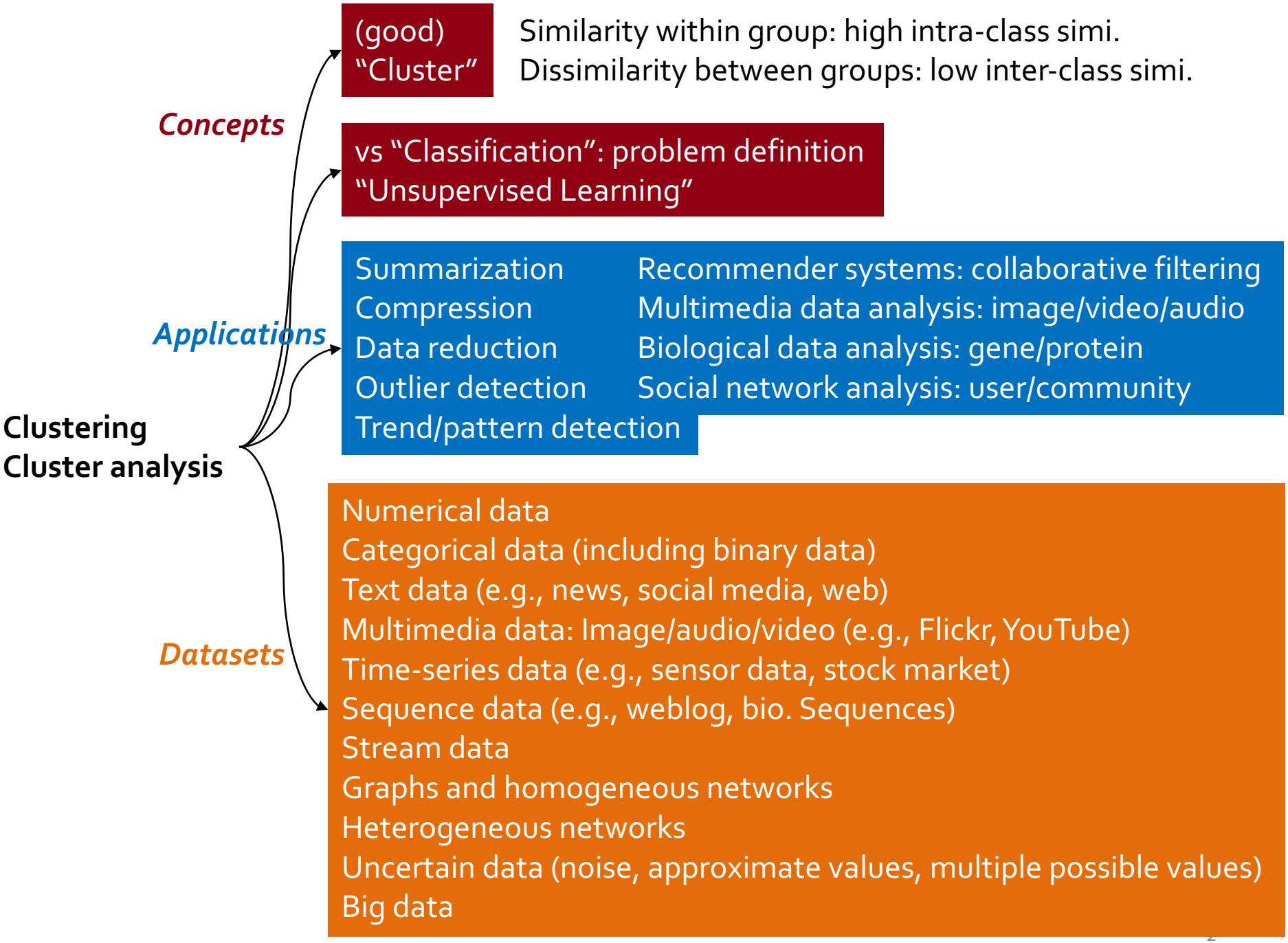
Chapter 10. Clustering: Concepts

Meng Jiang

Data Science


Clustering
Cluster analysis
Concepts
(good) "Cluster"
Similarity within group: high intra-class simi.
Dissimilarity between groups: low inter-class simi.
vs "Classification": problem definition
"Unsupervised Learning"
Applications
Summarization
Compression
Data reduction
Outlier detection
Trend/pattern detection
Recommender systems: collaborative filtering
Multimedia data analysis: image/video/audio
Biological data analysis: gene/protein
Social network analysis: user/community
Datasets
Numerical data
Categorical data (including binary data)
Text data (e.g., news, social media, web)
Multimedia data: Image/audio/video (e.g., Flickr, YouTube)
Time-series data (e.g., sensor data, stock market)
Sequence data (e.g., weblog, bio. Sequences)
Stream data
Graphs and homogeneous networks
Heterogeneous networks
Uncertain data (noise, approximate values, multiple possible values)
Big data

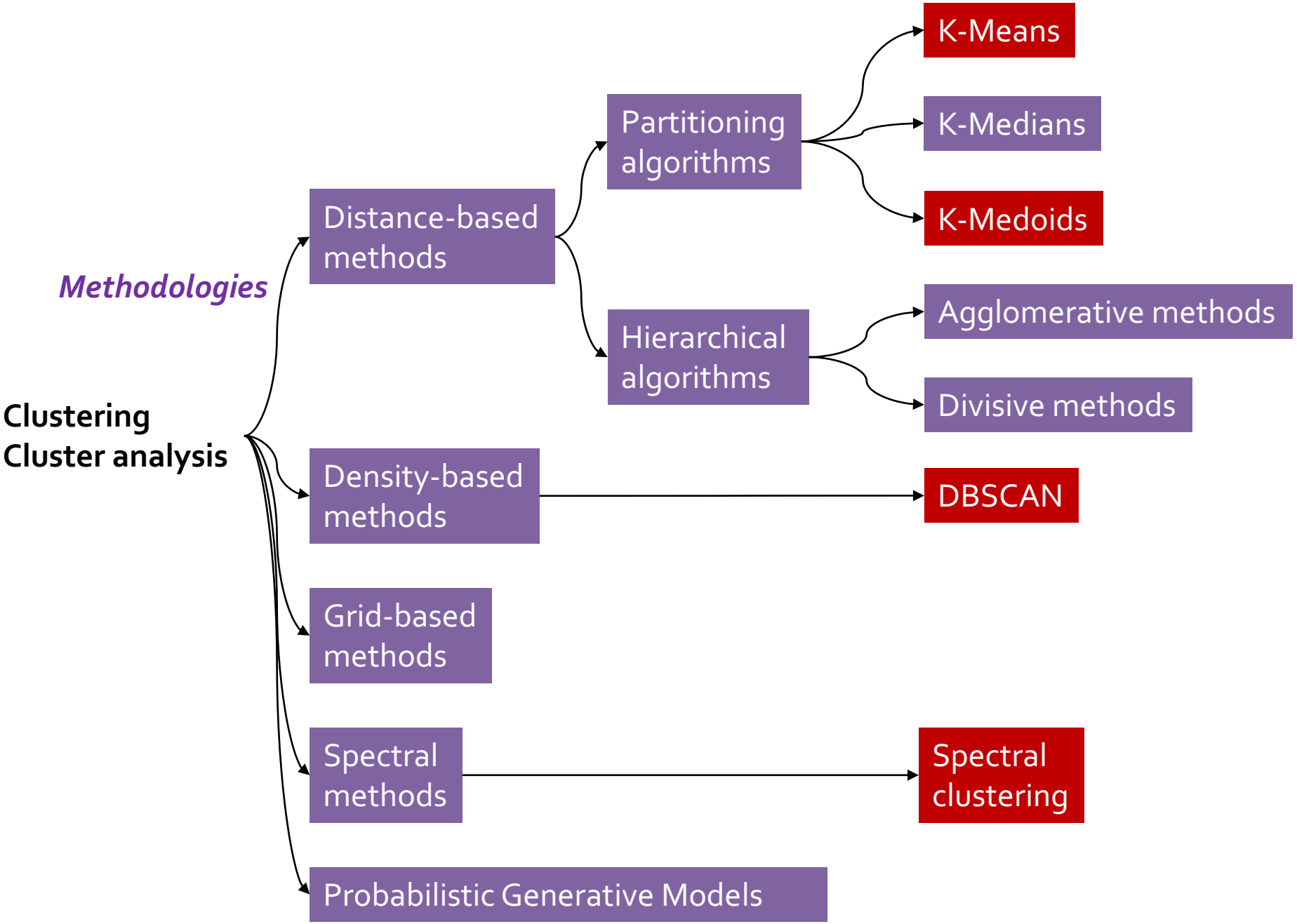
Clustering
Cluster analysis

Types of tasks

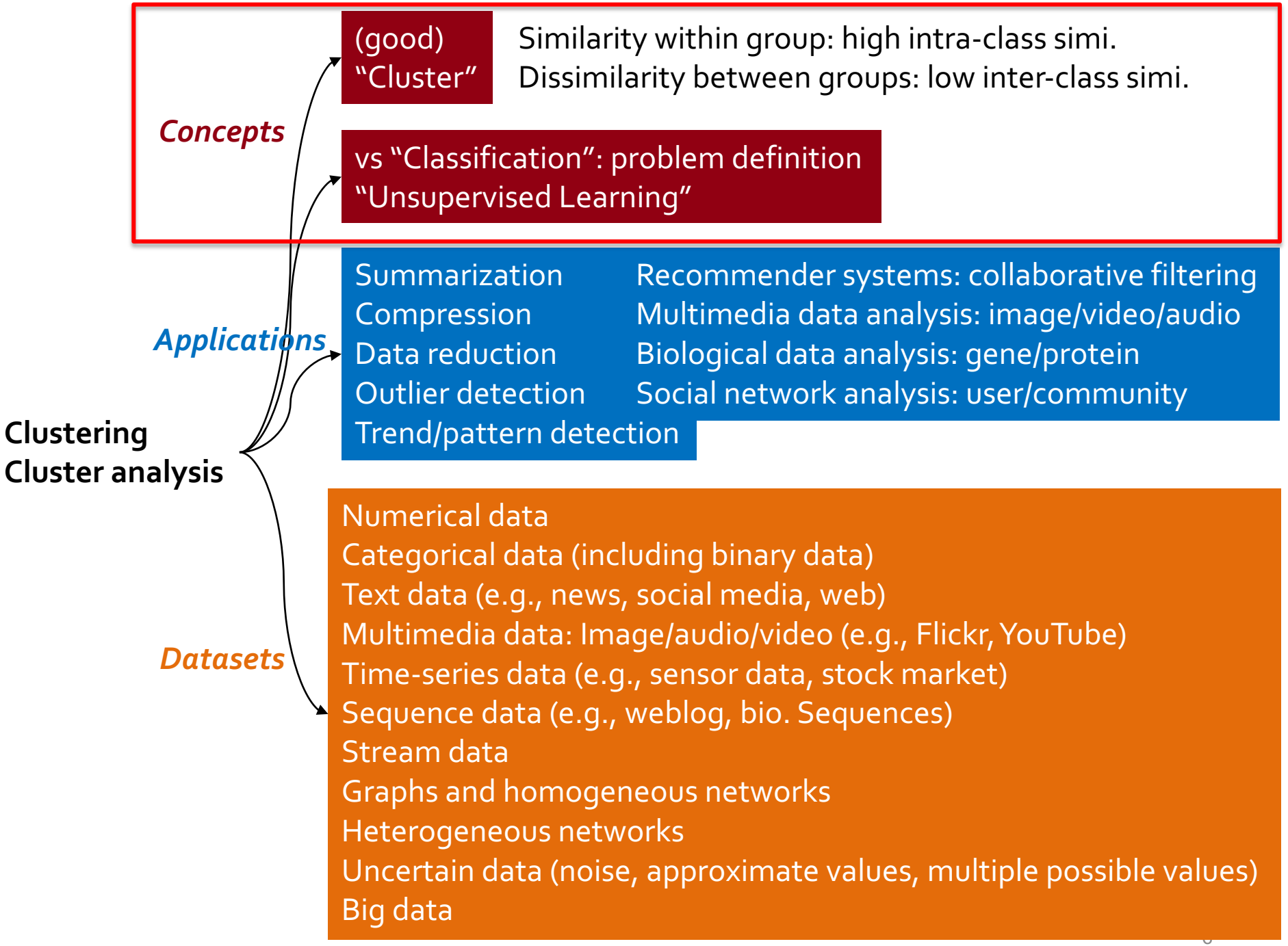
- Partitioning criteria
  - (Single-level) partitional clustering
  - (High-level) hierarchical clustering
- Separation of clusters
  - Exclusive clustering
  - Non-exclusive (overlapping) clustering
- Outlier vs Cluster
  - Complete clustering
  - Partial clustering
- Similarity measure
  - Distance-based (e.g., Euclidean)
  - Connectivity-based (e.g., density)
- Clustering space
  - Full space (for low dimensional clustering)
  - Subspaces (for high dimensional clustering)

Challenges

- Quality: Different types? Clusters with arbitrary shape? Noisy data?
- Scalability: All data samples. High dimensionality.
- Constraint-based clustering: User-given preferences?
- Interpretability and usability

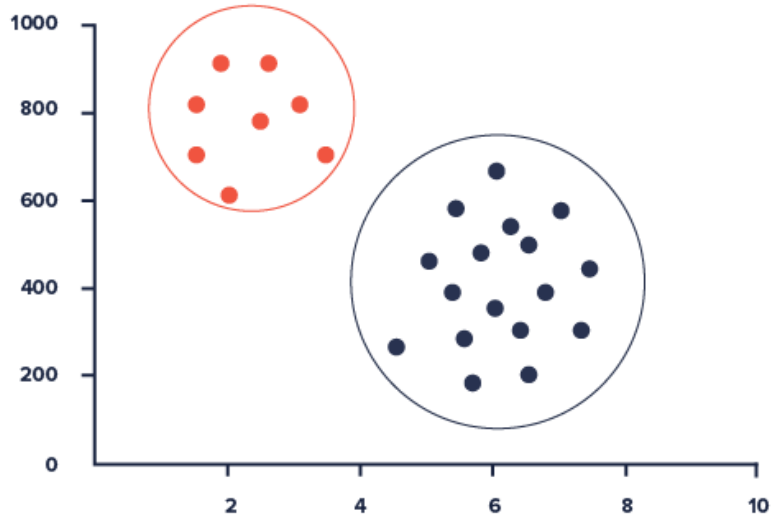

Methodologies
Clustering
Cluster analysis
Distance-based methods
Partitioning algorithms
K-Means
K-Medians
K-Medoids
Hierarchical algorithms
Agglomerative methods
Divisive methods
Density-based methods
DBSCAN
Grid-based methods
Spectral methods
Spectral clustering
Probabilistic Generative Models

Today's Lecture


Clustering
Cluster analysis
Concepts
(good) "Cluster"
Similarity within group: high intra-class simi.
Dissimilarity between groups: low inter-class simi.
vs "Classification": problem definition
"Unsupervised Learning"
Applications
Summarization
Compression
Data reduction
Outlier detection
Trend/pattern detection
Recommender systems: collaborative filtering
Multimedia data analysis: image/video/audio
Biological data analysis: gene/protein
Social network analysis: user/community
Datasets
Numerical data
Categorical data (including binary data)
Text data (e.g., news, social media, web)
Multimedia data: Image/audio/video (e.g., Flickr, YouTube)
Time-series data (e.g., sensor data, stock market)
Sequence data (e.g., weblog, bio. Sequences)
Stream data
Graphs and homogeneous networks
Heterogeneous networks
Uncertain data (noise, approximate values, multiple possible values)
Big data

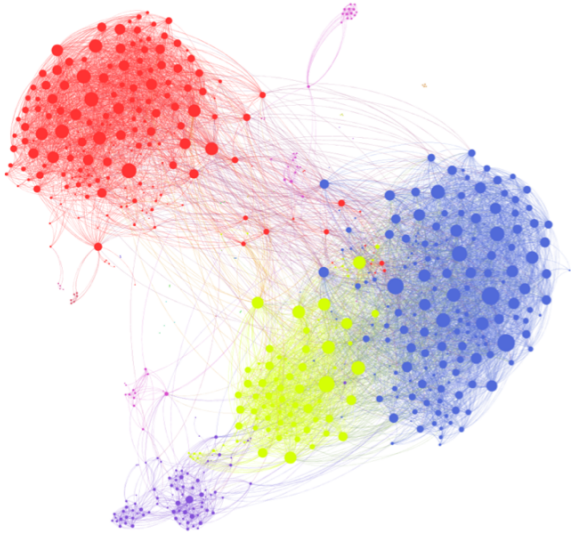
What is Cluster?

- A **cluster** is a collection of data objects which are
  - Similar to one another within the same group
  - Dissimilar to the objects in other groups

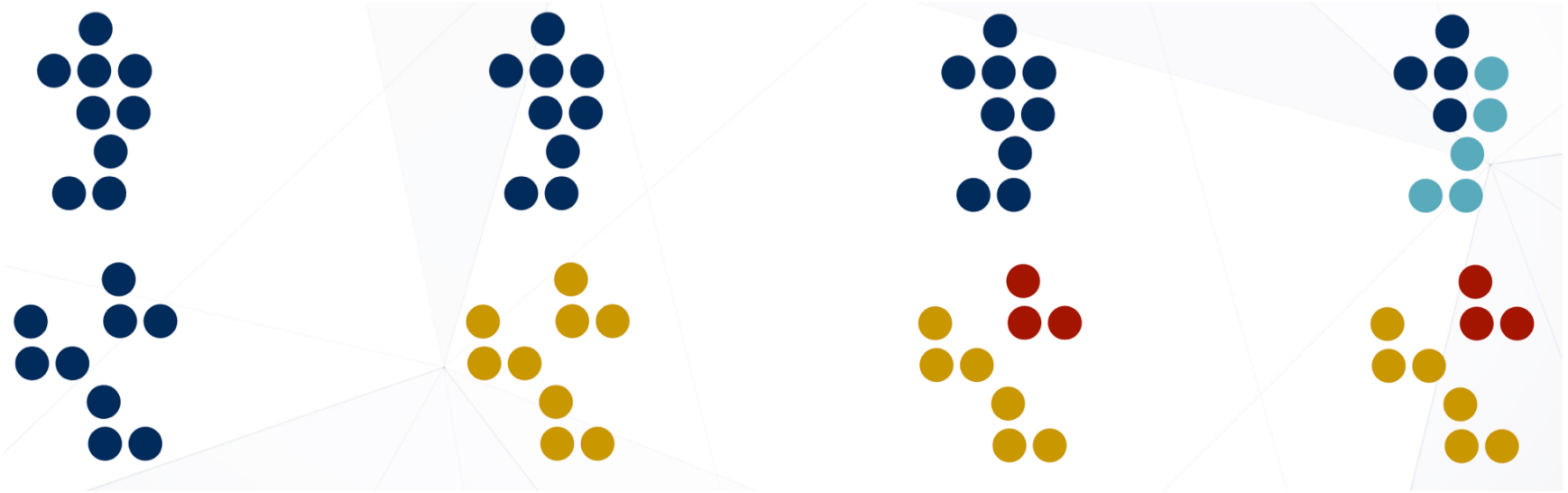


What is Cluster Analysis?

- Cluster Analysis (Clustering)
  - Def. 1. [PARTITIONING] Given a set of data points, **partition** them into a set of **"good clusters"**.
  - Def. 2. [NO LABEL] Grouping data objects based **only** on information found in the data **describing** these objects and their relationships (i.e., object/features or node/links).



What Defines a Good Cluster?



1-cluster 2-cluster 3-cluster 4-cluster

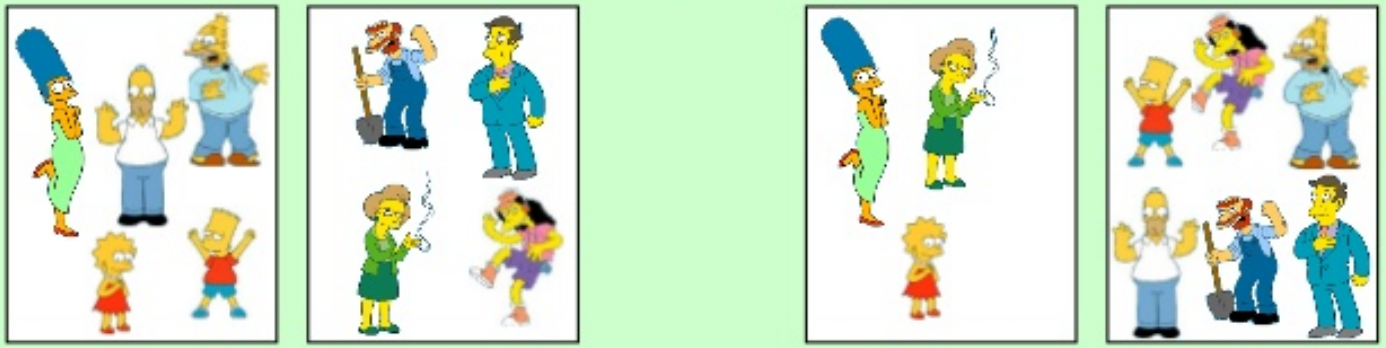
Good Clusters

- A good clustering method will produce high quality clusters which should have
  - **High intra-class similarity:** Cohesive within clusters
  - **Low inter-class similarity:** Distinctive between clusters
- Quality function?
  - There is usually a quality function that measures the "**goodness**" of a cluster
  - Depends on **similarity measure**
  - It is hard to define "similar enough" or "good enough"
    - The answer is typically highly **subjective**
    - Measured by ability to discover **hidden patterns**

Subjective…

What is a natural grouping among these objects?

Clustering is subjective



Similarity and Dissimilarity

- There exist many (dis)similarity measures and/or functions for different applications
- (Dis)similarity measure is critical for cluster analysis
- Similarity
  - Measure of "alikeness" of instances
  - Greater if more alike
  - E.g., Jaccard, Cosine, Pearson Corr., Density
- Dissimilarity
  - Measure of "unalikeness" of instances
  - Lower if more alike
  - Can be expressed as distance function
  - E.g., Euclidean

Formalizing Clustering

Let D denote a dataset containing N observations

$$D = \{\mathbf{x}_i \mid i = 1, 2, \ldots, N\}$$

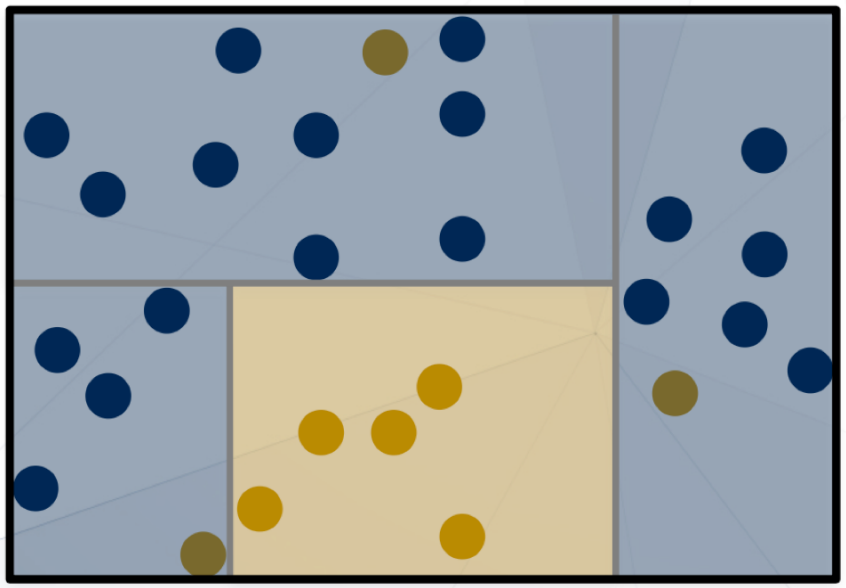
where each $\mathbf{x}_i$ corresponds to the set of features or explanatory variables – discrete or continuous (categorical or numerical) – of the i-th observation.

Then **clustering** is the task of learning a mapping of each feature set $\mathbf{x}$ into a previously undefined grouping.

Classification vs. Clustering

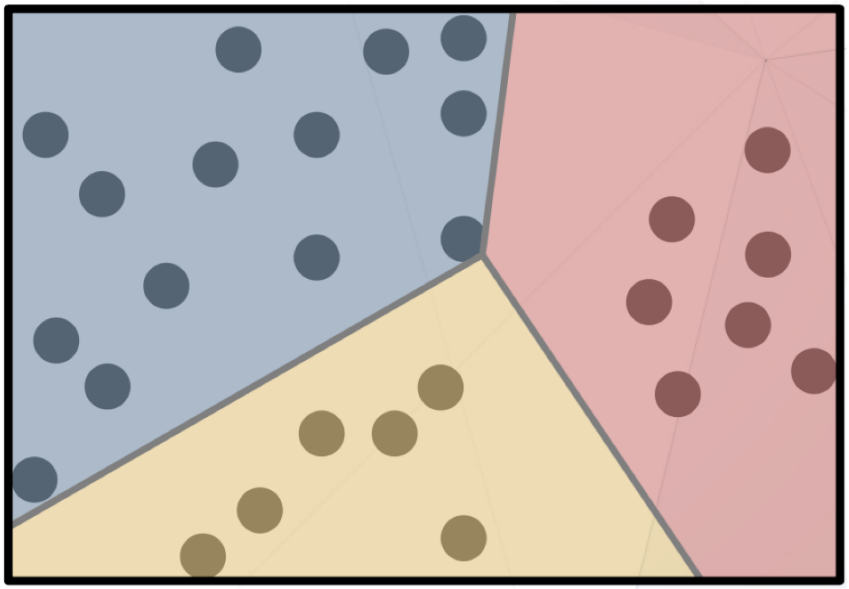
Classification:

Predicts instances classes from pre-defined set of classes.



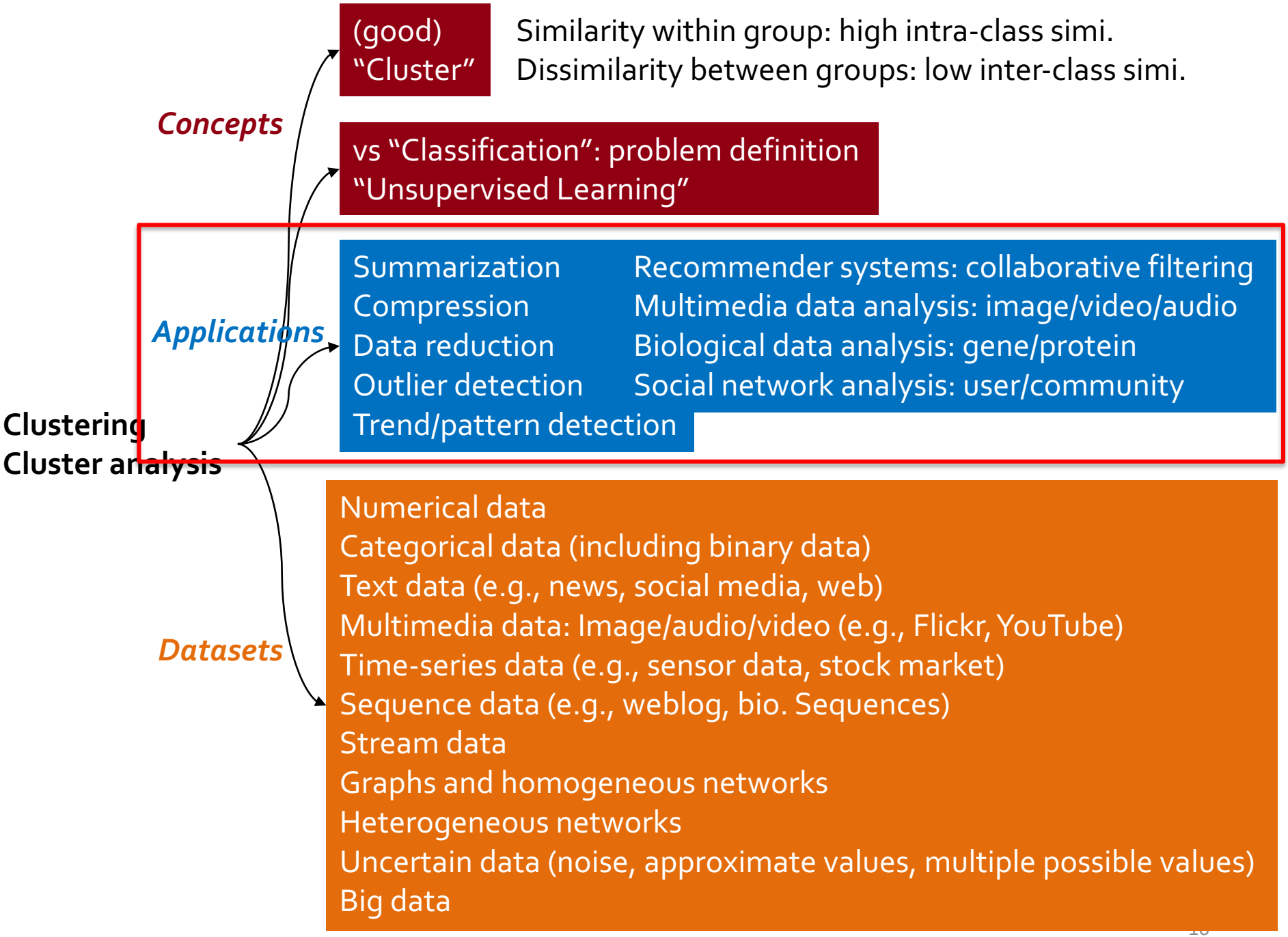
Clustering:

Finds "natural" grouping of instances with out class data.



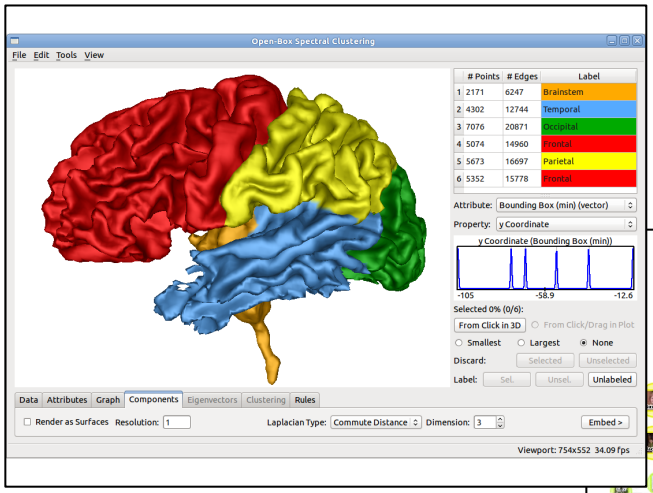
Unsupervised Learning

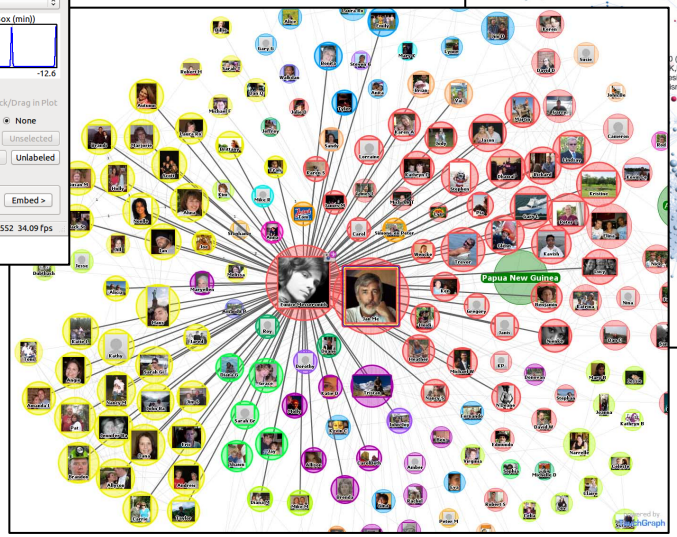
- Clustering
  - Input **set of features *X*** that can take any set of values (categorical or numerical). The goal is to identify a useful organization or grouping of the instances.
- Frequent pattern/association rule mining
  - Input **set of items X**. The goal is to discover a set of regularities, frequent itemsets, or rules between occurrences of items in the database or dataset.

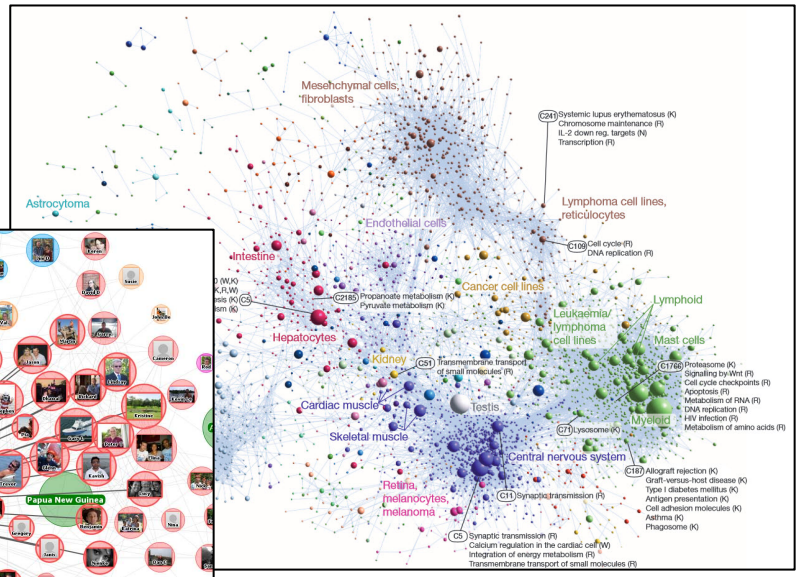

(good) "Cluster"
Similarity within group: high intra-class simi.
Dissimilarity between groups: low inter-class simi.
Concepts
vs "Classification": problem definition
"Unsupervised Learning"
Applications
Summarization
Compression
Data reduction
Outlier detection
Trend/pattern detection
Recommender systems: collaborative filtering
Multimedia data analysis: image/video/audio
Biological data analysis: gene/protein
Social network analysis: user/community
Clustering
Cluster analysis
Datasets
Numerical data
Categorical data (including binary data)
Text data (e.g., news, social media, web)
Multimedia data: Image/audio/video (e.g., Flickr, YouTube)
Time-series data (e.g., sensor data, stock market)
Sequence data (e.g., weblog, bio. Sequences)
Stream data
Graphs and homogeneous networks
Heterogeneous networks
Uncertain data (noise, approximate values, multiple possible values)
Big data

Applications

- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing/intermediate step for other algorithms



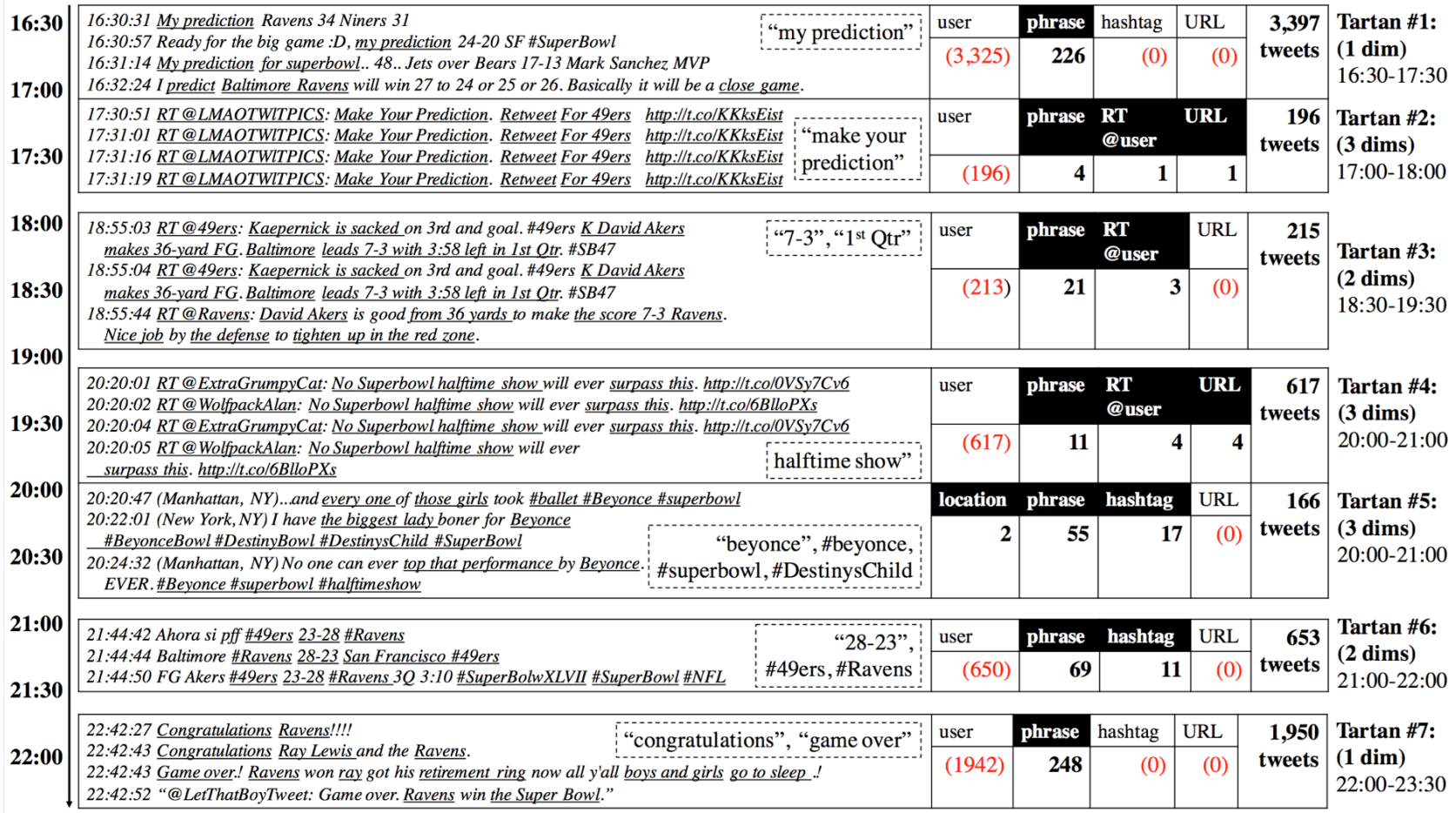




Clustering: Summarization

(to be introduced in details)

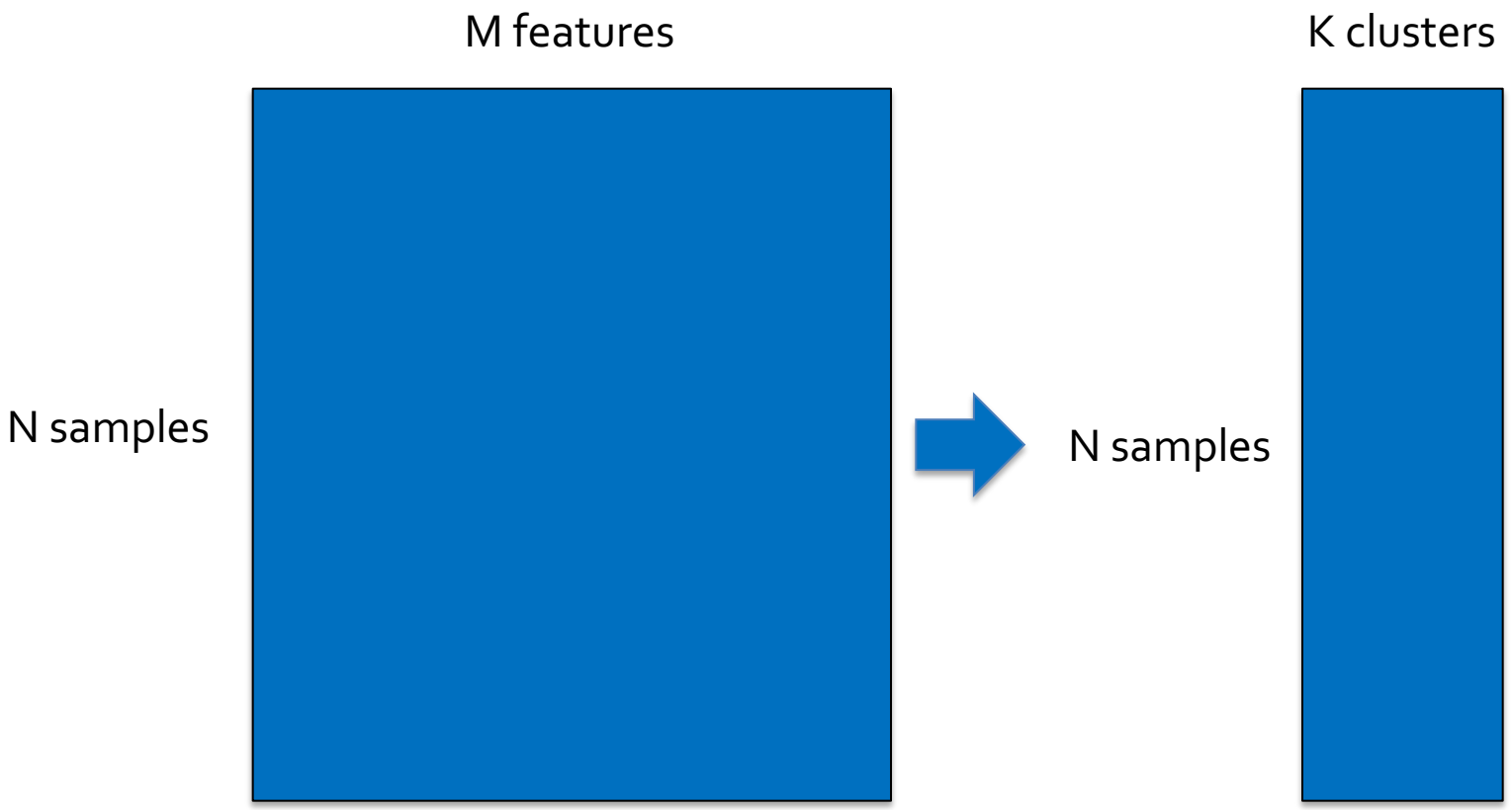
- Generating a **compact summary** of data for classification, pattern discovery, hypothesis generation and testing, etc.



Clustering: Compression

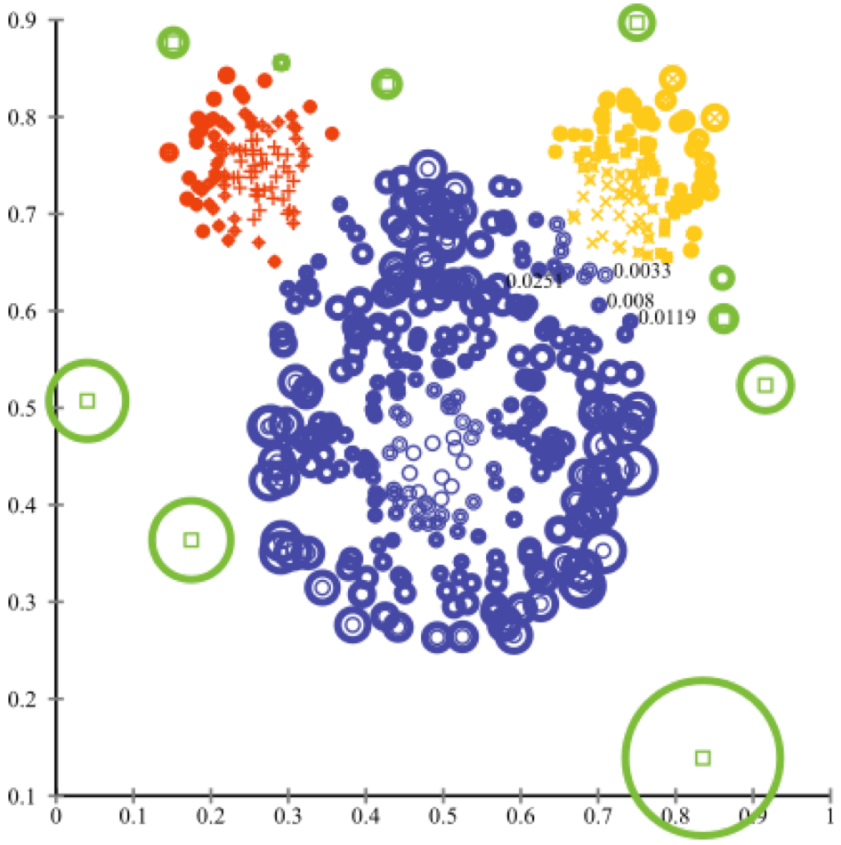
- Image compression

Clustering: Data/Dimension Reduction

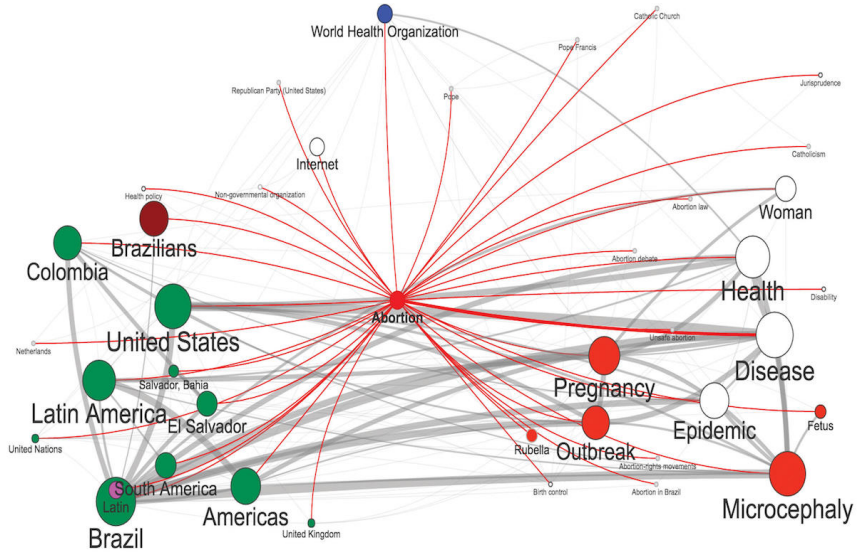


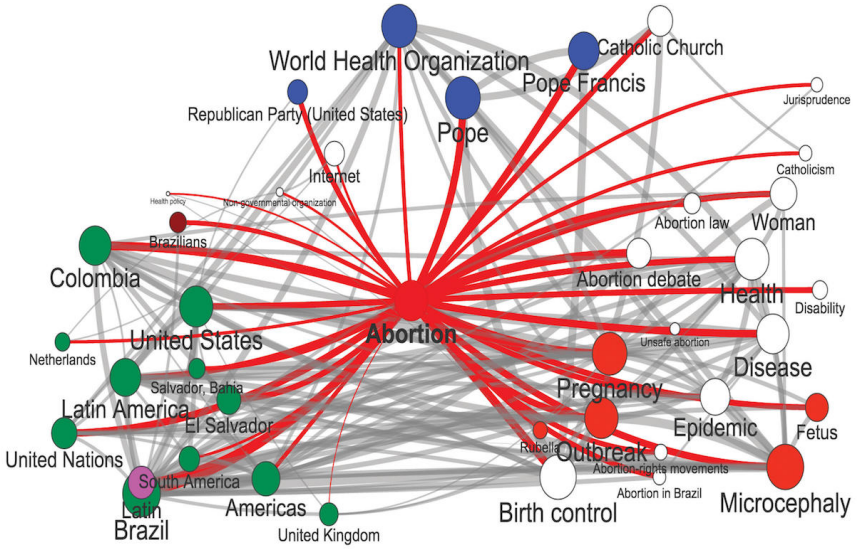
Clustering: Outlier Detection

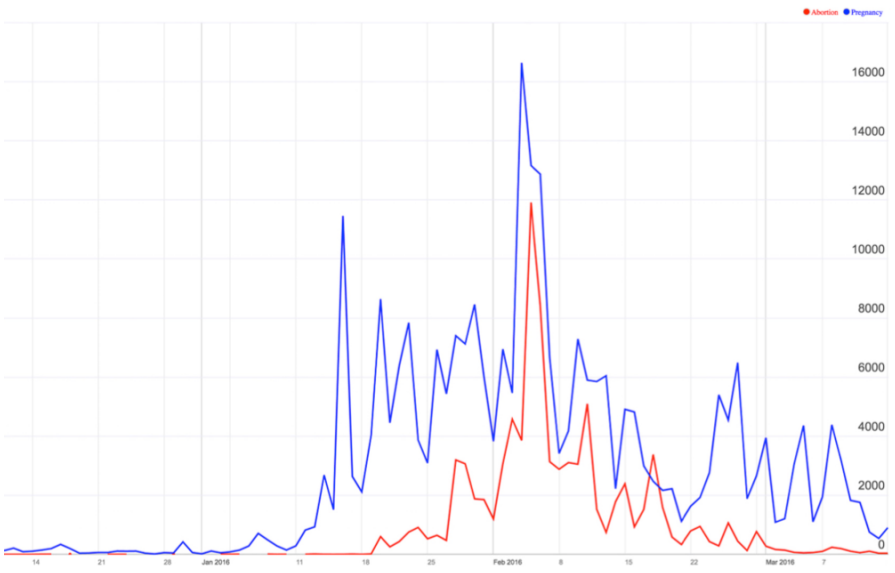
- Outliers—those "far away" from any cluster



Clustering: Trend Detection





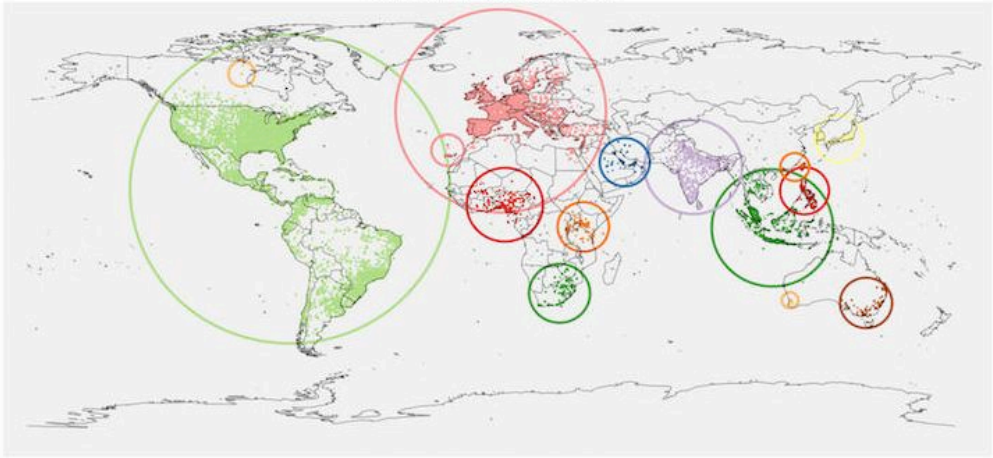


http://publichealth.jmir.org/2017/2/e22/

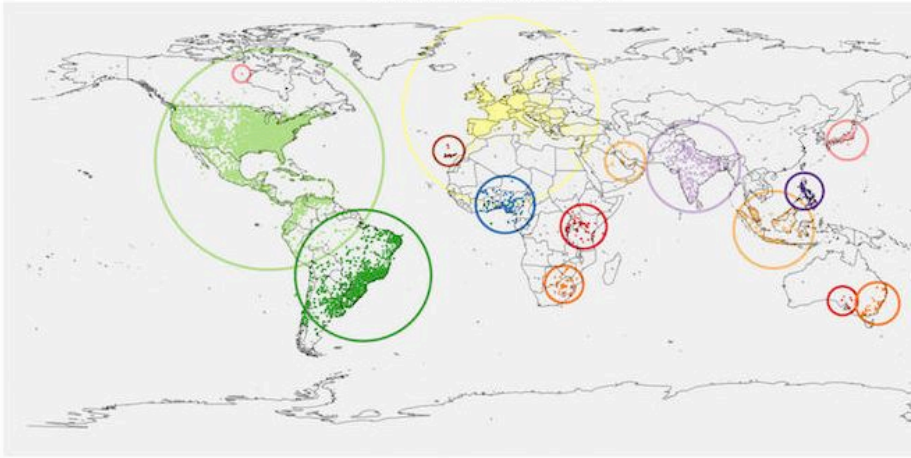
Clustering: Pattern Recognition

2016-02-01 ... 2016-02-07

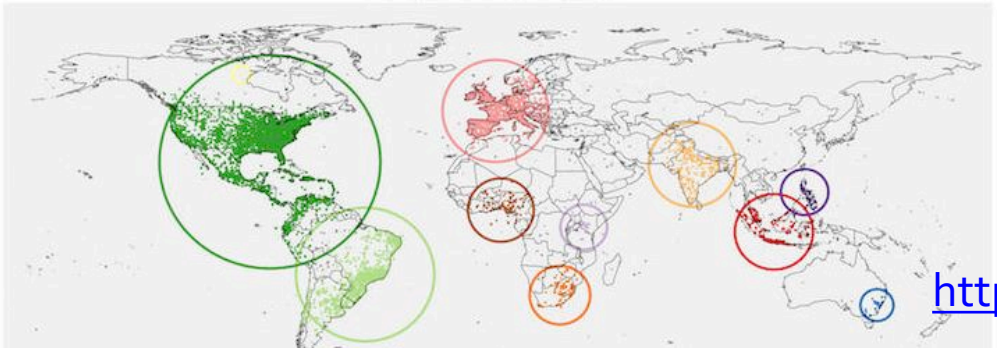
2016-02-08 ... 2016-02-14

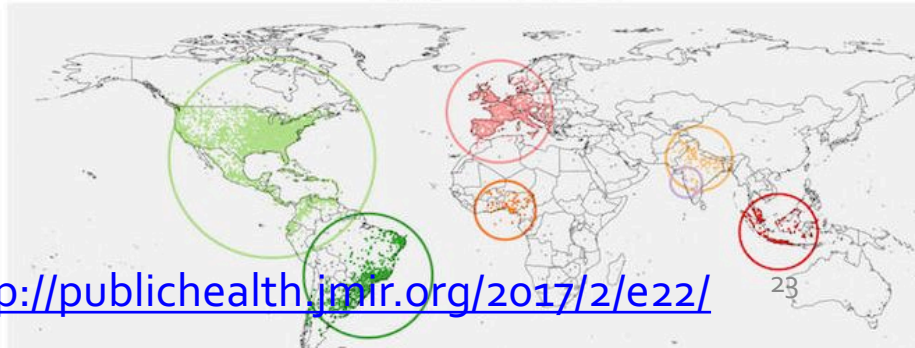


2016-02-15 ... 2016-02-21



2016-02-22 ... 2016-02-28

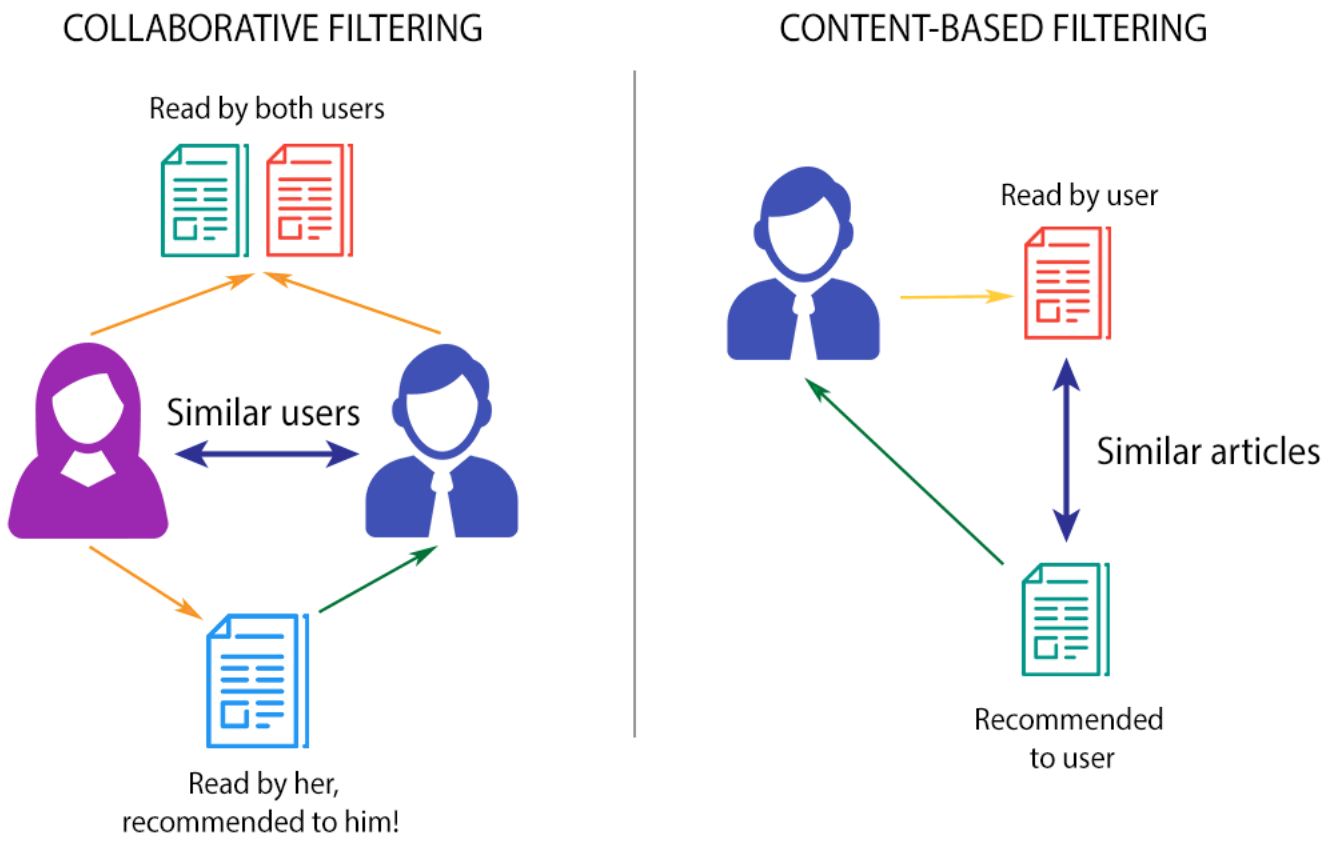




http://publichealth.jmir.org/2017/2/e22/

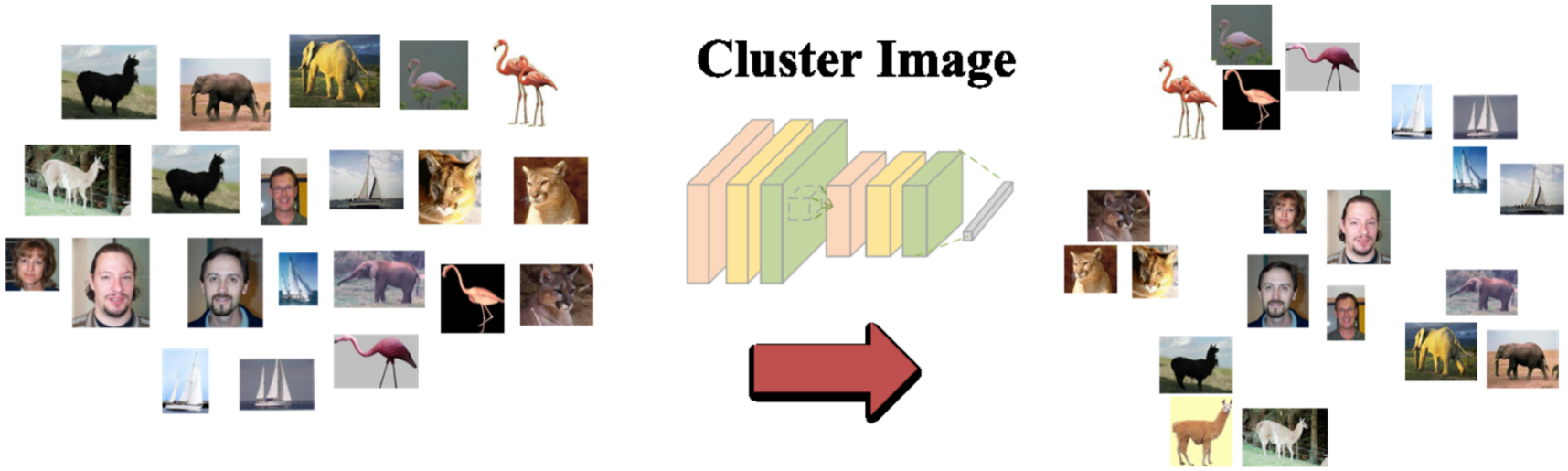
Clustering: Recommender Systems

- Find like-minded users or similar products



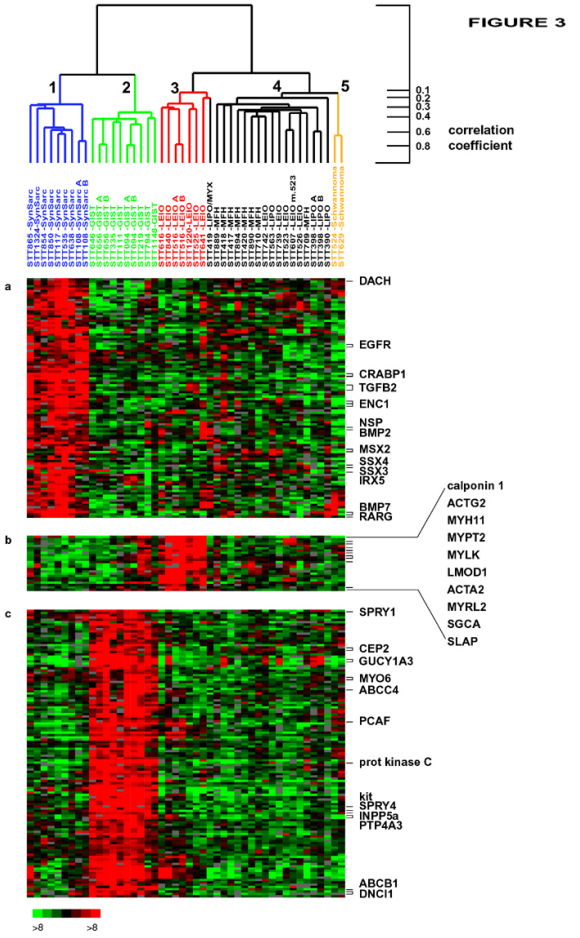
Clustering: Multimedia Data Analysis

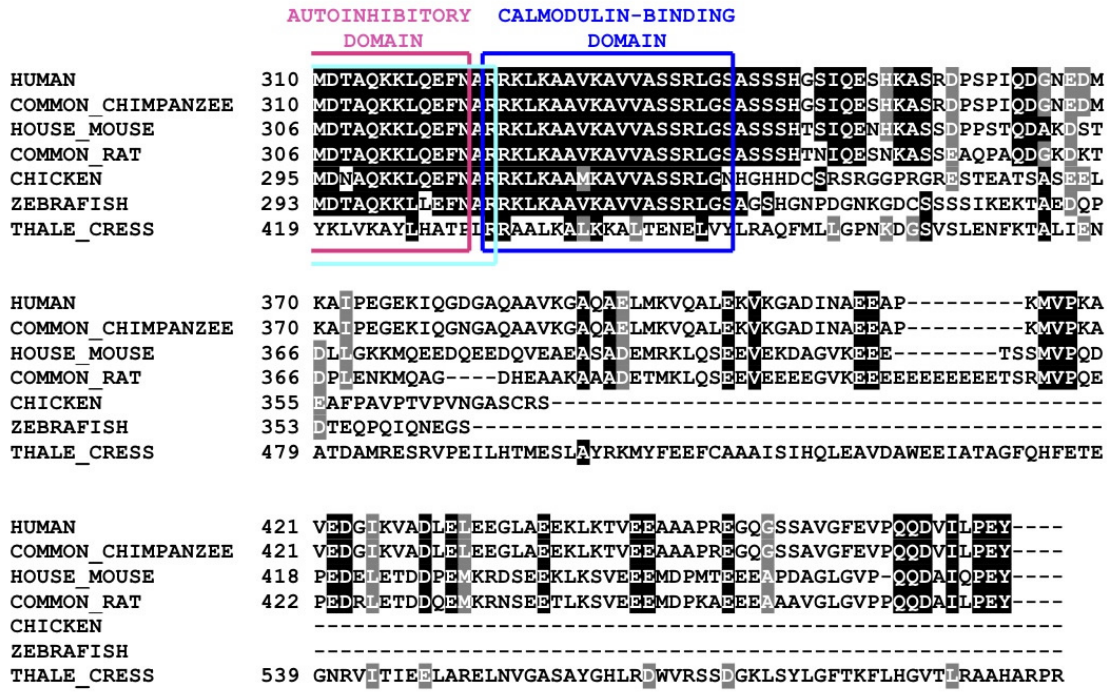
- Clustering images or video/audio clips



Clustering: Biological Data Analysis

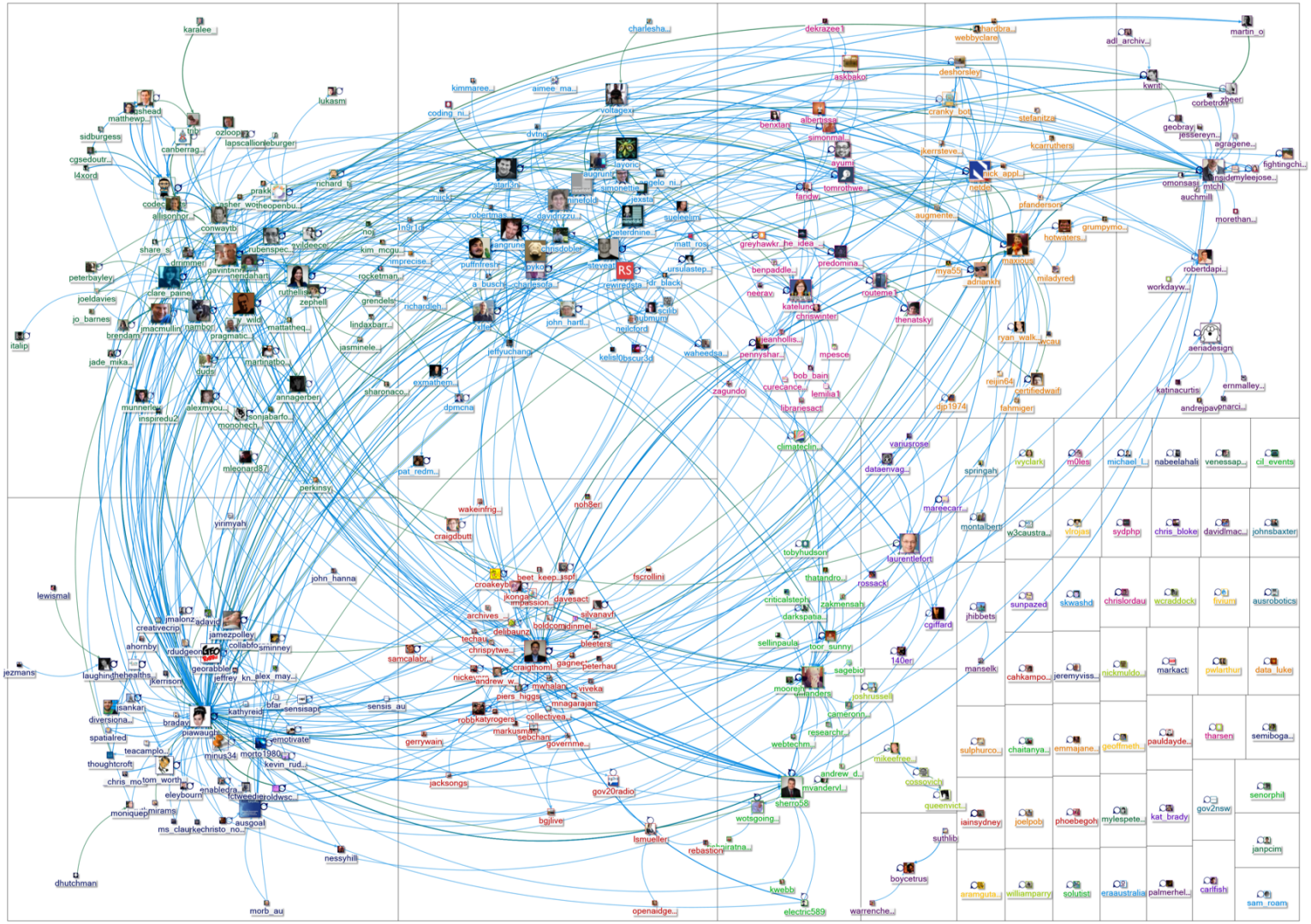
- Clustering gene/protein sequences





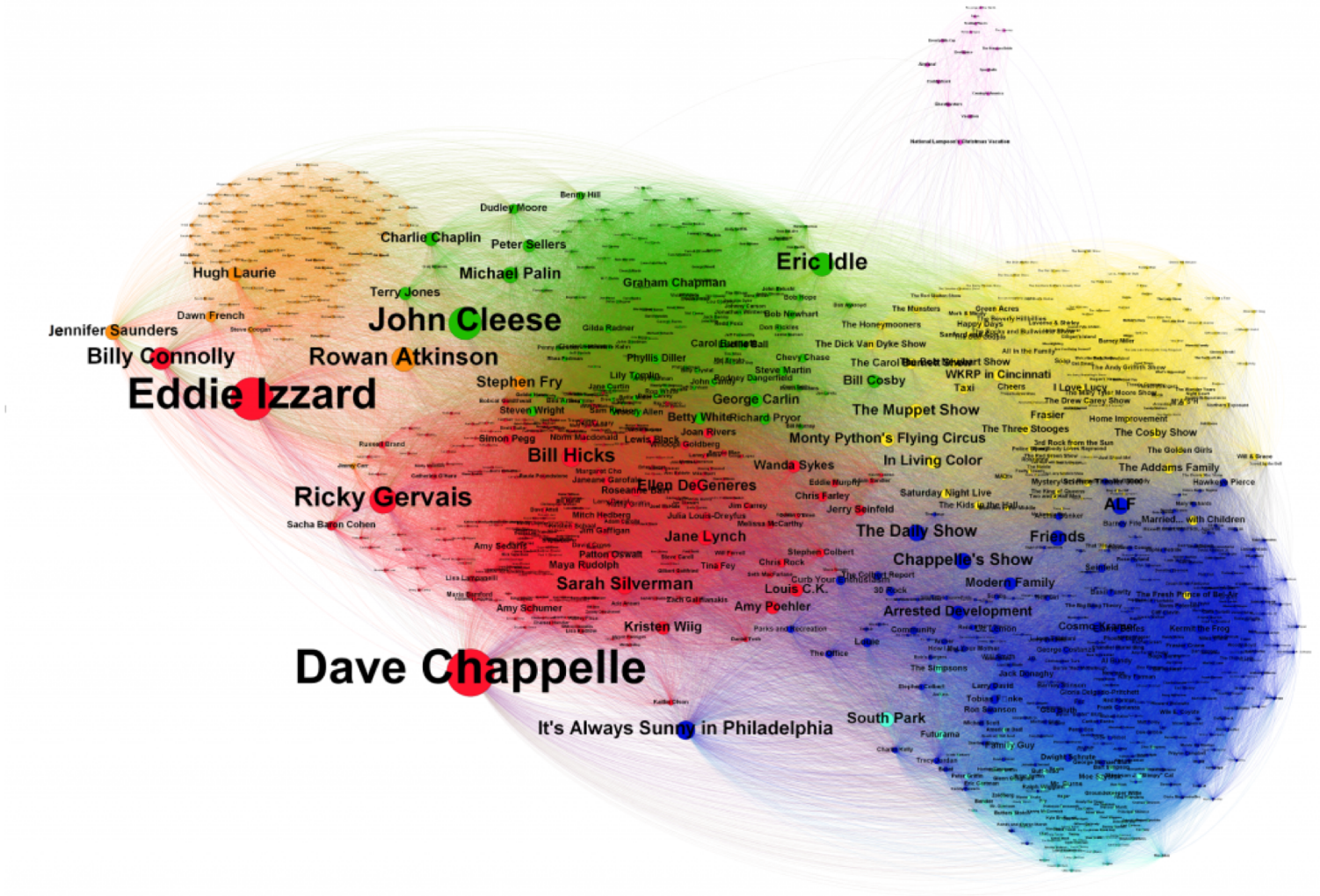
Clustering: Social Network Analysis

Social media network connections among Twitter users

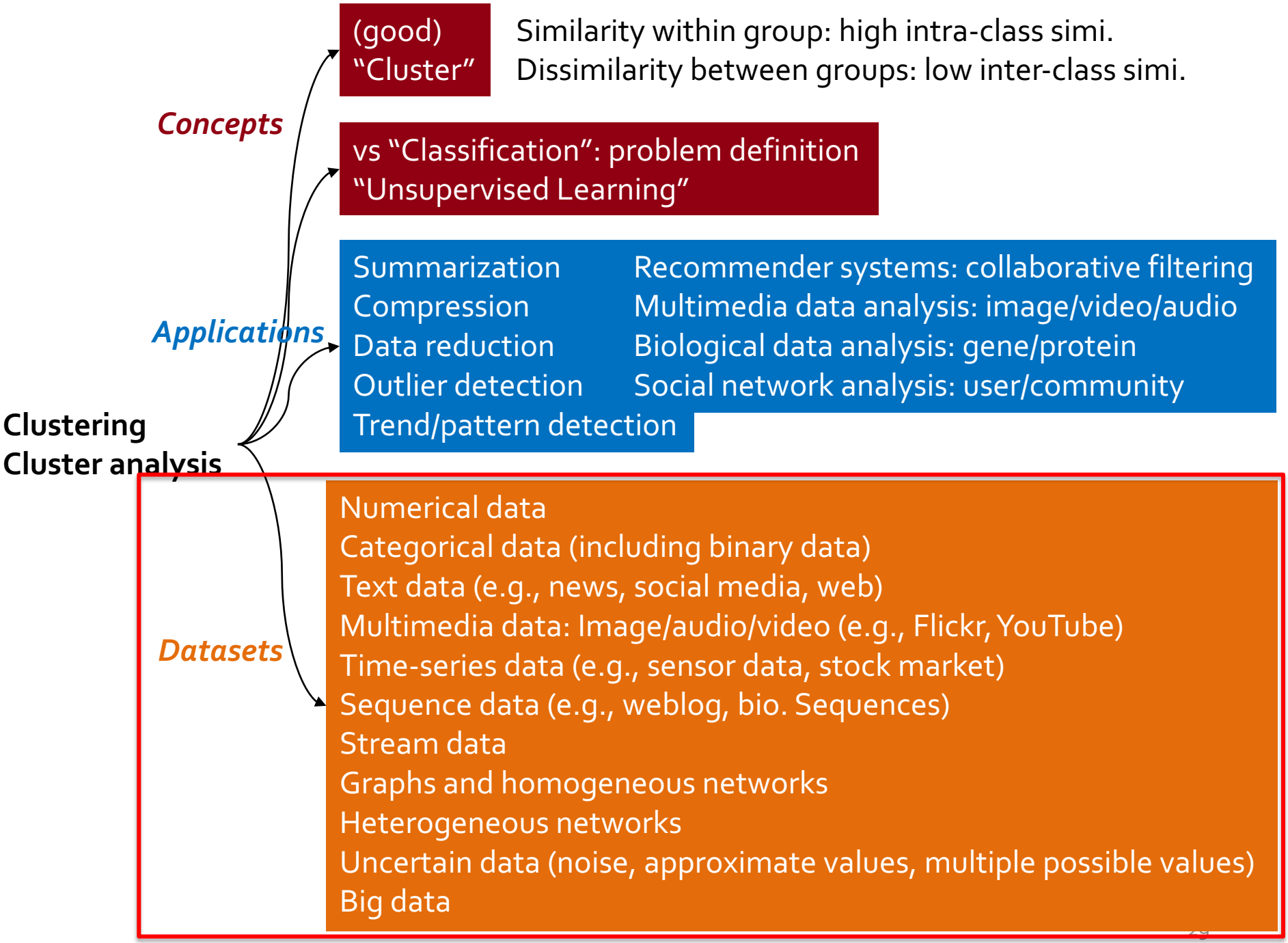


Created with NodeXL (http://nodexl.codeplex.com) from the Social Media Research Foundation (http://www.smrfoundation.org)

Clustering: Actor Clustering



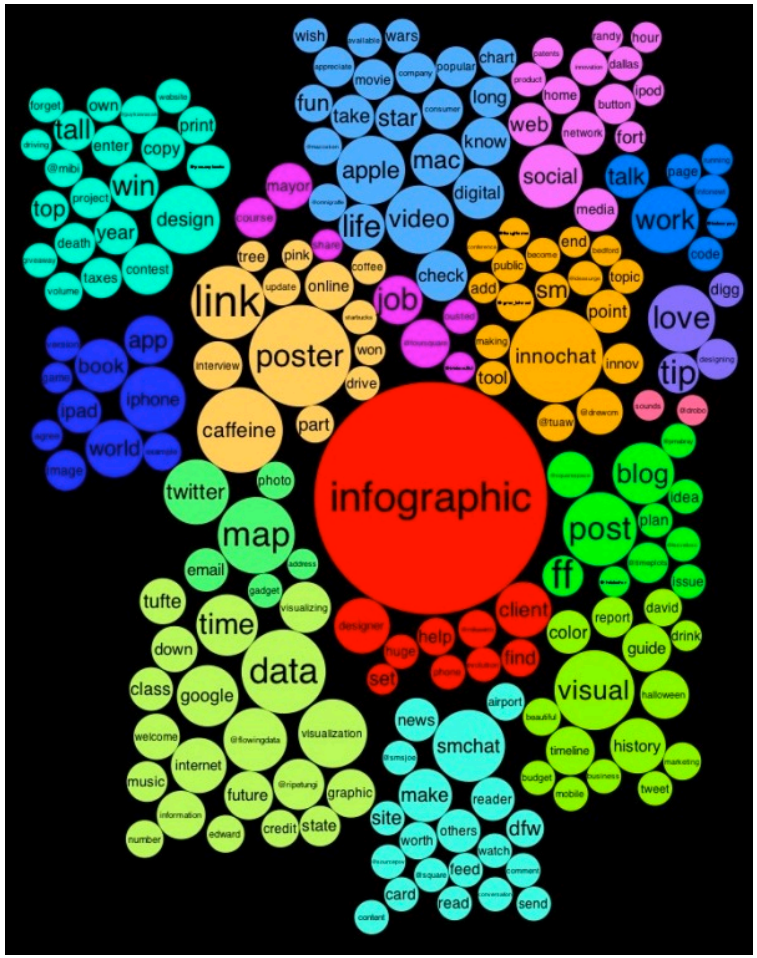
http://blog.ranker.com/tag/psychographics/#.Wq7VR5PwbOQ


(good) "Cluster"
Similarity within group: high intra-class simi.
Dissimilarity between groups: low inter-class simi.
Concepts
vs "Classification": problem definition
"Unsupervised Learning"
Applications
Summarization
Compression
Data reduction
Outlier detection
Trend/pattern detection
Recommender systems: collaborative filtering
Multimedia data analysis: image/video/audio
Biological data analysis: gene/protein
Social network analysis: user/community
Clustering
Cluster analysis
Datasets
Numerical data
Categorical data (including binary data)
Text data (e.g., news, social media, web)
Multimedia data: Image/audio/video (e.g., Flickr, YouTube)
Time-series data (e.g., sensor data, stock market)
Sequence data (e.g., weblog, bio. Sequences)
Stream data
Graphs and homogeneous networks
Heterogeneous networks
Uncertain data (noise, approximate values, multiple possible values)
Big data

Clustering Different Types of Data (I)

- **Numerical data**
  - Most earliest clustering algorithms were designed for numerical data
- **Categorical data** (including binary data)
  - Discrete data, no natural order (e.g., sex, race, zip-code, and market-basket)
- **Text data**: Popular in social media, Web, and social networks
  - Features: High-dimensional, sparse, value corresponding to word frequencies
  - Methods: Topic modeling
- **Multimedia data**: Image, audio, video (e.g., on Flickr, YouTube)
  - Multi-modal (often combined with text data)
  - Contextual: Containing both behavioral and contextual attributes
    - Images: Position of a pixel represents its context, value represents its behavior
    - Video and music data: Temporal ordering of records represents its meaning

Text Data Clustering



Clustering Different Types of Data (II)

- **Time-series data**: Sensor data, stock markets, temporal tracking, forecasting, etc.
  - Data are temporally dependent
  - Time: contextual attribute; data value: behavioral attribute
  - Correlation-based online analysis (e.g., online clustering of stock to find stock tickers)
  - Shape-based offline analysis (e.g., cluster ECG based on overall shapes)
- **Sequence data**: Weblogs, biological sequences, system command sequences
  - Contextual attribute: Placement (rather than time)
  - Similarity functions: Hamming distance, edit distance, longest common subsequence
  - Sequence clustering: Suffix tree; generative model (e.g., Hidden Markov Model)
- **Stream data**:
  - Real-time, evolution and concept drift, single pass algorithm
  - Create efficient intermediate representation, e.g., micro-clustering

Sequence Data Clustering

CM: ΨKxE/D

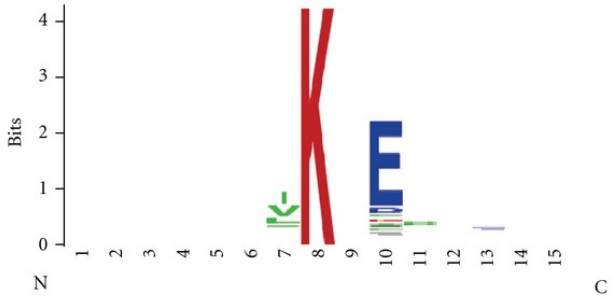
ICM: E/DxKΨ

PDSM: ΨKxE/DxxSP

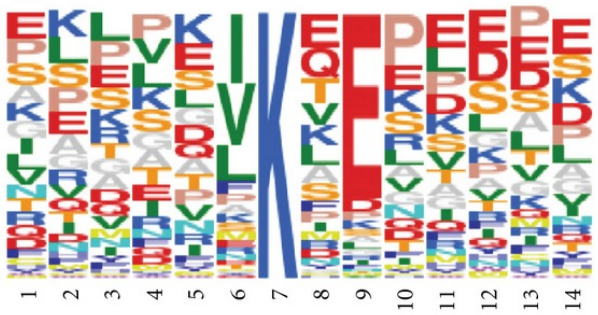
NDSM: ΨKxExxEEEE

HCSM: ΨΨΨKxE

(a)



(b)

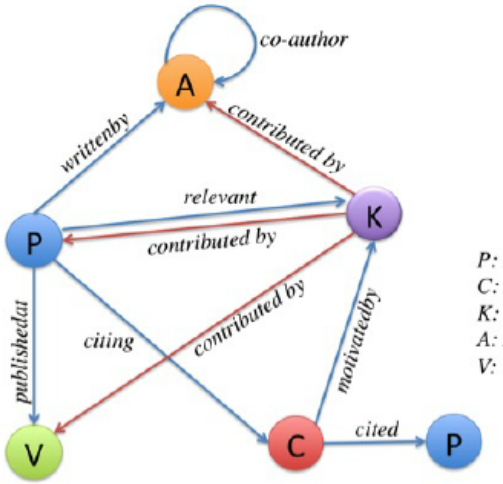


(c)

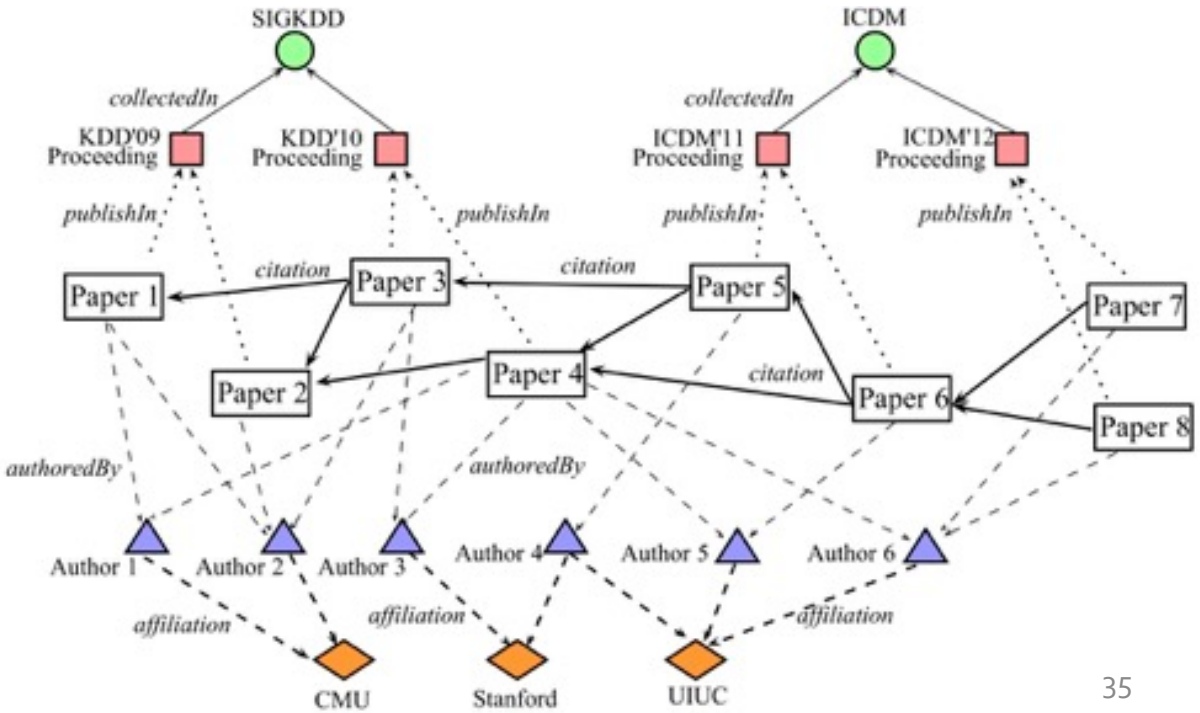
Clustering Different Types of Data (III)

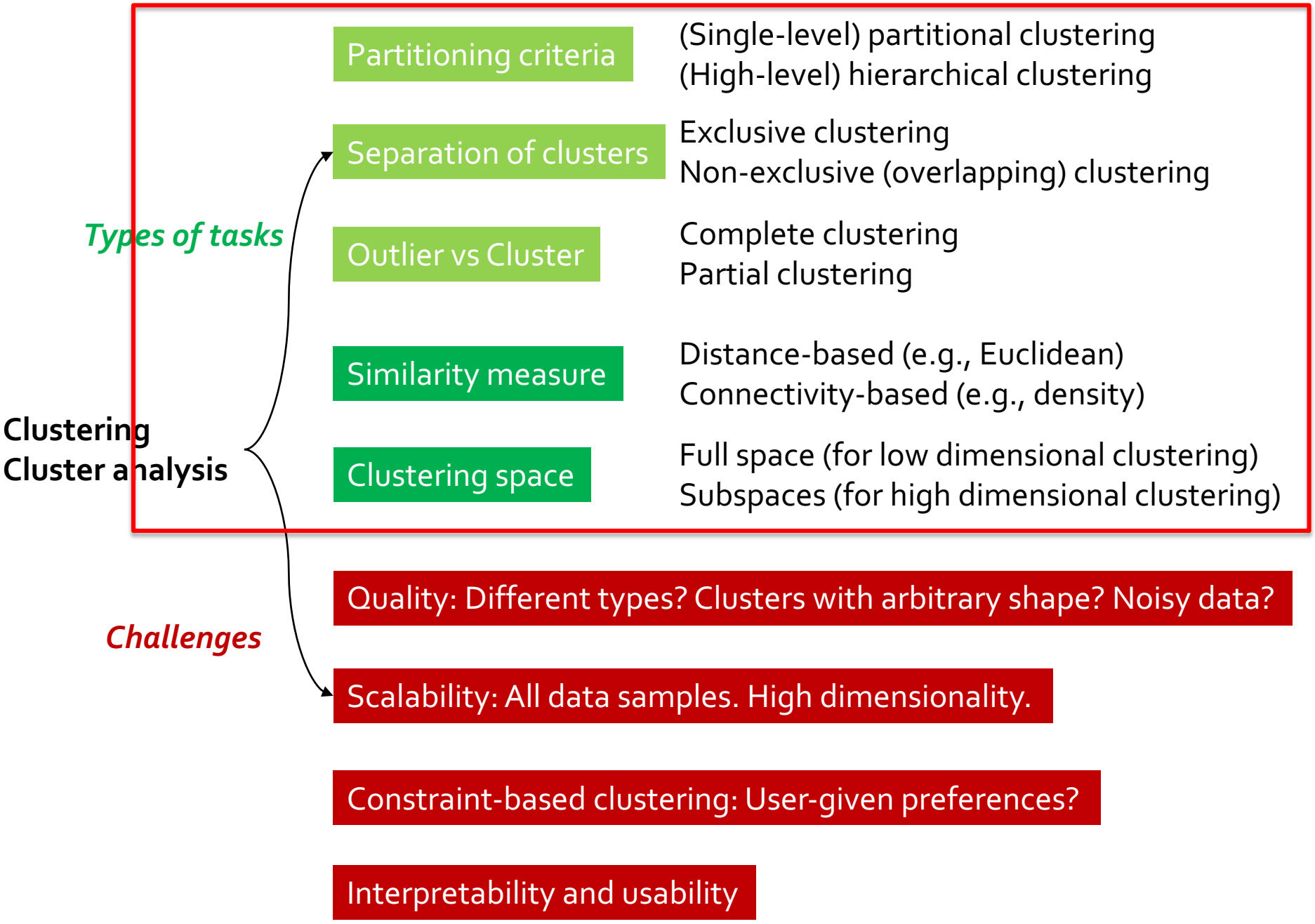
- **Graphs and homogeneous networks**
  - Every kind of data can be represented as a graph with similarity values as edges
  - Methods: Generative models; combinatorial algorithms (graph cuts); spectral methods; non-negative matrix factorization methods
- **Heterogeneous networks**
  - A network consists of multiple typed nodes and edges (e.g., bibliographical data)
  - Clustering different typed nodes/links together (e.g., NetClus)
- **Uncertain data**: Noise, approximate values, multiple possible values
  - Incorporation of probabilistic information will improve the quality of clustering
- **Big data**: Model systems may store and process very big data (e.g., weblogs)
  - Ex. Google's MapReduce framework
    - Use *Map* function to distribute the computation across different machines
    - Use R*educe* function to aggregate results obtained from the Map step

Heterogeneous Network Clustering



P: Paper
C: Citation
K: Keyword (topic)
A: Author
V: Venue



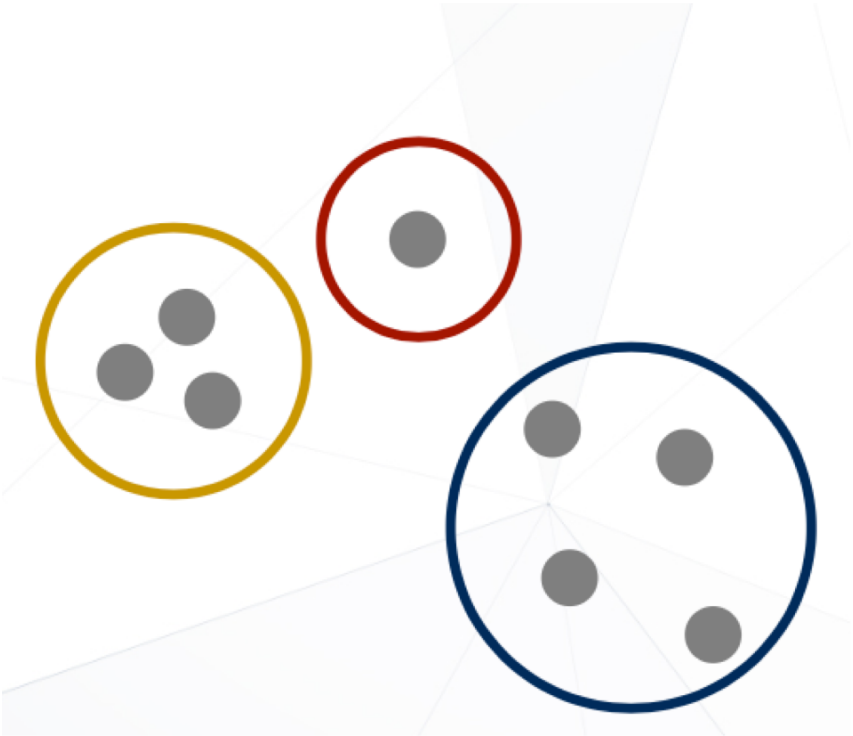

Clustering
Cluster analysis
Types of tasks
Partitioning criteria
(Single-level) partitional clustering
(High-level) hierarchical clustering
Separation of clusters
Exclusive clustering
Non-exclusive (overlapping) clustering
Outlier vs Cluster
Complete clustering
Partial clustering
Similarity measure
Distance-based (e.g., Euclidean)
Connectivity-based (e.g., density)
Clustering space
Full space (for low dimensional clustering)
Subspaces (for high dimensional clustering)
Challenges
Quality: Different types? Clusters with arbitrary shape? Noisy data?
Scalability: All data samples. High dimensionality.
Constraint-based clustering: User-given preferences?
Interpretability and usability

Considerations for Cluster Analysis

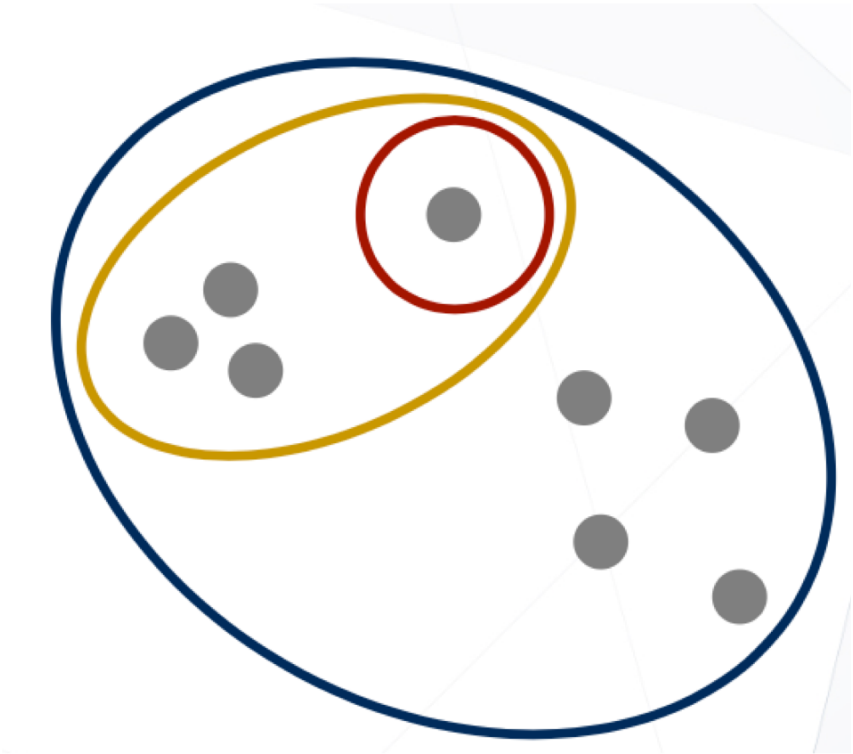
- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)

Partitioning Criteria

Partitional clustering



Hierarchical clustering

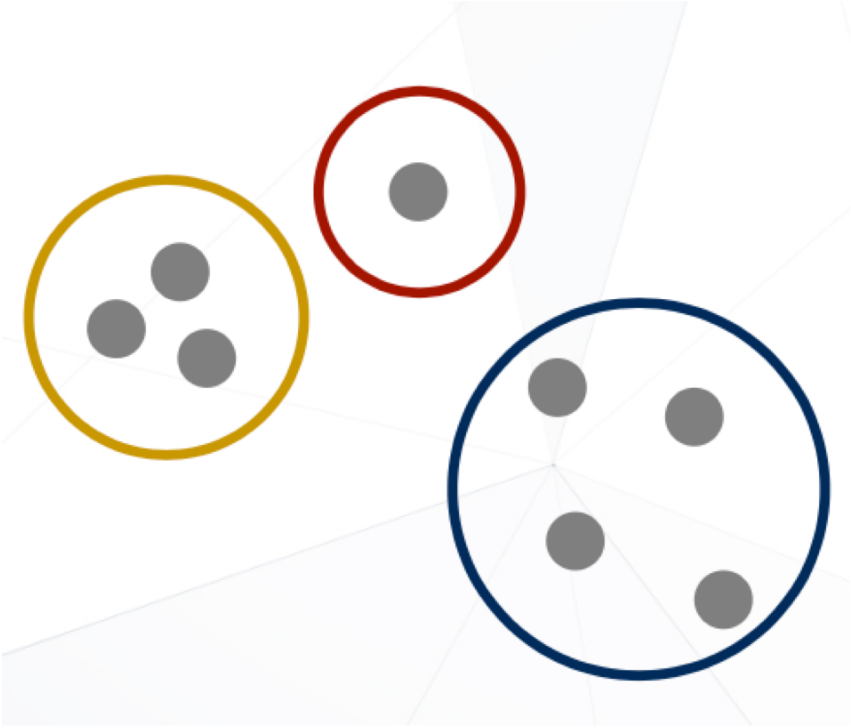


Considerations for Cluster Analysis

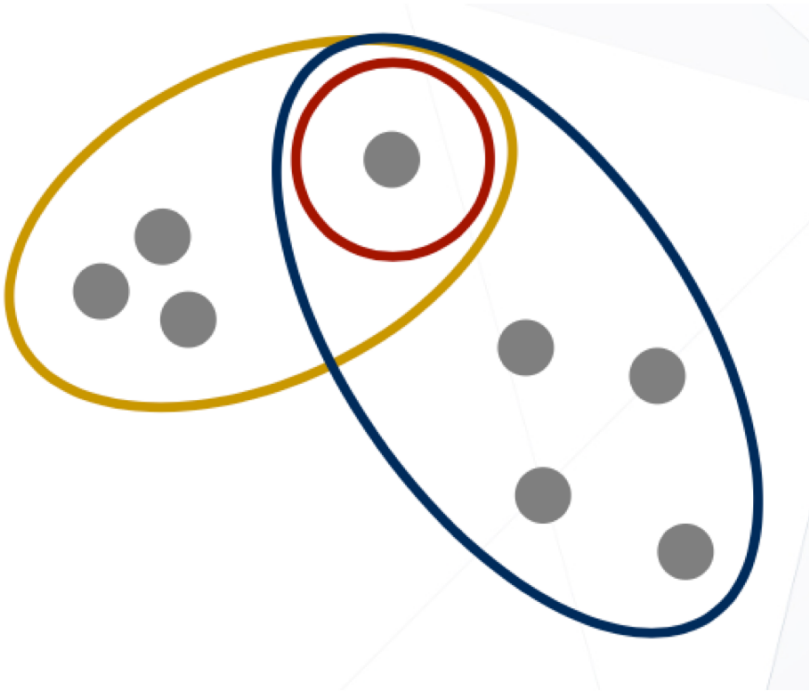
- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)

Separation of Clusters

Exclusive clustering



Overlapping clustering

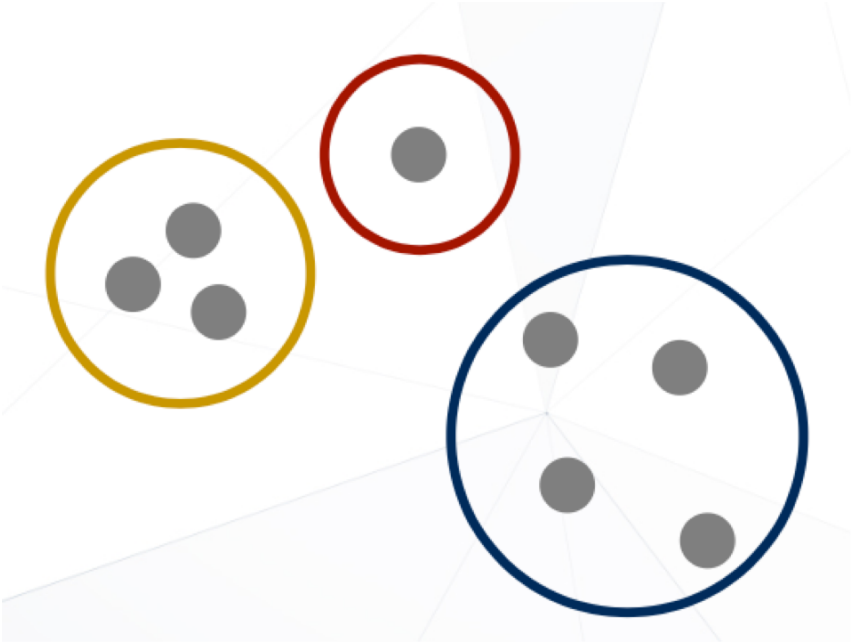


Considerations for Cluster Analysis

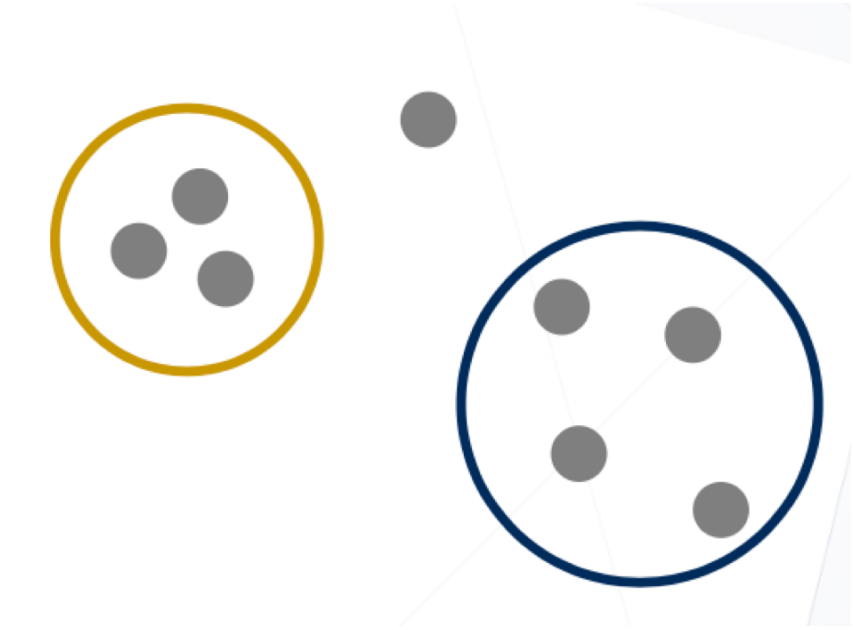
- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- **Outlier vs. Cluster**
  - Complete (e.g., every item has one cluster) vs. partial (e.g., one item may not belong to any cluster)

Outlier vs. Cluster

Complete clustering



Partial clustering



Considerations for Cluster Analysis

- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- **Outlier vs. Cluster**
- **Similarity measure**
  - Distance-based (e.g., Euclidean, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- **Clustering space**
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Clustering
Cluster analysis

Types of tasks

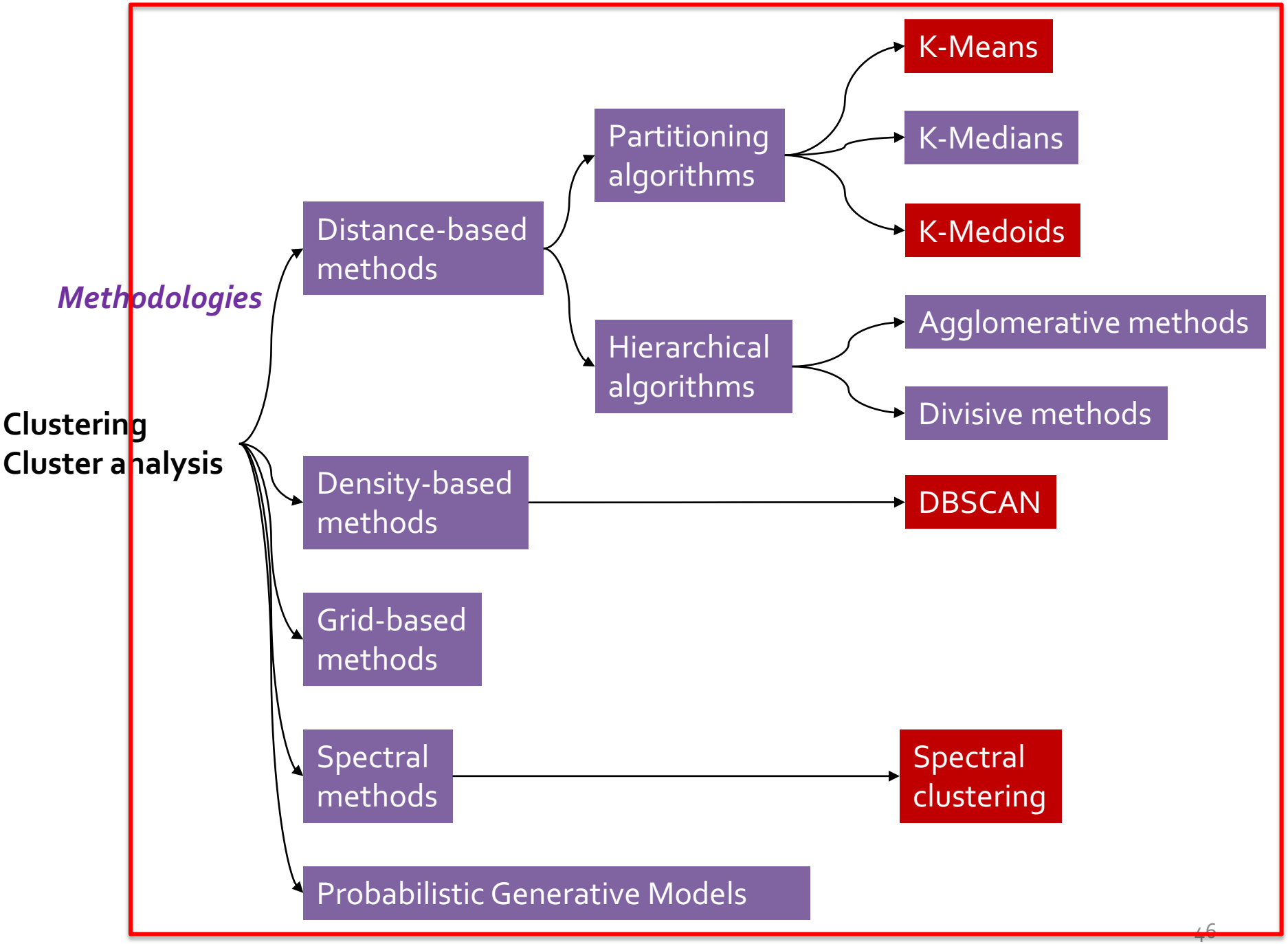
- Partitioning criteria
  - (Single-level) partitional clustering
  - (High-level) hierarchical clustering
- Separation of clusters
  - Exclusive clustering
  - Non-exclusive (overlapping) clustering
- Outlier vs Cluster
  - Complete clustering
  - Partial clustering
- Similarity measure
  - Distance-based (e.g., Euclidean)
  - Connectivity-based (e.g., density)
- Clustering space
  - Full space (for low dimensional clustering)
  - Subspaces (for high dimensional clustering)

Challenges

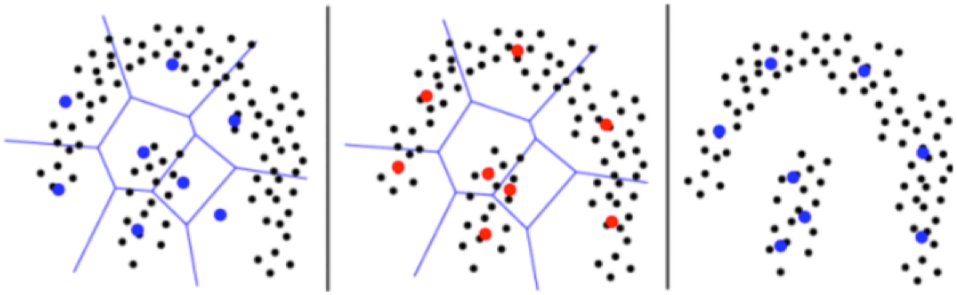
- Quality: Different types? Clusters with arbitrary shape? Noisy data?
- Scalability: All data samples. High dimensionality.
- Constraint-based clustering: User-given preferences?
- Interpretability and usability

Requirements and Challenges

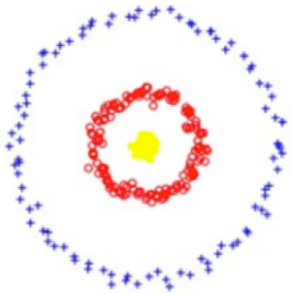
- Quality
  - Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
  - Discovery of clusters with arbitrary shape
  - Ability to deal with noisy data
- Scalability
  - Clustering all the data instead of only on samples
  - High dimensionality
  - Incremental or stream clustering and insensitivity to input order
- Constraint-based clustering
  - User-given preferences or constraints; domain knowledge; user queries
- Interpretability and usability


Methodologies
Clustering
Cluster analysis
Distance-based methods
Partitioning algorithms
K-Means
K-Medians
K-Medoids
Hierarchical algorithms
Agglomerative methods
Divisive methods
Density-based methods
DBSCAN
Grid-based methods
Spectral methods
Spectral clustering
Probabilistic Generative Models

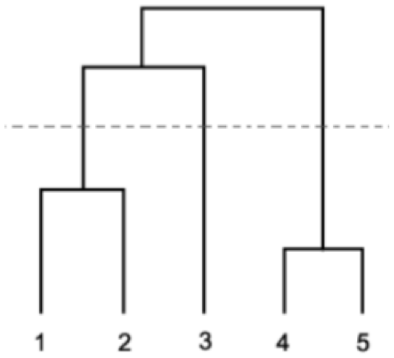
Methodologies



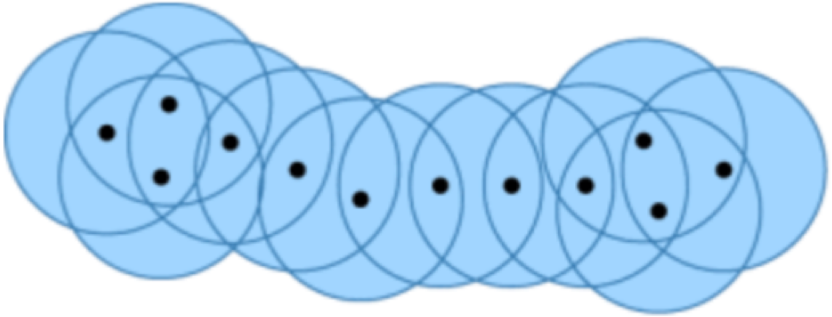
K Partitioning (Distance-based)



+ Kernel Tricks

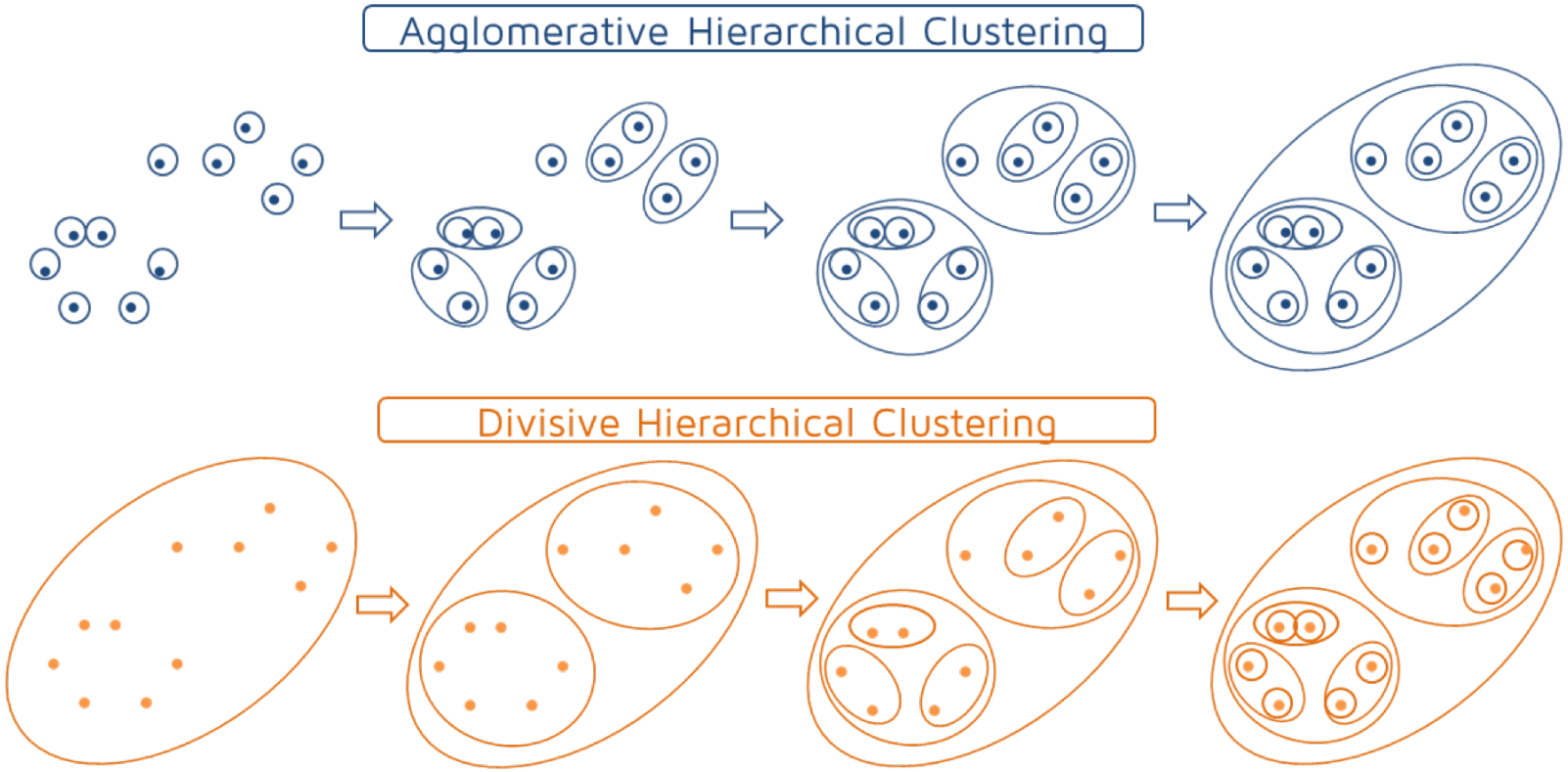


Hierarchical (Distance-based)

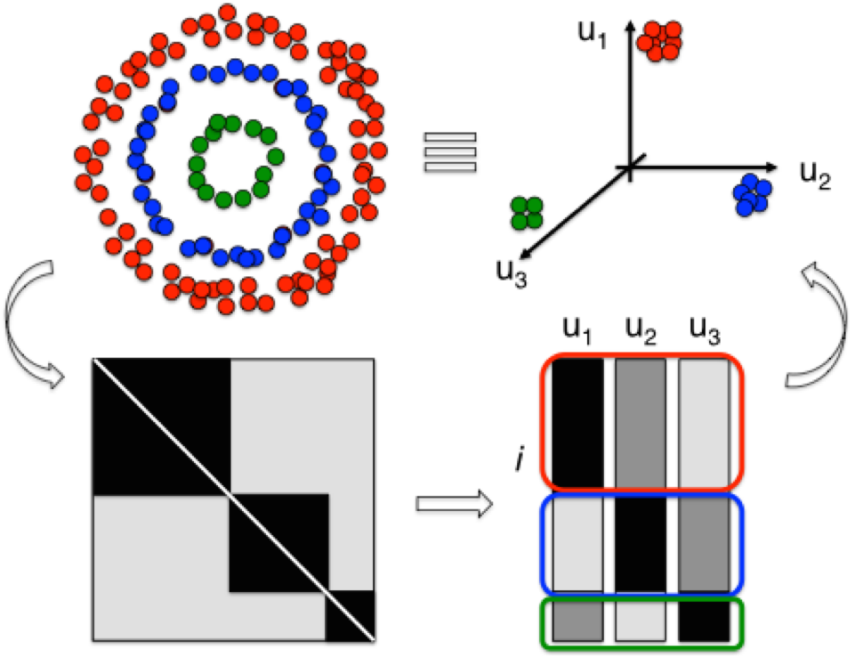


DBSCAN (Density-based)

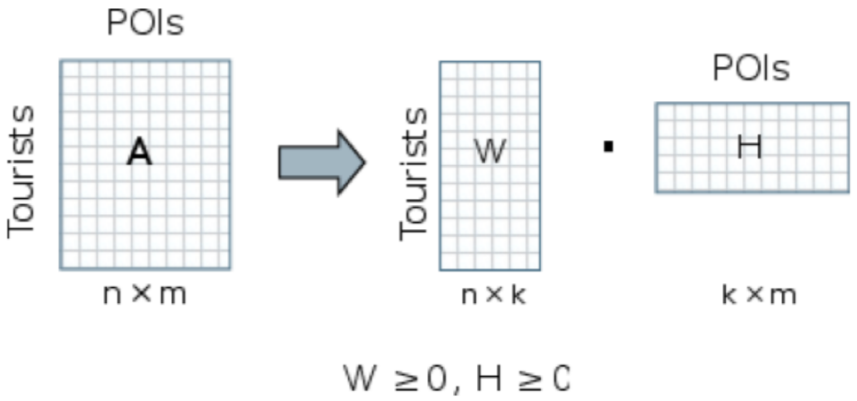
Methodologies



Methodologies

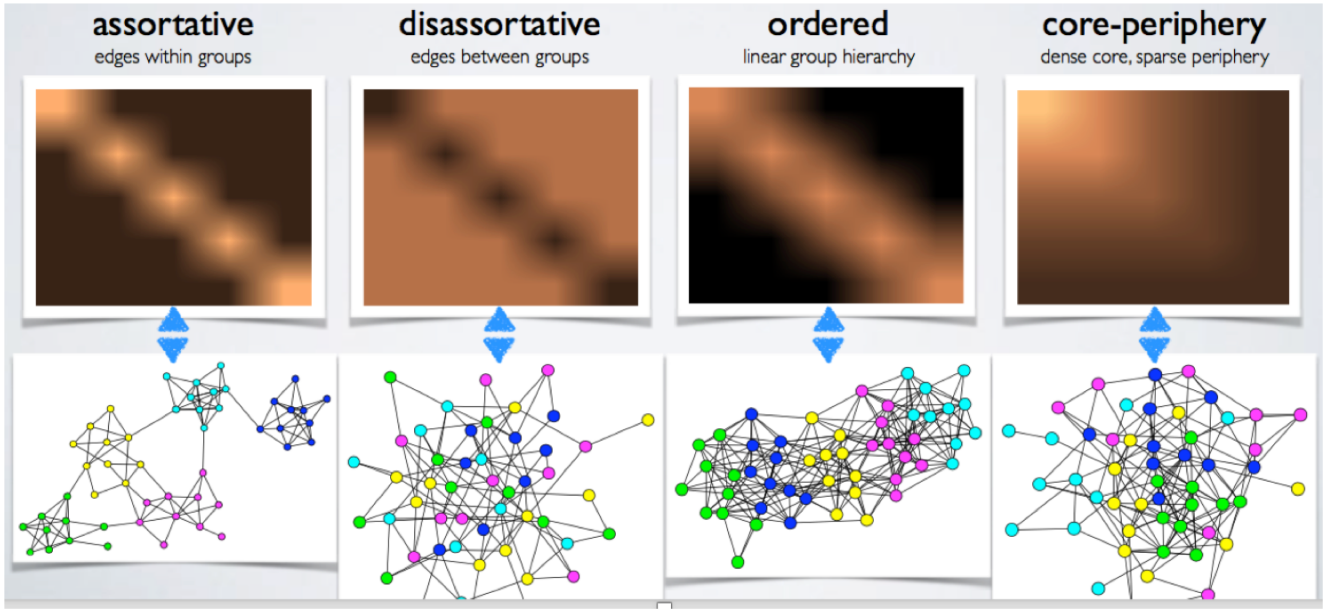


Spectral clustering



Nonnegative Matrix Factorization (NMF)

Methodologies



Probabilistic Graphical Models (based on Expectation-Maximization algorithms)

Methodologies

Probabilistic Graphical Models (based on Expectation-Maximization algorithms)

$$\Pr(A \mid z, \theta, x) = \prod_{ij} f(A_{ij} \mid \theta_{\mathcal{R}(z_i, z_j)}, x_i, x_j)$$

A_{ij} : value of adjacency

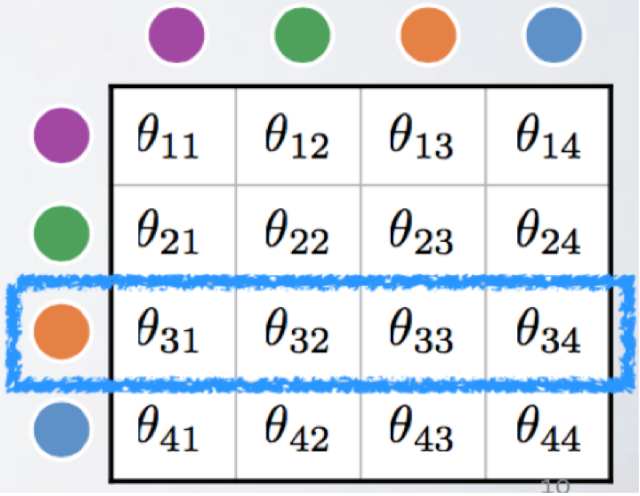
$\mathcal{R}$: partition of adjacencies

f : probability function

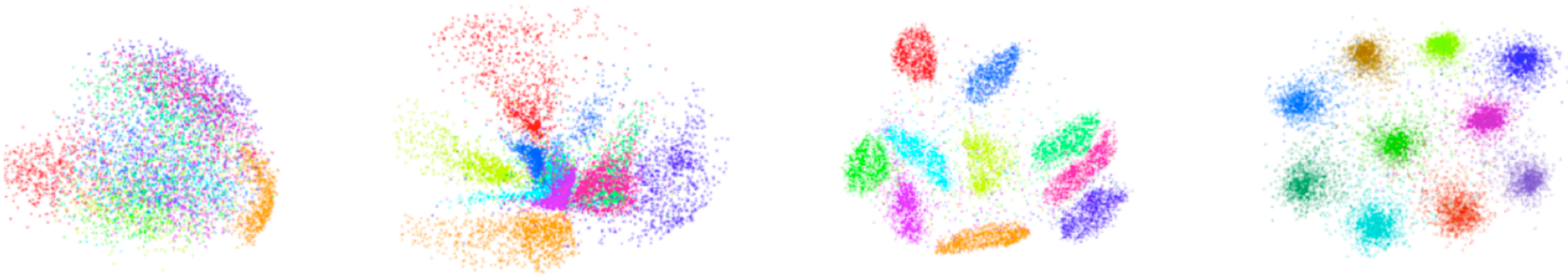
$\theta_{a,*}$: pattern for a-type adjacencies

x : node attributes (metadata)

Binomial = simple graphs
Poisson = multi-graphs
Normal = weighted graphs
etc.



Network Representation Learning (Network Embedding)



Colors: ground-truth classes

Given a N-node network,
Learn k-dimensional representation
(i.e., k features, k-length vector) for each node.

If K-Means is able to generate better clustering
on the feature space, the representations are better.

References

- Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011 (Chapters 10 & 11)
- Charu Aggarwal and Chandran K. Reddy (eds.). Data Clustering: Algorithms and Applications. CRC Press, 2014
- Mohammed J. Zaki and Wagner Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990
- Charu Aggarwal. An Introduction to Clustering Analysis. in Aggarwal and Reddy (eds.). Data Clustering: Algorithms and Applications (Chapter 1). CRC Press, 2014

CatchTartan:

Representing and Summarizing Dynamic Multicontextual Behaviors

Meng Jiang, Christos Faloutsos, Jiawei Han



What is Tartan?

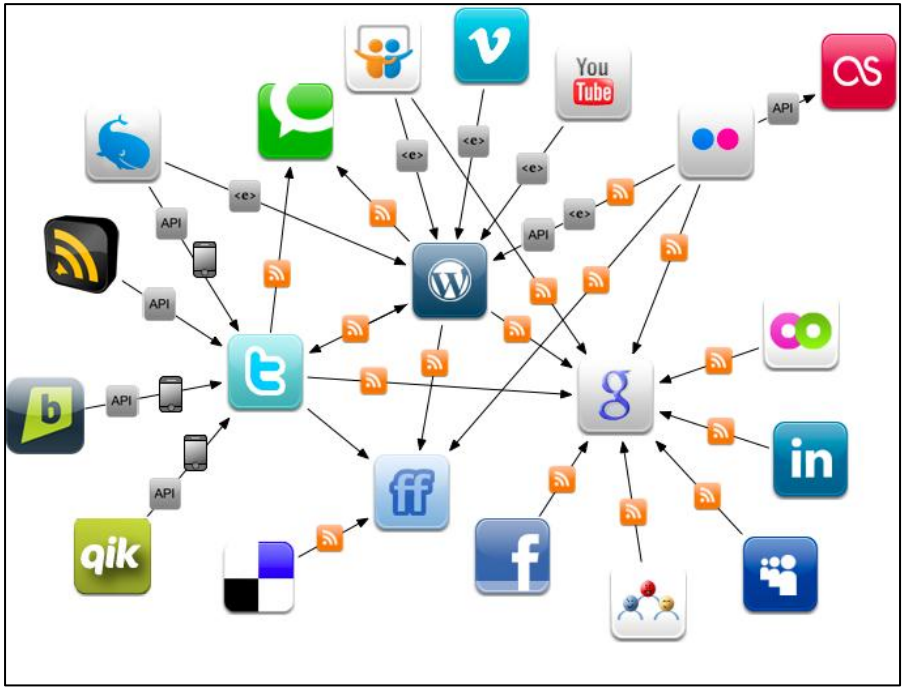
Visited CMU in 2012-13

Watched lots of Tartans' games…

Why We Talk about Behavior Today?

Physical Environment

Online Environment



The human behaviors are broadly and deeply recorded in an unprecedented level.

Given the behavioral data (e.g., DBLP data, tweets)

2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Return behavioral summaries (e.g., research trends, events)

1997 2000 2003 2006 2009 2012

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	**7** SIGIR VLDB TKDE	**7** "information retrieval" "data integration" "text classification"	**68** p56743[1] p62995 p76869	**32** 2003- 2007

[1] "A language modeling approach to information retrieval"

Venue	Keyword	#Paper
5 ICML NIPS …	**6** "reinforcement learning" "machine learning"	**40** 1997- 2002

Author	Venue	Cited	#Paper
6 Jiawei Han Xifeng Yan	**1** SIG- MOD	**1** p76095[2]	**22** 2004- 2010

[2] "Frequent subgraph discovery"

Venue	Keyword	#Paper
3 ICDM AAAI TKDE	**1** "anomaly detection"	**25** 2005- 2013

Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal…	**6** KDD ICDM ICDE TKDE …	**12** "large graphs" "data streams" "evolving data" "evolving graphs" …	**70** 2006- 2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng…	**5** SIGIR WWW WSDM CIKM…	**3** "web search" "click-through data" "sponsored search"	**12** p82630[3] p116290 p103899 p106191…	**32** 2006- 2013

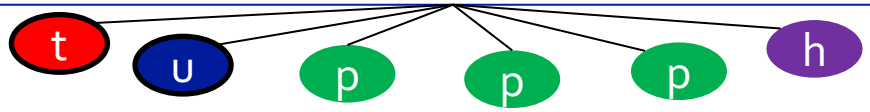
[3] "Optimizing search engines using clickthrough data"

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan…	**3** KDD PAKDD AAAI	**6** "transfer learning" "data mining" "localization models"	**17** 2007- 2010

Behaviors: Dynamic and Multicontextual

- Tweeting behavior

20:03:09 @ebekahwsm
this better be the best halftime show ever in the history of halftimes shows. ever. #SuperBowl



Contextual factors:

Dynamic

One-guaranteed value

Empty (set of) *value*

Set value

Empty (set of) *value*

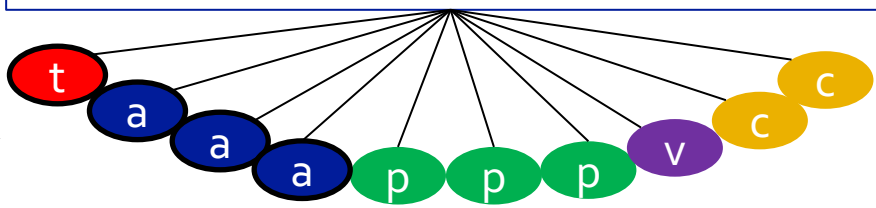
Time slice	User	Location	Phrase	Hashtag	URL
20:00-20:30	@ebekahwsm	∅	{best halftime show, in the history, halftimes shows}	{#SuperBowl}	∅

Behaviors: Dynamic and Multicontextual

- Publishing-paper behavior

2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Contextual factors:



Dynamic

Set value

One-guaranteed value

Set value

Set value

Time slice	Author	Venue	Keyword	Cited papers
2009	{P. Melville, W. Gryc, R. Lawrence}	SIGKDD	{sentiment analysis, lexical knowledge, text classification}	{p81623, p84395, p95393, p95409, p99073, p116349 …}

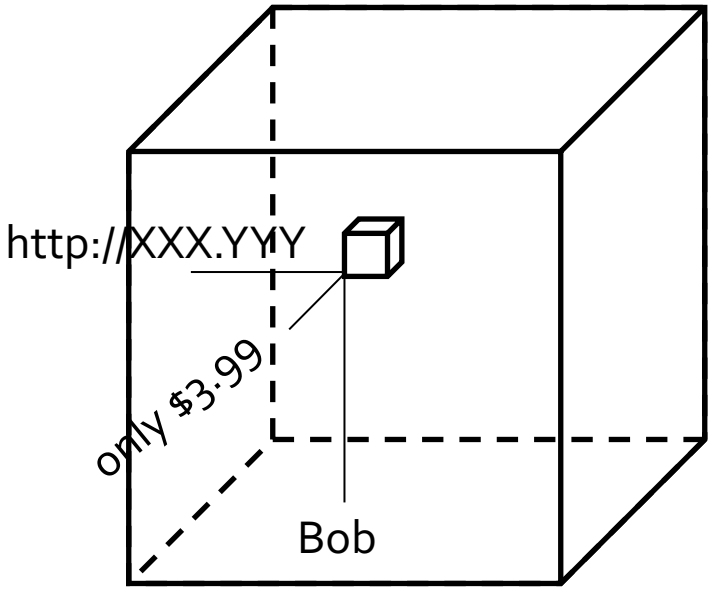
Summarizing Behaviors

- *Dynamic:* taking a set of consecutive time slices
- *Multicontextual:* taking a set of dimensions and a set of dimensional values in each dimension

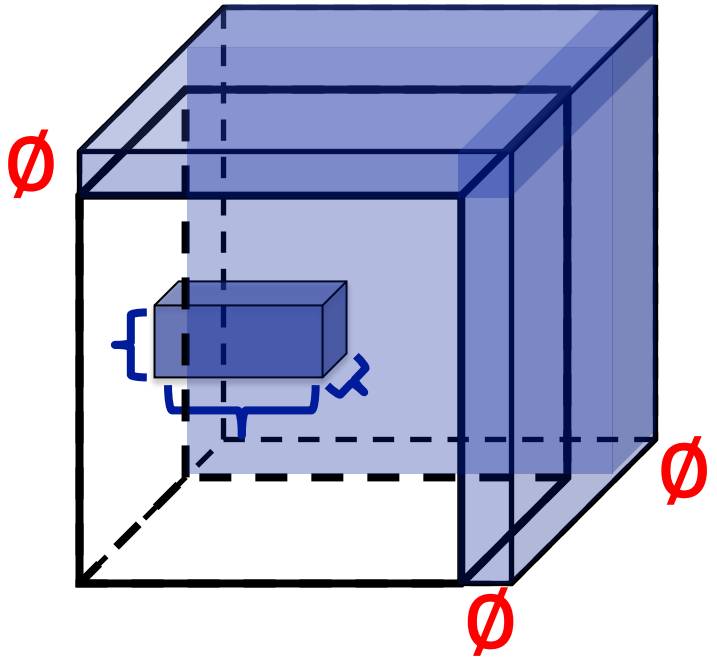
Term	Definition
Dimension	The type of a contextual factor (e.g., location, phrase; author, keyword)
(Dimensional) value	The contextual factor in the dimension
Time slice	The period for consecutive behaviors
Behavior	A set of dimensions, a set of values in each dimension, a time slice for the timestamp

Tensor Fails

- Tensor - modeling multidimensions: FEMA (KDD'14), CrossSpot (ICDM'15)

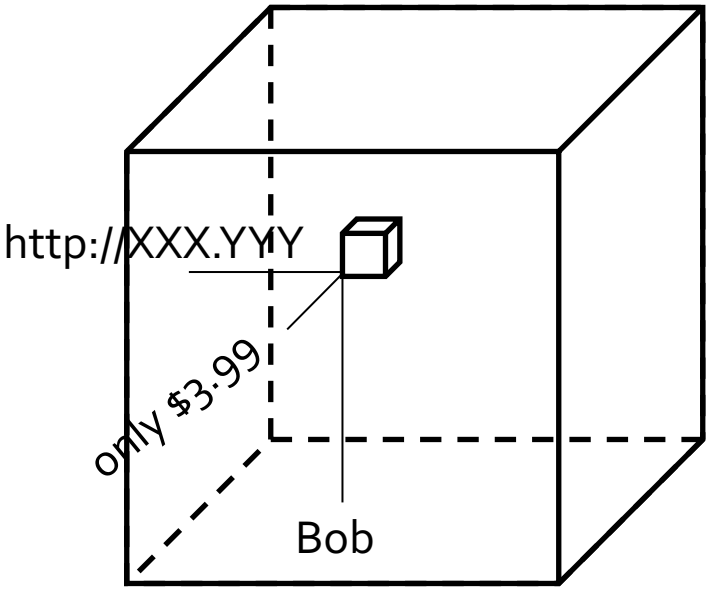


- **Representation: (multicontextual)**
  - **Empty values?**

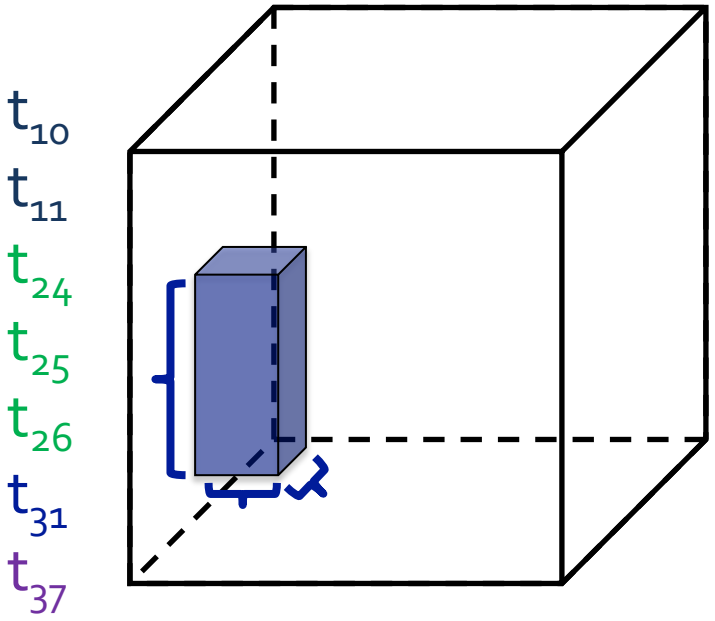


Tensor Fails (cont.)

- Tensor - modeling multidimensions: FEMA (KDD'14), CrossSpot (ICDM'15)

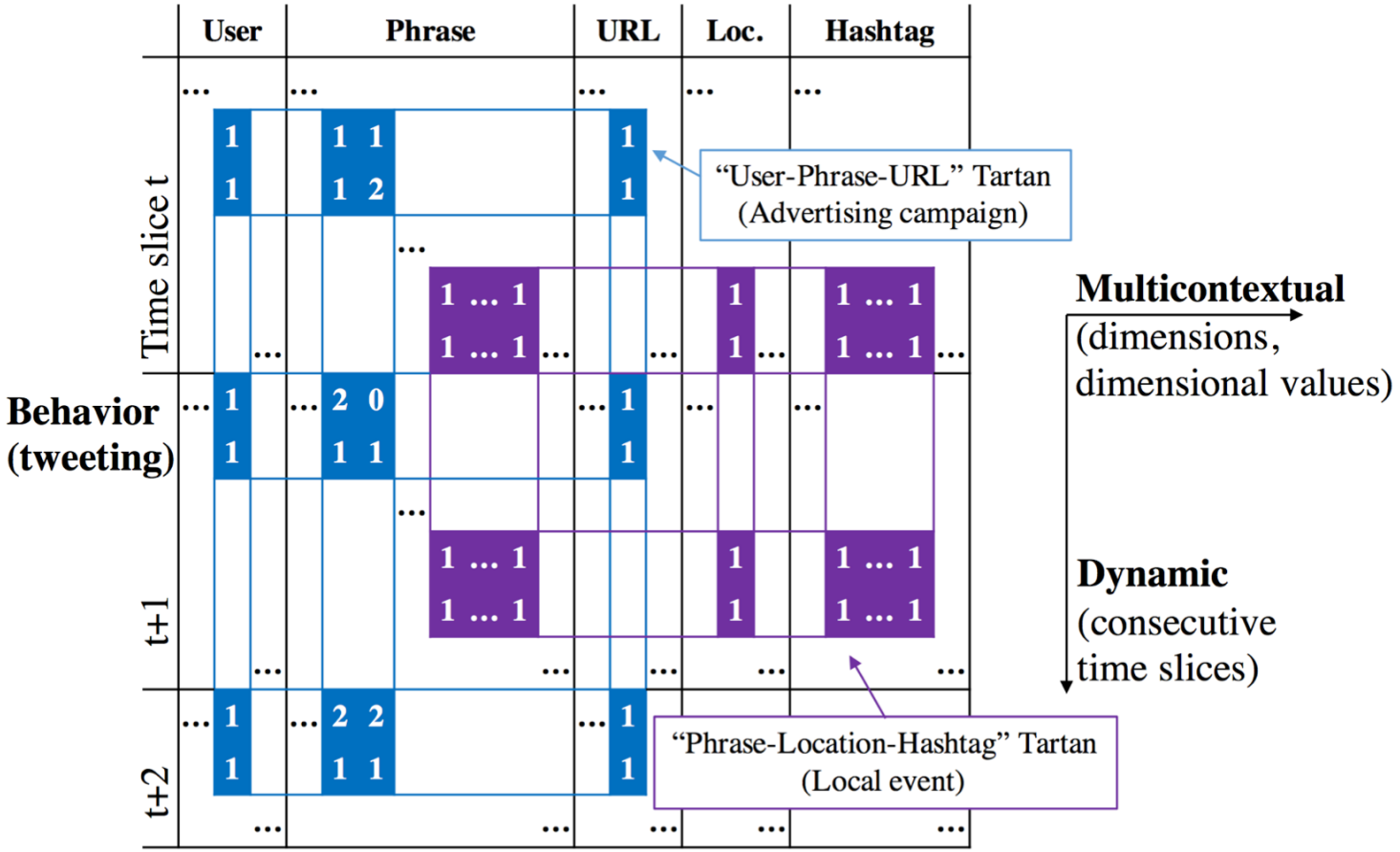


- **Summarization: (dynamic)**
  - **Temporal patterns?**



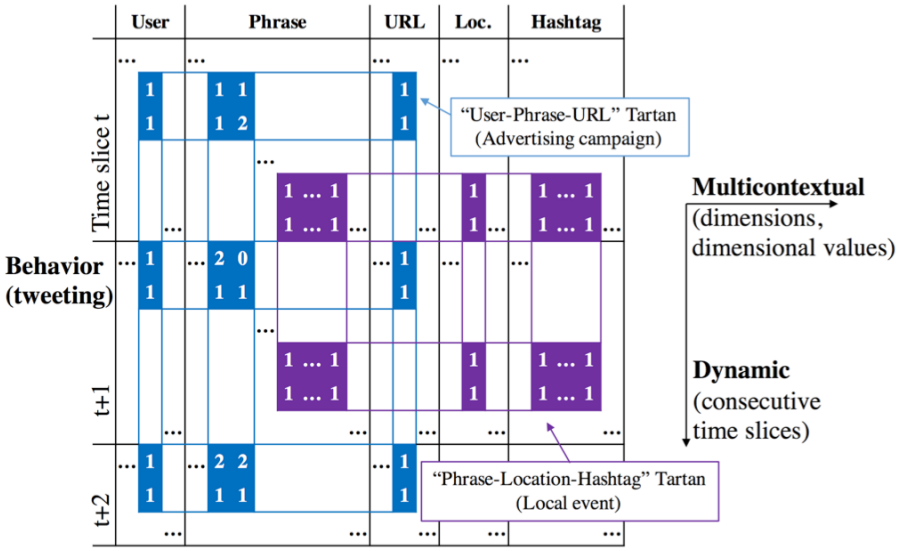
Our Representations for Behavior and Behavioral Summary

- **Behavior: "Two-level matrix"**
- **Behavioral summary: "Tartan"**



The Problem of Behavioral Summarization

PROBLEM 1 (BEHAVIORAL SUMMARIZATION). ***Given*** *the behavioral data (a two-level matrix)* $\mathcal{X} = \{D, N_d|_{d=1}^{D}, T, E^{(t)}|_{t=1}^{T}\}$, ***find*** *a list of behavioral summaries (Tartans)* $\tilde{\mathcal{A}} = \{\ldots, \mathcal{A}, \ldots\}$ *ordered by a principled metric function* $f(\mathcal{A}, \mathcal{X})$ *which defines how well the sets of meaningful dimensions, values, time slices and behaviors are partitioned and how well the meaningful subset of data is summarized, where* $\mathcal{A} = \{\mathcal{D}, \mathcal{V}_d|_{d\in\mathcal{D}}, \mathcal{T}, \mathcal{B}^{(t)}|_{t\in\mathcal{T}}\}$.



CATCHTARTAN

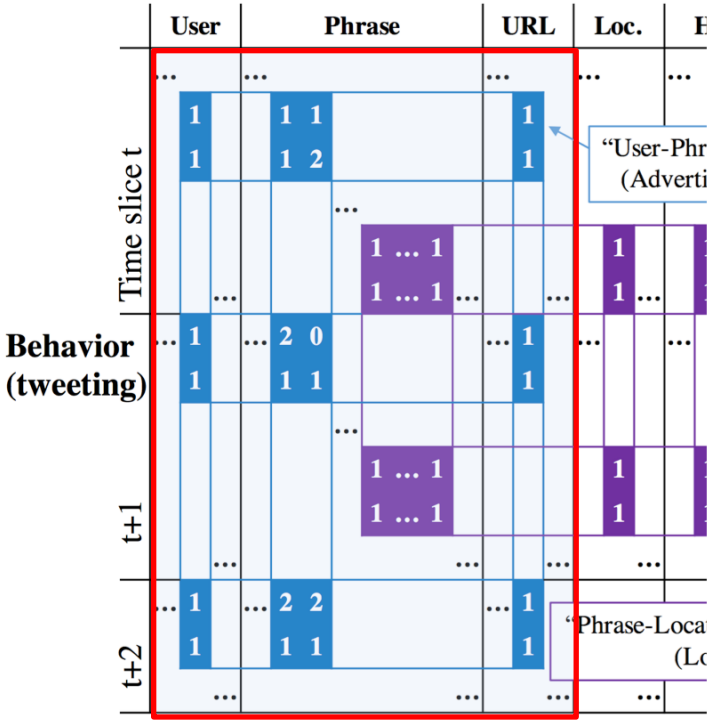
- Employing a lossless encoding scheme
  - The *Minimum Description Length* (MDL) principle
  - Estimating the number of bits that encoding the Tartan can save from merging the meaningful pattern into the encoding of the data

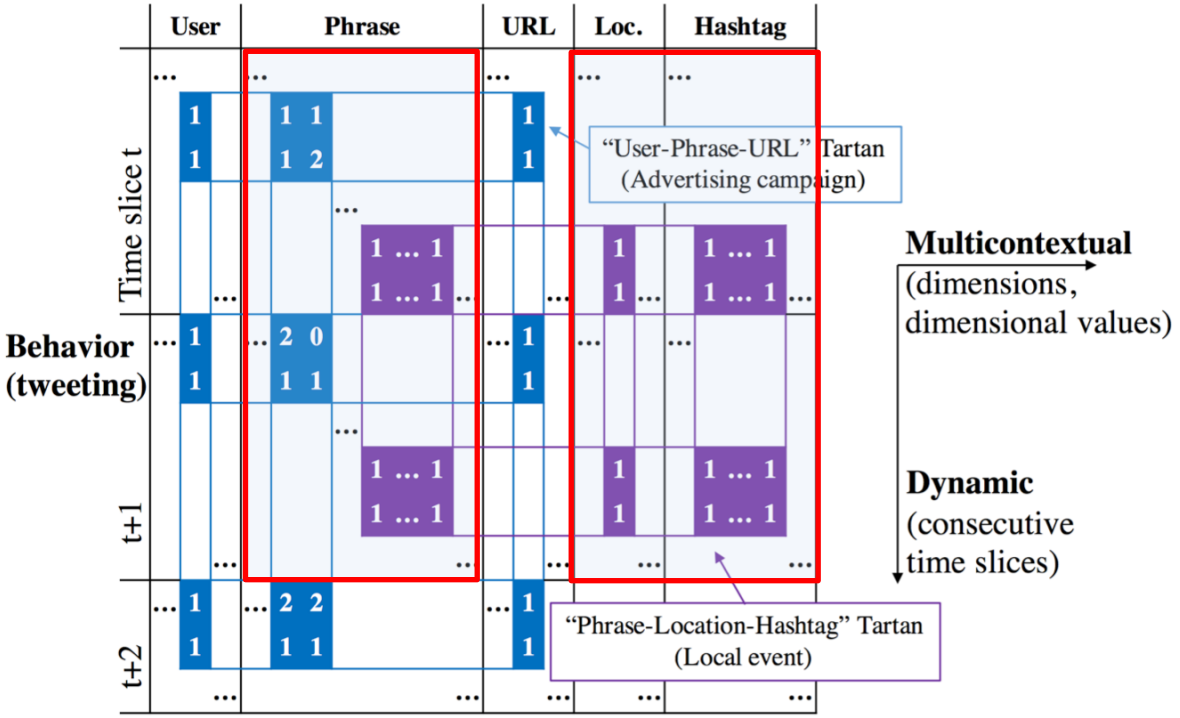
	FSG [18]	GRAPH-CUBE [33]	EVENT-CUBE [29]	MDC [21]	BOW [6]	FEMA [9]	COM2 [2]	CROSS-SPOT [8]	GRAPH-SCOPE [27]	VOG [15]	TIME-CRUNCH [26]	**CATCH-TARTAN**
Principled scoring								√	√	√	√	√
Parameter-free	√				√		√	√	√	√	√	√
Multidimensional		√	√	√	√	√	√	√				√
Multicontextual												√
Timestamp value			√	√		√	√	√	√		√	√
Dynamics							√		√		√	√

Objective Function to Maximize

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}).$$

Tartan **Data** **First-level matrix** **Individual entries**





$$\mathcal{X}^{\mathcal{A}} = \{\mathcal{X}_d^{(t)}(b,i) | d \in \mathcal{D}, t \in \mathcal{T}, i \in \{1,\ldots,N_d\}, b \in \{1,\ldots,E^{(t)}\}\}.$$

Objective Function to Maximize (cont.)

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}).$$

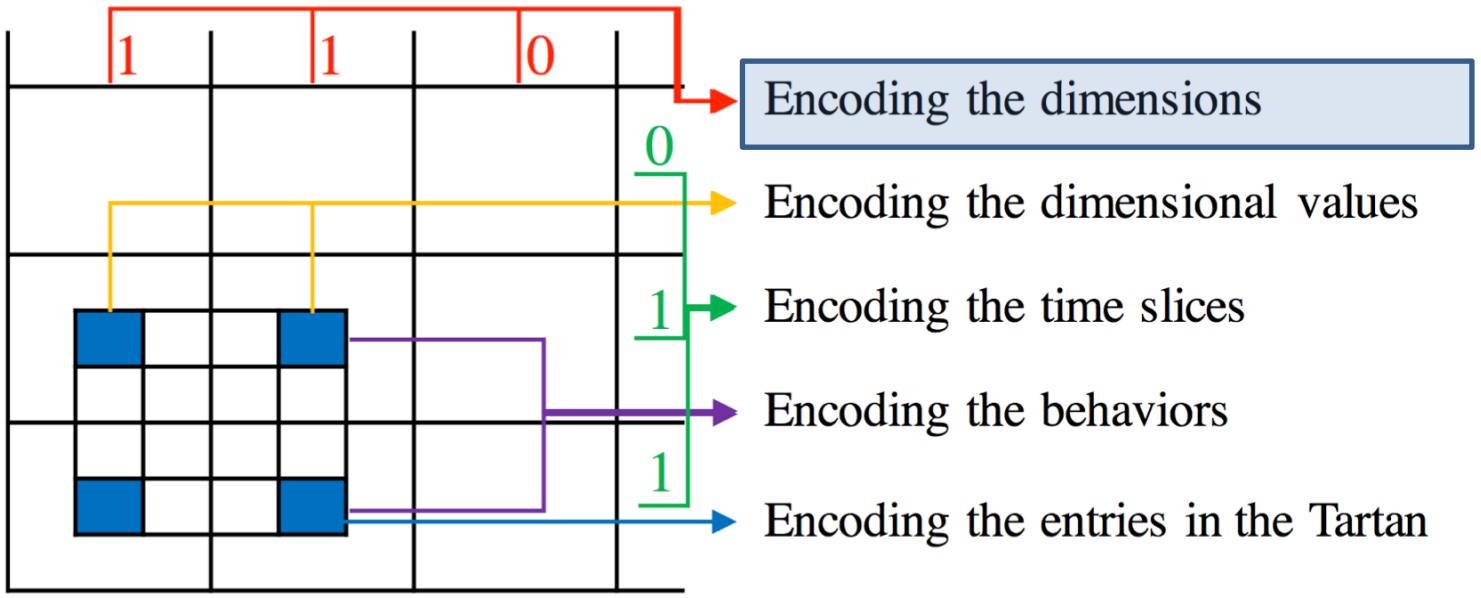
$$\begin{aligned} V &= \left(\sum_{d \in \mathcal{D}} N_d\right)\left(\sum_{t \in \mathcal{T}} E^{(t)}\right). \\ C &= \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \{1,\ldots,E^{(t)}\}, i \in \{1,\ldots,N_d\}} \mathcal{X}_d^{(t)}(b,i). \end{aligned}$$

$$\begin{aligned} L(\mathcal{X}^{\mathcal{A}}) &= g(V + C, C) + L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) \\ &\quad + \sum_{d \in \mathcal{D}} \log^* N_d + \sum_{t \in \mathcal{T}} \log^* E^{(t)}. \end{aligned}$$

$$L(\mathcal{A}) = L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{V}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + L_{\mathcal{B}}(\mathcal{A}) + L_{\mathcal{A}}(\mathcal{A}).$$

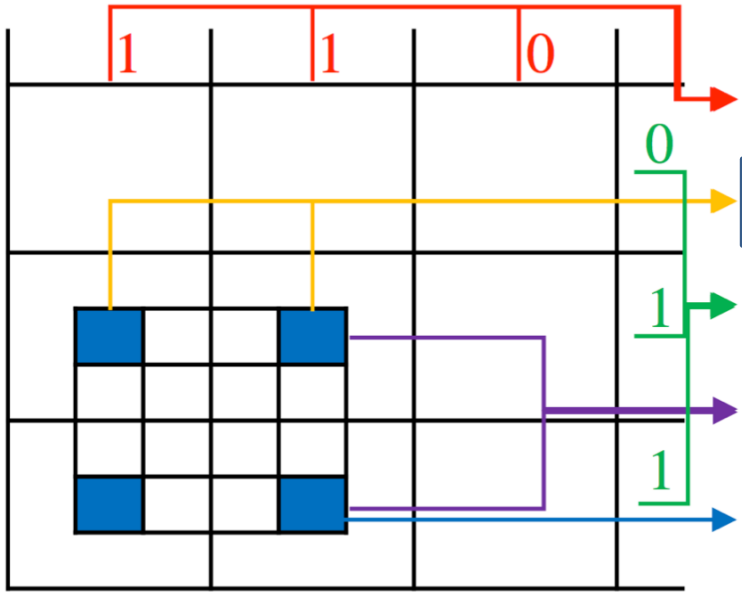
$$L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}) = g(V + C - v - c, C - c);$$

Encoding the Tartan: Dimensions



$$
\begin{aligned}
H_{\mathcal{D}}(X) &= -\sum_{x\in\{0,1\}} P(X=x)\log P(X=x) \\
&= -\left(\frac{D^{\mathcal{A}}}{D}\log\frac{D^{\mathcal{A}}}{D} + \frac{D-D^{\mathcal{A}}}{D}\log\frac{D-D^{\mathcal{A}}}{D}\right). \\
L_{\mathcal{D}}(\mathcal{A}) &= \log^* D + \log^* D^{\mathcal{A}} + D\cdot H_D(X) \\
&= \log^* D + \log^* D^{\mathcal{A}} + g(D, D^{\mathcal{A}}),
\end{aligned}
$$

Encoding the Tartan: Dimensional Values



Encoding the dimensions

Encoding the dimensional values

Encoding the time slices

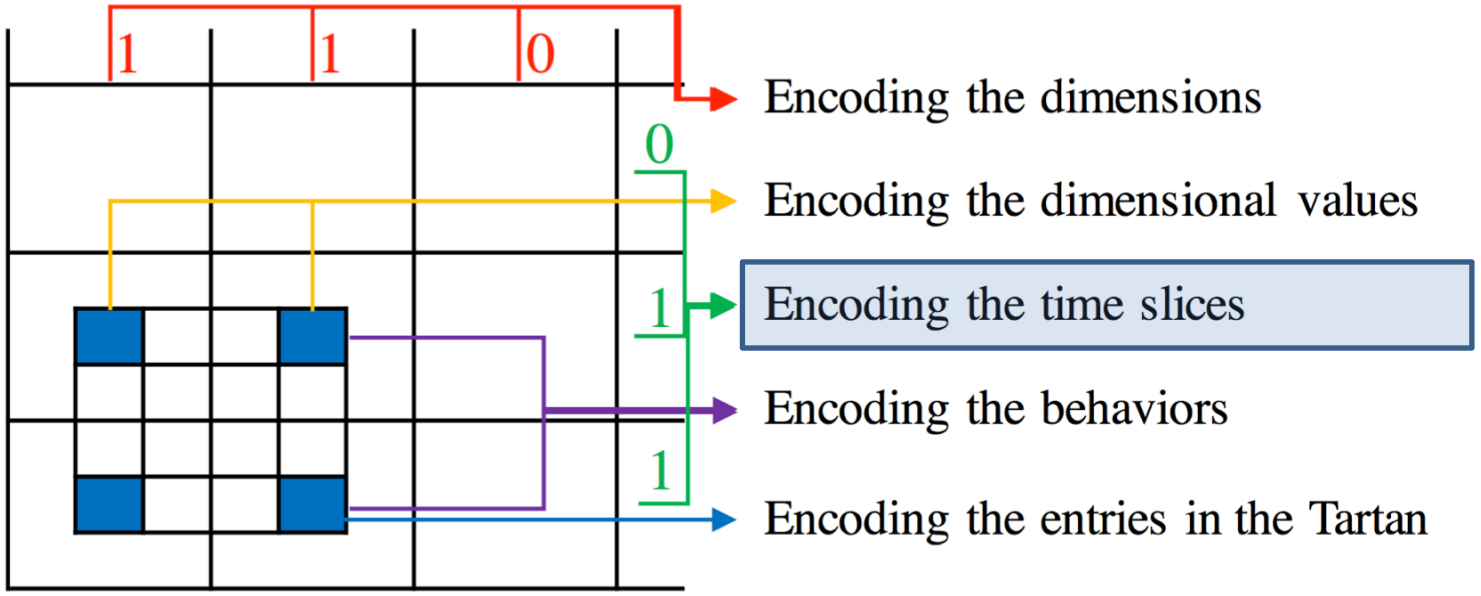
Encoding the behaviors

Encoding the entries in the Tartan

$$H_{\mathcal{V}_d}(X) = -\left(\frac{n_d}{N_d}\log\frac{n_d}{N_d} + \frac{N_d - n_d}{N_d}\log\frac{N_d - n_d}{N_d}\right).$$

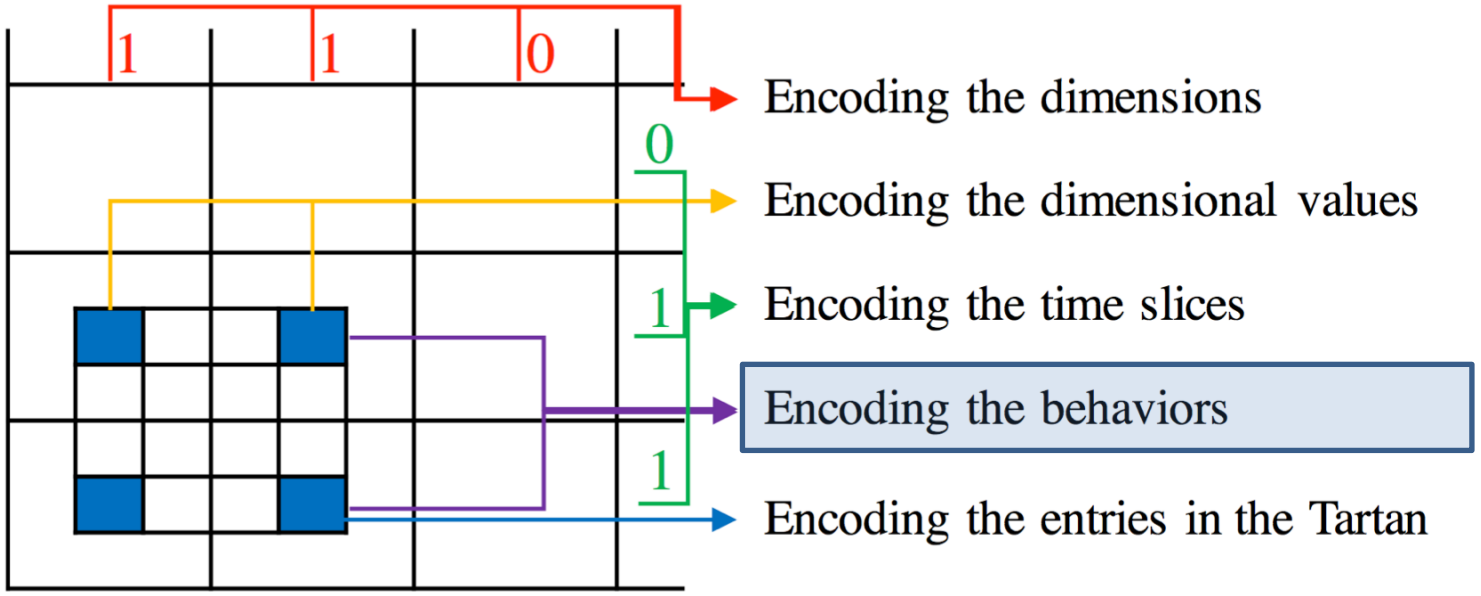
$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding the Tartan: Time Slices



$$L_{\mathcal{T}}(\mathcal{A}) \quad = \quad \log^* T + \log^* T^{\mathcal{A}} + \log^* t_{start}$$

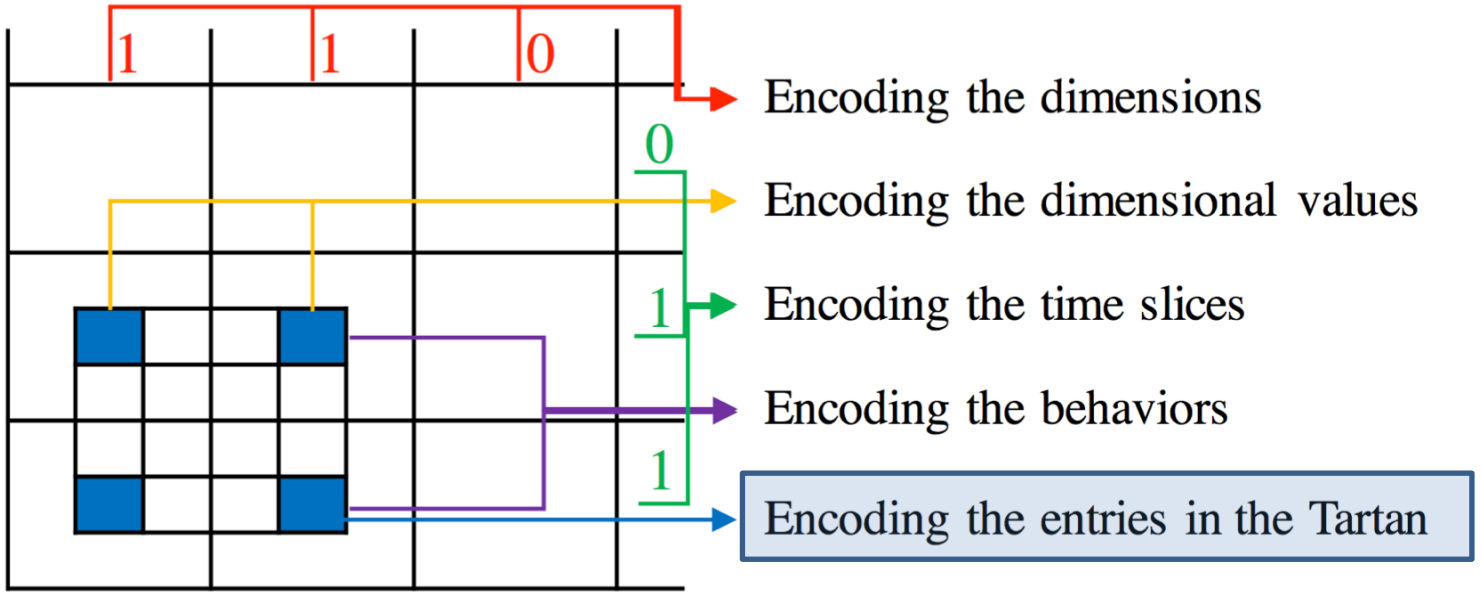
Encoding the Tartan: Behaviors



$$H_{\mathcal{B}^{(t)}}(X) = -\left(\frac{e^{(t)}}{E^{(t)}} \log \frac{e^{(t)}}{E^{(t)}} + \frac{E^{(t)}-e^{(t)}}{E^{(t)}} \log \frac{E^{(t)}-e^{(t)}}{E^{(t)}}\right).$$

$$L_{\mathcal{B}}(\mathcal{A}) = \sum_{t\in\mathcal{T}} \left(\log^* E^{(t)} + \log^* e^{(t)} + g(E^{(t)}, e^{(t)})\right).$$

Encoding the Tartan: Entries



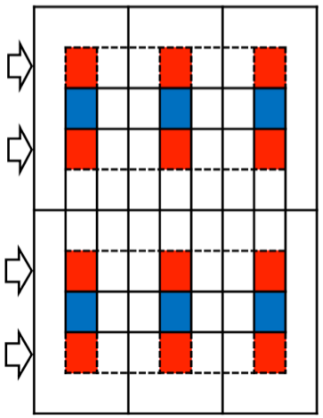
$$v = \left(\sum_{d\in\mathcal{D}} n_d\right)\left(\sum_{t\in\mathcal{T}} e^{(t)}\right).$$

$$c = \sum_{d\in\mathcal{D},t\in\mathcal{T}} \sum_{b\in\mathcal{B}^{(t)},i\in\mathcal{V}_d} \mathcal{X}_d^{(t)}(b,i).$$

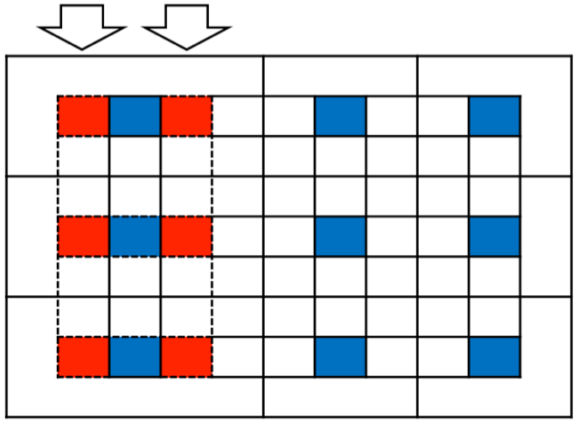
$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c}\log\frac{c}{v+c} + \frac{v}{v+c}\log\frac{v}{v+c}\right).$$

$$L_{\mathcal{A}}(\mathcal{A}) = (v+c)H_{\mathcal{A}}(X) = g(v+c,c).$$

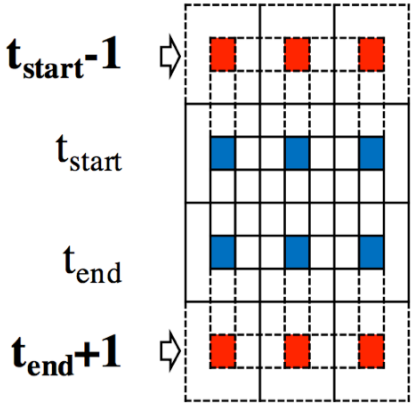
Greedy Search for the Local Minimum



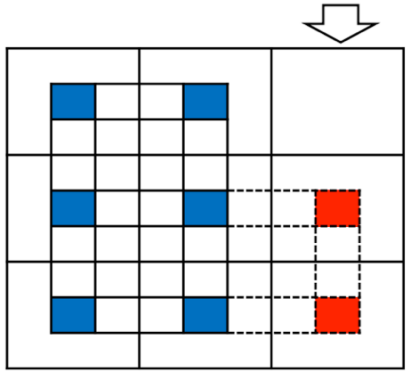
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.



(d) Update the set of dimensions.

Time complexity:

$$\mathcal{O}(\sum_d N_d \log N_d + \sum_t E^{(t)} \log E^{(t)})$$

Qualitative Analysis: DBLP data

1997 2000 2003 2006 2009 2012

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	**7** SIGIR VLDB TKDE	**7** "information retrieval" "data integration" "text classification"	**68** p56743[1] p62995 p76869	**32** 2003- 2007

[1] "A language modeling approach to information retrieval"

Venue	Keyword	#Paper
5 ICML NIPS …	**6** "reinforcement learning" "machine learning"	**40** 1997- 2002

Author	Venue	Cited	#Paper
6 Jiawei Han Xifeng Yan	**1** SIG- MOD	**1** p76095[2]	**22** 2004- 2010

[2] "Frequent subgraph discovery"

Venue	Keyword	#Paper
3 ICDM AAAI TKDE	**1** "anomaly detection"	**25** 2005- 2013

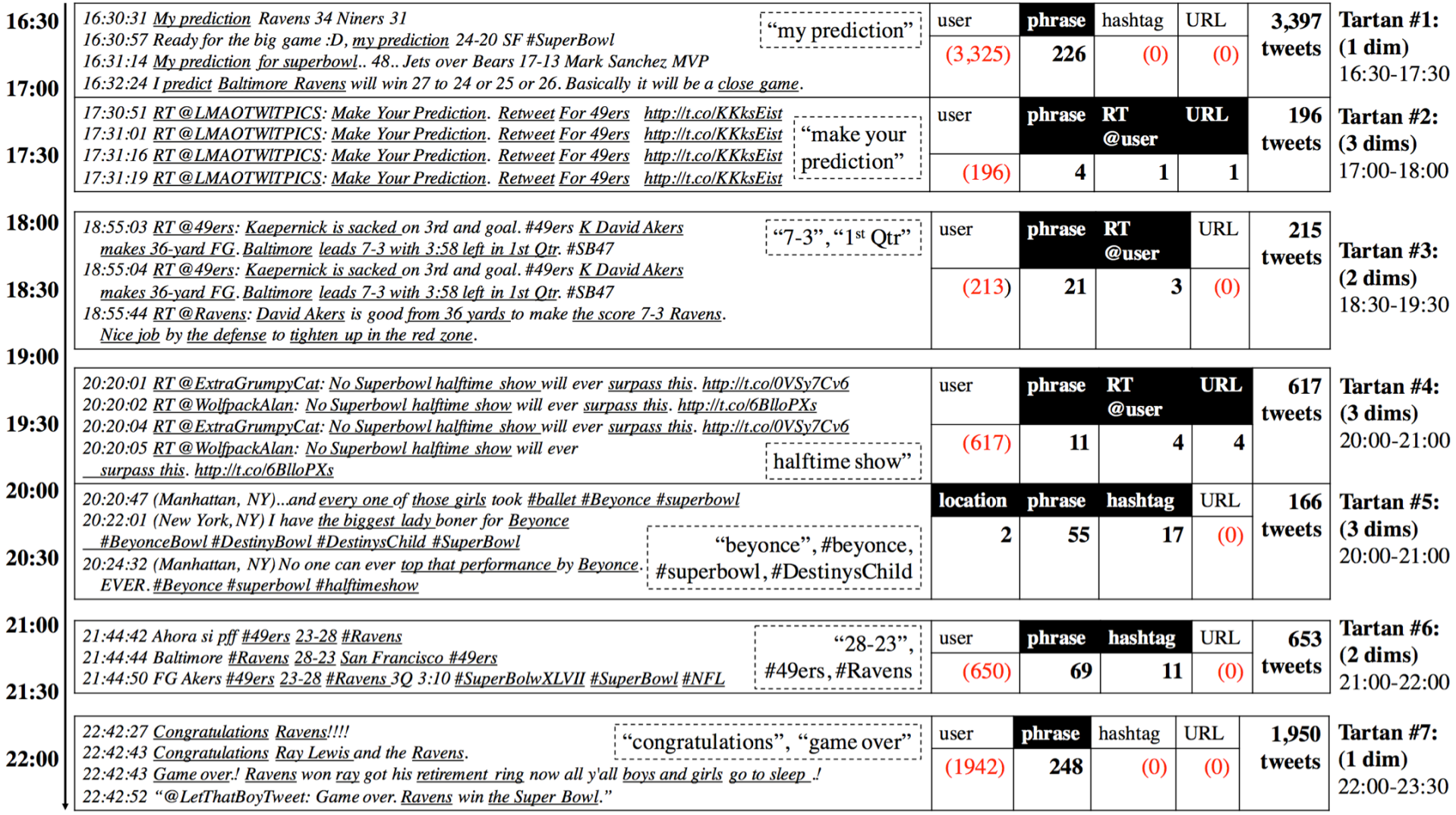
Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal…	**6** KDD ICDM ICDE TKDE …	**12** "large graphs" "data streams" "evolving data" "evolving graphs" …	**70** 2006- 2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng…	**5** SIGIR WWW WSDM CIKM…	**3** "web search" "click-through data" "sponsored search"	**12** p82630[3] p116290 p103899 p106191…	**32** 2006- 2013

[3] "Optimizing search engines using clickthrough data"

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan…	**3** KDD PAKDD AAAI	**6** "transfer learning" "data mining" "localization models"	**17** 2007- 2010

Qualitative Analysis: Super Bowl 2013



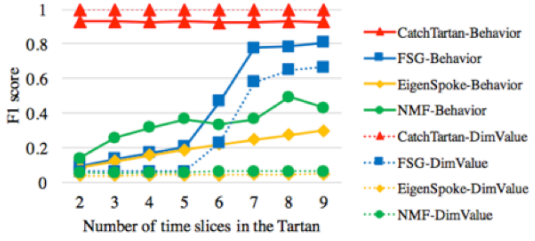
Quantitative Analysis: Accuracy and Efficiency in Synthetic Experiments

- Tartan distribution

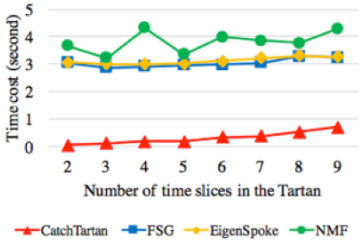
1. $T^{\mathcal{A}} \in [2,\ 9]$, the number of consecutive time slices in the Tartan $\mathcal{A}$, 4 as default;
2. $e^{(t)} \in [100,\ 2,000]$, the number of behaviors in the time slice, $1,000$ as default;
3. $D^{\mathcal{A}} \in [2,\ 9]$, the number of dimensions in $\mathcal{A}$, 3 as default;
4. $n_d \in [50,\ 200]$, the number of values per dimension in $\mathcal{A}$, 100 as default;
5. $\rho \in [1,\ 10]$, the average number of values per dimension in the behaviors, 3 as default;

- Data distribution

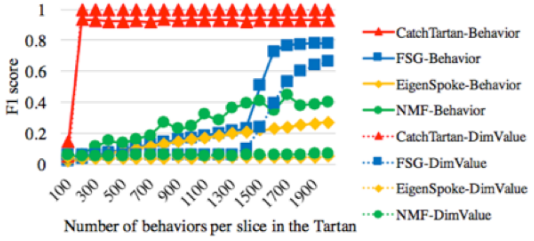
6. $T \in [5, 30]$, the total number of time slices in the dataset, 10 as default;
7. $E^{(t)} \in [1,000,\ 10,000]$, the number of behaviors per time slice in the dataset, $5,000$ as default;
8. $N_d \in [1,000,\ 2,000]$, the number of values per dimension in the data, $1,000$ as default.



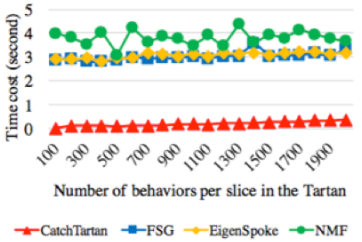
(a) F1 score vs $T^{\mathcal{A}}$.



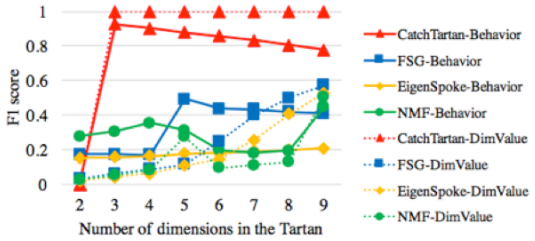
(b) Time cost vs $T^{\mathcal{A}}$.



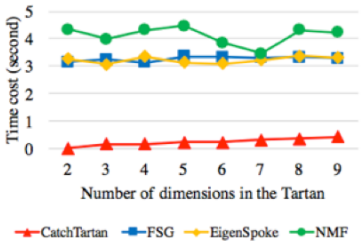
(c) F1 score vs $e^{(t)}$.



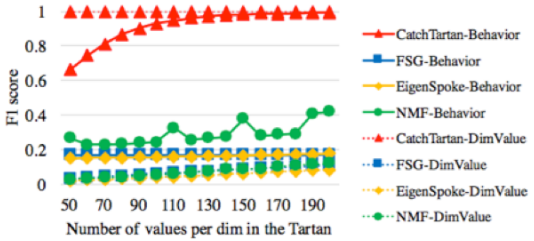
(d) Time cost vs $e^{(t)}$.



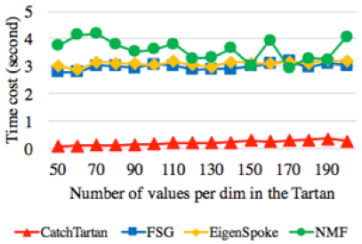
(e) F1 score vs $D^{\mathcal{A}}$.



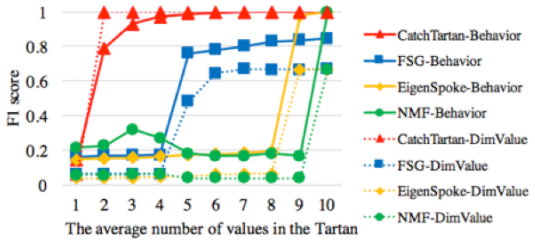
(f) Time cost vs $D^{\mathcal{A}}$.



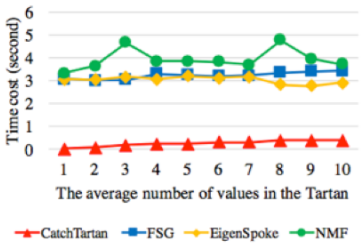
(g) F1 score vs n_d.



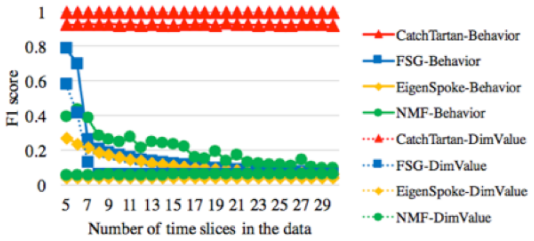
(h) Time cost vs n_d.



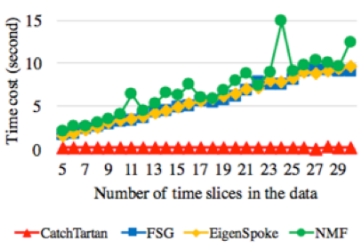
(i) F1 score vs ρ.



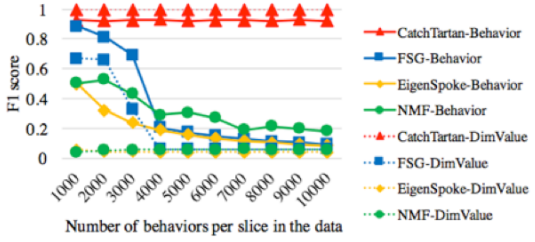
(j) Time cost vs ρ.



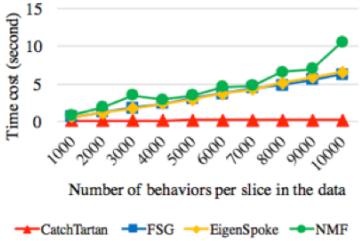
(k) F1 score vs T.



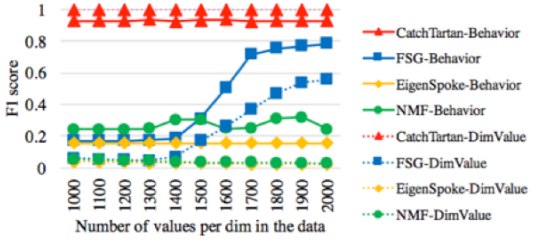
(l) Time cost vs T.



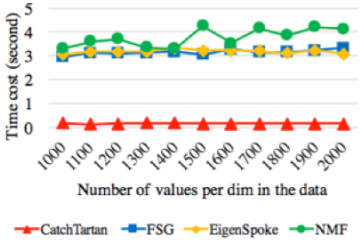
(m) F1 score vs $E^{(t)}$.



(n) Time cost vs $E^{(t)}$.



(o) F1 score vs N_d.



(p) Time cost vs N_d.

Summary

- Novel representations
  - Behavior: "two-level matrix" vs. tensor
  - Behavioral summary: "Tartan" vs. dense block
- A new summarization algorithm
  - Principled-scoring and Parameter-free: Objective function based on Minimum Description Length
  - Scalable: Greedy search for local optimum
- Effectiveness, discovery and efficiency