Data Mining & Machine Learning

Arizona State University

Keeping Pace with Big Data
- A Data Mining Perspective

Huan Liu

Data Mining and Machine Learning Lab

Arizona State University, Tempe, AZ

http://www.public.asu.edu/~huanliu

NSF Workshop on Big Data Analytics for Infrastructure and Building Resilience and Sustainability, Beijing, China Sept 19-20, 2014

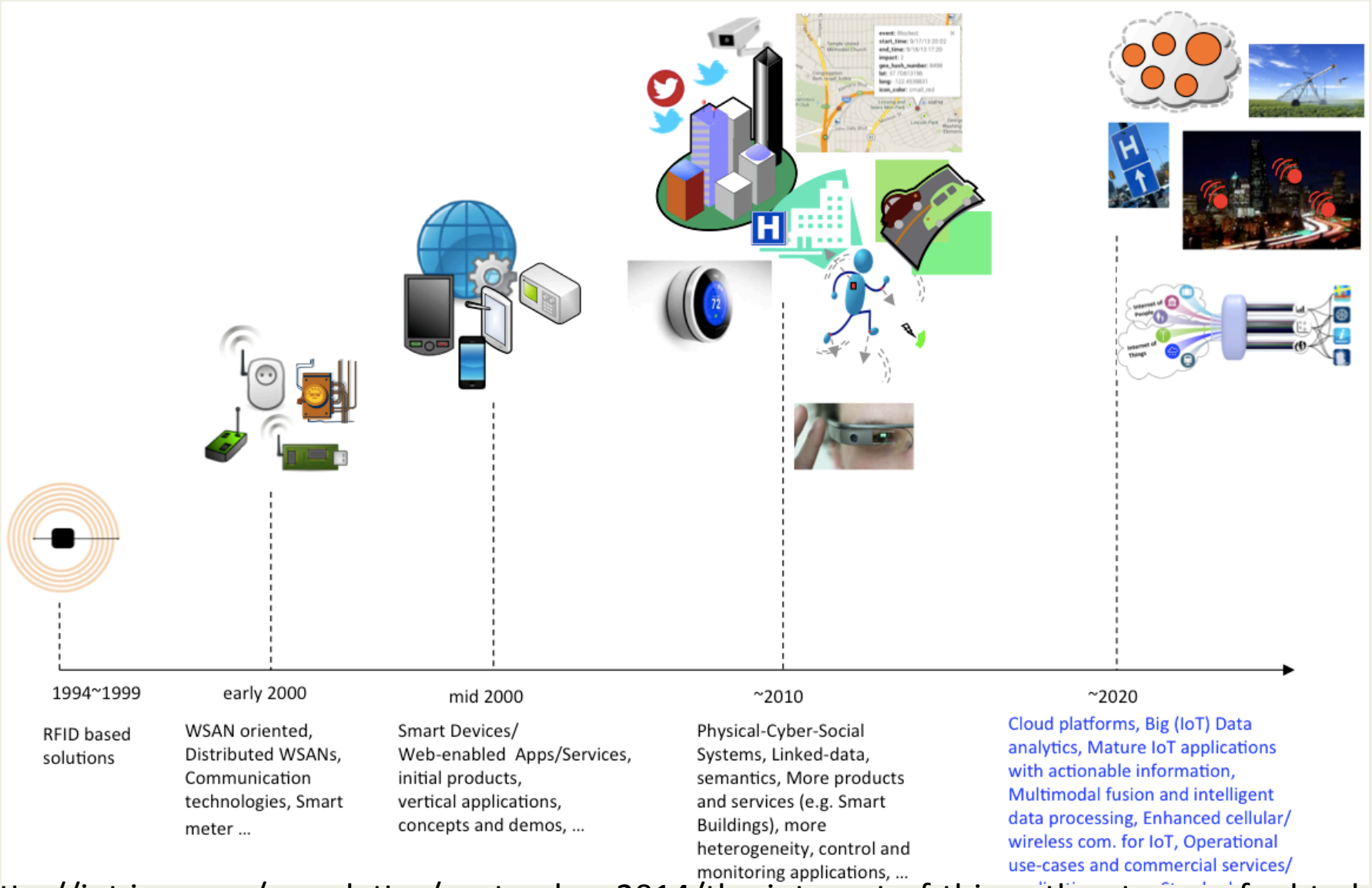
Concluding Remarks

- Big Data is a good problem to have
- Data mining is one way of approaching it
- Together, we can harness it for better sci & eng

Keeping Pace with Big Data

- Big data is not a new problem, but a persistent one
  - Why now?
    - We're overwhelmed, start appreciating data value, and data is generated ubiquitously (we're part of the problem)
  - We have been dealing with it since we had data
    - Feature selection, as an example, to battle data explosion (mainly for attribute-value data)
- Big data will only become bigger
  - Ubiquitous and fast growing linked data in the age of social media
    - Example continued, Feature selection for linked data
- Big data is a good problem to have
  - And, many a time, big data may not be big enough

Data will only become bigger



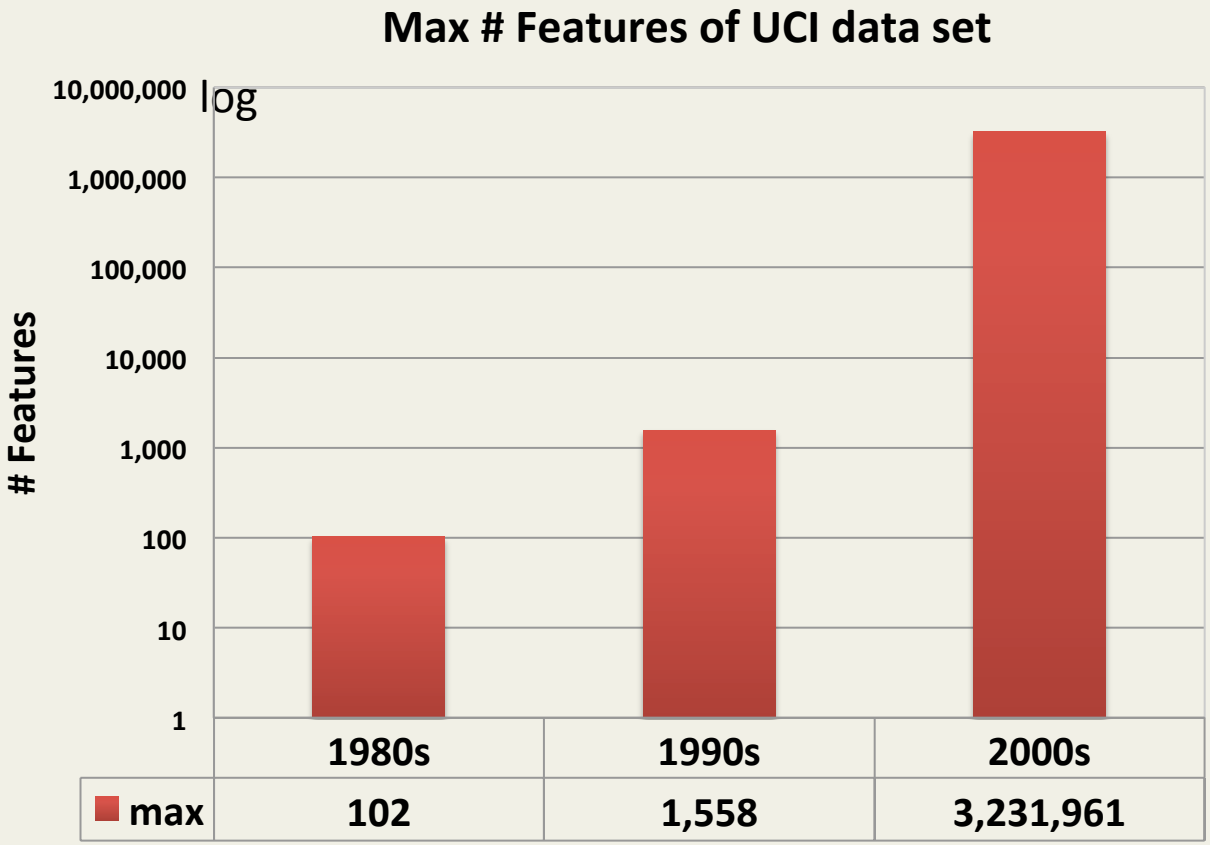
http://iot.ieee.org/newsletter/september-2014/the-internet-of-things-the-story-so-far.html

Begin with Attribute-Value Data

- It is the most familiar form of data we encounter
  - Tables in Excel, Databases, …
  - Data is conveniently collected everywhere
- Some typical challenges
  - Data overload (increasing in both *width* and *length*)
  - Data is collected for various reasons
  - Data accumulates at an unprecedented speed
  - Data itself does not offer any insight, but has potential
- To make sense of massive amounts of data is to focus: using only relevant data
  - Data preprocessing is an important part of machine learning and data mining
  - Feature selection is an effective approach to downsizing data

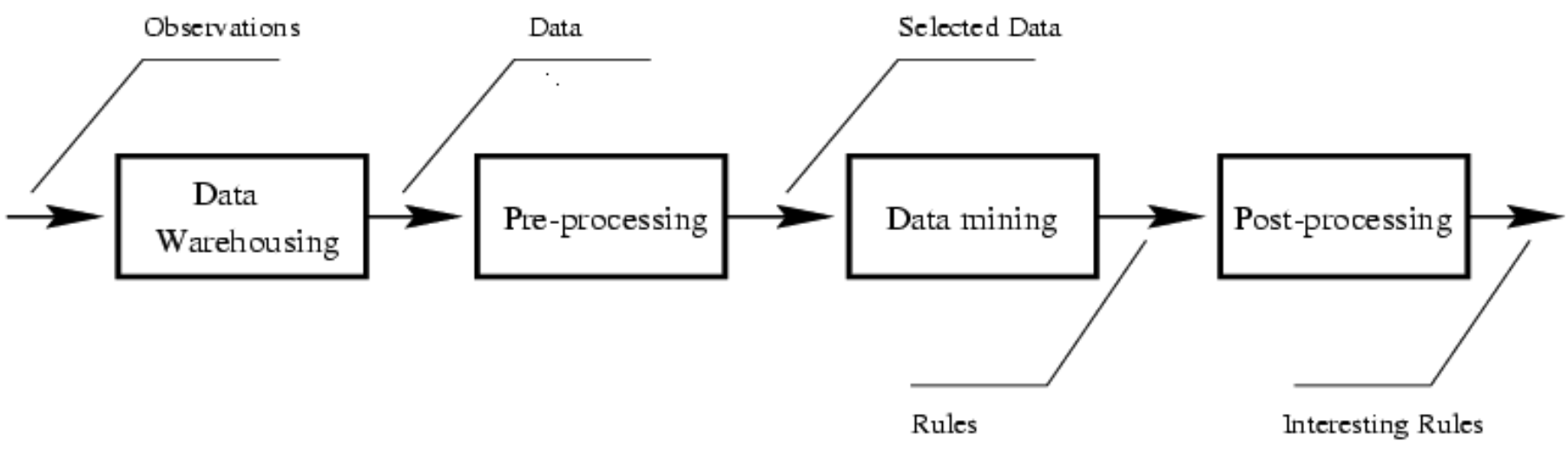
Massive Data and High Dimensionality

- Dimensionality of data has increased exponentially



A General Model of KDD

- Knowledge Discovery and Data Mining



- Data mining
  - Applying analytical methods and tools to discover actionable patterns, construct statistical or predictive models, and identify relationships among massive data

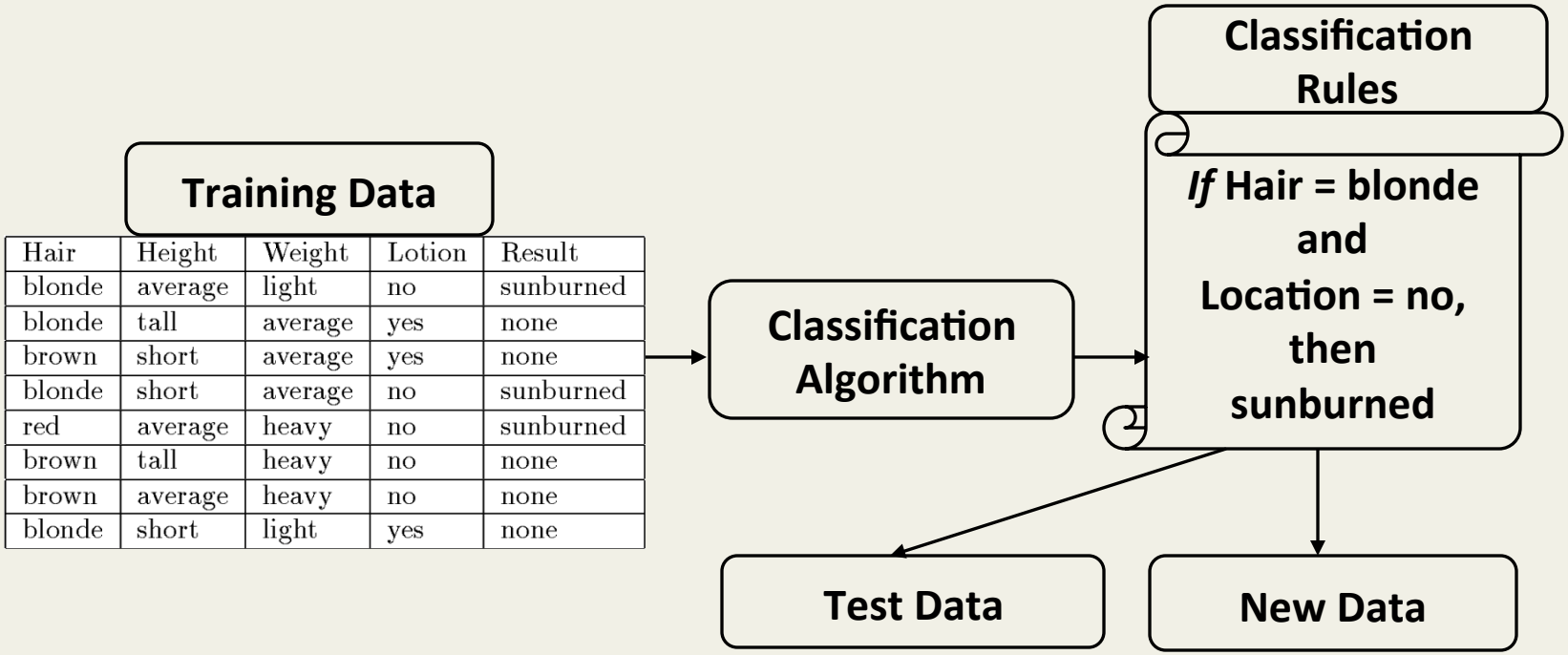
Why Feature Selection?

- Most machine learning and data mining techniques may not be effective for high-dimensional data
  - Curse of Dimensionality
  - Query accuracy and efficiency degrade rapidly as the dimensionality increases.

- The intrinsic dimensionality may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.

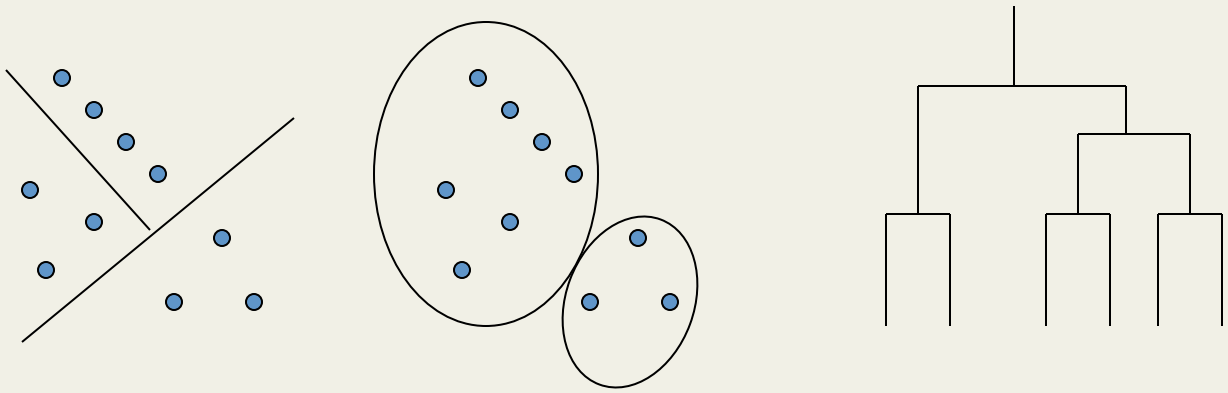
Classification

- A process of predicting the classes of unseen instances based on patterns learned from available instances
- Supervised learning with labeled data



Clustering

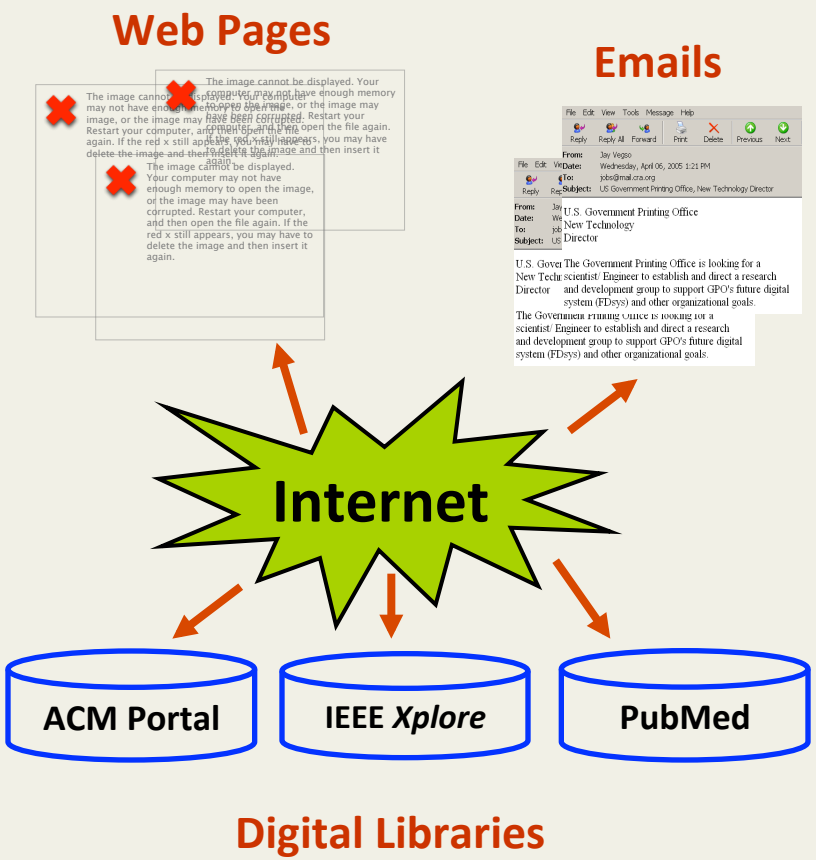
- A process of grouping objects (or instances) into *clusters* so that objects are similar to one another within a cluster but dissimilar to objects in other clusters
- Unsupervised learning with unlabeled data
- Clustering tasks

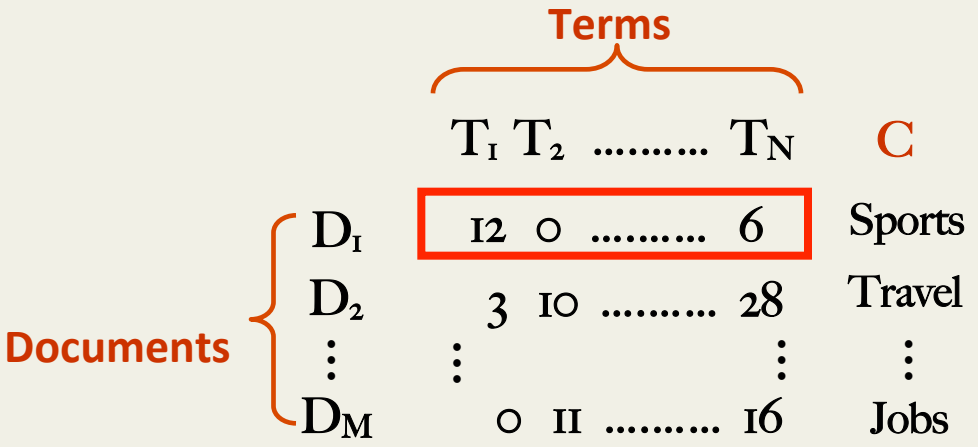


Applications of Feature Selection

- Customer relationship management
- Text mining and visual analytics
- Image retrieval
- Microarray data analysis and protein classification
- Face recognition and handwritten digit recognition
- Intrusion detection
- Social media and social networking apps

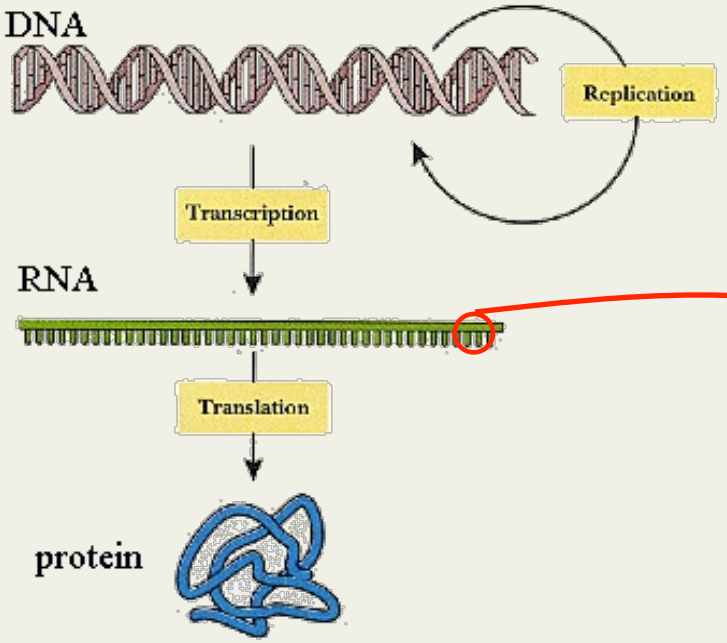
Online Document Classification

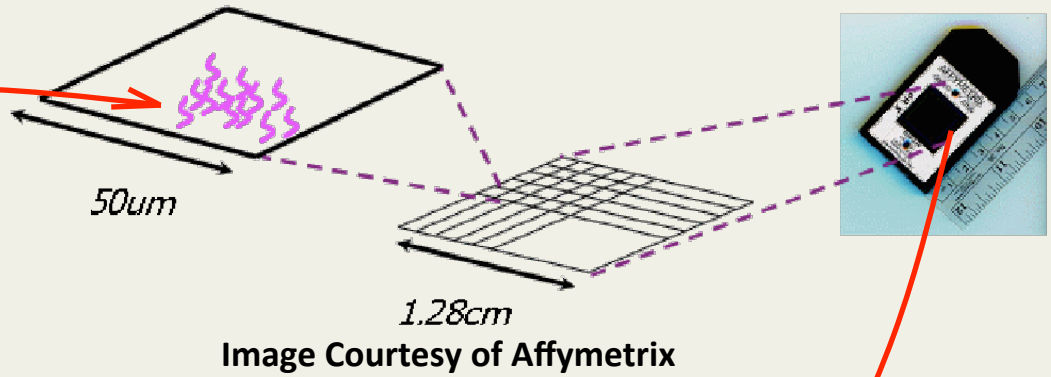




- **Task:** To classify unlabeled documents into categories
- **Challenge:** thousands of terms
- **Solution:** to apply dimensionality reduction

Gene Expression Microarray Analysis





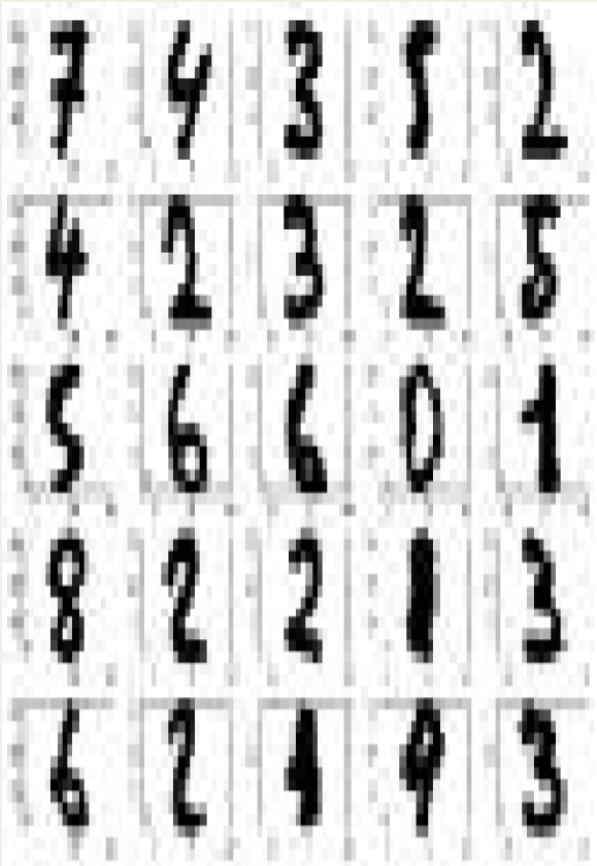
- **Task:** To classify novel samples into known disease types (disease diagnosis)
- **Challenge:** hundreds of thousands of genes, but a few samples
- **Solution:** Feature Selection

Sample \ Gene	M23197_at	U66497_at	M92287_at	. . .	Class
Sample 1	261	88	4778	. . .	ALL
Sample 2	101	74	2700	. . .	ALL
Sample 3	1450	34	498	. . .	AML
.	.	.	.	. . .	.
.	.	.	.	. . .	.
.	.	.	.	. . .	.

Expression Microarray Data Set

Other Types of High-Dimensional Data

Face images



Handwritten digits

Evaluation Measures for Ranking and Selecting Features

- The goodness of a feature/feature subset is dependent on measures
- Various measures
  - Information measures
  - Distance measures
  - Dependence measures
  - Consistency measures
  - Accuracy measures

Information Measures

- Entropy of variable X

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i))$$

- Entropy of X after observing Y

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j))$$

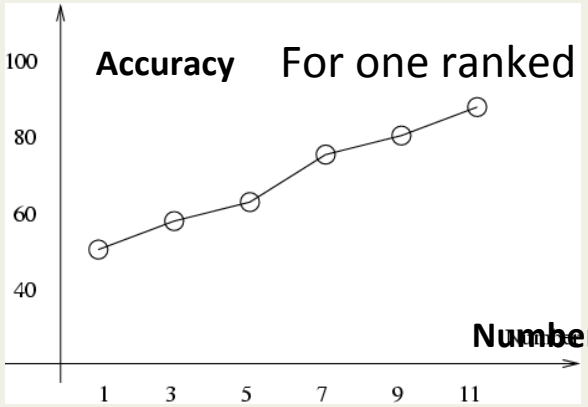
- Information Gain

$$IG(X|Y) = H(X) - H(X|Y)$$

How to Validate Selection Results

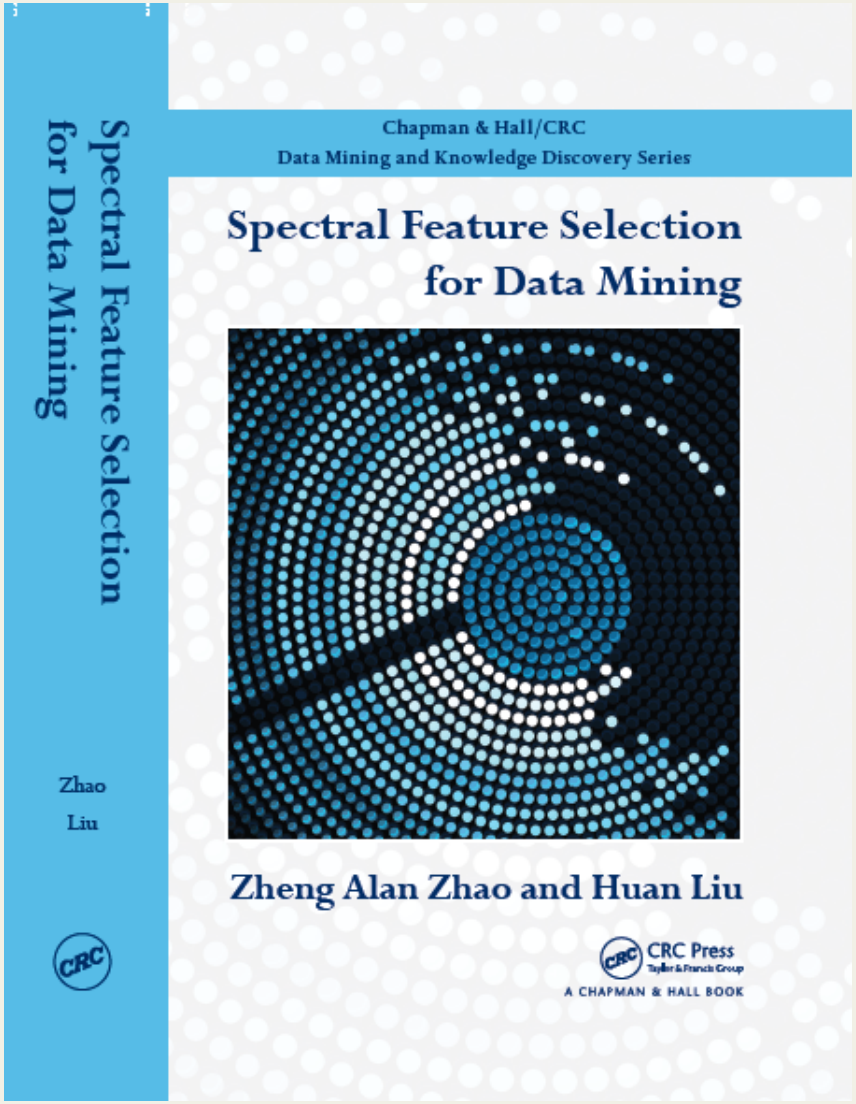
- Direct evaluation (if we know *a priori* …)
  - Often suitable for artificial data sets
  - Based on prior knowledge about data
- Indirect evaluation (if we don't know …)
  - Often suitable for real-world data sets
  - Based on **a)** number of features selected, **b)** performance on selected features (e.g., predictive accuracy, goodness of resulting clusters), and **c)** speed

Methods for Result Evaluation



- Learning curves
  - For results in the form of a ranked list of features
- Before-and-after comparison
  - For results in the form of a minimum subset
- Comparison using different classifiers
  - To avoid learning bias of a particular classifier
- Repeating experimental results
  - For non-deterministic results

A Recent Book for Further Information



- Six Chapters
  1. Data of High Dimensionality and Challenges
  2. Univariate Formulation of Spectral Feature Selection (SFS)
  3. Multivariate Formulations
  4. Connections to Existing Algorithms
  5. Large-Scale SFS
  6. Multi-Source SFS

  Algorithms with software are available at dmml.asu.edu/sfs

Data Mining & Machine Learning

Arizona State University

From Attribute-Value Data to Linked Data

- We are living in an increasingly connected world

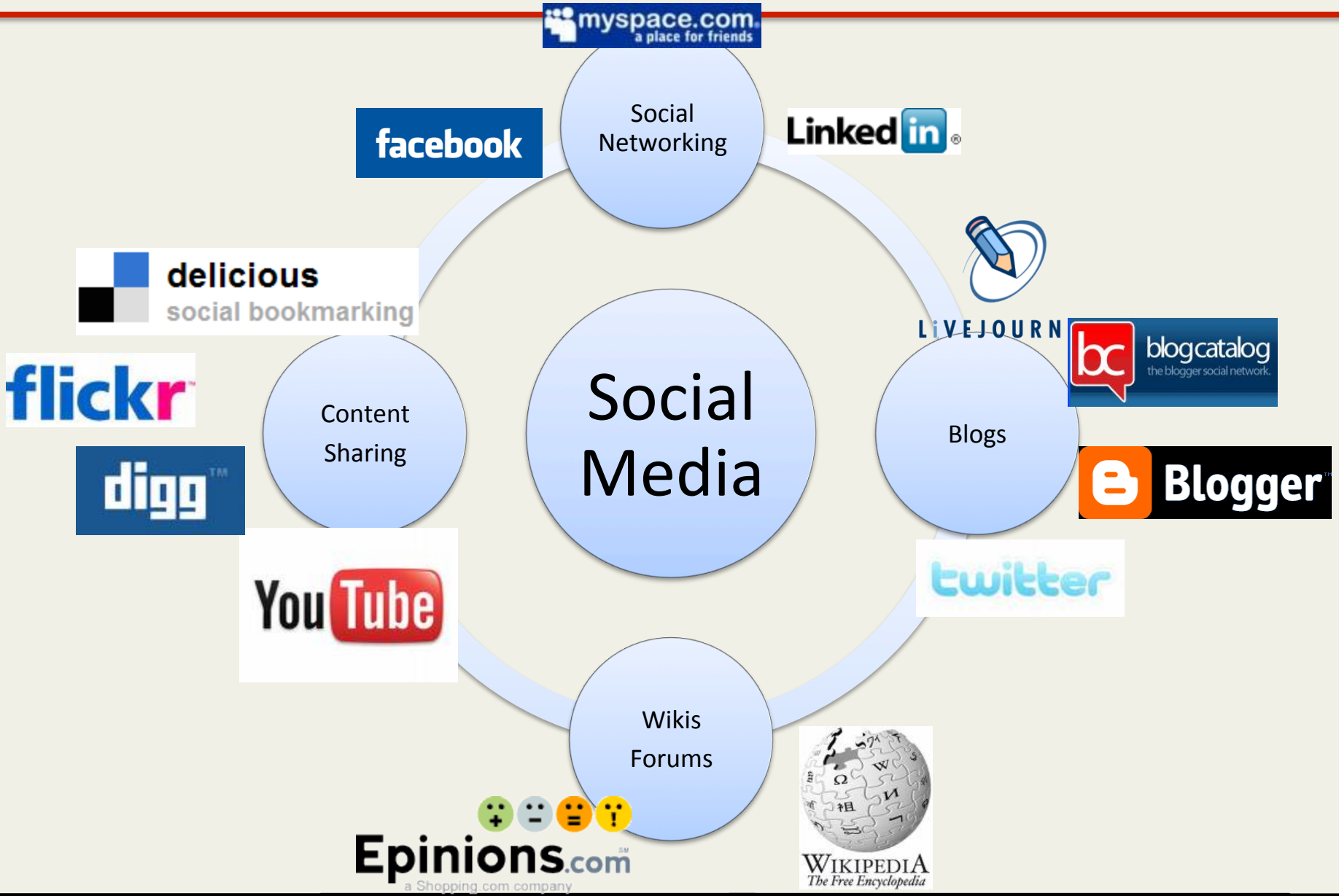
Traditional Media and Data

Broadcast Media
One-to-Many

Communication Media
One-to-One

Traditional Data

Linked Data in the Age of Social Media



Social Media: Many-to-Many

- Everyone can be a media outlet or producer
- Disappearing communication barrier
- Distinct characteristics
  - User generated content: Massive, dynamic, extensive, instant, and noisy
  - Rich user interactions: Linked data
  - Collaborative environment: Wisdom of the crowd
  - Many small groups: The long tail phenomenon; and
  - Attention is hard to get

Noise Removal Fallacy in Social Media

- We often learn that:
  - Noise should be removed before data mining; and
  - "99% Twitter data is useless."
    - "Had eggs, sunny-side-up, this morning"
- Can we remove noise as we usually do in DM?
- What is left after noise removal?
  - Twitter data can be rendered useless after conventional noise removal
- As we are certain there is noise in data and there is a peril of removing it, what can we do?

Linked Data and Attribute-Value Data

- They exist for different purposes
  - Relations, Connections, or Links
  - Properties, Content, etc.
- Classic machine learning and data mining methods assume "independent, identically distributed" or i.i.d. property for attribute-value data
- Additional challenges with the confluence of attribute-value and linked data
  - User-generated
  - Large
  - Noisy, short, incomplete
  - Unstructured, or free form

Feature Selection for Social Media Data

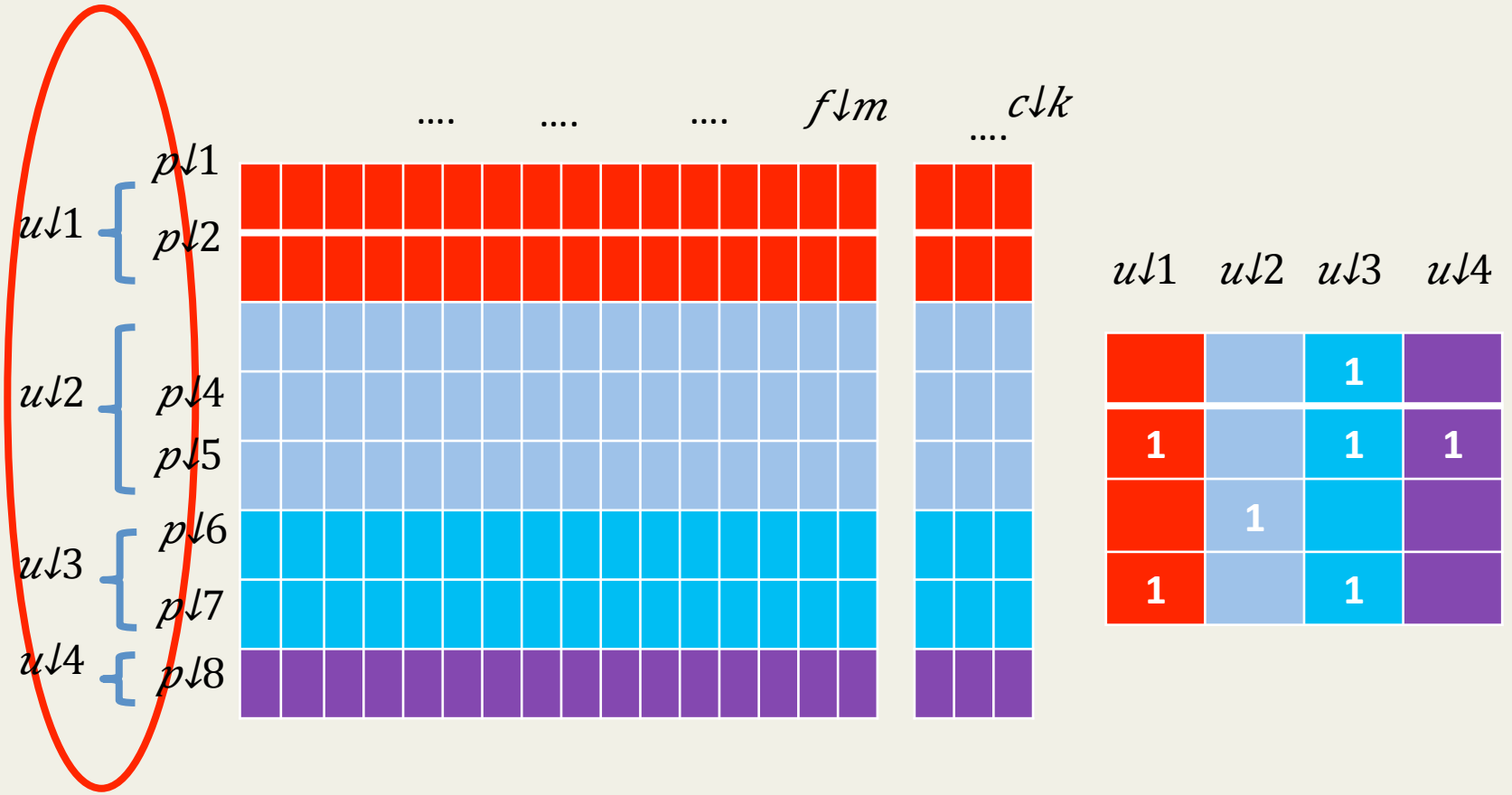
- Massive and high-dimensional social media data poses unique challenges to data mining tasks
  - Scalability
  - Curse of dimensionality
- Social media data is inherently linked
  - A key difference between social media data and attribute-value data

Jiliang Tang and Huan Liu. ``Feature Selection with Linked Data in Social Media'', SIAM International Conference on Data Mining (SDM), 2012.

Feature Selection of Social Media Data

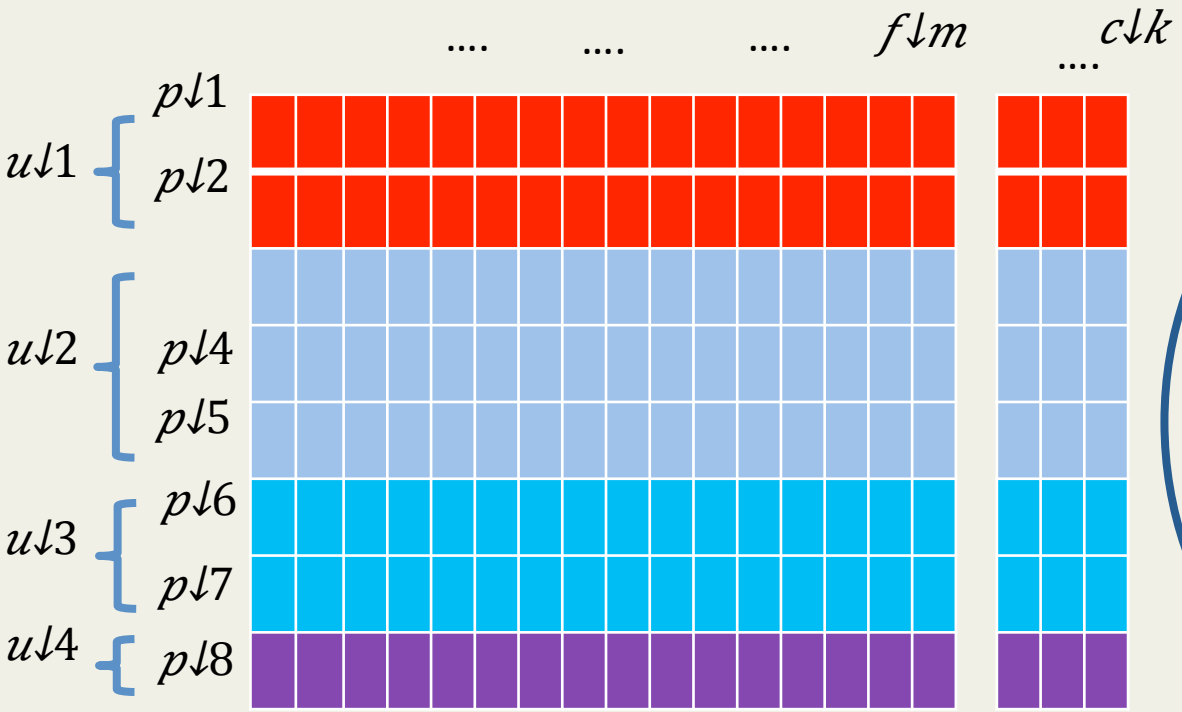
- Feature selection has been widely used to prepare large-scale, high-dimensional data for effective data mining
- Traditional feature selection algorithms deal with only “flat" data (*attribute-value data*).
  - Independent and Identically Distributed (i.i.d.)
- We need to take advantage of linked data for feature selection

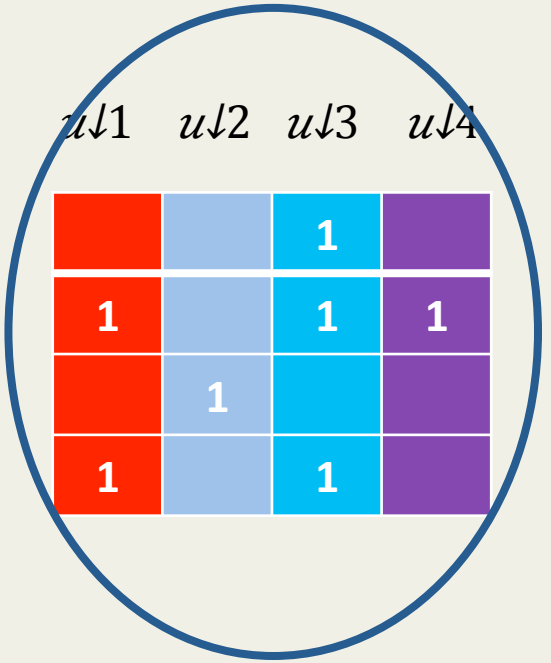
Representation for Social Media Data



User-post relations

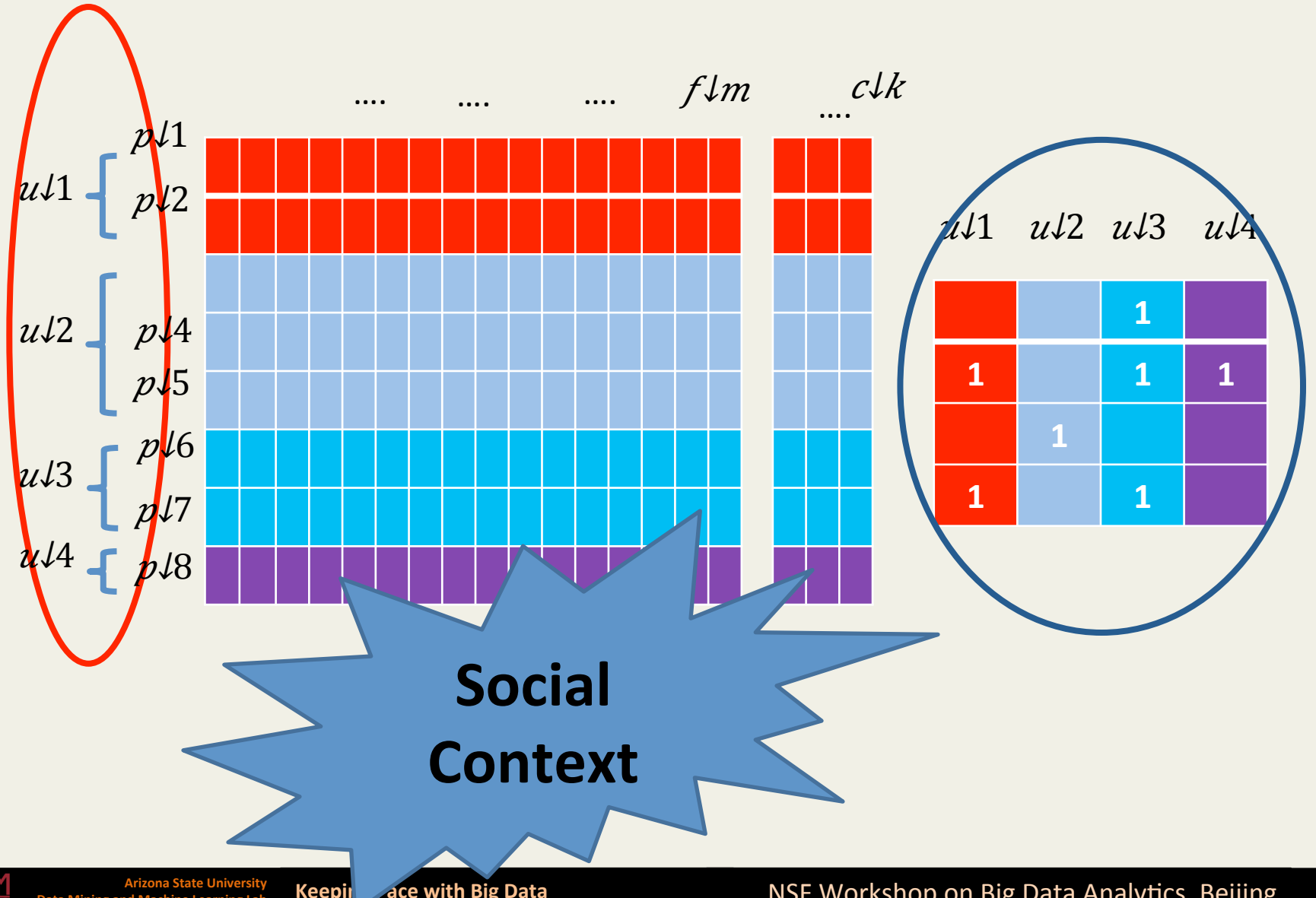
Representation for Social Media Data





User-user relations

Representation for Social Media Data



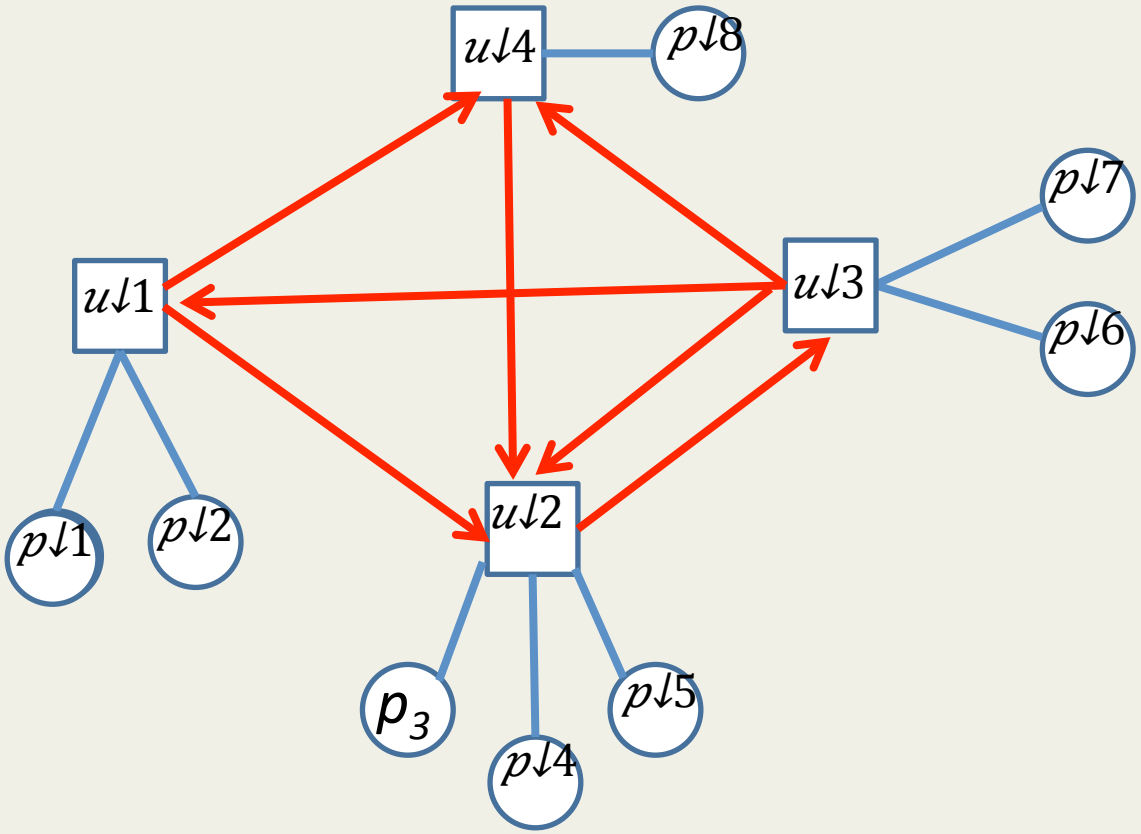
Problem Statement

- Given labeled data X and its label indicator matrix Y, the dataset F, its social context including user-user following relationships *S* and user-post relationships *P*,
- Select ***k*** most relevant features from ***m*** features on dataset *F* with its social context *S* and *P*

How to Use Link Information

- The new question is how to proceed with additional information for feature selection
- Two basic technical problems
  - Relation extraction: What are distinctive relations that can be extracted from linked data
  - Mathematical representation: How to use these relations in feature selection formulation
- Do we have theories to guide us in this effort?

Relation Extraction



1. CoPost
2. CoFollowing
3. CoFollowed
4. Following

Relations, Social Theories, Hypotheses

- Social correlation theories suggest that the four relations may affect the relationships between posts
- Social correlation theories
  - Homophily: People with similar interests are more likely to be linked
  - Influence: People who are linked are more likely to have similar interests
- Thus, four relations lead to four hypotheses for verification

Modeling CoFollowing Relation

- Two co-following users have similar topics of interests

Users' topic interests

$$\hat{T}(u_k) = \frac{\sum_{f_i \in F_k} T(f_i)}{|F_k|} = \frac{\sum_{f_i \in F_k} W^T f_i}{|F_k|}$$

$$\min_{\mathrm{W}} \| \mathrm{X}^T \mathrm{W} - \mathrm{Y} \|_F^2 + \alpha \| \mathrm{W} \|_{2,1} + \beta \sum_u \sum_{u_i, u_j \in N_u} \| \hat{T}(u_i) - \hat{T}(u_j) \|_2^2$$

Evaluation Results on Digg

Table 3: Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	**58.82**	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	**59.43**	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	**63.32**
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	**66.19**
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	**58.90**	57.76	57.01	**58.90**
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	**65.02**
	200	59.97	57.37	60.07	64.36	**67.33**	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	**69.52**	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	**60.77**
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	**65.74**
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	**71.32**
	300	61.47	62.35	64.77	69.58	77.86	71.40	70.50	**78.65**
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	**61.51**	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	**65.65**
	200	60.49	62.78	65.18	66.87	**69.75**	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	**73.01**	70.99	69.50	72.64

Evaluation Results on Digg

Table 3: Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	**58.82**	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	**59.43**	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	**63.32**
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	**66.19**
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	**58.90**	57.76	57.01	**58.90**
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	**65.02**
	200	59.97	57.37	60.07	64.36	**67.33**	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	**69.52**	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	**60.77**
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	**65.74**
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	**71.32**
	300	61.47	62.35	64.77	69.58	77.86	71.40	70.50	**78.65**
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	**61.51**	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	**65.65**
	200	60.49	62.78	65.18	66.87	**69.75**	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	**73.01**	70.99	69.50	72.64

Summary

- LinkedFS is evaluated under varied circumstances to understand how it works.
  - Link information can help *feature selection for social media data.*
- Unlabeled data is more often in social media, unsupervised learning is more sensible, but also more challenging.

Jiliang Tang and Huan Liu. `` Unsupervised Feature Selection for Linked Social Media Data'', the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2012.

Jiliang Tang, Huan Liu. ``Feature Selection with Linked Data in Social Media'', SIAM International Conference on Data Mining, 2012.

Looking Ahead

- New, rich data sources like social media present challenges and opportunities
  - Feature selection is shown here for illustration
- Challenges abound
  - Data collection (sampling bias, is data enough?)
  - Data preparation (what is noise?)
  - Pattern discovery (content, context, networks)
  - Evaluation (when without ground truth)
- Big data allows more opportunities for researchers of different disciplines to conduct collaborative research

Thank You …

- For this opportunity to share our research
- Acknowledgments
  - Grants from NSF, ONR, and ARO, among others
  - DMML members and project leaders
  - Collaborators

Concluding Remarks

- Big Data is a good problem to have
- Data mining is one way of approaching it
- Together, we can harness it for better sci & eng