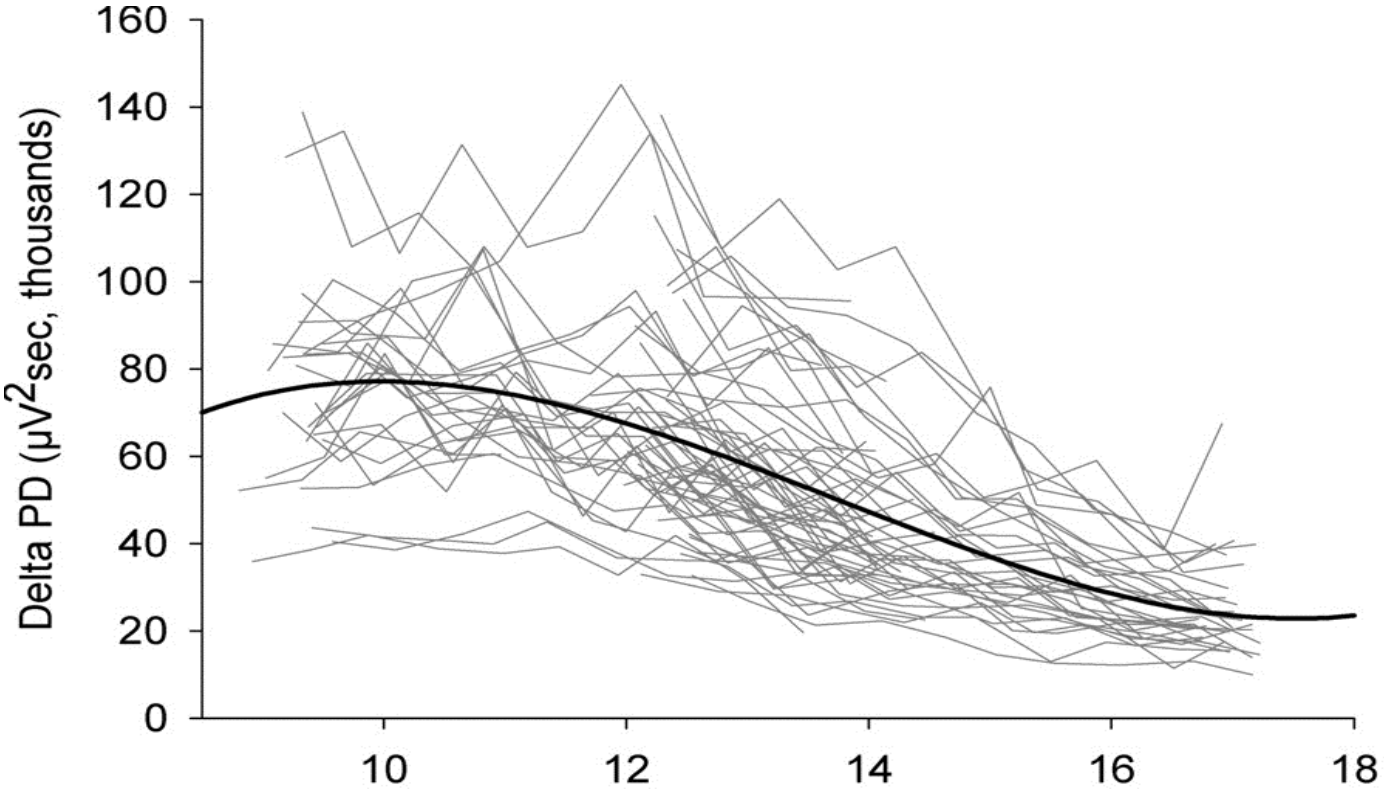
person	year	income	age	sex
1	2001	1600	23	1
1	2002	1500	24	1
2	2001	1900	41	2
2	2002	2000	42	2
2	2003	2100	43	2
3	2002	3300	34	1

Approach 1: Trajectory Analysis

- Treat each subject as a **trajectory**
  - Measurements of one variable over time
  - Goal: explore the pattern of trajectory
- Challenges
  - Parametric model usually could not capture the structure
  - missing data

Adolescent Decline in Non-rapid Eye Movement (NREM)

- Longitudinal data of 59 adolescents
- Estimate the average trajectory (cubic spline for following estimation)

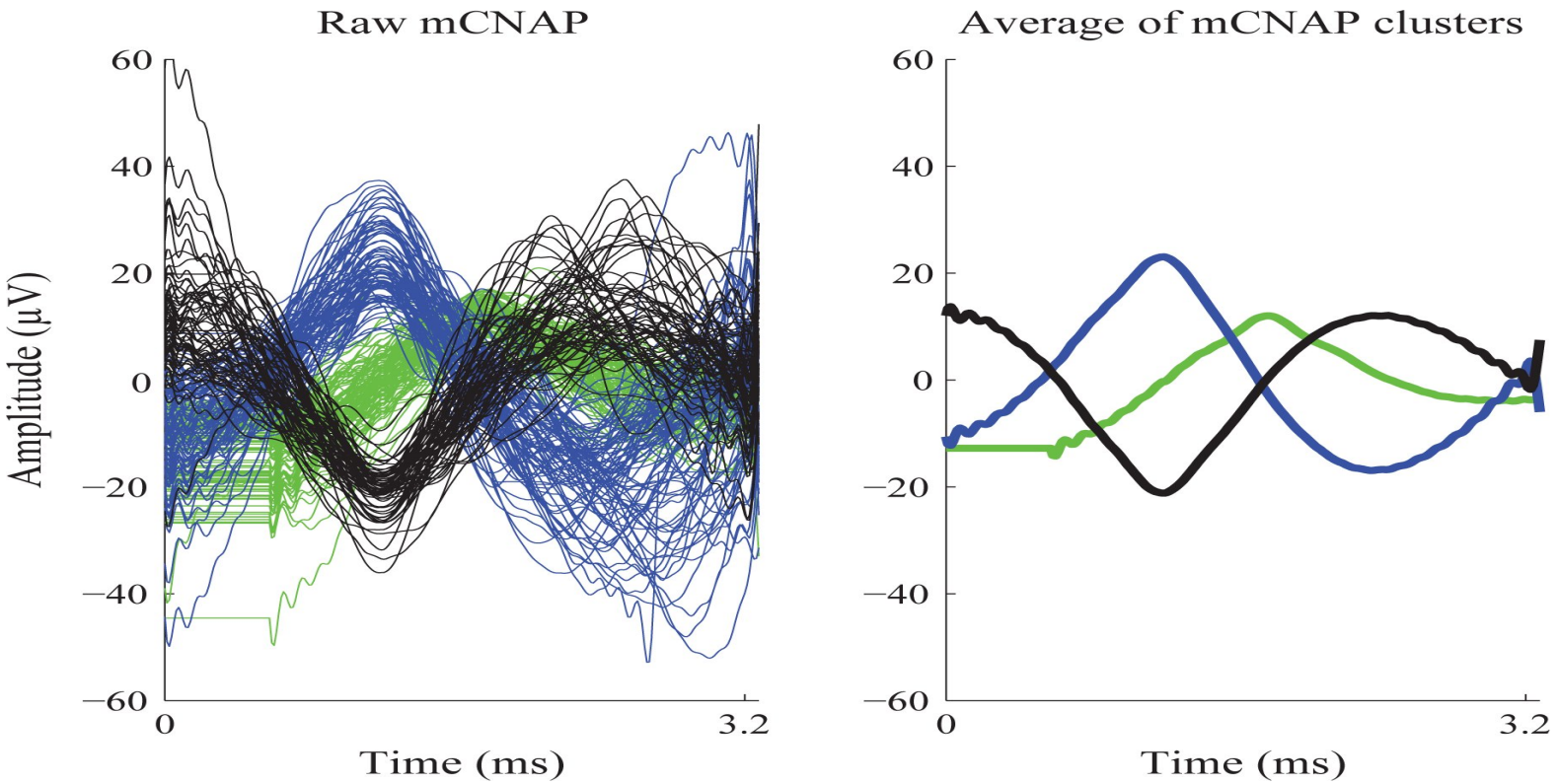


Miniature Compound Nerve Action Potentials (mCNAP)

- To investigate muscle nerve signals responsible for hand grasping
  - Longitudinal neural signals: wavelet trajectory
  - Different waveform features correspond to different phases of grasping
  - Wave features: Miniature compound nerve action potentials (mCNAP)

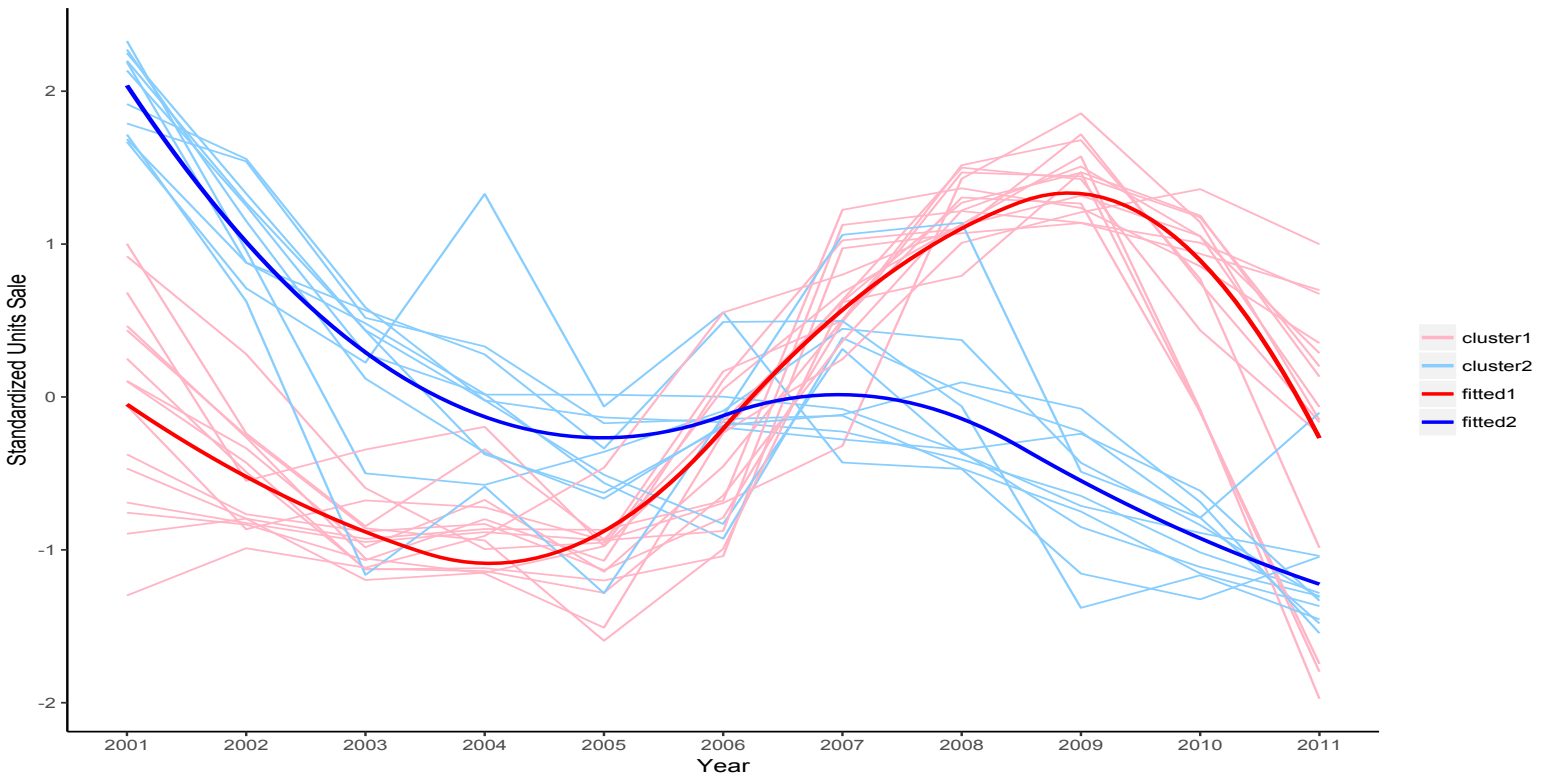
Clustering of Trajectory

- Subgroup longitudinal signals based on their trajectories' patterns



Clustering of Trajectory

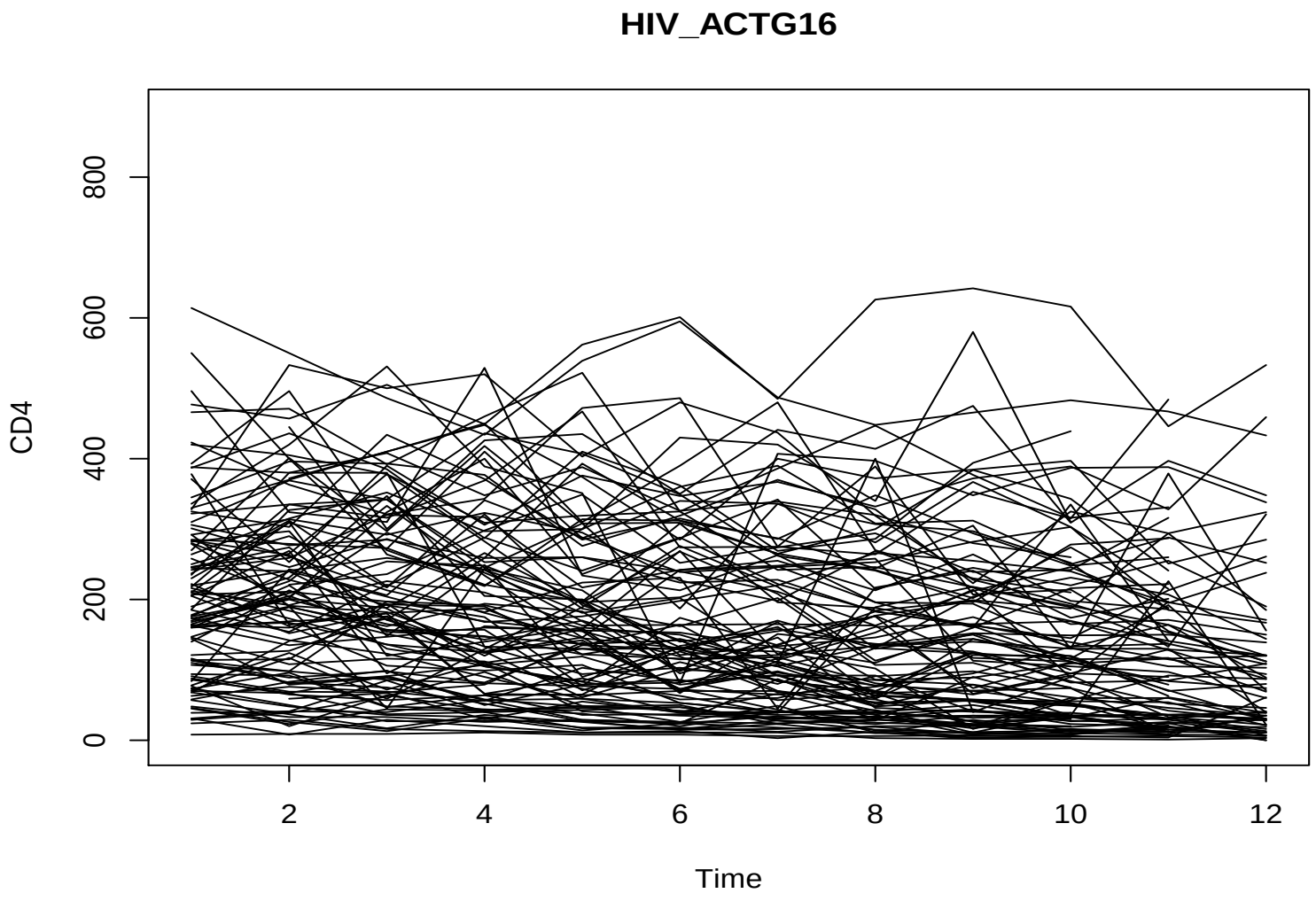
- Recent paper by Xiaolu Zhu and Qu (2018): Cluster analysis of longitudinal profiles with subgroups for grocery products
- Cluster 1: beer/ale/alcoholic, cider, coffee, cold cereal, frozen dinners/entrees, frozen pizza ...
- Cluster 2: blades, cigarettes, deodorant, diapers, facial tissue, paper towels, photography supplies ...



Approach 2: Vector Data Analysis

- Treat as point observation (the same as the vector data)
  - Explore association between the response and the predictors
  - **Correlation** exists within the same individual

Longitudinal Data: HIV Data



HIV Data (ACTG Study 16)

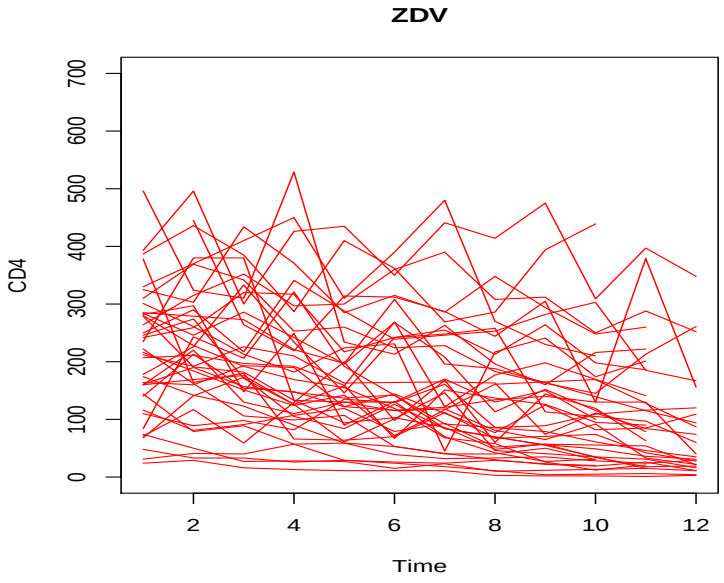
- Harvard AIDS clinical trial group: longitudinal data study
- 830 patients with repeated measurements over 14 time points
- **Unbalanced data**: the number of observations from each patient varies from 1 to 14
- https://actgnetwork.org/clinical-trials/access-published-data

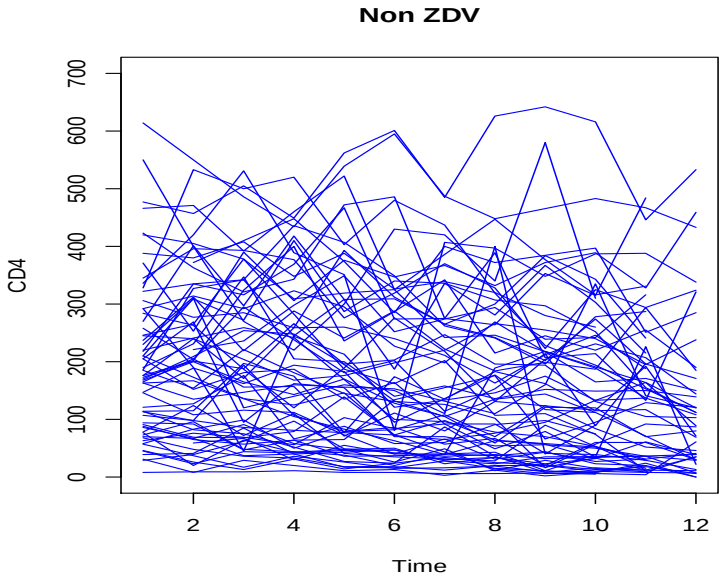
Variables in ACTG Study 16

- **Response variable**: CD4 counts (the larger, the better)
- **Explanatory variable**:
  - Demographic information: e.g., Age, Gender, Race
  - Medical measurement: e.g., Blood pressure
  - Medical treatment (drug): ZDV=1: treatment group, ZDV=0: control group
  - Time, interactions of time and other variables

Goal1: Average Treatment Effect

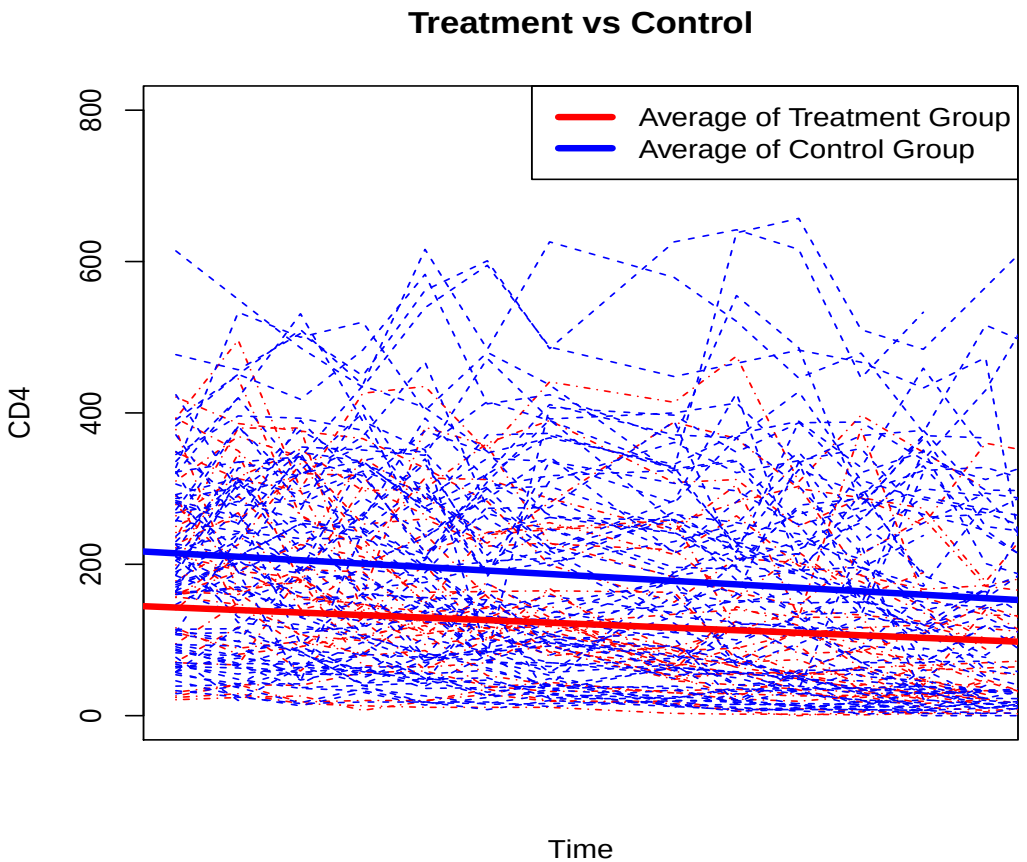
- Does the treatment have effect on patients' CD4 counts in average?





Goal1: Average Treatment Effect

- Treatment effect over time

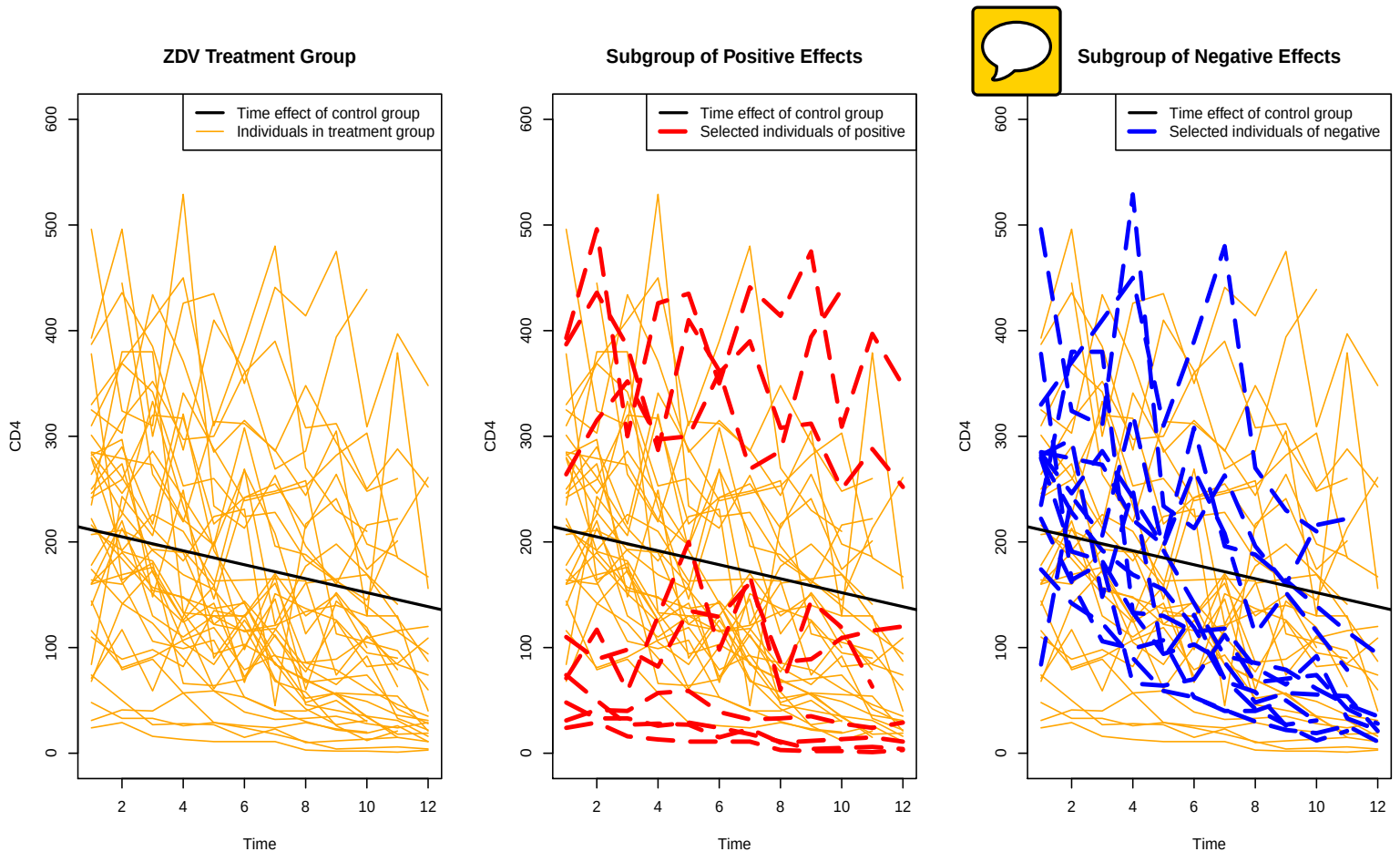


Goal of Study 1:Subgroup Treatment Effect

- Different individuals have different responses to the treatment

  1. Positive effect: decrease more slowly (not decrease) than control group
  2. Negative effect: decrease more rapidly than control group
  3. No effect: the same as the control group

Subgrouping

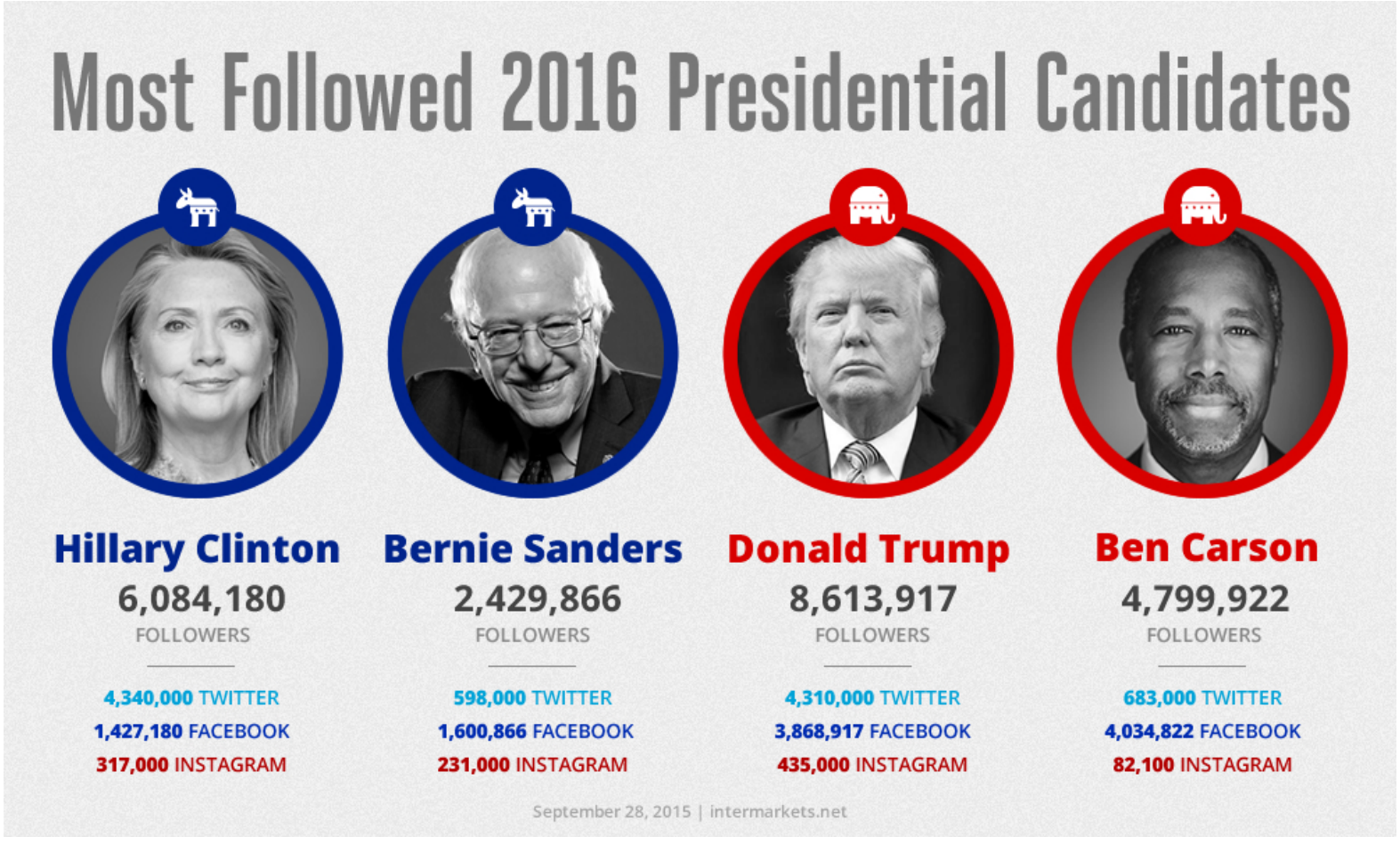


Other Applicable Methods for Subgrouping

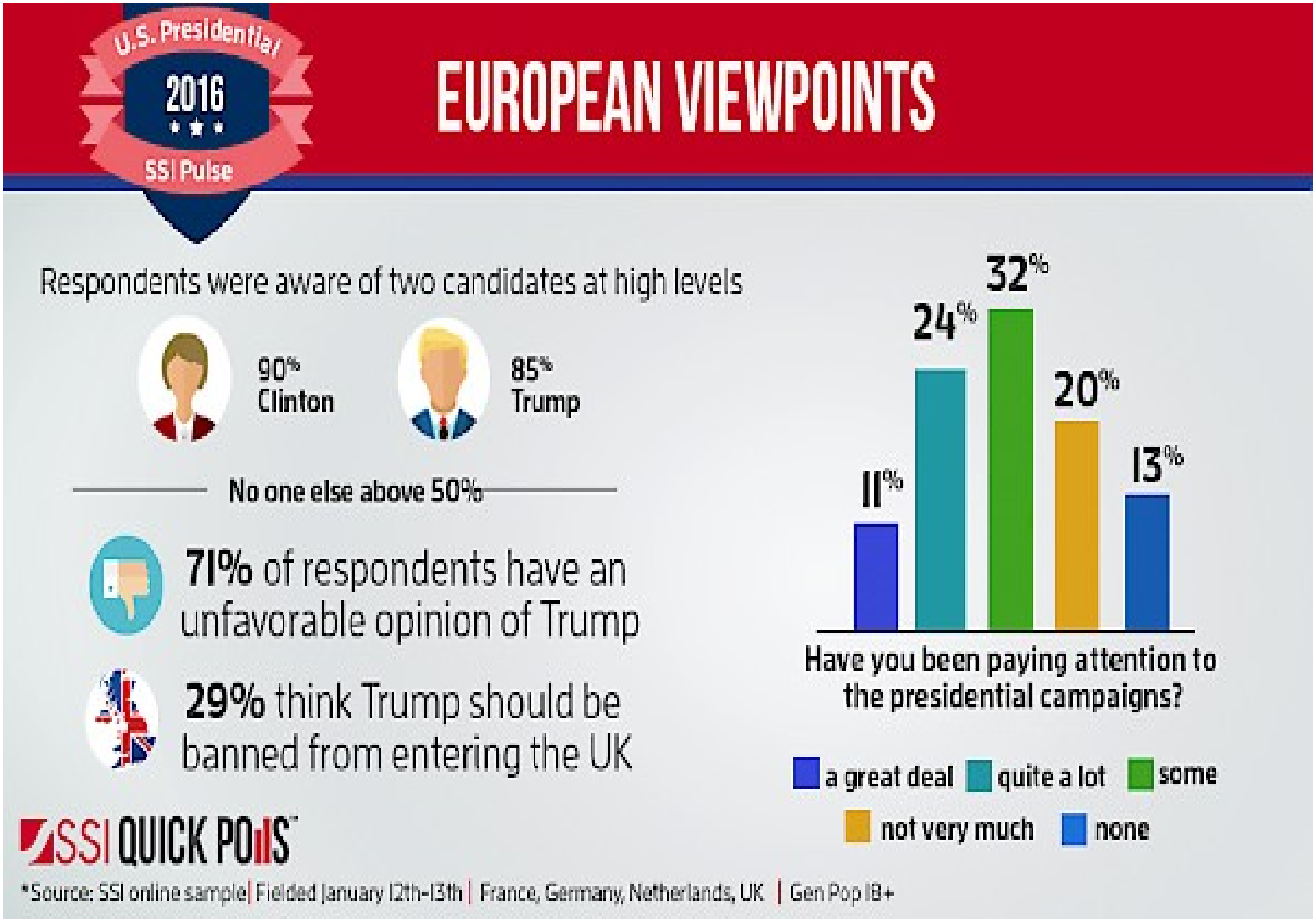
- Classification and clustering
- Model assessment and model selection
- Tree methods and random forests
- Mixture modeling

Survey Data: Presidential Election

- 2016 US Presidential Election data
  https://www.kaggle.com/benhamner/2016-us-election
- Why is the prediction of 2016 presidential election totally skewed?



Survey Company's Busiest Season!



Survey Data

- Demographic Information: e.g., Age, Gender, Race, Income, Education
- Opinion Question: **categorical variable**
  1. Ordinal variable: "How conservative do you think Trump is?"
  2. Nominal Variable: "Which candidate do you support?"
- Text opinion from blogs: unstructured data

Survey Data

- Goal
  - Summarize the opinions: Bar chart, Pie graph, percentages
  - Explore the association between the predictors and the opinion
- Challenges
  - High missing rate, informative missing
  - Response not reliable: ordinal ranking is subjective

2008 Presidential Election Data

- 2008 election survey (2007-2008, Associated Press-Yahoo! News Poll) (http://www.knowledgenetworks.com/GANP/election2008/index.html)
- 1200 survey participants were measured over 9 waves
- Selected survey question: "*How much interest do you have in election campaign news?*"
  - Ordinal responses: "interest" level was measured in 1 to 5 (strong - weak)
  - Repeated measurements of each participant at 9 time waves
  - Predictors include Time, Gender, Race, Income, Education, etc.

Ordinal Response

- Participants appear to have two subgroups regarding their interests in election (Tang and Qu, 2016).

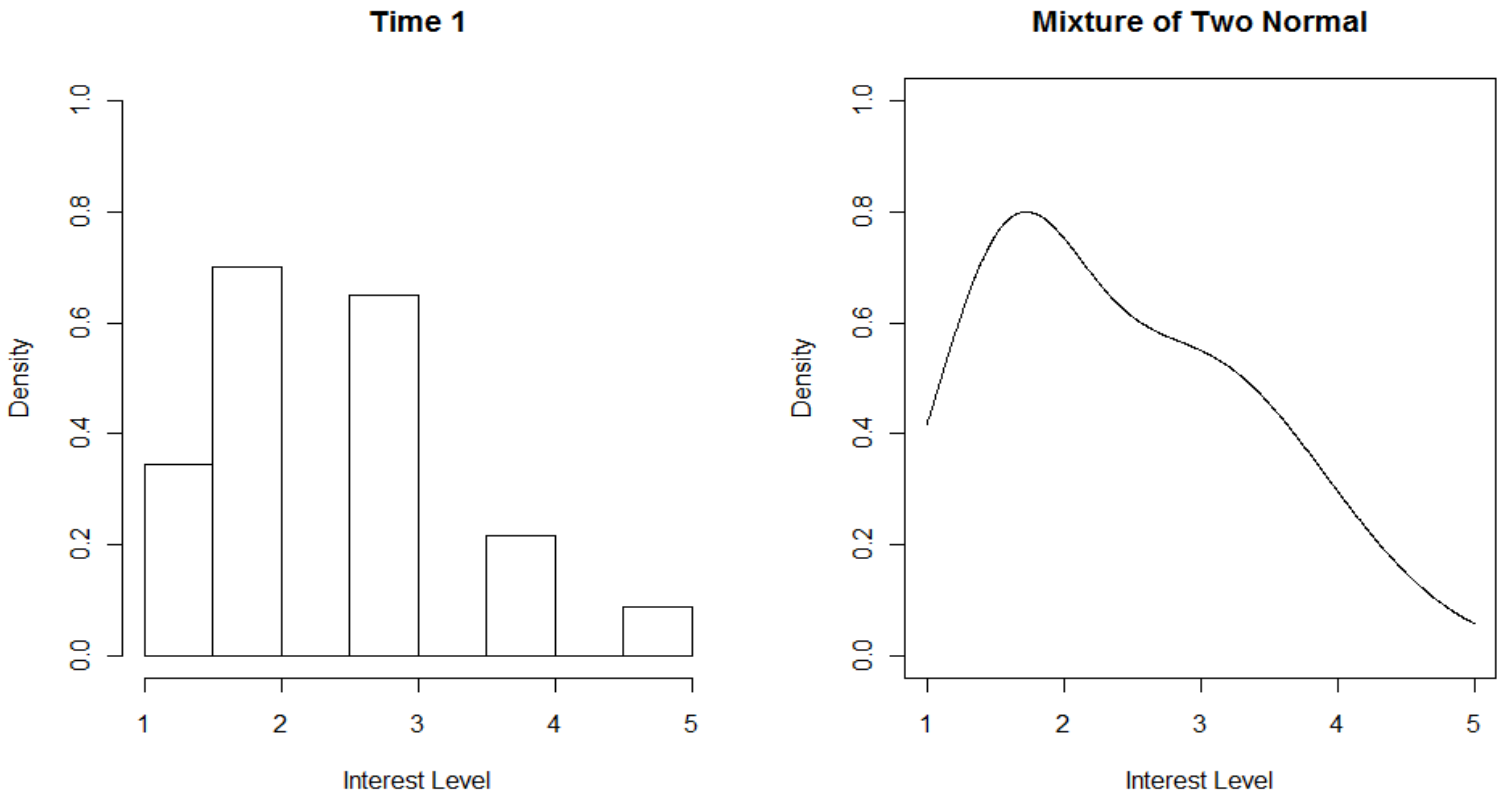
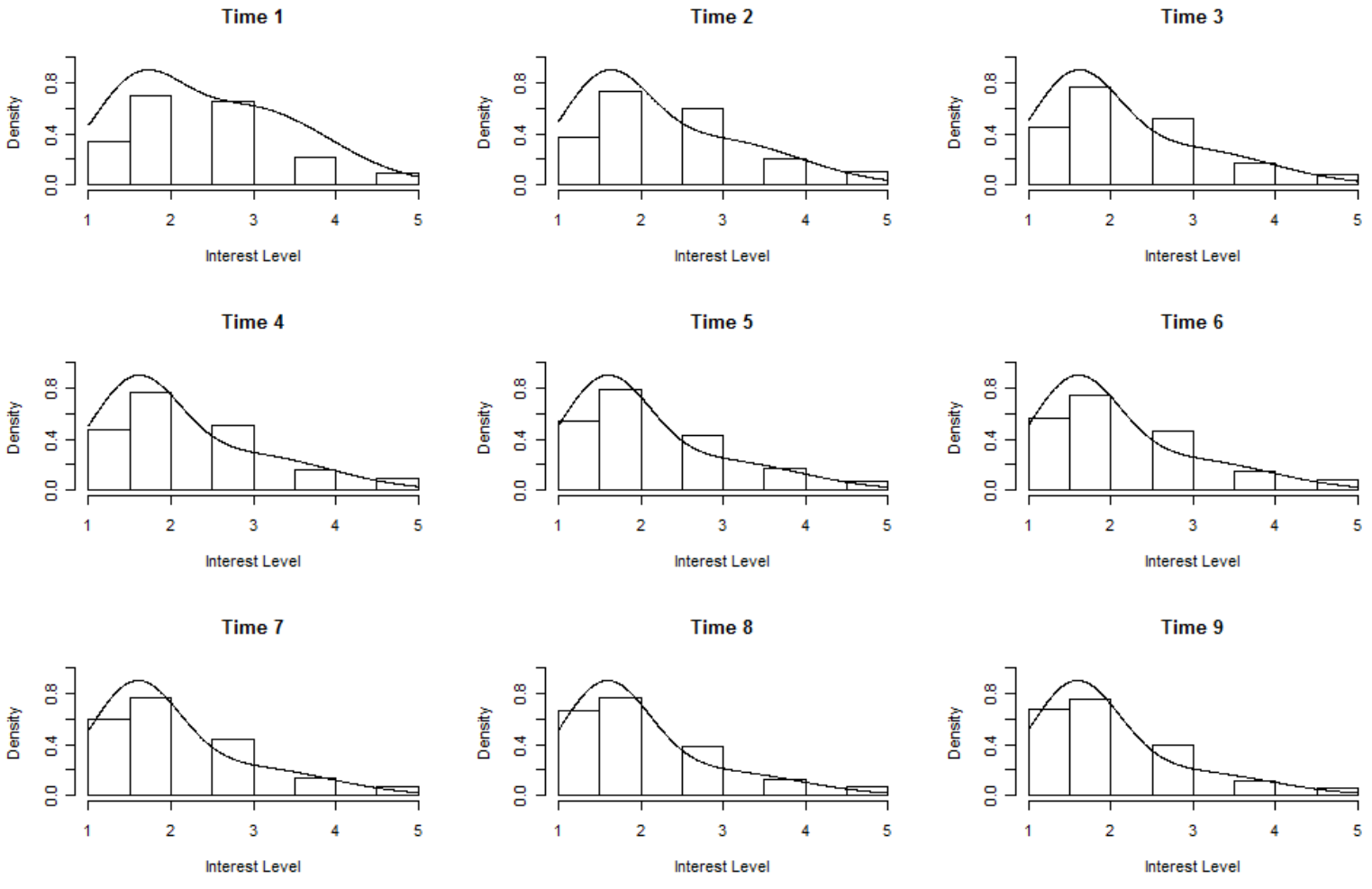


Figure: Histogram of responses at time 1 and an approximate mixture of normal density.

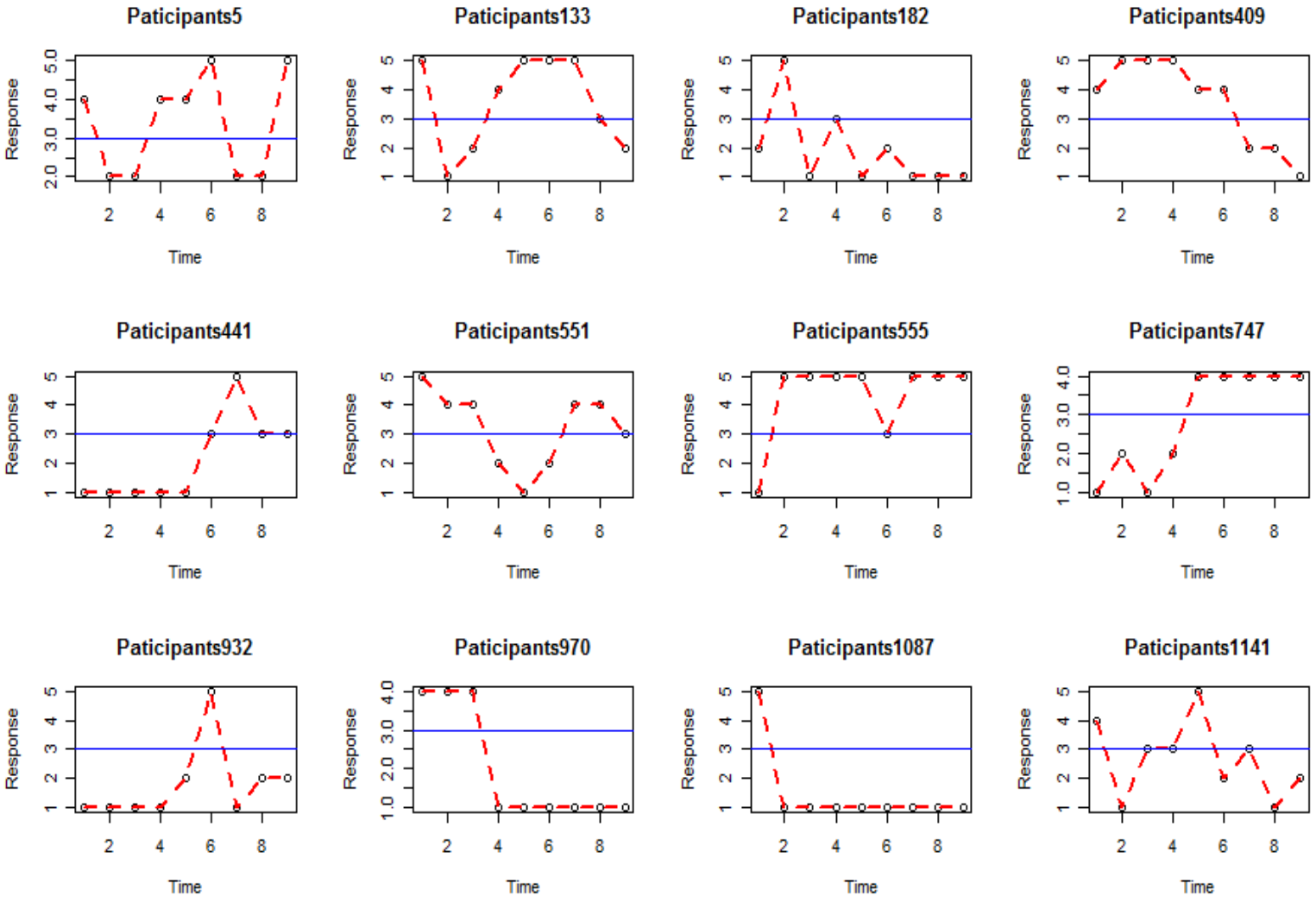
Patten Changing Over Time

- Participants become more interested when election is approaching



Subgroup Memberships' Change for Some Participants

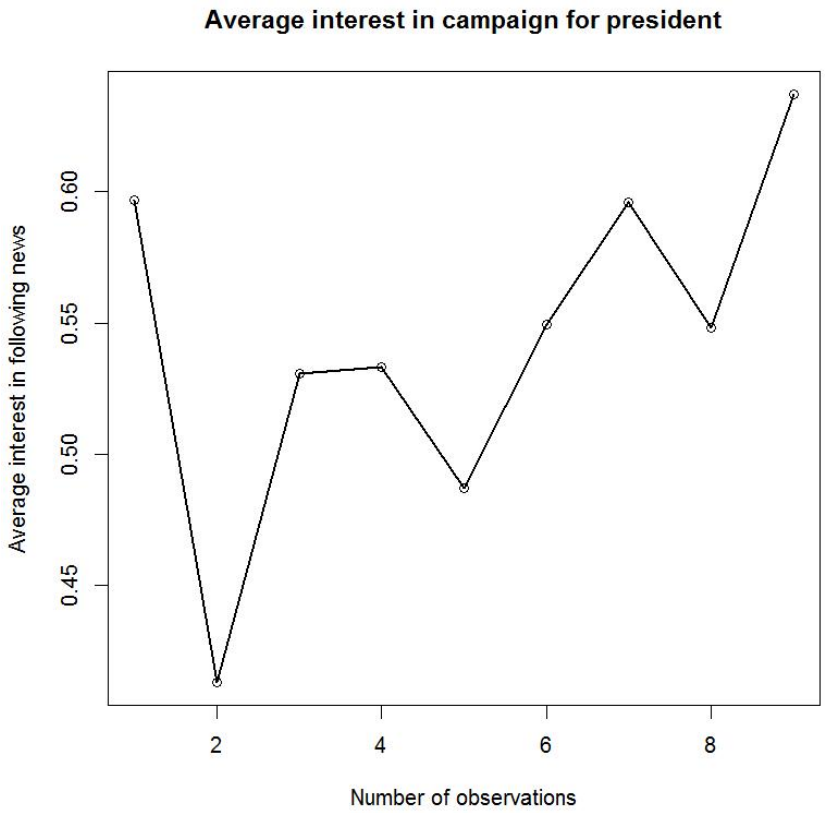
- A median score of 3 is chosen as cut off



Challenges

- Great data attrition (e.g., participants drop out): self-selected
- Participants who stayed longer show higher interest (Bi and Qu, 2017)
- Refreshment samples are drawn: baseline information not available

Missing Data and Refreshment Samples



- Respondents with higher interest tend to participate more
- The last wave refreshment samples (with one observation and close to the election) show higher interest
- Data could be missing not at random
- How to correct estimation bias?

Goal of Study

- Investigate predictors' effects on participant's responses
- Incorporate the longitudinal information
- Predict what kind of participants are more likely to vote
- Mixture modeling: one parametric model is not adequate

Recommender systems

- A system that recommend items to users
- Track users' preferences and make personalized predictions
- **Direct applications:**
  - movies, music, restaurants recommendations, on-line shopping
- **Broad applications:**
  - product sales forecasting (store & manufacturer)
  - election prediction (county & candidate)
  - computer-aided diagnosis (personalized medicine)
    (ongoing work with Northwestern University School of Medicine)

How recommendation works?

- Recommend similar items (Netflix.com)

How recommendation works?

- Recommend items frequently bought together (Amazon.com)

Matrix Data

- A matrix of users' ratings over items
- Each row represents a user; each column represent an item
- A utility matrix with 8 ratings looks like:

					
Tom	?	4	5	?	?
Jerry	?	?	?	3	1
Denny	2	5	?	?	?
Sarah	?	?	5	?	?
Edwin	?	?	?	?	4

Challenges

- Data are extremely high-volume (1M to 1B ratings)
- Extremely sparse
- Dynamic: New users registered everyday; new items released everyday

MovieLens Data

- 71,567 users over 10,681 movies (Harper and Konstan, 2016)
- 10,000,054 observed ratings out of 764,407,127 possible ratings (1% observation rate)
- Datasets are still constantly updating
- `http://grouplens.org/datasets/movielens/`

How the data look like?

- A heat map of the original data:

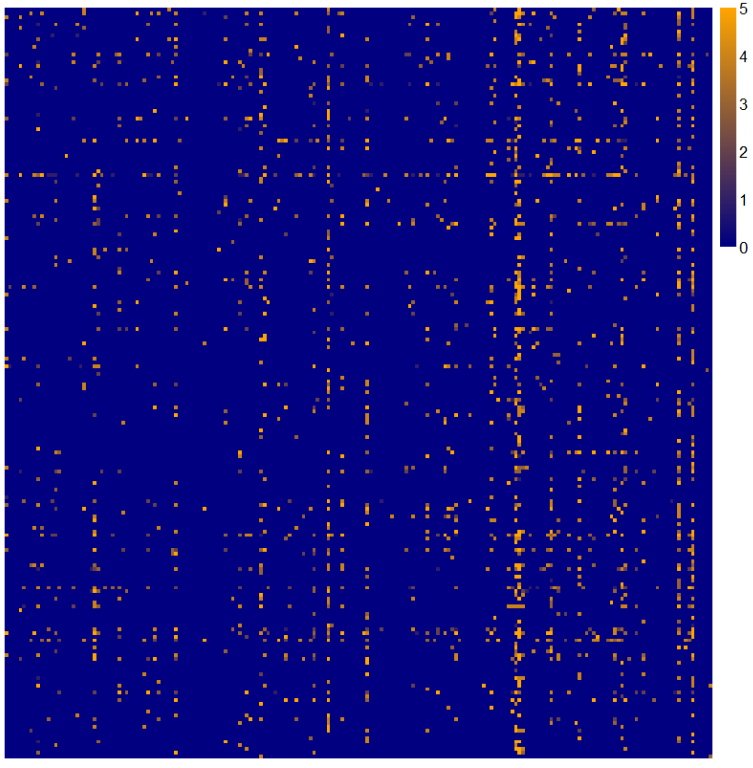
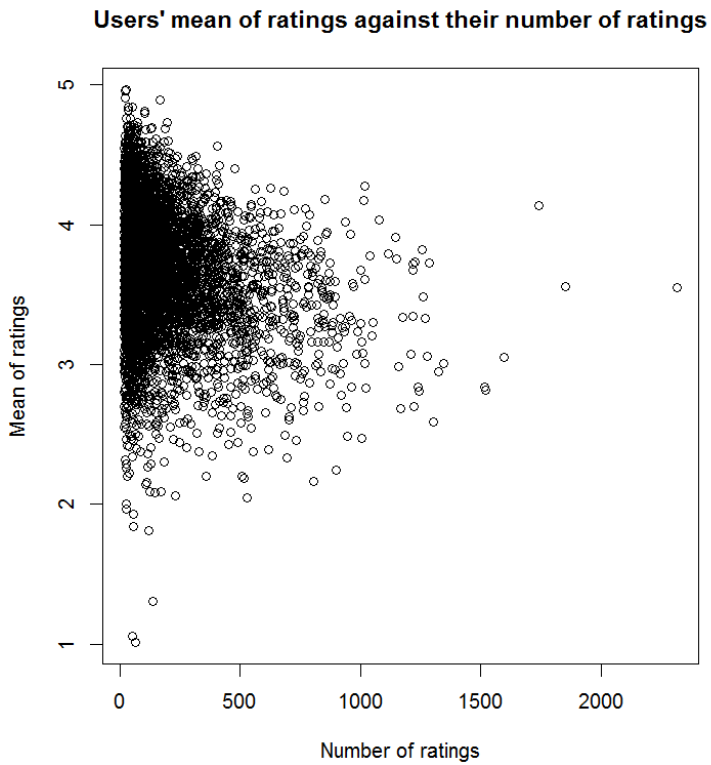
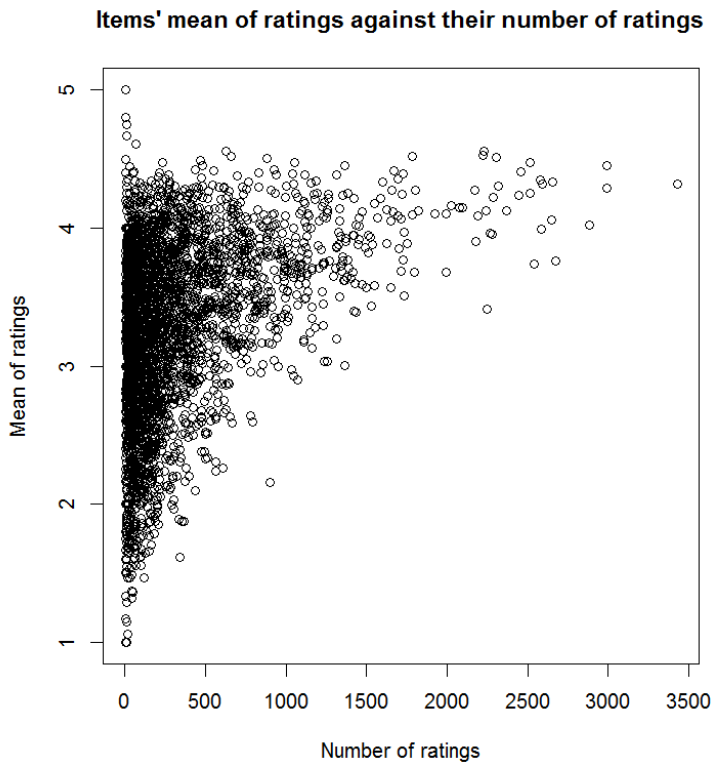


Figure: A random 200×200 sub-matrix of MovieLens 1M data

Missing Data





- The number of ratings are highly associated with the mean of ratings

Goal of Studies

- Predict users' rankings on unrated items
- Complete the data matrix
- Recommend personalized items of interests to users

Possible Solutions

- Regression methods
- Classification
- Other machine learning techniques
- Matrix factorization, matrix completion, boosting, ensemble methods

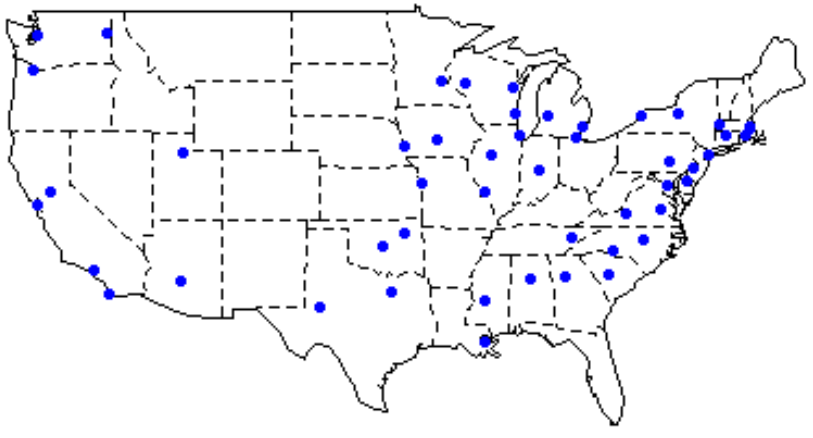
Tensor Data: Sales Data

- Data from IRi company (IRi Data, Bronnenberg et al., 2008)
- Keep tracking of transaction history from nationwide drug stores and grocery stores
- Discover popular merchandise and make predictions

IRI Marketing Data: Description & Features

- 7202 grocery stores
- 83 chains
- 31 categories of products
- 50 markets/locations
- Collected over years 2001-2011

Participating Markets/Locations



Features:

1. Store variability
2. Product variability
3. Time-varying behavior

How the data look like?

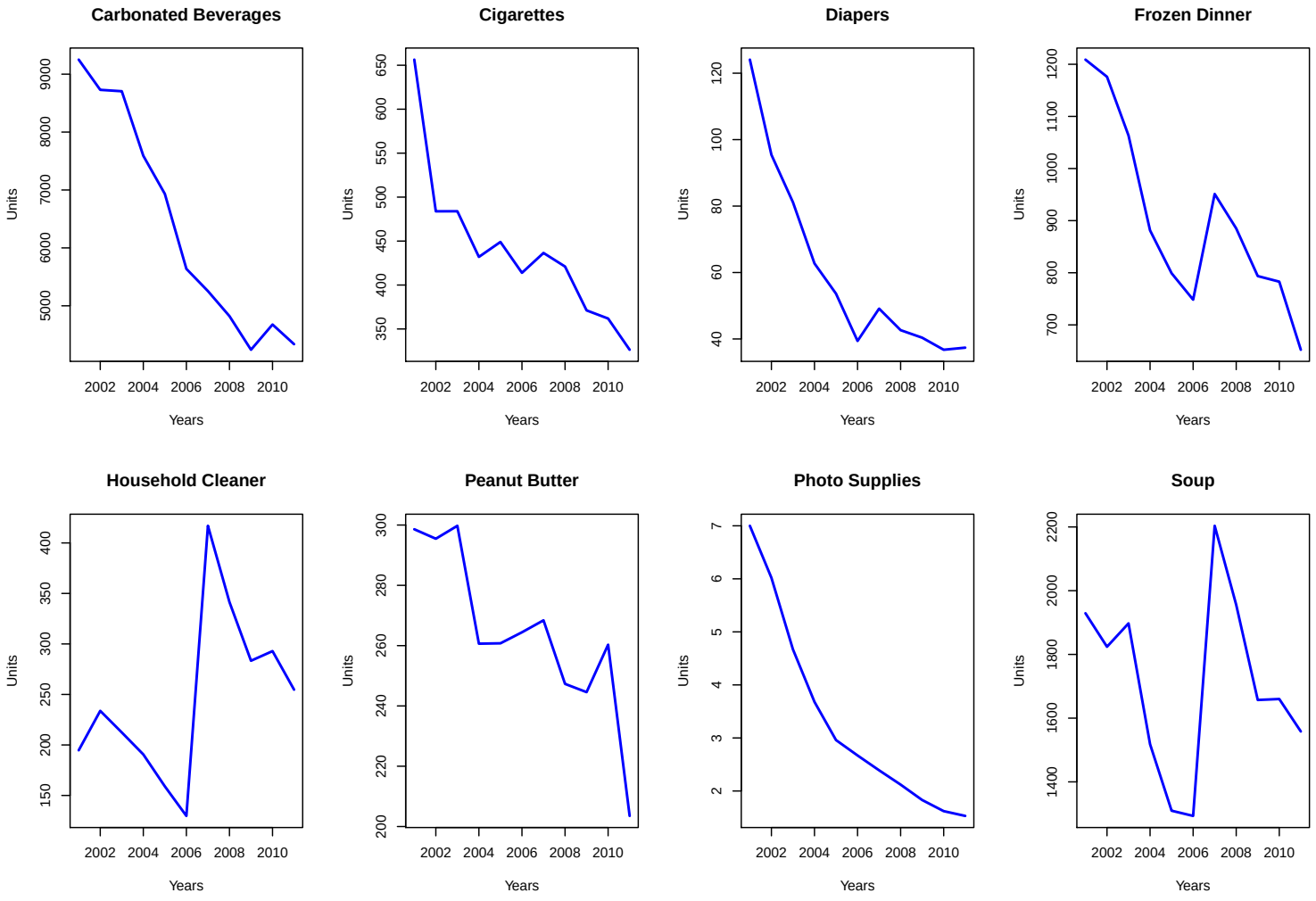
- A snapshot of the original data:

IRI_KEY	WEEK	SY	GE	VEND	ITEM	UNITS	DOLLARS	F	D	PR
681530	1373	0	1	28400	4874	2	1.98	NONE	0	0
681530	1373	0	1	28400	4853	7	6.93	NONE	0	0
681530	1373	0	1	28400	4361	20	40.00	A	0	1
681530	1373	0	1	28400	4852	1	0.99	NONE	0	0
681530	1373	0	1	28400	4363	5	10.00	A	0	1
681530	1373	0	1	28400	4854	3	2.97	NONE	0	0
681530	1373	0	1	28400	4855	1	0.99	NONE	0	0
681530	1373	0	1	28400	4365	8	16.00	A	0	1

- From left to the right:
  - Musked store number, time of transaction, system code, vendor code, total unit sales, total dollar sales, advertisements, display, price reduction

One-Store Example

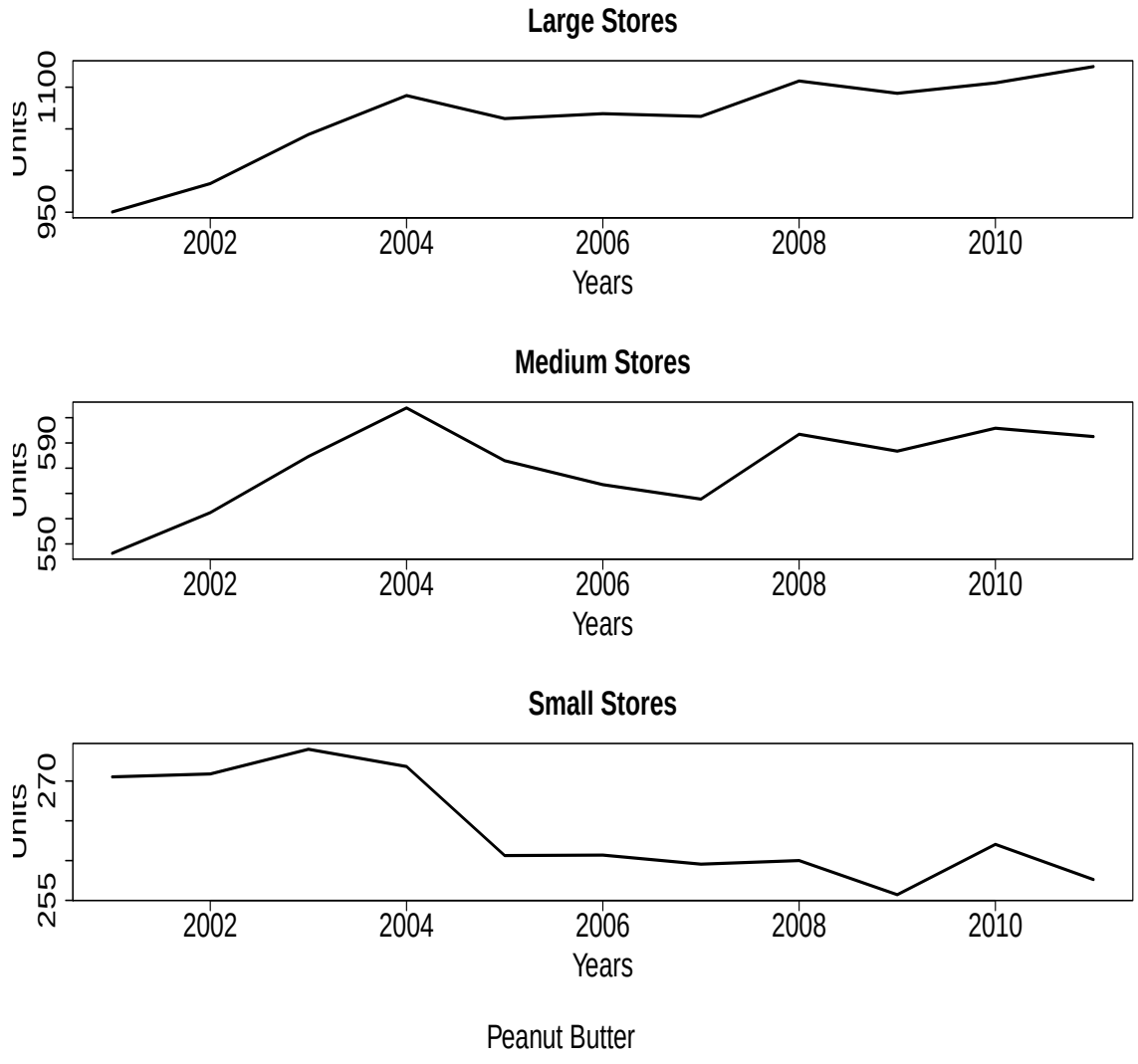
- Longitudinal data, one grocery store in New York



IRI Marketing Data: Features

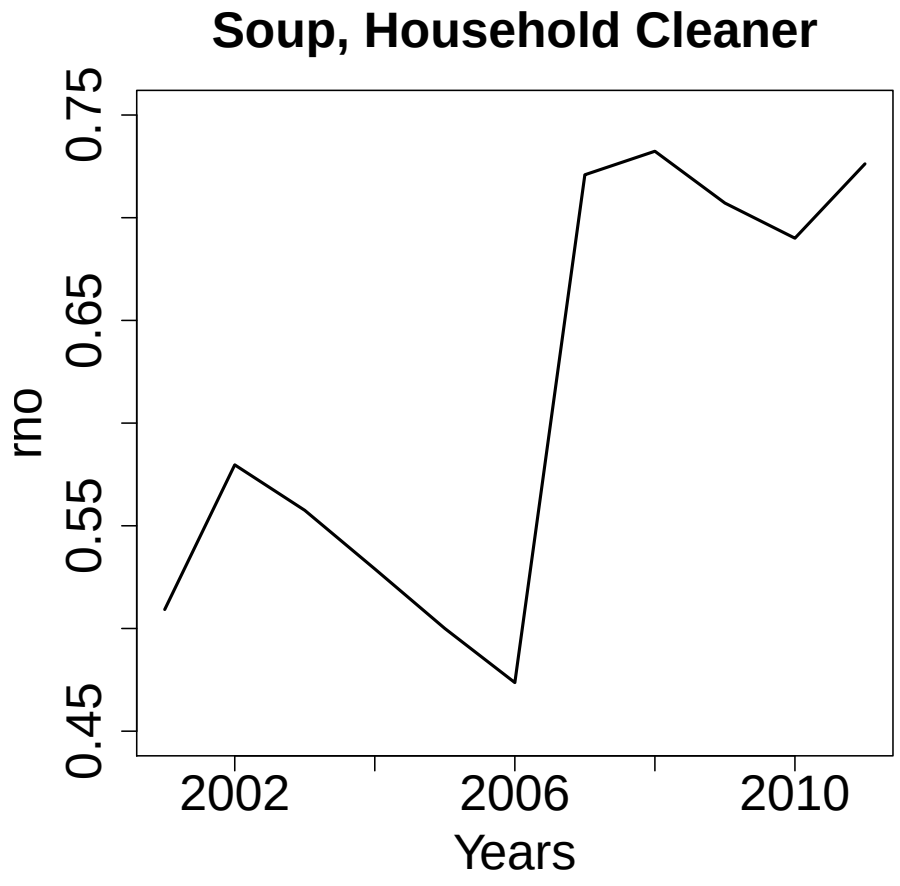
Store variability:

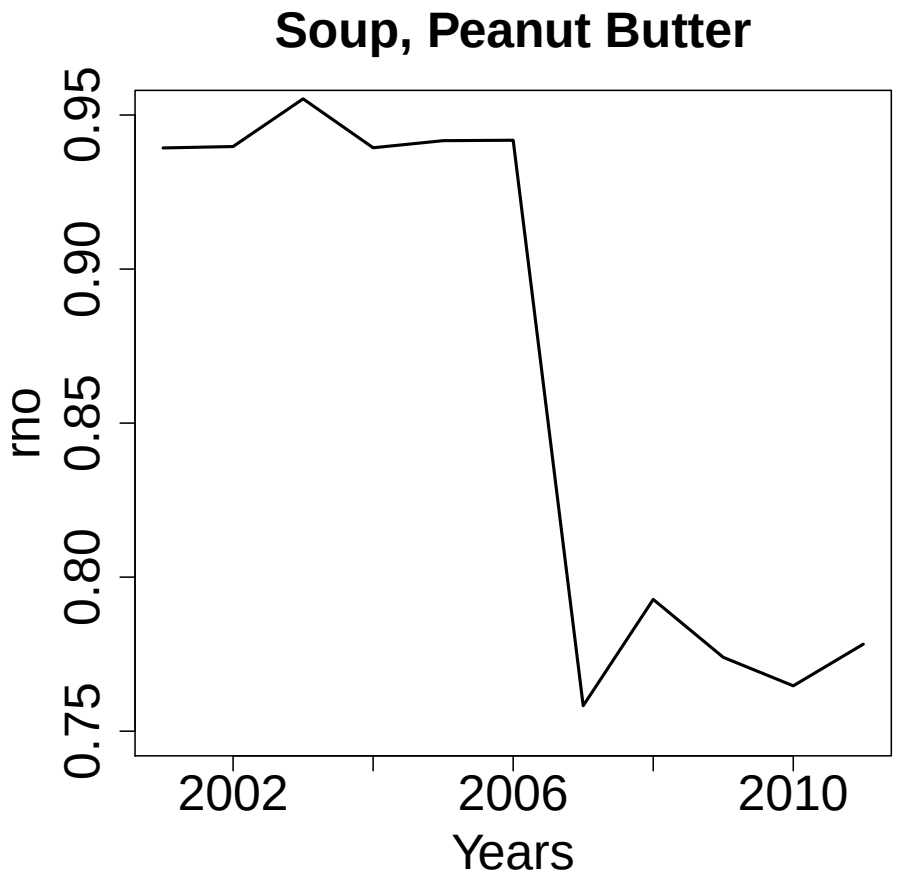
- Store size
- Markets
- Supply chain
- Years
- Local demographics
- Store strategy



IRI Marketing Data

- Time-varying correlations among products over time





Tensor Data

- Tensor:
  - Multi-dimensional array
  - More informative and complicated than matrix data
- Tensor structure:
  - User-Item-Context information
  - In IRI data: Store-Merchandise-Time
  - $\geq$ 4-dimensional tensors also exist
  - Decomposition using structure information

Context-Aware Recommender Systems

- Recommender systems may contain additional contextual information
  - Time, location, companion, promotion strategies, etc.
- Context-aware recommender systems
  - Incorporate contextual information to improve predictions
- Tensor/multidimensional array: a powerful tool for data integration

Tensor Representation

- Each mode of a tensor corresponds to *user*, *item*, or a *contextual variable*

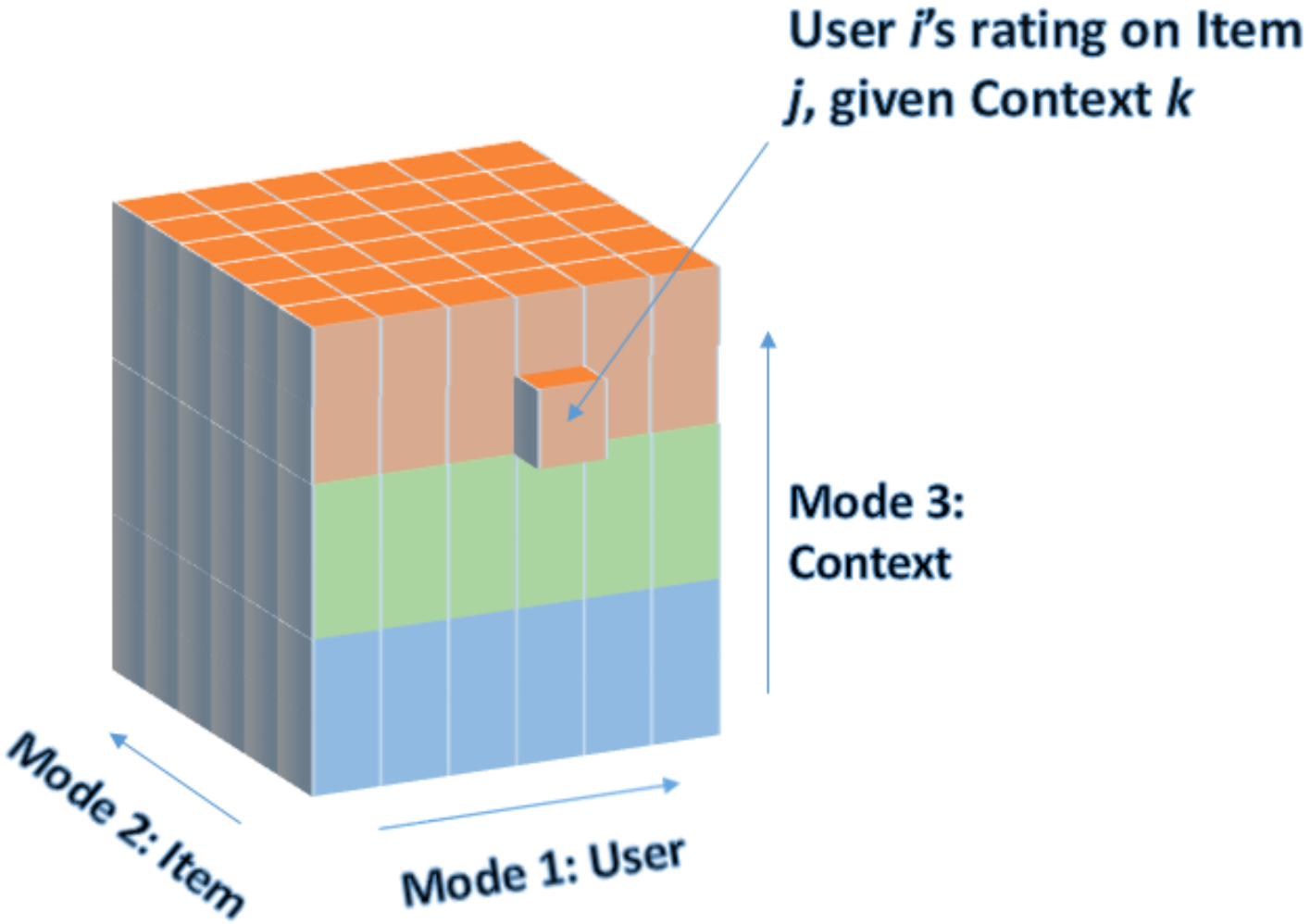


Figure: An illustration of a third-order tensor

Challenges

- Large size
  - 7202 Stores
  - 32003 Products
  - 130 Gigabytes
- Forecasting
  - Prediction of future events
  - Heterogeneity among stores and merchandises
- Sparsity
  - Some items are rarely sold,
    e.g., old-generation diapers, frozen pizza in drug stores

Goals of the Study

- Inference
  - Investigate the trajectories of nationwide product sales
- Prediction
  - Forecast the unit sales of each merchandise from each store in the following years
  - Select popular merchandises for each store

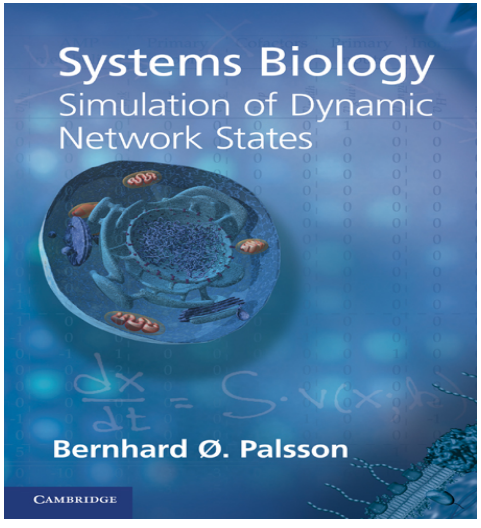
Network Data

- Model the connections among subjects
- Understand underlying factors that influence connectivity
- Predict potential connections

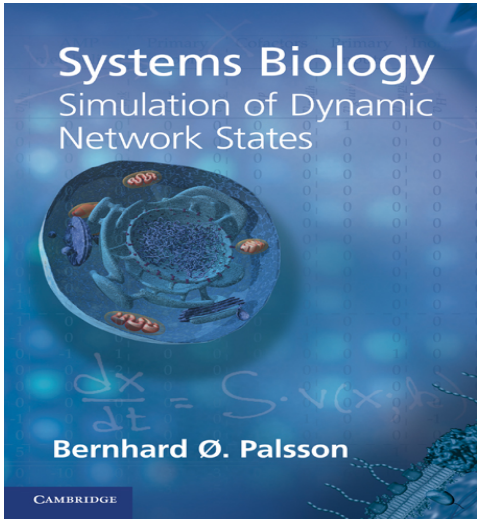
Examples of Real-Life Networks

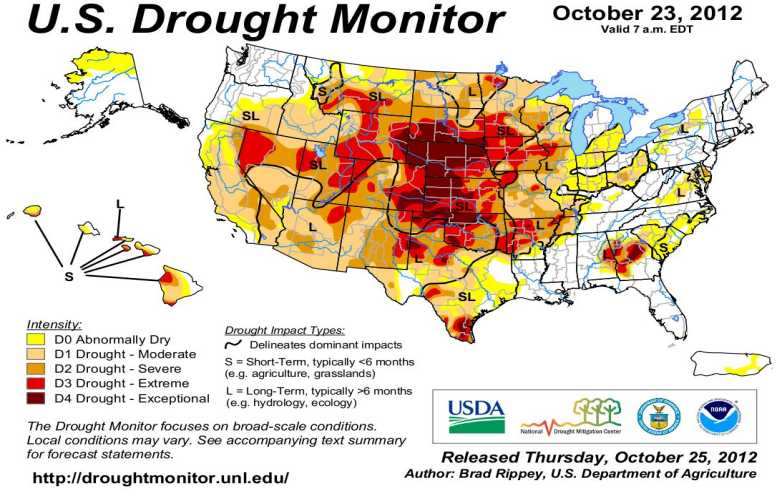
Examples of Real-Life Networks

Examples of Real-Life Networks



Examples of Real-Life Networks





Example of Network in Brain Imaging

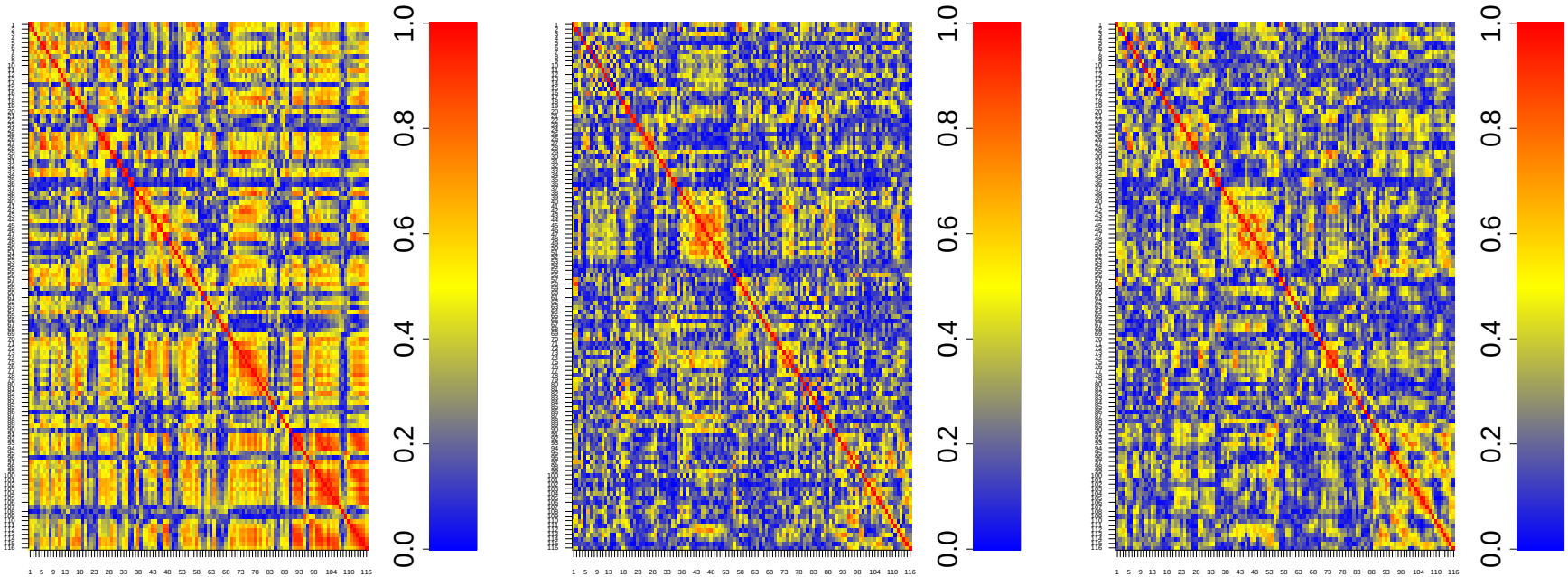
Figure: Dynamic changes of associations among different regions of interest in brain over 3 time points.

ADHD Data

- Children's attention deficit hyperactivity disorder (ADHD) fMRI data from ADHD-200 test samples
- Contain 116 regions of interest in brain over 74 time points from 78 patients's fMRI
- Model dynamic changes of brain connectivities over time
- Better understand how ADHD patients' brains function and react to different stimulants
- `http://www.nitrc.org/frs/?group%20id=383`

Correlation

Figure: Heat map of correlation matrix for 116 regions of interest at times $t = 18, 41$ and 69 for ADHD-200 data

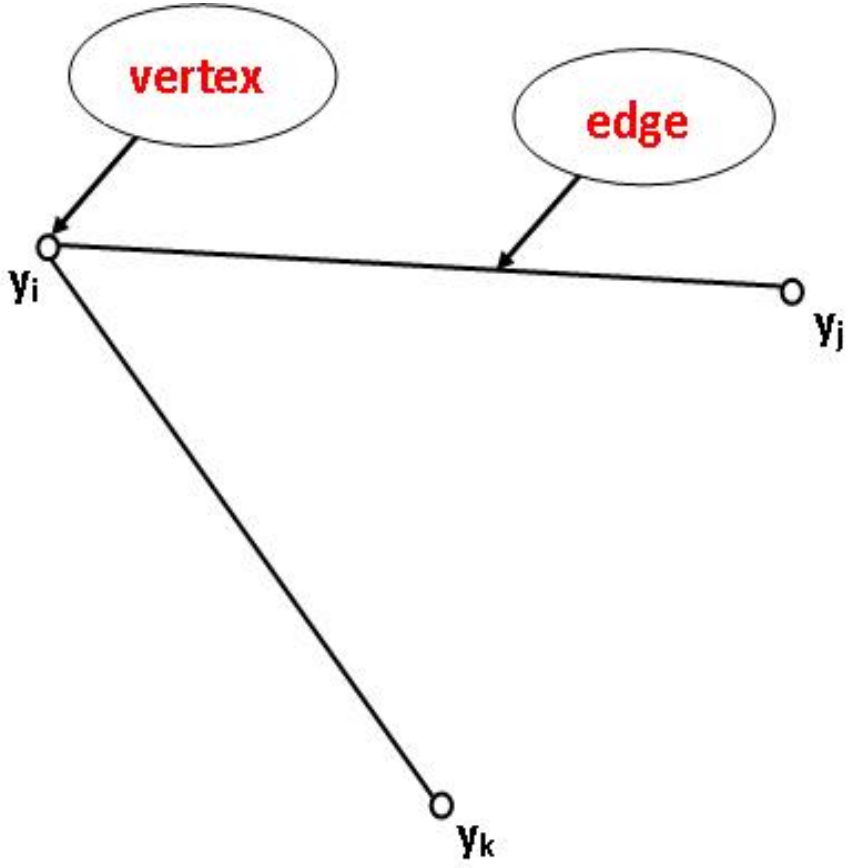


Modeling through Network

Network definition

A network is defined as an **undirected graph**,

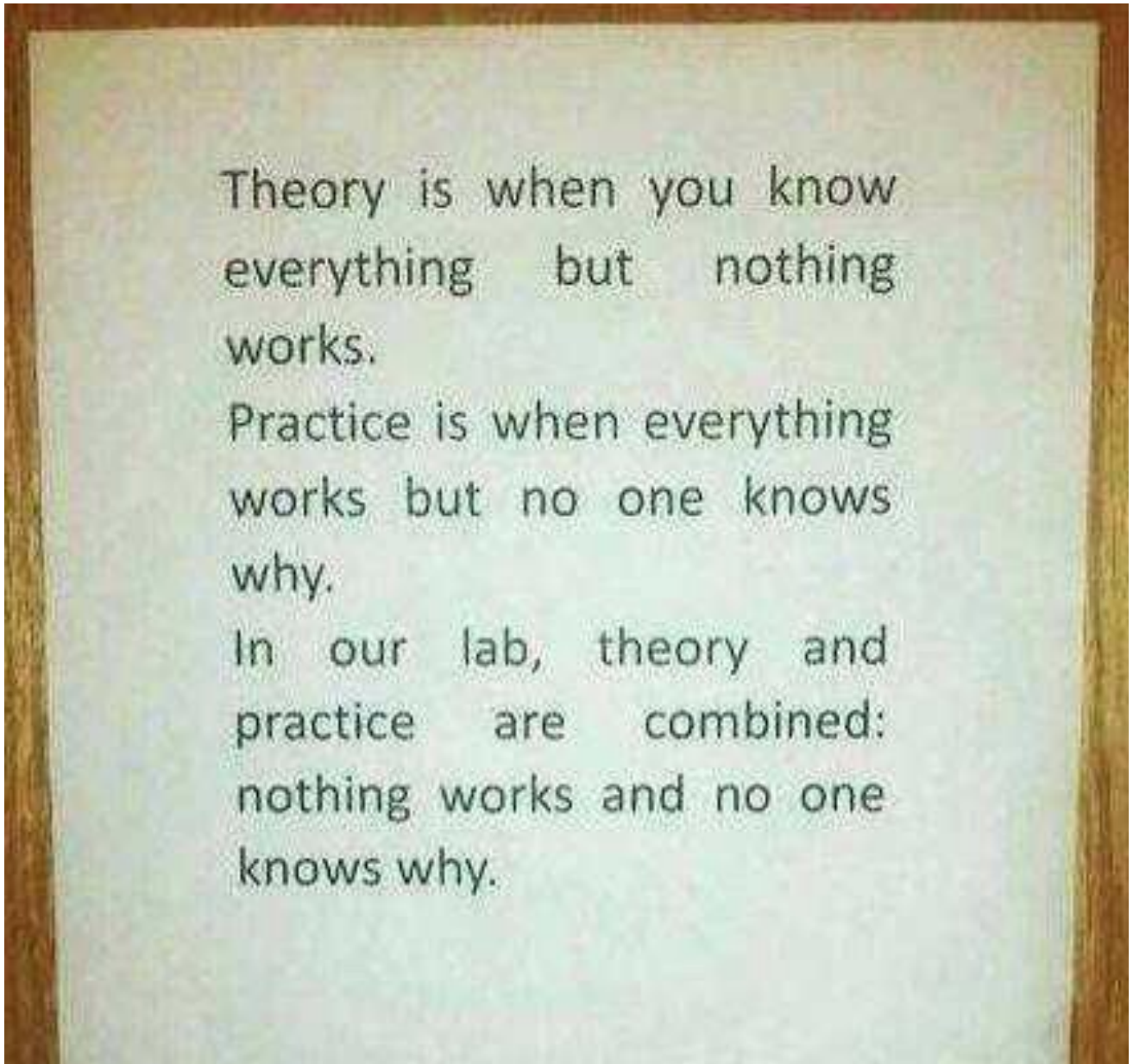
- p vertices represent the p random variables;
- The edge connects y_i and y_j if the correlation between y_i and y_j is nonzero



Possible Research Directions

- Dynamic associations among networks over time
- Detect hubs of network
- Community detection
- Computationally intensive
- Require effective optimization and machine learning tools

Text mining


Theory is when you know everything but nothing works.
Practice is when everything works but no one knows why.
In our lab, theory and practice are combined: nothing works and no one knows why.

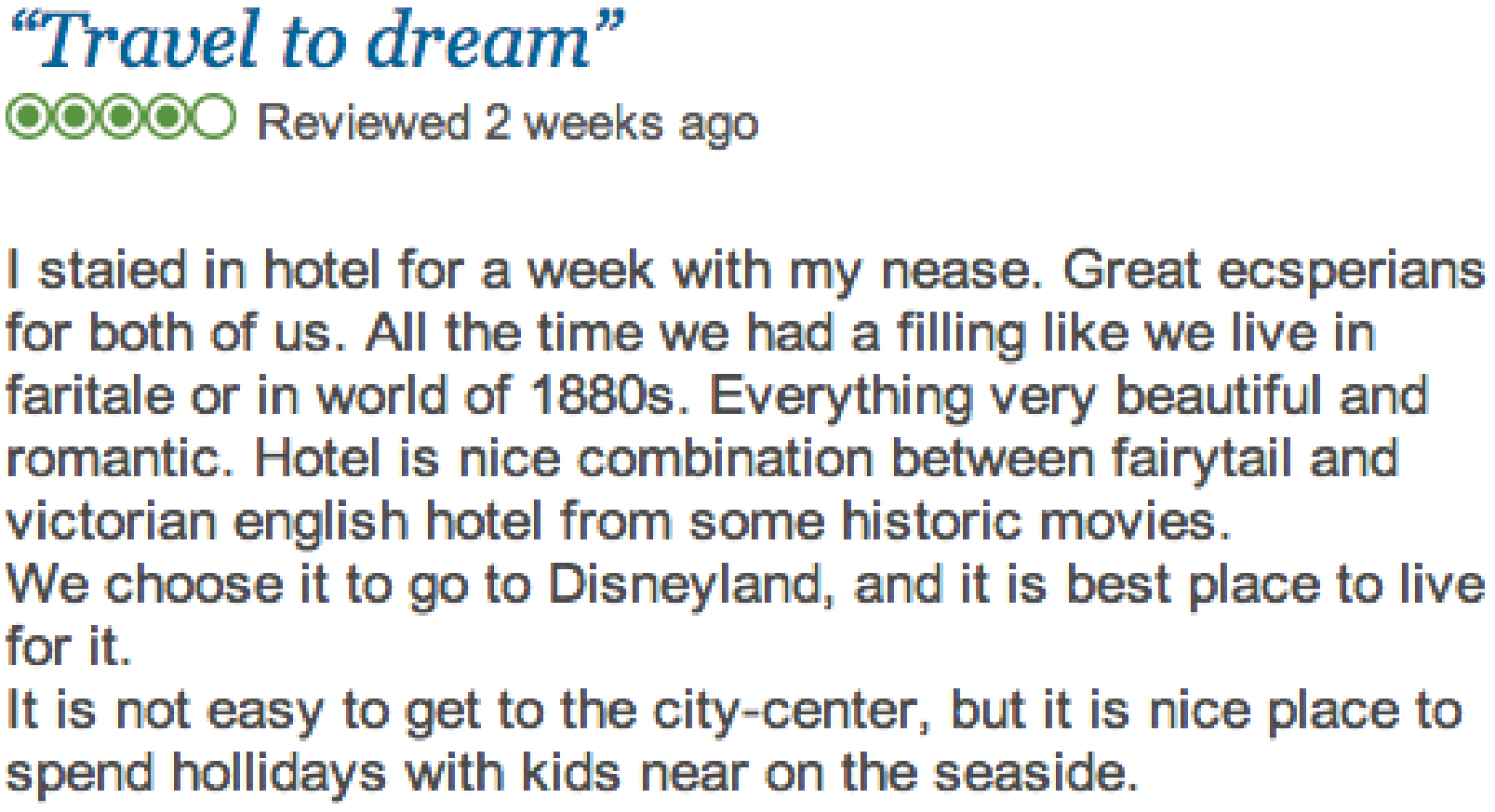
Text mining

- Text Mining = Data Mining+ Text Data
- Data Mining: Information retrieval, natural language processing, applied machine learning algorithm
- Text data: Scientific literature, emails, tweets, news articles, software documents, blogs, webpage

Unstructured data

Refers to data that lacks of well-defined structure; many types of data are unstructured.

- Textual documents: customer review, speech, ...
- Multimedia files: graph, audio, video, ...



"Travel to dream"

Reviewed 2 weeks ago

I staied in hotel for a week with my nease. Great ecsperians for both of us. All the time we had a filling like we live in faritale or in world of 1880s. Everything very beautiful and romantic. Hotel is nice combination between fairytail and victorian english hotel from some historic movies.
We choose it to go to Disneyland, and it is best place to live for it.
It is not easy to get to the city-center, but it is nice place to spend hollidays with kids near on the seaside.

Text mining: Summarization in word cloud

Challenges

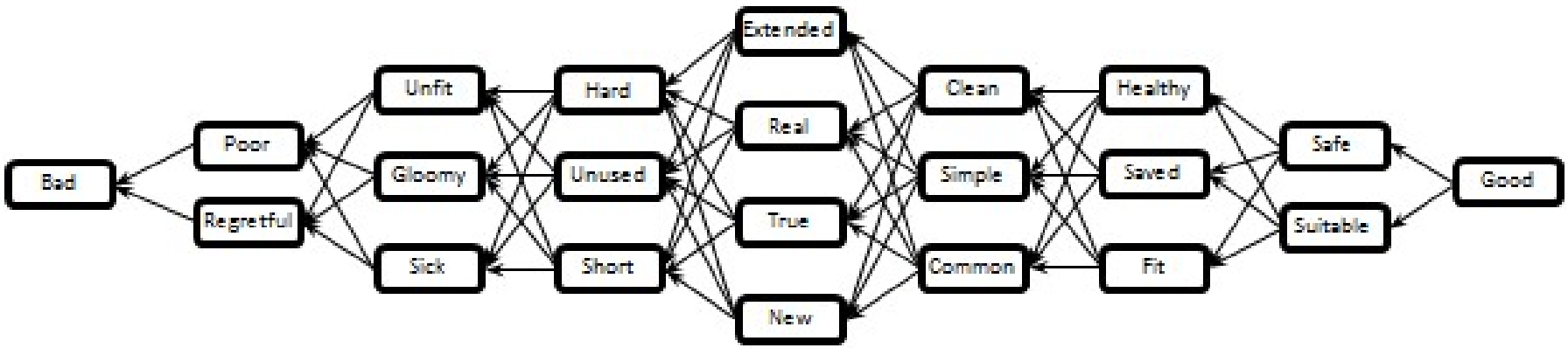
- Challenges for analyzing data with unstructured predictors.
  - Numericalization is necessary. However, this may result in loss of information.
  - Size of unstructured data could be BIG for both the sample size n and the dimension p.
- Classification with unstructured predictors is under-studied
- Statistics literature is limited for classification

Text data

- Text analytics: Extracting high-quality information from text
  - High quality: relevance, novelty, and interest of particular groups
- Goals: text tagging, sentiment analysis, segmentation, summarization, abstraction, etc.
- An interdisciplinary research of Linguistics, Statistics, and Computer Science
- Might require literacy in the targeted natural language (e.g., Chinese, Spanish, etc.)

An example: Sentiment analysis

- Sentiment analysis identifies sentiment opinion of a textual document towards certain event.
- Vast literature in computer linguistics, natural language processing, and social science.
- Requires large-scale sentiment lexicons



Real application: TripAdvisor.com dataset

- Customer review data from `http://times.cs.uiuc.edu/~wang296/Data/`.
- Data: Many review documents. For each text review, a rating score from 1 to 5 is assigned to indicate the degree of satisfaction.
- Dictionary: 6,800 positive and negative sentiment words (verbs, adjectives and adverbs).
- WordNet distance is used to construct the graph for word sentiment strengths. E.g., "Interact" $\rightarrow$ "Communicate" $\rightarrow$ "Inform" indicates that the positive strengths decrease (Wang et al., 2016).

Top 30 sentiment words

Positive	loved, awesome, appreciated, enjoy, love, recommended, helped, comfort, well, celebrate, appreciate, intelligent, humor
Negative	ignorant, scratch, bait, creeps, appalling, beg, crafty, stuck, insulting, pitiful, tainted, false, rubbish, cramp, scam, desolate

- Intend to estimate the weights associated with sentiment words
- Through the appearance of sentiment words and their weights to identify positive or negative opinions

Another example: Text segmentation

- The process of dividing text into meaningful units, disregarding grammar:
  - Word segmentation
  - Sentence segmentation
  - Topic segmentation
  - Other segmentations
- Sensitive to the targeted language, for example:
  - English: ignore punctuation and capitalization.
  - Chinese, Eastern Asian languages: segmentation is necessary

Chinese segmentation

- Five categories:
  - lexical words (LW), morphologically derived words (MDW), factoids (FT), named entities (NE) and new words (NW)

category	subcategory	examples
LW	lexical word	学生，照片，约会
MDW	affixation	老师们
	reduplication	马马虎虎
	splitting	吃了饭
	merging	上下文
	head+particle	拿出来
FT	date & time	5月3日，六月五日，12点半，三点二十分
	number & fraction	一千零二十四，4897，60%，百分之一，1/6
	email & website	johnson@email.com, www.google.com
NE	person name	张三，约翰
	location name	北京，上海
	organization name	长城，大都会博物馆
NW	new word	吐槽，非典

Real application: Chinese segmentation

Peking University corpus

- Training set: $161,212$ sentences and 1.1 million words
- Test set: $14,922$ sentences and 17 thousand words
- Domains: politics, economics, culture, law, science and technology, sport, military and literature (Shu et al., 2017)

Real application: Chinese segmentation

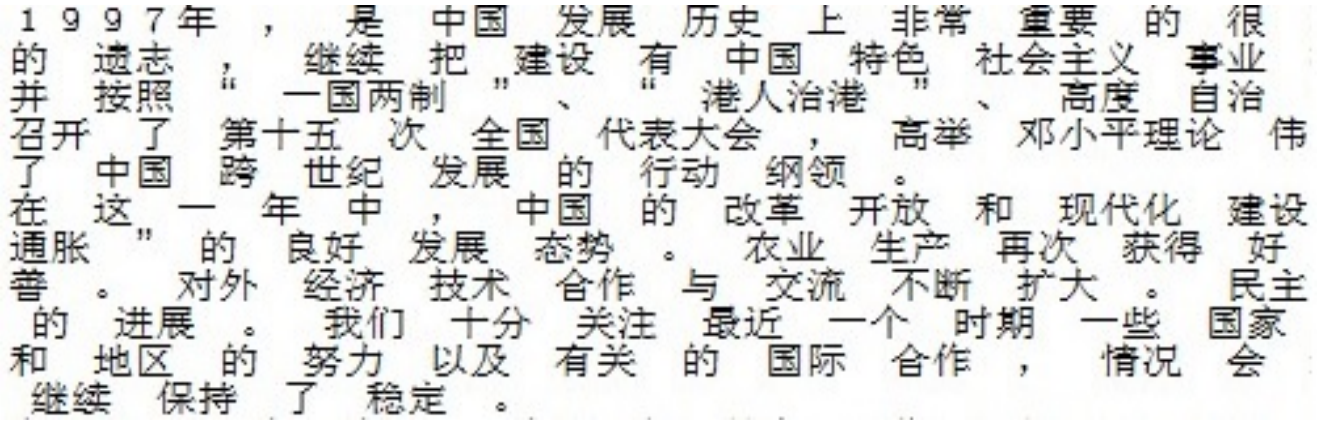


Figure: Fragments of the PK training set.

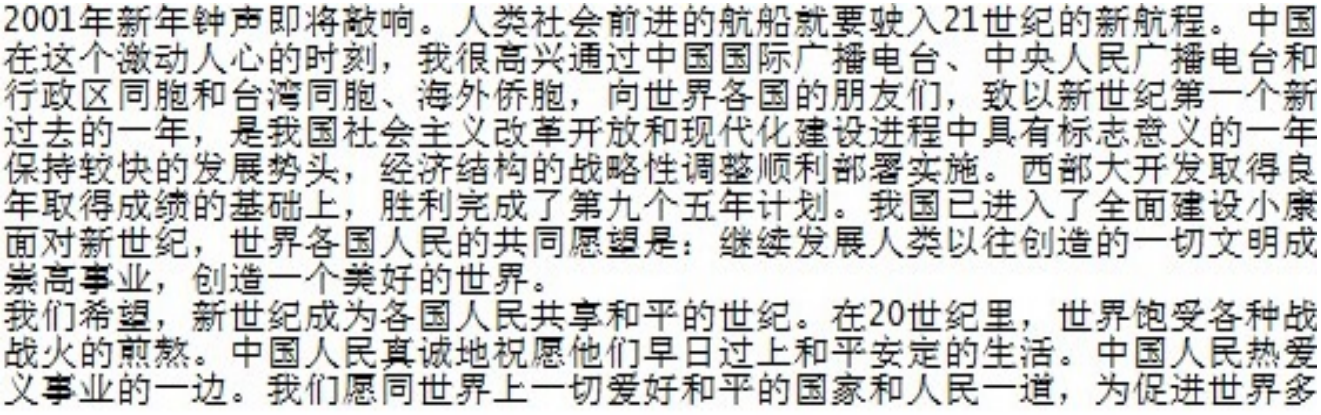


Figure: Fragments of the PK test set.

Imaging Data

- Imaging data: pattern recognition, image inpainting

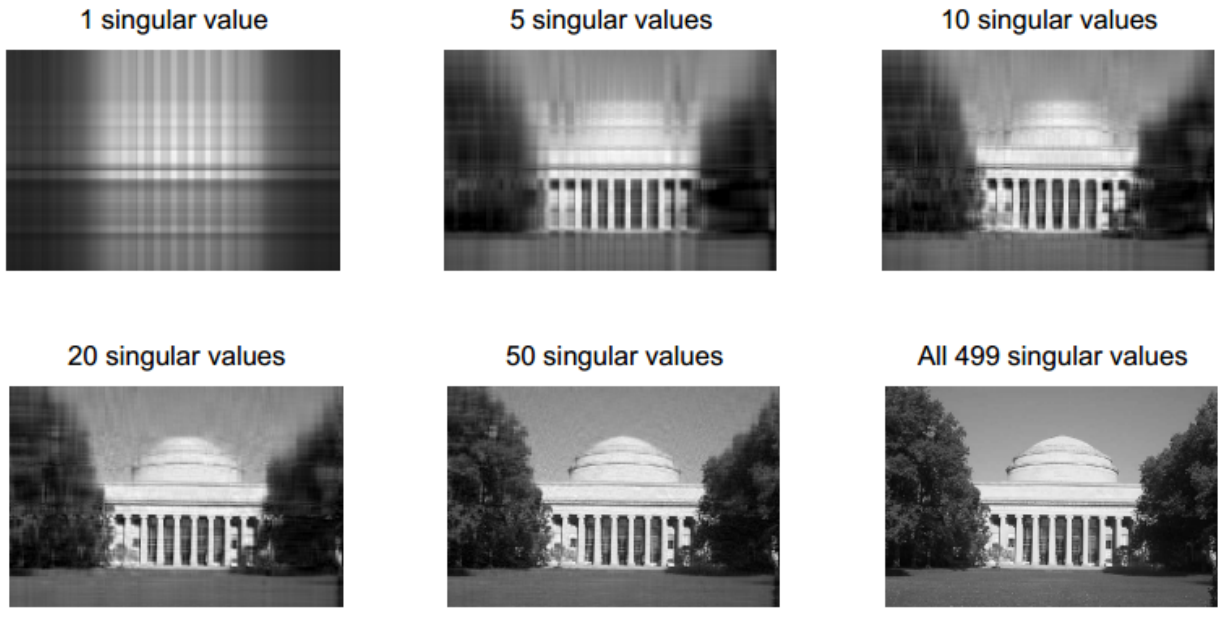


Figure: Image Compression using SVD

Source: `http://web.mit.edu/18.06/www/Fall03/svd.pdf`

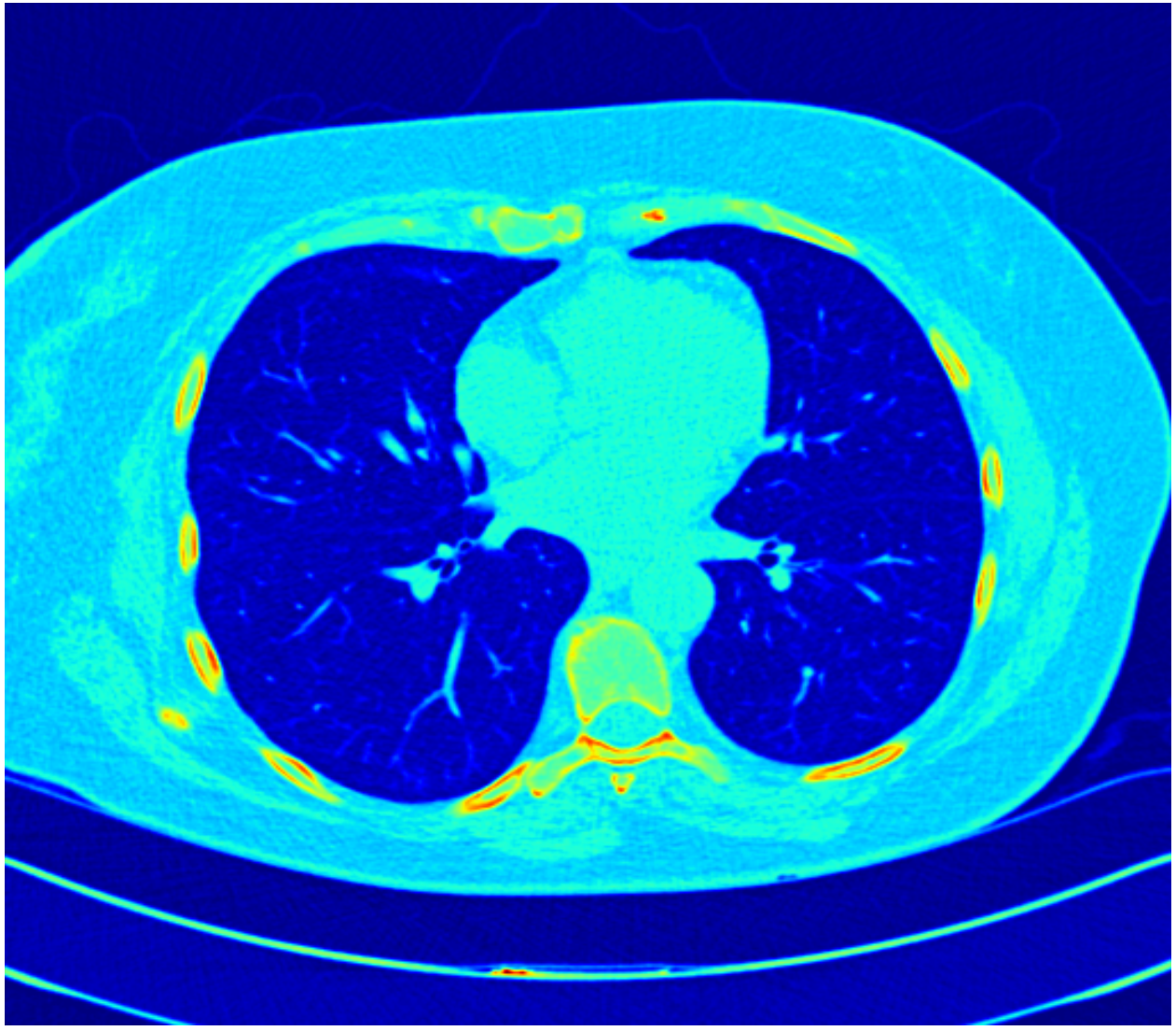
Brain Imaging Data

- President Obama launched the Brain Initiative in 2013
- https://www.whitehouse.gov/BRAIN
- ABIDE: Autism Brain Imaging Data Excahnge
- http://fcon_1000.projects.nitrc.org/indi/abide/
- The Alzheimers Disease Neuroimaging Initiative (ADNI)
- Study the progression of Alzheimers disease
- MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors for the disease
- http://adni.loni.usc.edu/

Data Competition on Lung Cancer

- CT scan data: image data
- The competition task is to create an automated method capable of determining whether or not the patient will be diagnosed with lung cancer within one year of the date the scan was taken
- The ground truth labels were confirmed by pathology diagnosis
- The images in this dataset come from many sources and will vary in quality.
- Older scans were imaged with less sophisticated equipment
- The stage 2 data are more recent and have higher quality than the stage 1 data
- Your algorithm should perform well across a range of image quality

Sample Lung MRI Imaging



Other Types of Data

- Implicit data: Internet browse/click history, fraud detection
- Voice data: voice recognition, segmentation
- Video data: security data, capture suspects' image and movement

Competition data

- Check www.kaggle.com for competition data

References

Bi, X. and Qu, A. (2017). A mixed-effects estimating equation approach to nonignorable missing longitudinal data with refreshment samples. *Statistica Sinica*, to appear.

Bronnenberg, B. J., Kruger, M. W., and Mela, C. F. (2008). Database paper-the iri marketing data set. *Marketing science*, 27(4):745–748.

Harper, F. M. and Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4).

Shu, X., Wang, J., Shen, X., and Qu, A. (2017). Words segmentation in chinese language processing. *Statistics and Its Interface*, 10:165–173.

Tang, X. and Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, 25:1117–1137.

Wang, J., Shen, X., Sun, Y., and Qu, A. (2016). Classification with unstructured predictors and an application to sentiment analysis. *Journal of the American Statistical Association*, 111:1242–1253.