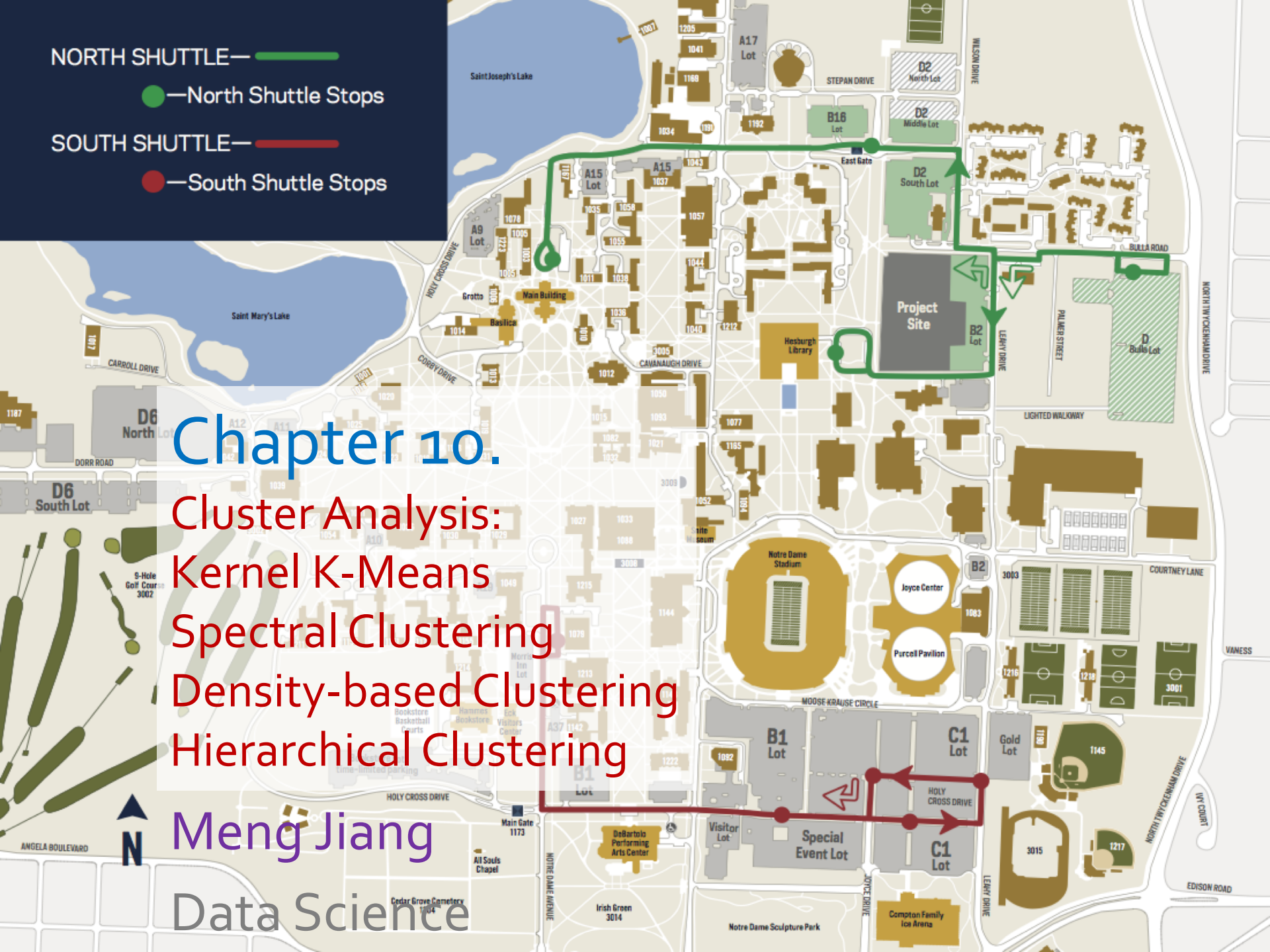
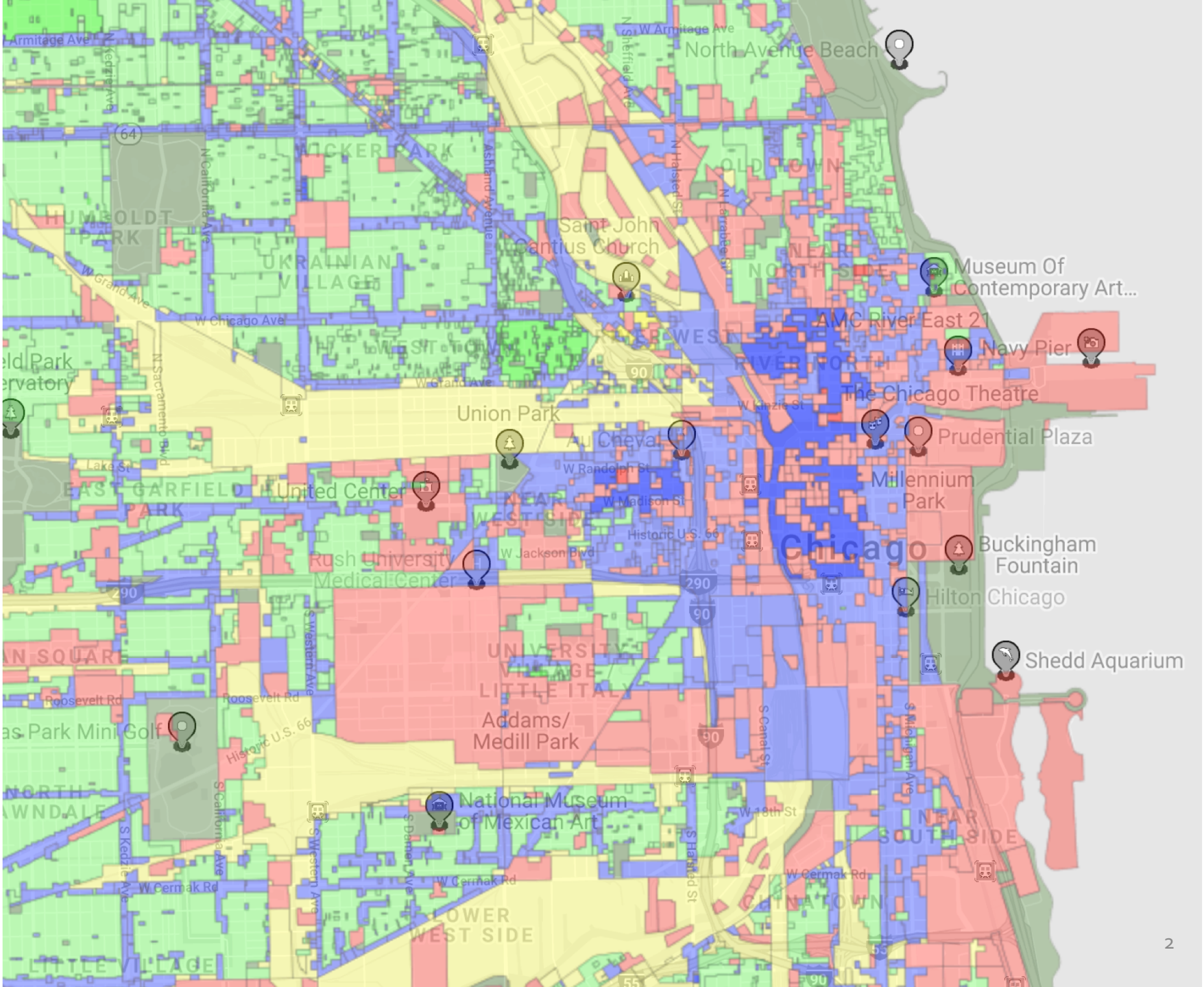
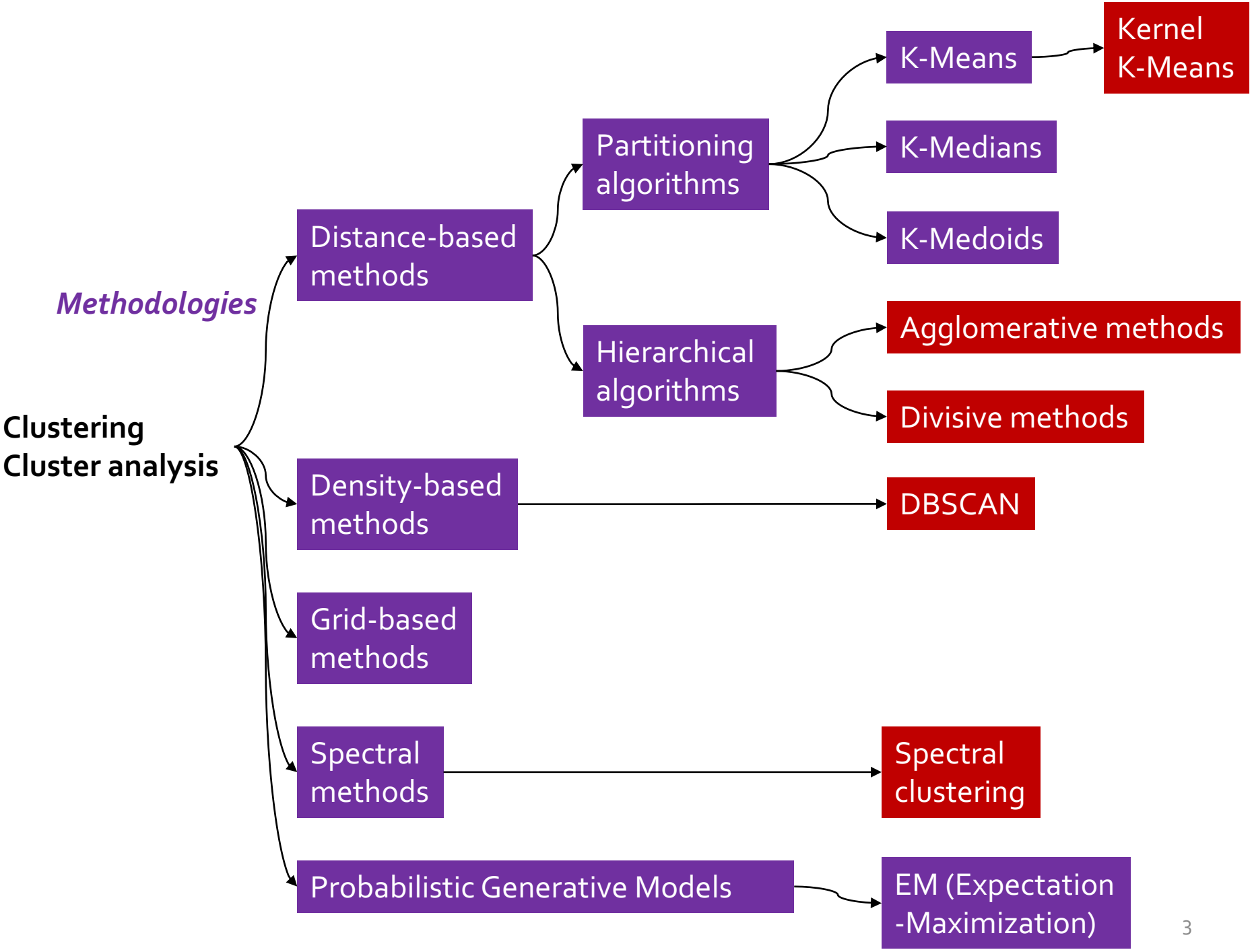

NORTH SHUTTLE—
—North Shuttle Stops
SOUTH SHUTTLE—
—South Shuttle Stops
Project Site
Chapter 10.
Cluster Analysis:
Kernel K-Means
Spectral Clustering
Density-based Clustering
Hierarchical Clustering
Meng Jiang
Data Science


North Avenue Beach
Saint John Cantius Church
Museum Of Contemporary Art...
AMC River East 21
Navy Pier
The Chicago Theatre
Prudential Plaza
Union Park
United Center
Millennium Park
Chicago
Buckingham Fountain
Hilton Chicago
Shedd Aquarium
Rush University Medical Center
Addams/ Medill Park
National Museum of Mexican Art
WICKER PARK
UKRAINIAN VILLAGE
OLD TOWN
NEAR NORTH SIDE
RIVER NORTH
NEAR WEST SIDE
EAST GARFIELD PARK
UNIVERSITY VILLAGE
LITTLE ITALY
NEAR SOUTH SIDE
CHINATOWN
LOWER WEST SIDE
LITTLE VILLAGE
W Armitage Ave
W Grand Ave
W Chicago Ave
W Randolph St
W Madison St
W Jackson Blvd
W Cermak Rd
W 18th St
Roosevelt Rd
Lake St
Historic U.S. 66
Historic U.S. 66
N Halsted St
N Sheffield Ave
N Ashland Ave
N California Ave
N Sacramento Blvd
S Western Ave
S Damen Ave
S Canal St
S Michigan Ave
S Halsted St
S Kedzie Ave
90
290
64
55


Methodologies
Clustering
Cluster analysis
Distance-based methods
Partitioning algorithms
K-Means
Kernel 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
EM (Expectation -Maximization)

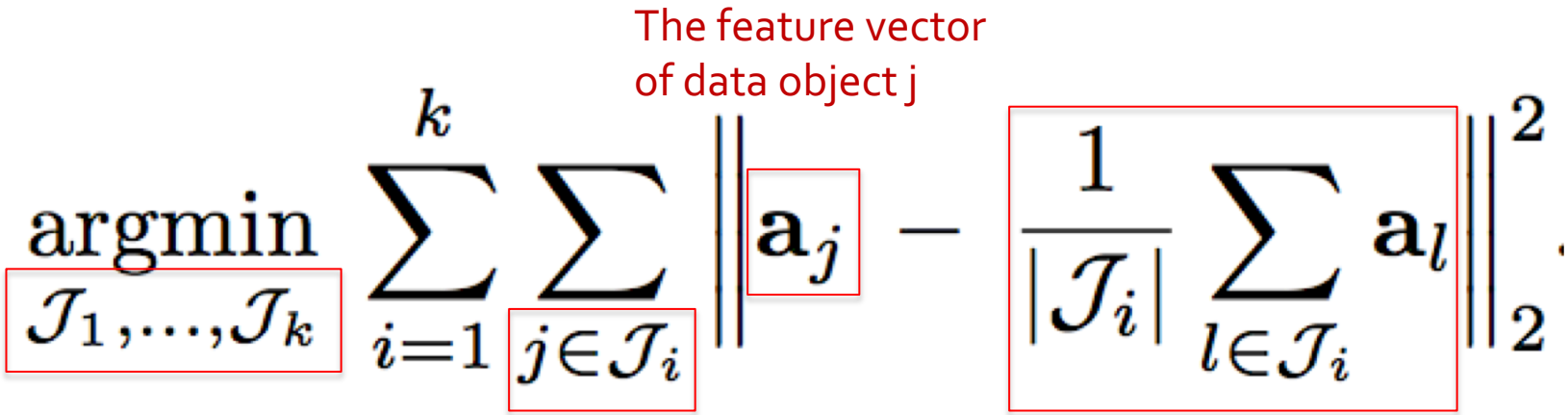
Outline

- **Kernel K-Means**
- Spectral Clustering
- Density-based Clustering: DBSCAN
- Hierarchical Clustering: Agglomerative and Divisive

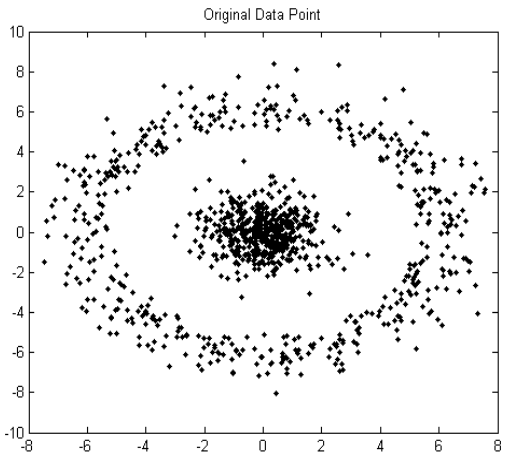
SSE for K-Means Clustering

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \mathbf{a}_j - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \mathbf{a}_l \right\|_2^2$$

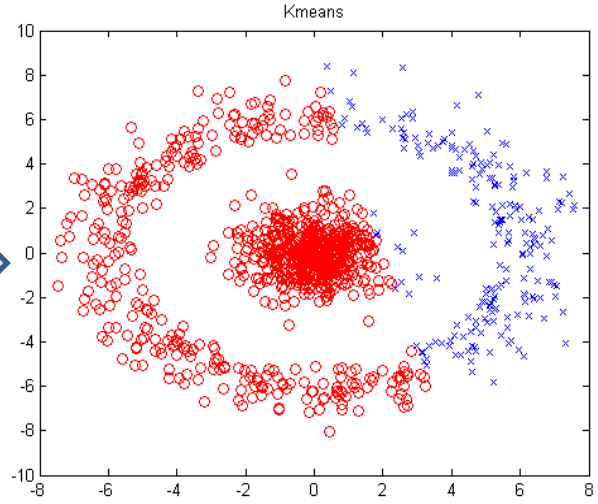
SSE for K-Means Clustering



K-Means' Limitation on Cluster Shape



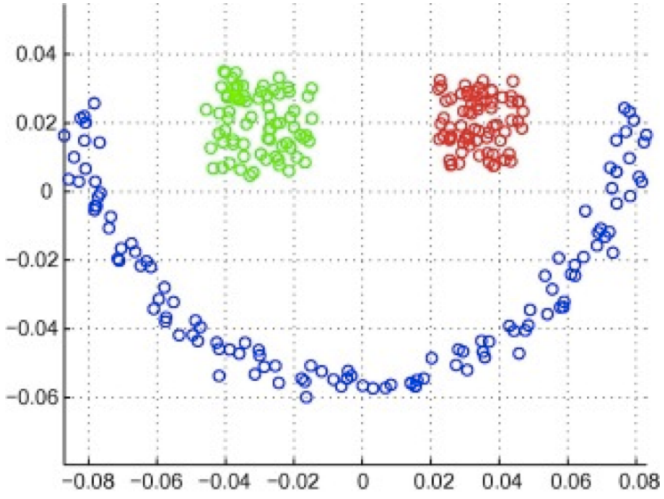
The original data set



The result of K-Means clustering

Kernel K-Means

- Perform k-means, but in a **different feature space**
- Conceptually, use $\varphi(x_i)$, instead of x_i, as the points you are clustering
- $\varphi(x_i)$ is a vector-valued function of x_i
- BUT you **never need to compute $\varphi(x_i)$**
- You do need to compute and store **n x n kernel matrix** generated from the kernel function on the original data



- Computational complexity is higher than K-Means

Kernel Tricks

Pick yourself a kernel. Common choices are

- Polynomial kernel: $K(x_i,x_j) = (x_i^T x_j + c)^d$
  - c is a free parameter (can tune for performance)
  - d is the desired degree of the polynomial
- RBF kernel: $K(x_i,x_j) = \exp(-\gamma|x_i-x_j|^2)$
  - $\gamma = 1/2\sigma^2$

Recall that $K(x_i,x_j) = \varphi(x_i)^T\varphi(x_j)$
The kernel trick allows you to avoid computing $\varphi(x_i)$

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \cdots$$

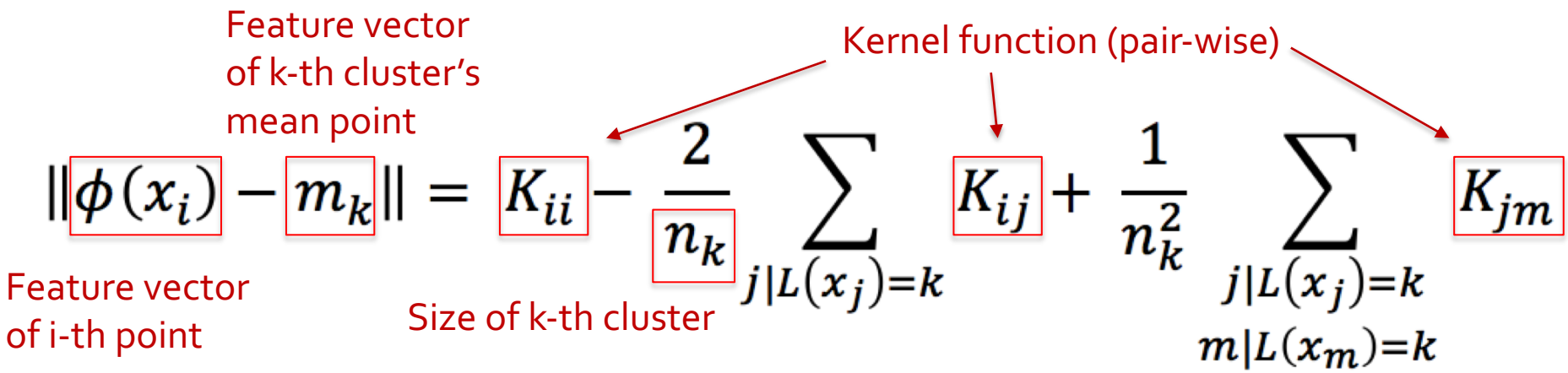
Error/Distance Computation Using Kernel Matrix

- The distance between data object and cluster's representative point (mean point)

$$\|\phi(x_i) - m_k\| = K_{ii} - \frac{2}{n_k} \sum_{j|L(x_j)=k} K_{ij} + \frac{1}{n_k^2} \sum_{\substack{j|L(x_j)=k \\ m|L(x_m)=k}} K_{jm}$$

Error/Distance Computation Using Kernel Matrix

- The distance between data object and cluster's representative point (mean point)



SSE of Kernel K-Means

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \mathbf{a}_j - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \mathbf{a}_l \right\|_2^2$$

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \boldsymbol{\phi}(\mathbf{a}_j) - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \boldsymbol{\phi}(\mathbf{a}_l) \right\|_2^2$$

$$\kappa(\mathbf{a}_i, \mathbf{a}_j) = \langle \boldsymbol{\phi}(\mathbf{a}_i), \boldsymbol{\phi}(\mathbf{a}_j) \rangle.$$

RBF: Example

- Gaussian radial basis function (RBF) kernel:

$$K(X_i, X_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2} \qquad K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$$

- Suppose there are 5 original 2-dimensional points:
  - x_1(0, 0), x_2(4, 4), x_3(-4, 4), x_4(-4, -4), x_5(4, -4)
- If we set σ to 4, we will have the following points in the kernel space, e.g.,

$$\left|\left|x_1 - x_2\right|\right|^2 = (0-4)^2 + (0-4)^2 = 32,$$

therefore, $K(x_1, x_2) = e^{-\frac{32}{2 \cdot 4^2}} = e^{-1}$

RBF: Example (cont.)

Original Space

RBF Kernel Space ($\sigma = 4$)

	x	y	$K(x_i, x_1)$	$K(x_i, x_2)$	$K(x_i, x_3)$	$K(x_i, x_4)$	$K(x_i, x_5)$
x_1	0	0	1	$e^{-\frac{4^2+4^2}{2\cdot 4^2}} = e^{-1}$	e^{-1}	e^{-1}	e^{-1}
x_2	4	4	e^{-1}	1	e^{-2}	e^{-4}	e^{-2}
x_3	-4	4	e^{-1}	e^{-2}	1	e^{-2}	e^{-4}
x_4	-4	-4	e^{-1}	e^{-4}	e^{-2}	1	e^{-2}
x_5	4	-4	e^{-1}	e^{-2}	e^{-4}	e^{-2}	1

Poly: Example

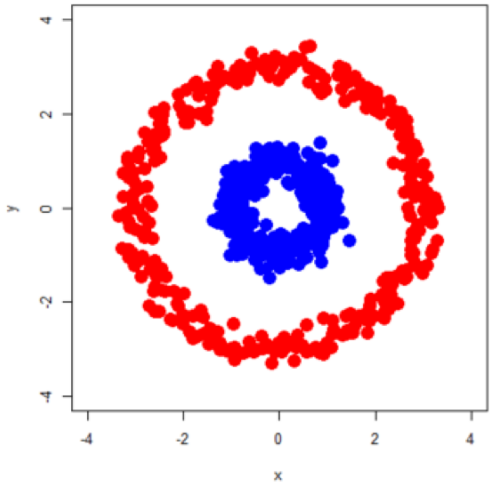
- Polynomial kernel of degree h=2:

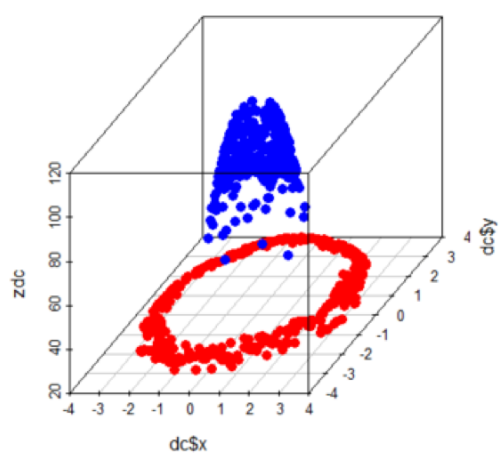
$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \boldsymbol{X}_i \cdot \boldsymbol{X}_j{}^2 \rightarrow \phi(x, y) = (x^2, \sqrt{2}xy, y^2)$$

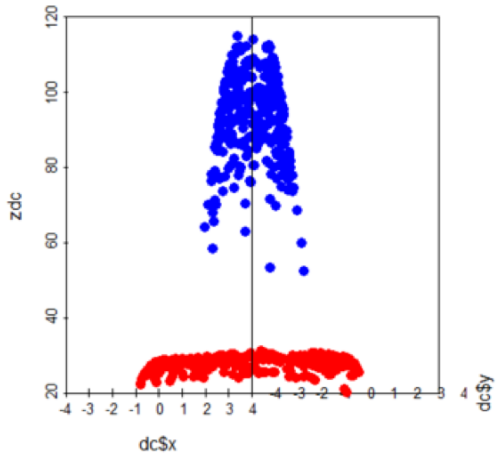
- Suppose there are 5 original 2-dimensional points:
  - $x_1(0, 0)$, $x_2(4, 4)$, $x_3(-4, 4)$, $x_4(-4, -4)$, $x_5(4, -4)$

	x	y	$\Phi_1 = x^2$	$\Phi_2 = xy$	$\phi_3 = y^2$	$K(x_i, x_1)$	$K(x_i, x_2)$	$K(x_i, x_3)$	$K(x_i, x_4)$	$K(x_i, x_5)$
x_1	0	0	0	0	0	0	0	0	0	0
x_2	4	4	16	$16\sqrt{2}$	16	0	32^2	0	32^2	0
x_3	-4	4	16	$-16\sqrt{2}$	16	0	0	32^2	0	32^2
x_4	-4	-4	16	$16\sqrt{2}$	16	0	32^2	0	32^2	0
x_5	4	-4	16	$-16\sqrt{2}$	16	0	0	32^2	0	32^2

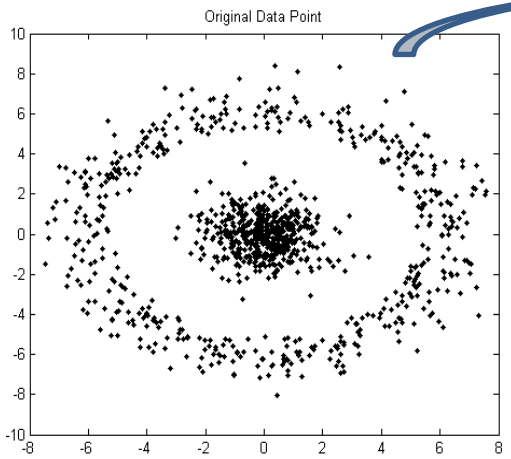
Poly: Example (cont.)



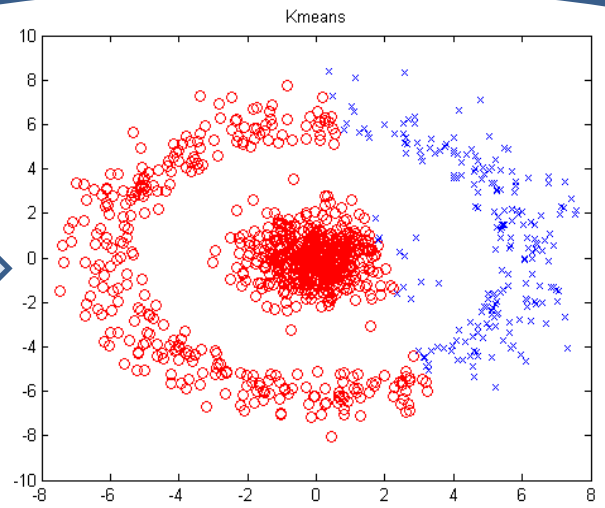




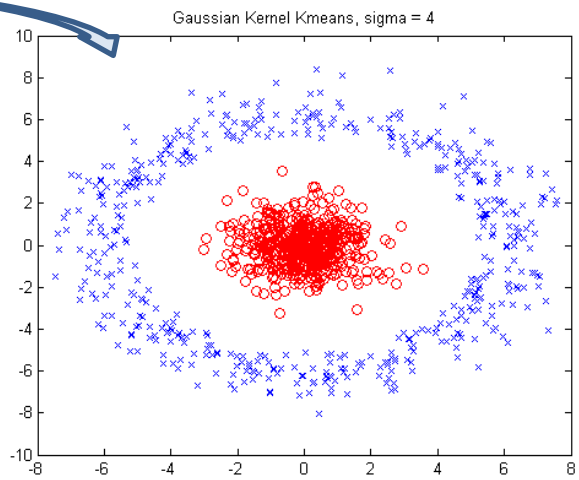
Results



The original data set



The result of K-Means clustering

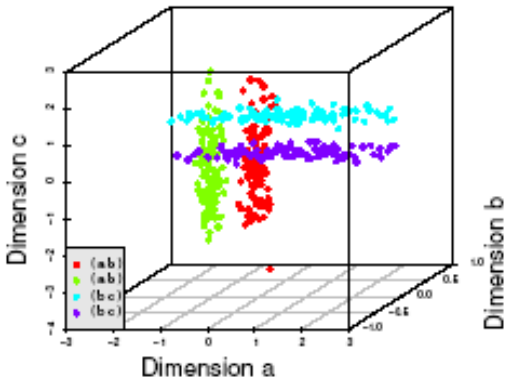


The result of Gaussian Kernel K-Means clustering

- The above data set cannot generate quality clusters by K-Means since it contains non-convex clusters
- Gaussian RBF Kernel transformation maps data to a kernel matrix K for any two points x_i, x_j: $K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$ and Gaussian kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$
- K-Means clustering is conducted on the mapped data, generating quality clusters

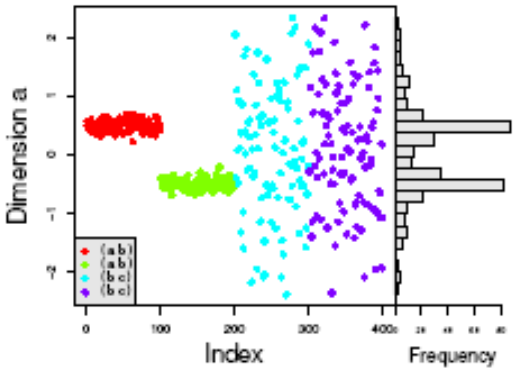
Outline

- Kernel K-Means
- **Spectral Clustering**
- Density-based Clustering: DBSCAN
- Hierarchical Clustering: Agglomerative and Divisive

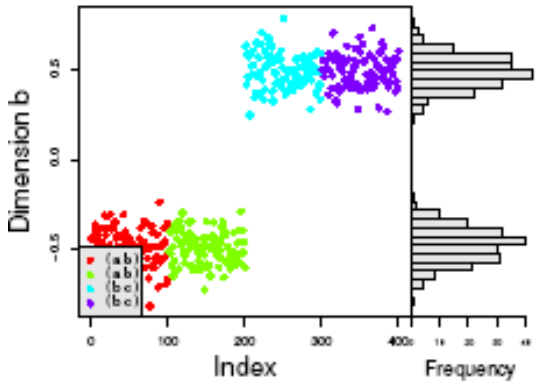


Why Subspace Clustering?

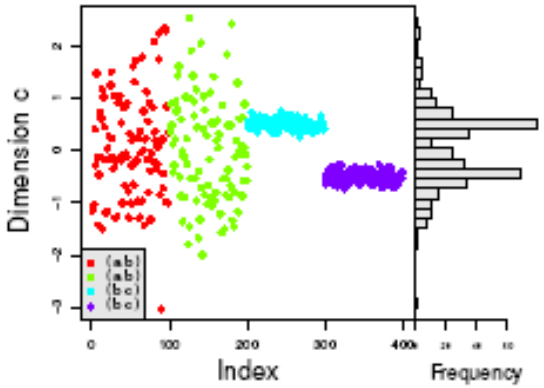
- Clusters may exist only in some subspaces
- Subspace-clustering: find clusters in all the subspaces



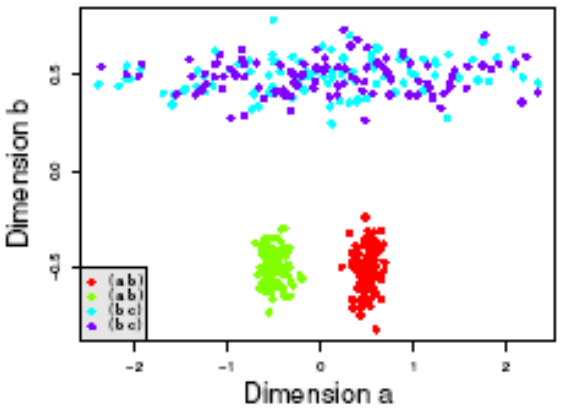
(a) Dimension a



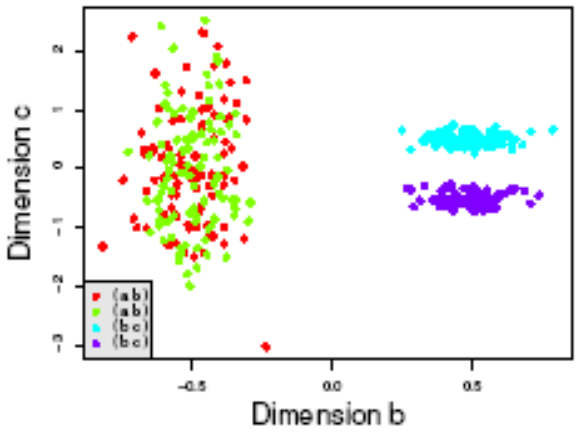
(b) Dimension b



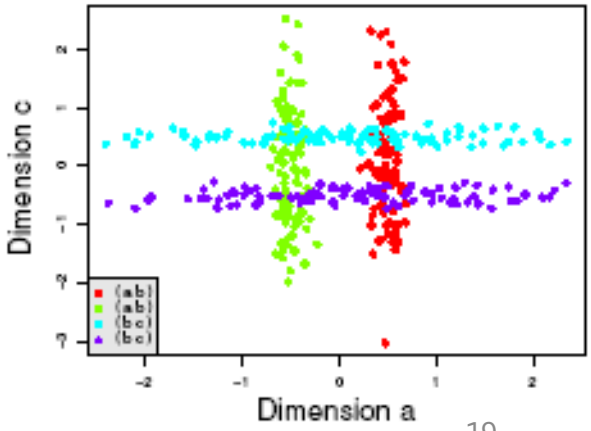
(c) Dimension c



(a) Dims a & b

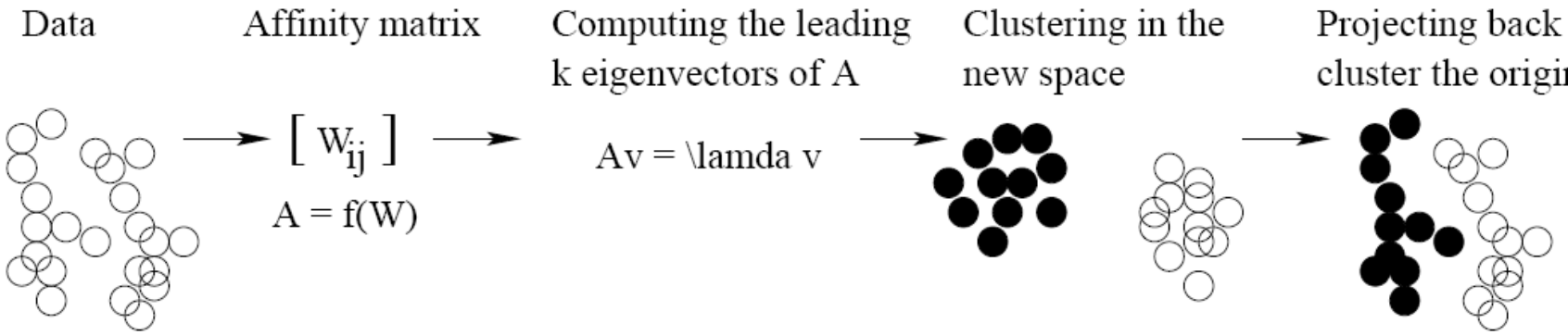


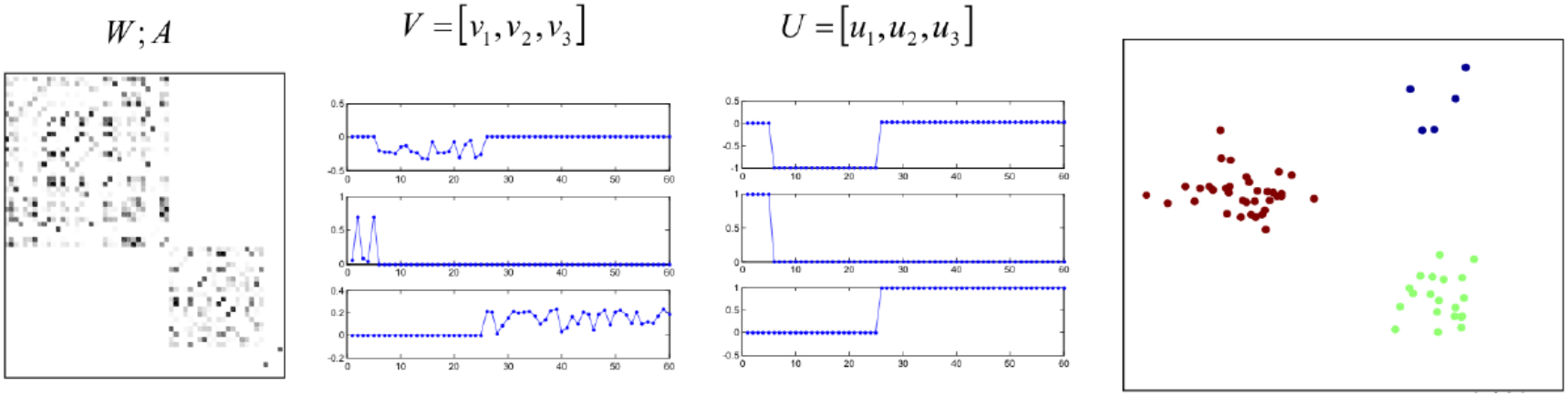
(b) Dims b & c



(c) Dims a & c

Spectral (Subspace) Clustering





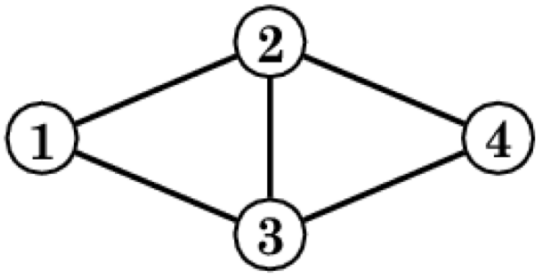
- Spectral clustering: Effective in tasks like image processing
- Scalability challenge: Computing eigenvectors on a large matrix is costly
- Can be combined with other clustering methods

Spectral Clustering: Algorithm

- Input: n data objects and m features, number of clusters k
- Step 1: Build n-by-n similarity graph W (or called Laplacian matrix)
  - D: degree matrix
  - A: adjacency matrix
  - L = D-A

$$L_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$



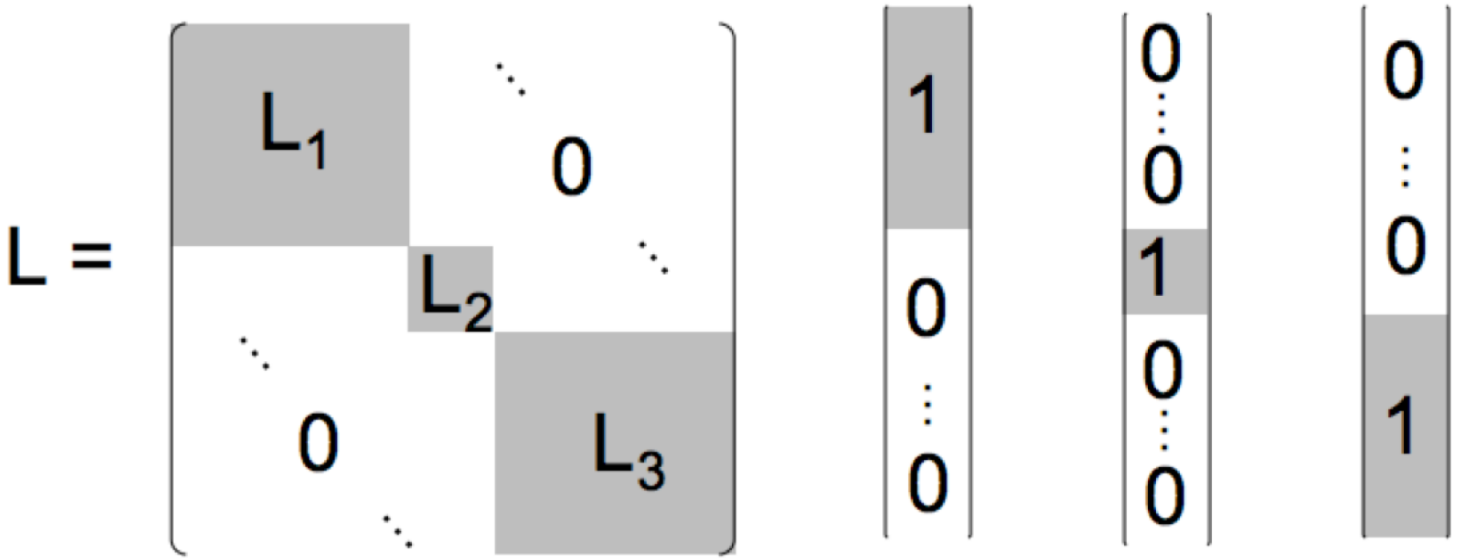
$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$L = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

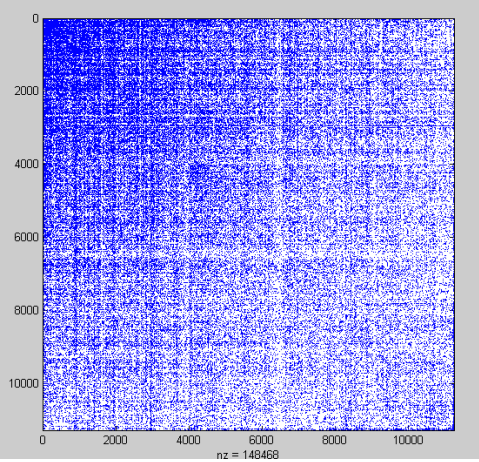
Spectral Clustering: Algorithm (cont.)

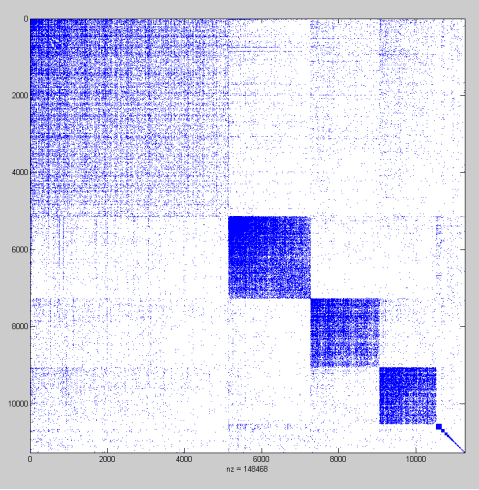
Step 2: Compute the first k eigenvectors v_1, …, v_k of the matrix; Build the n-by-k matrix V with eigenvectors as columns

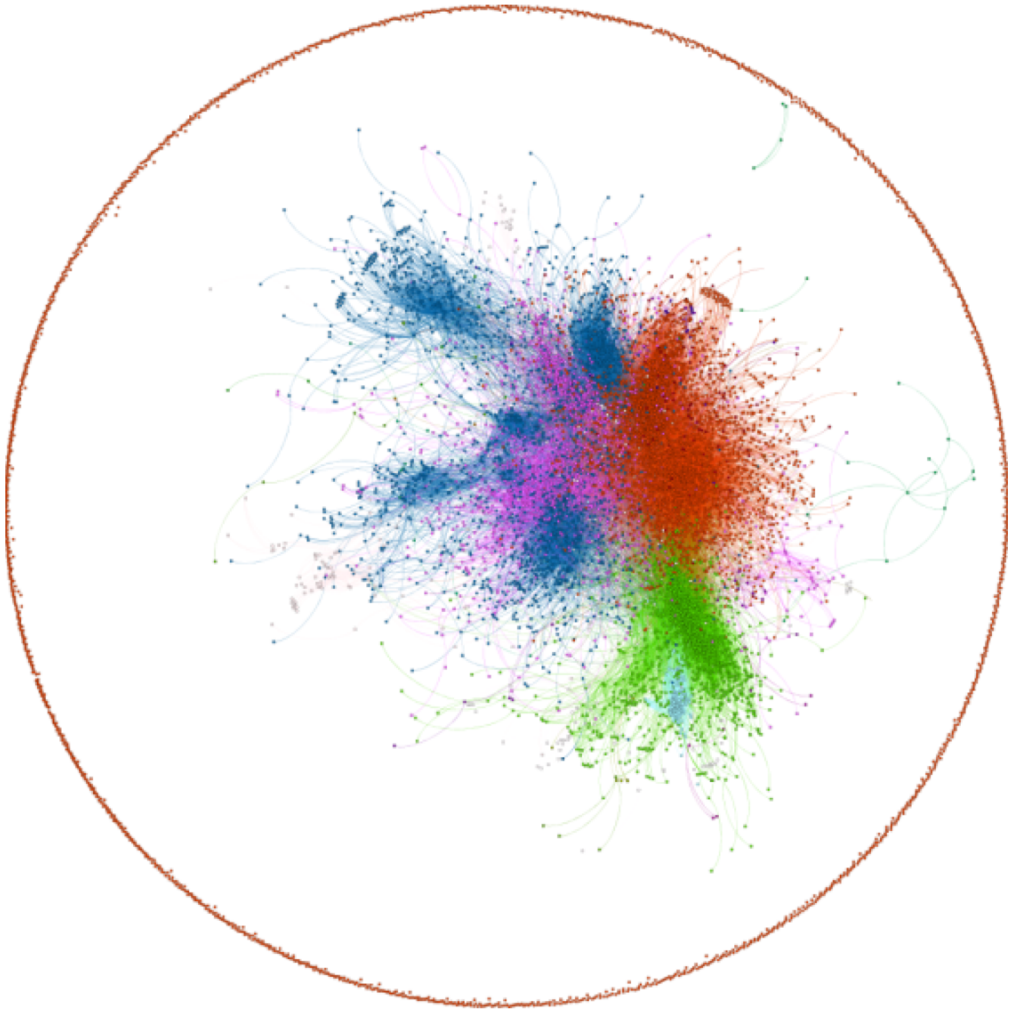
Step 3: Interpret the rows of V as new data points in k-dimensional feature space; Cluster the points with the K-Mearns algorithms

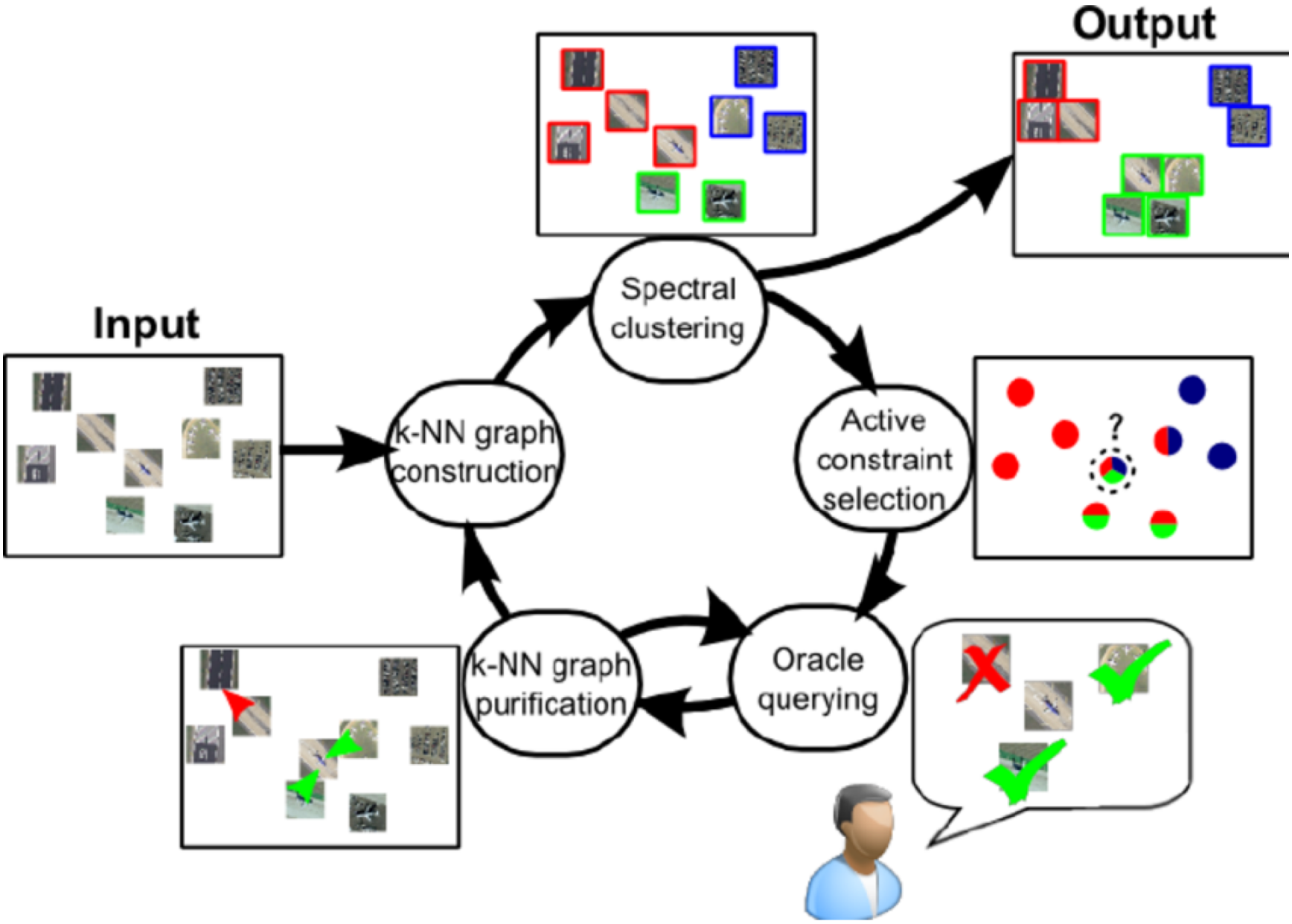


Spy Plot (MATLAB)








Output
Input
Spectral clustering
k-NN graph construction
Active constraint selection
?
k-NN graph purification
Oracle querying

Outline

- Kernel K-Means
- Spectral Clustering
- **Density-based Clustering: DBSCAN**
- Hierarchical Clustering: Agglomerative and Divisive

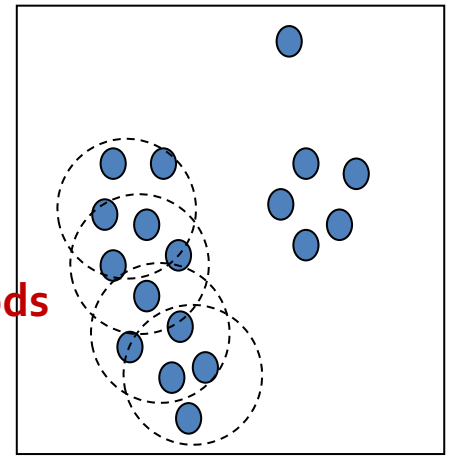
Compare Inputs for Clustering

- Kernel K-Means
  - Input: Objects and their features
  - Issue: Arbitrary shape of clusters in original feature space
  - Idea: Kernel feature space
- Spectral Clustering
  - Input: Objects and their features:
  - Issue: High dimensionality – too many features
  - Idea: Low-dimensional spectral subspace
- Density-based Clustering
  - Input: Objects and their distances
  - Issue: Sometimes no feature value + Arbitrary shape + Outliers
  - Idea?

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

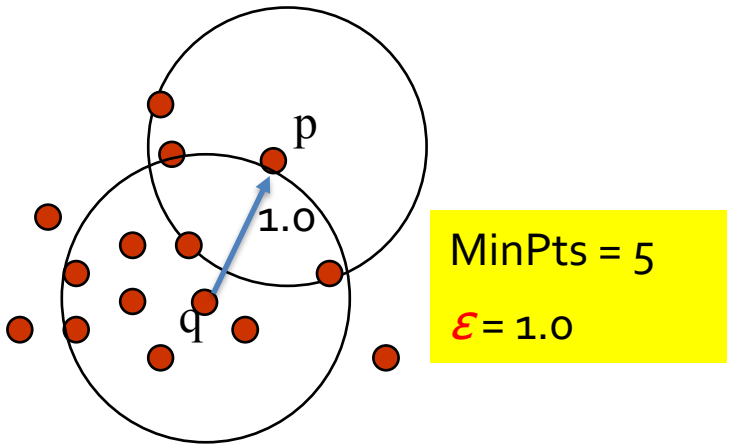
DBSCAN

- DBSCAN (M. Ester, H.-P. Kriegel, J. Sander, X. Xu, KDD'96)
  - Discovers clusters of arbitrary shape: Density-Based Spatial Clustering of Applications with Noise
- 2014 KDD test-of-time award
- Idea: "Finds core samples of high density and expands clusters from them (by neighborhoods)"
- A ***cluster*** is defined as a **maximal** set of **density-connected** points

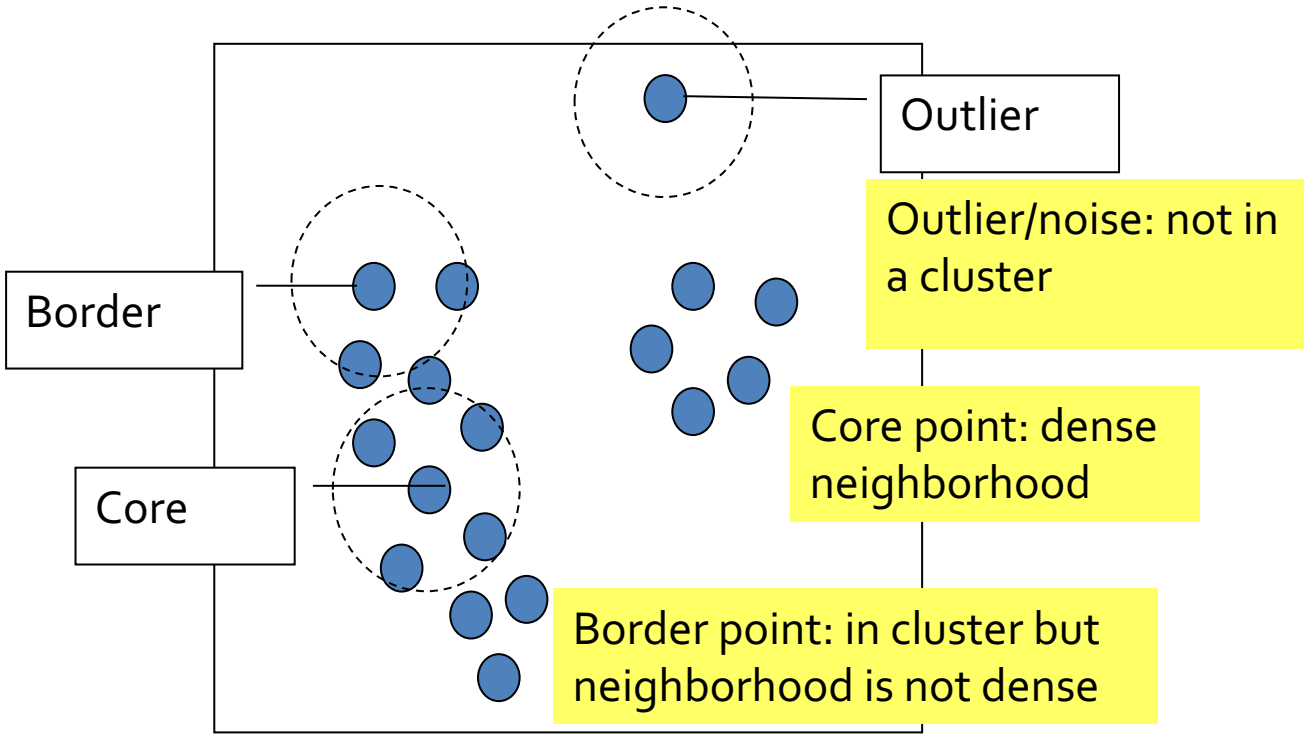


Two Parameters

- ε: Maximum radius of the neighborhood
- *MinPts*: Minimum number of points in the Eps-neighborhood of a point
- The ε-neighborhood of a point q:
  - $N_{\varepsilon}(q)$: {p belongs to D | dist(p, q) ≤ ε}

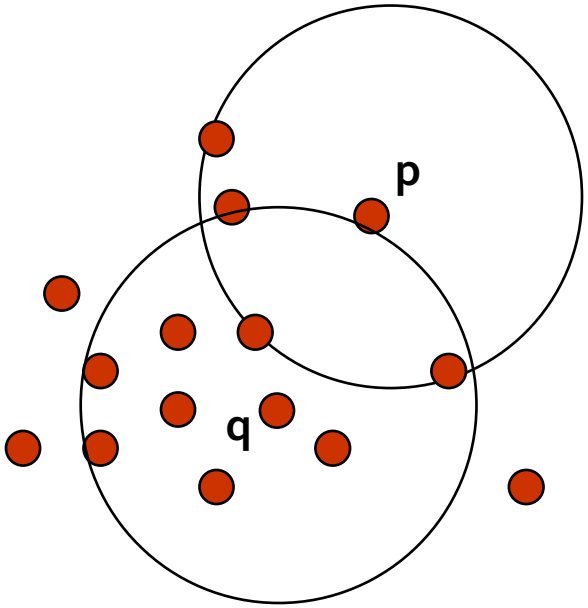


Three Kinds of Points



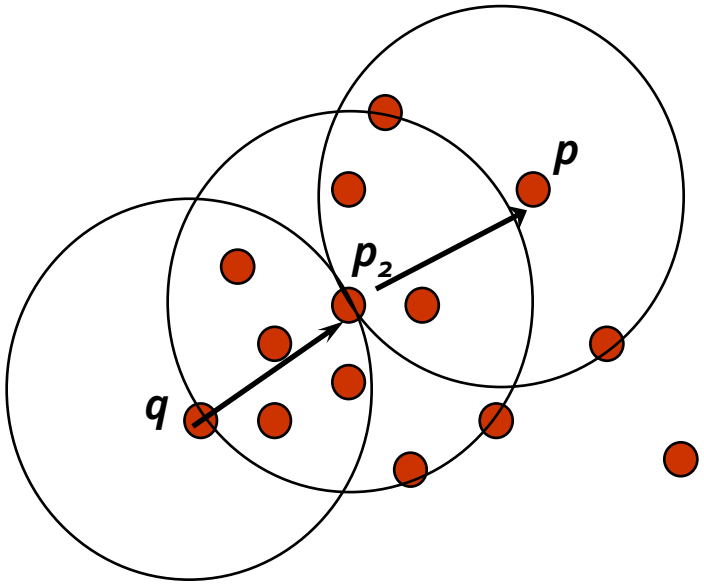
Three Concepts

- **C1) Directly density-reachable**:
  - A point *p* is directly density-reachable from a point *q* if
    - *p* belongs to q's ε-neighborhood $N_{\varepsilon}(q)$
    - **and** q is a core point: $|N_{\varepsilon}(q)| \geq MinPts$ (dense neighborhood)



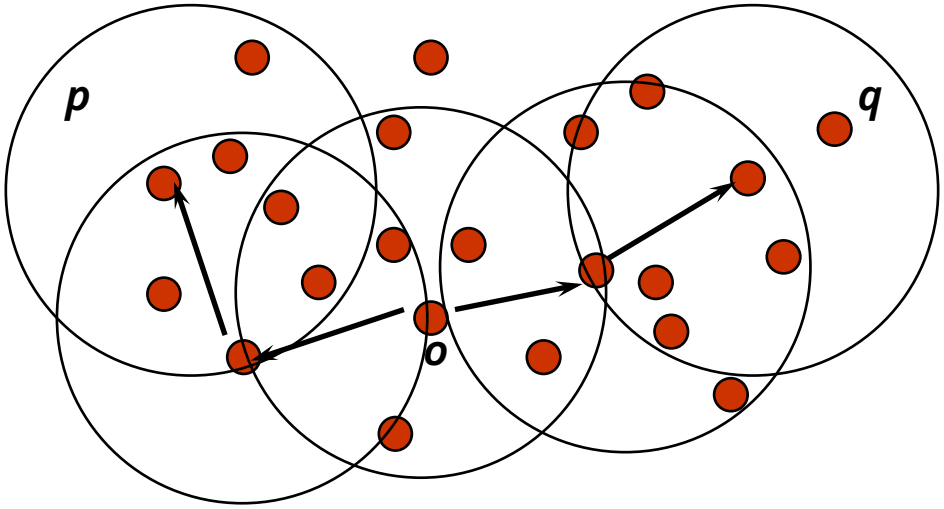
Three Concepts (cont.)

- **C2) Density-reachable**:
  - A point *p* is density-reachable from a point *q* if there is a chain of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is *directly density-reachable* from p_i



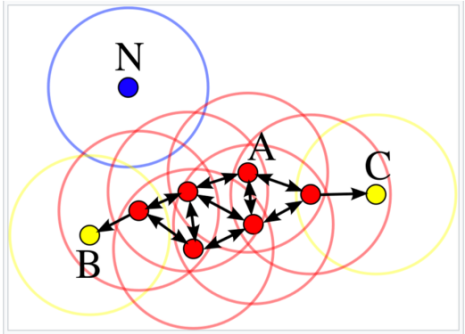
Three Concepts (cont.)

- **Density-connected:**
  - A point *p* is density-connected to a point *q* if there is a point *o* such that *p* and *q* are **both** *density-reachable* from *o* w.r.t. *ε* and *MinPts*



DBSCAN: Algorithm

- Specify *ε* and *MinPts* (guess)
- Arbitrarily select a point *p*
- Retrieve all points *density-reachable* from *p*
  - If *p* is a core point, a cluster is formed
  - If *p* is a border point, no points are density-reachable from *p*, and DBSCAN visits the next point of the database
- Continue until *all* of the points have been processed



In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

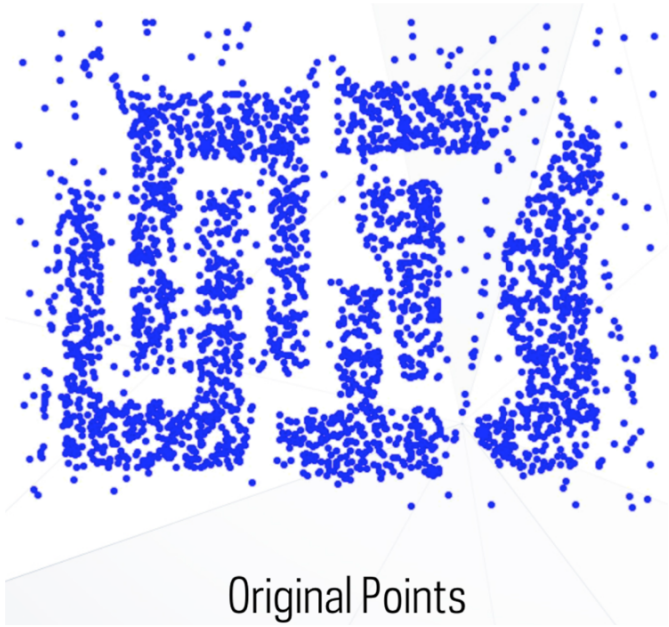
DBSCAN: Properties

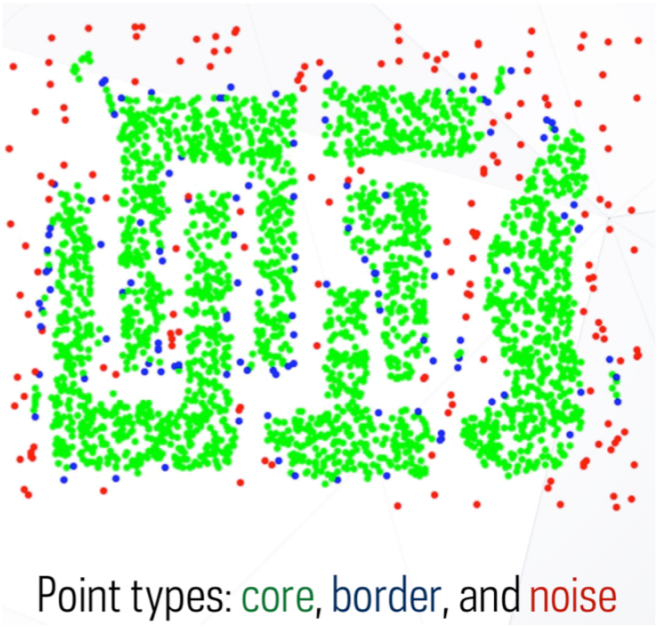
- **Maximality**: if p in a cluster, and q is density-reachable from p, then q is in the same cluster
- **Connectivity**: any pair of points p, q in a cluster are density-connected
- **Computational complexity**:
  - If a spatial index is used, the computational complexity of DBSCAN is *O(nlogn)*, where n is the number of database objects
  - Otherwise, the complexity is *O(n2)*

DBSCAN: Advantages

- Resistant to noise (outliers)
- Arbitrary cluster shape is OK
- One scan (only examine the local region to justify density)

Figure from Ackerman's site





DBSCAN: Disadvantage

- Sensitive to the Setting of Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

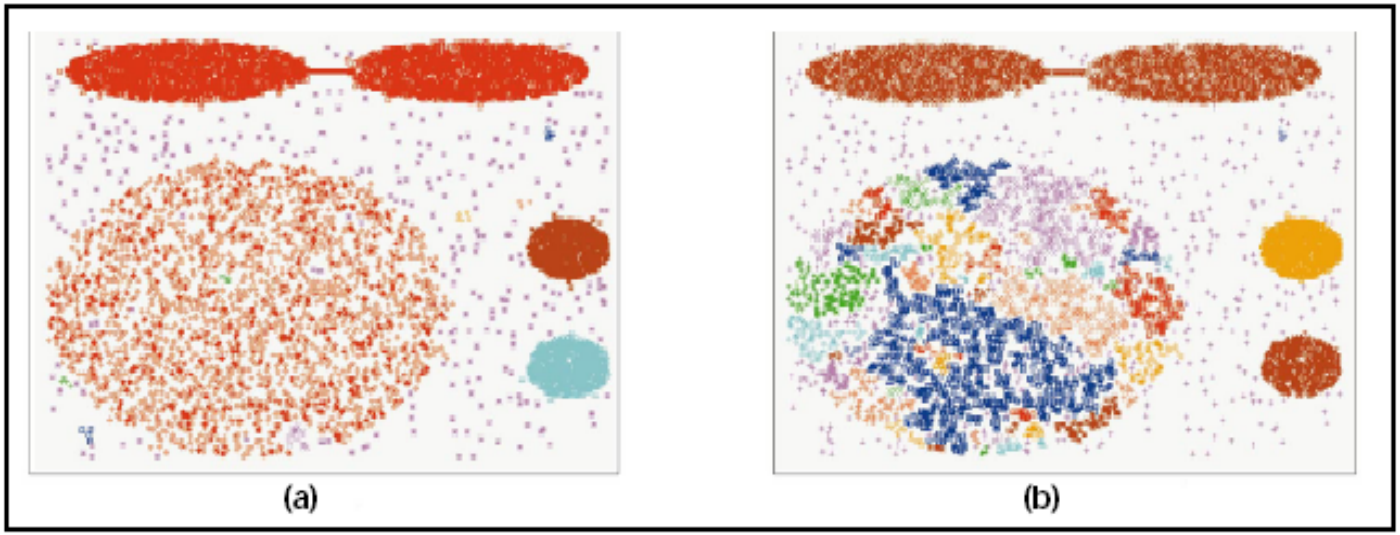
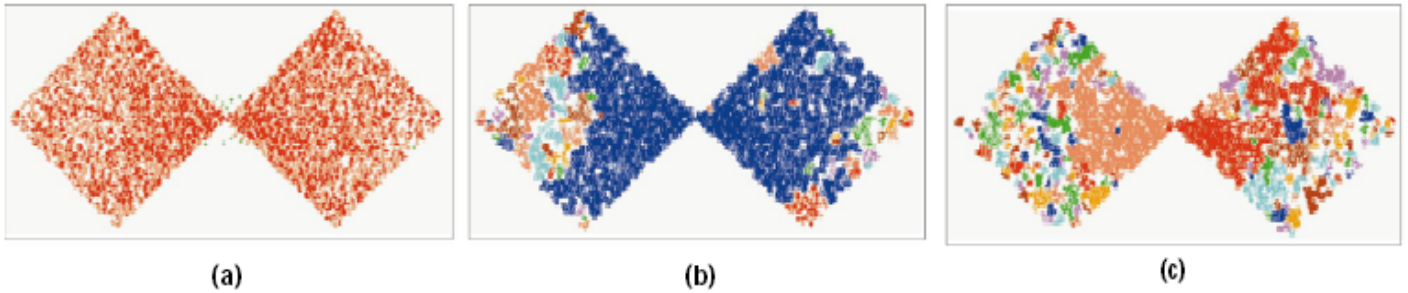


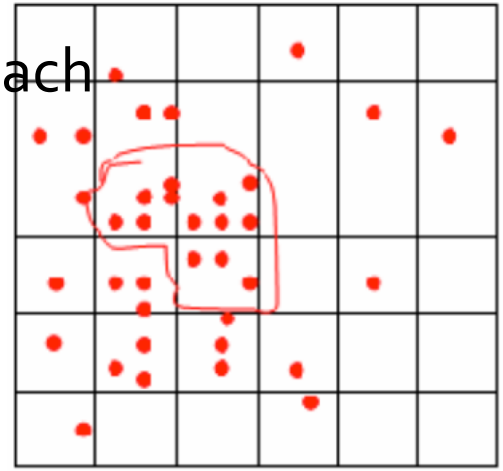
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



Ack. Figures from G. Karypis, E.-H. Han, and V. Kumar, *COMPUTER*, 32(8), 1999

More Methods

- Density-Based Clustering
  - DBSCAN: A Density-Based Clustering Algorithm
    - Ester, et al. (KDD'96)
  - OPTICS: Ordering Points To Identify Clustering Structure
    - Ankerst, et al (SIGMOD'99)
- Grid-Based Clustering
  - STING: A Statistical Information Grid Approach
    - Wang et al. (VLDB'97)
  - CLIQUE: Grid-Based Subspace Clustering
    - Agrawal, et al. (SIGMOD'98)

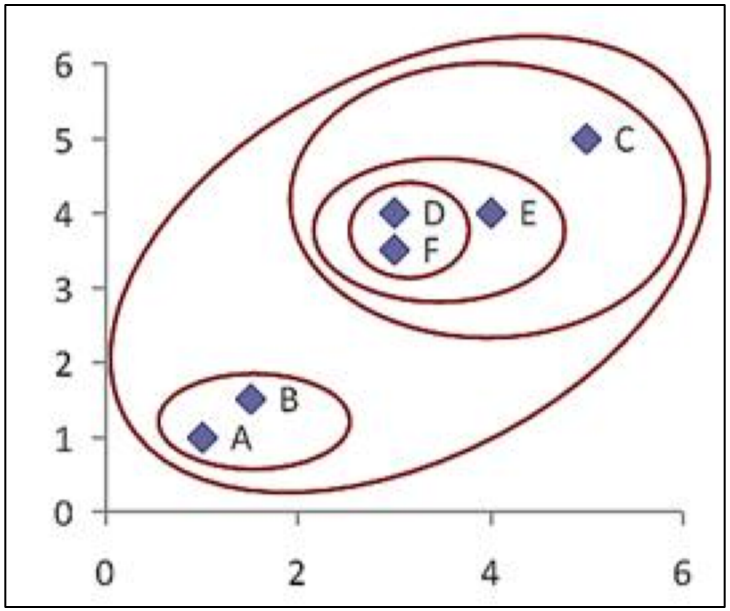


Outline

- Kernel K-Means
- Spectral Clustering
- Density-based Clustering: DBSCAN
- **Hierarchical Clustering: Agglomerative and Divisive**

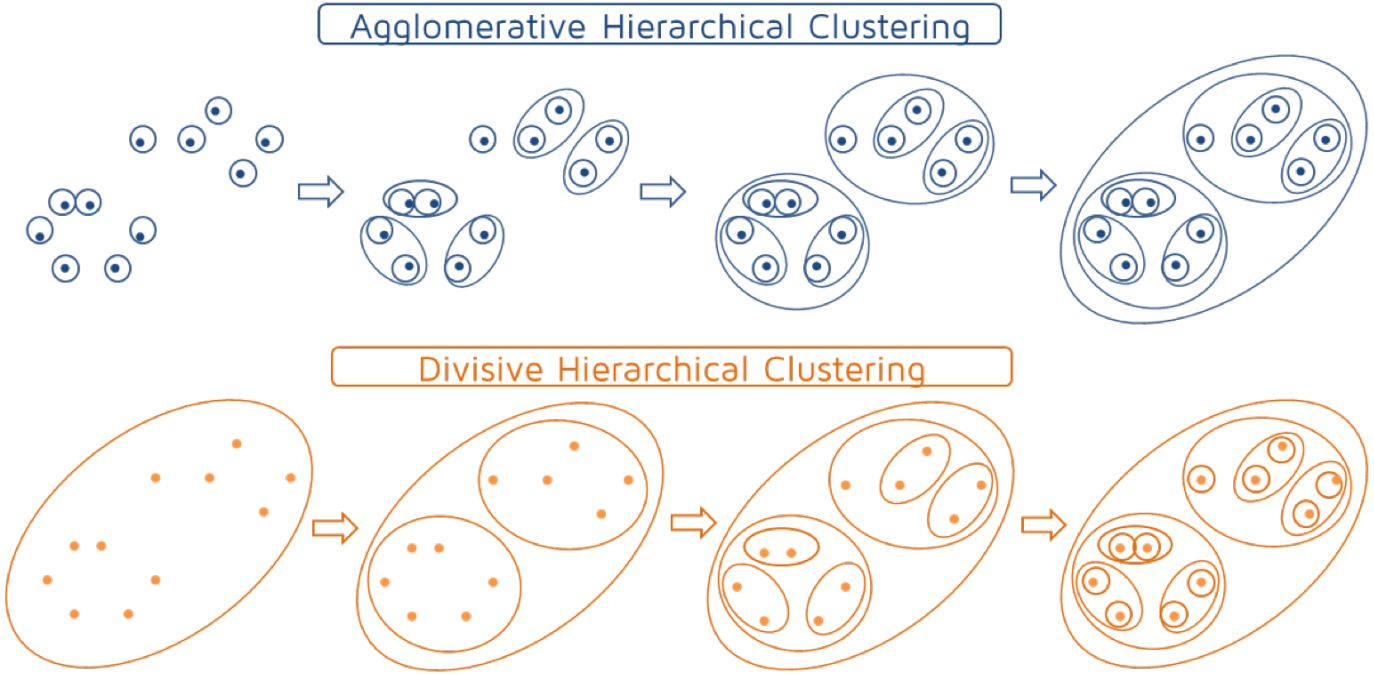
Hierarchical Clustering

- Suppose you want clusters-within-clusters.
- Why? You might suspect that the data reflects a hierarchical process and want to recover the hierarchy (it might matter more than the data)



Two Basic Approaches

- Agglomerative (bottom-up)
  - Start with each item in its own cluster, then merge the clusters according to some criterion, until only one cluster is present.
- Divisive (top-down)
  - Start with one, divide, end with each in its own cluster



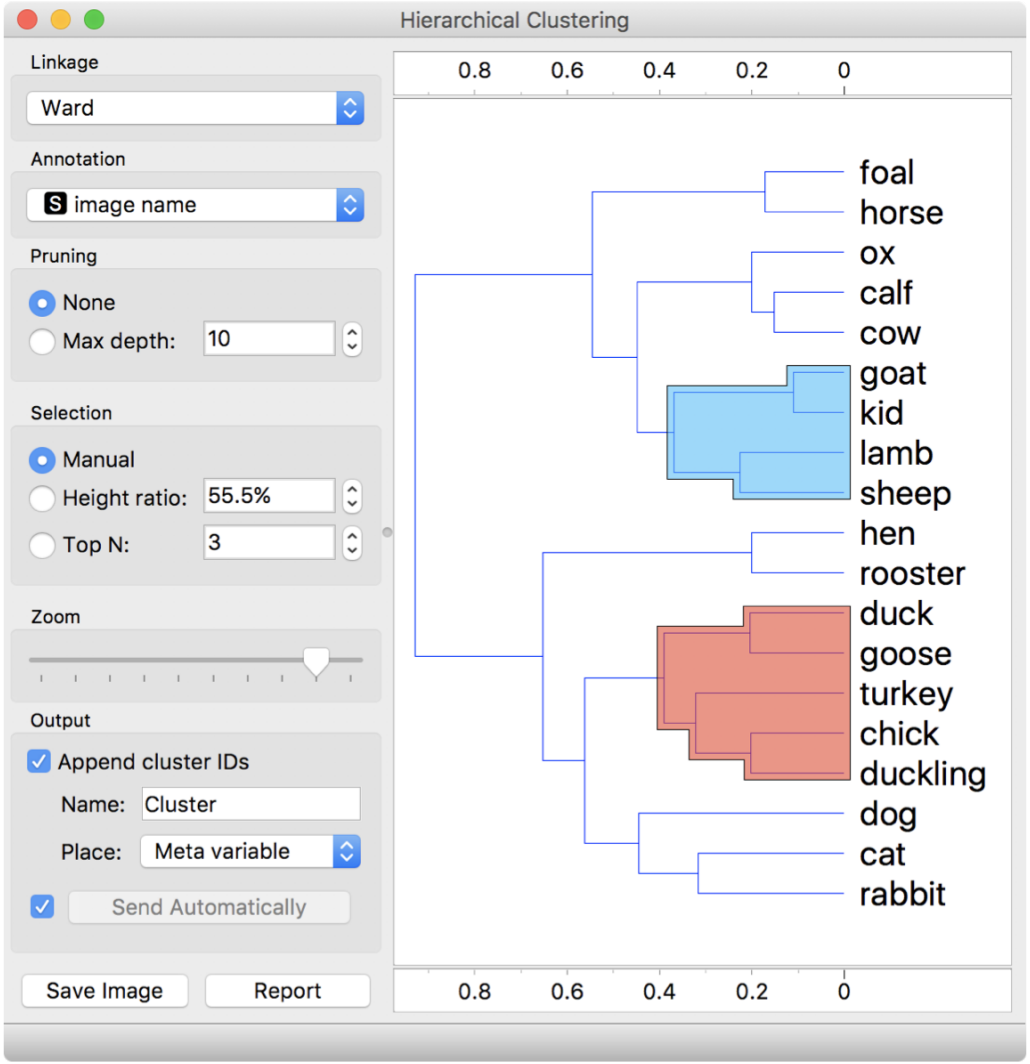
Two Basic Approaches (cont.)

- In either case, can "stop early" with an intermediate number of clusters
- In both cases, some notion of "similarity" or "dissimilarity" drives the merges/splits. This is based exclusively on a similarity or dissimilarity measure.

Representing Hierarchical Clustering: Dendrogram

- Membership versus similarity
- See merging happening at various levels

- Figure: Wikipedia (Orange software)



Research: Clustering for Summarization on Social Media Data

CatchTartan:

Representing and Summarizing Dynamic Multicontextual Behaviors

Meng Jiang, Christos Faloutsos, Jiawei Han



What is Tartan?

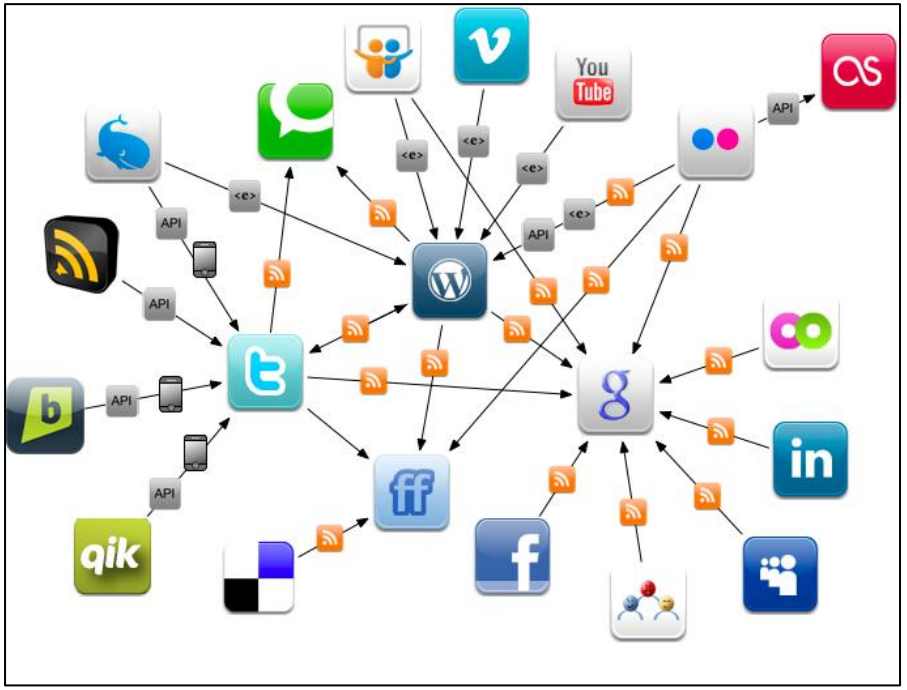
Visited CMU in 2012-13

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

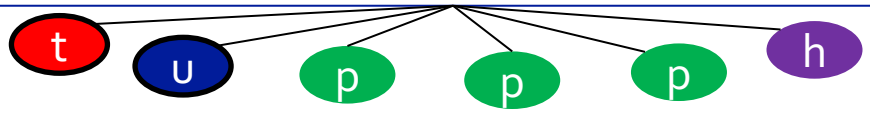
[3] "Optimizing search engines using clickthrough data"

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*

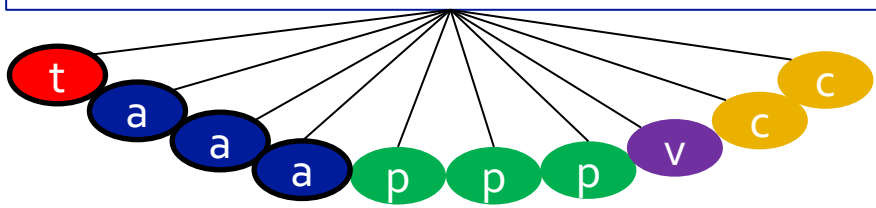
Time slice	User	Location	Phrase	Hashtag	URL
20:00-20:30	@ebekahwsm	∅	{best halftime show, in the history, halftimes shows}	{#SuperBowl}	∅

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 …}

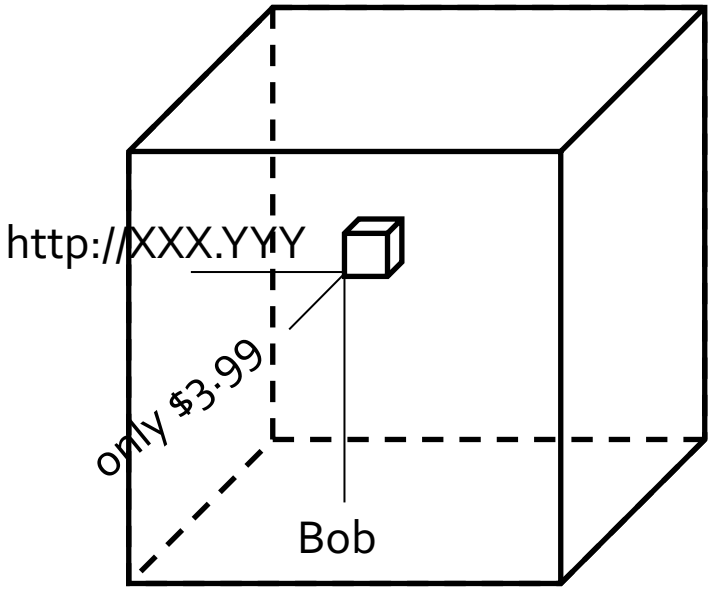
Summarizing Behaviors

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

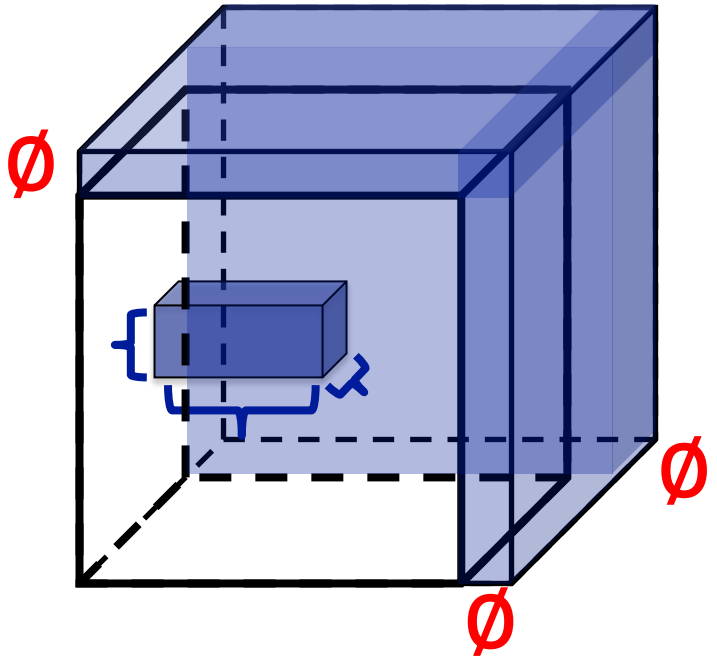
Term	Definition
Dimension	The type of a contextual factor (e.g., location, phrase; author, keyword)
(Dimensional) value	The contextual factor in the dimension
Time slice	The period for consecutive behaviors
Behavior	A set of dimensions, a set of values in each dimension, a time slice for the timestamp

Tensor Fails

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

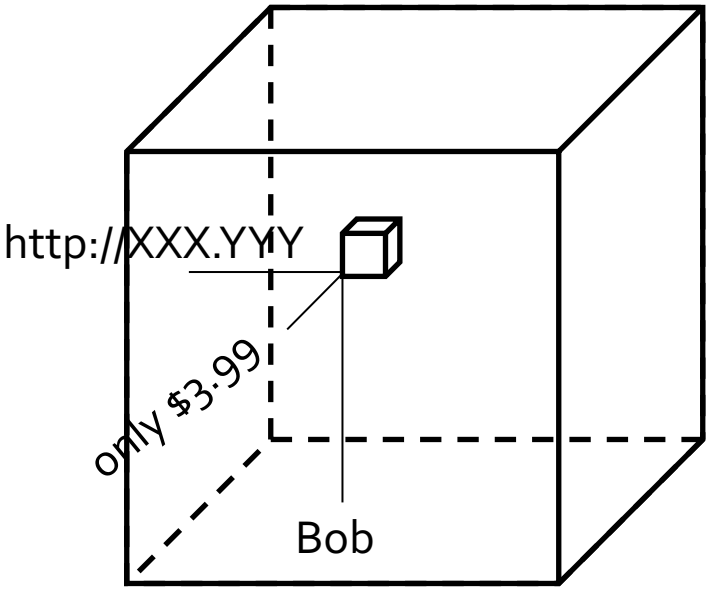


- **Representation: (multicontextual)**
  - **Empty values?**

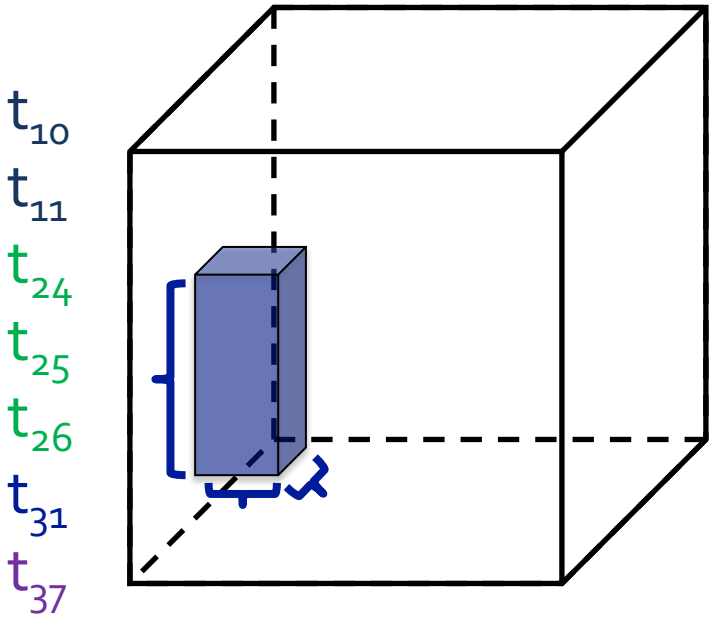


Tensor Fails (cont.)

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

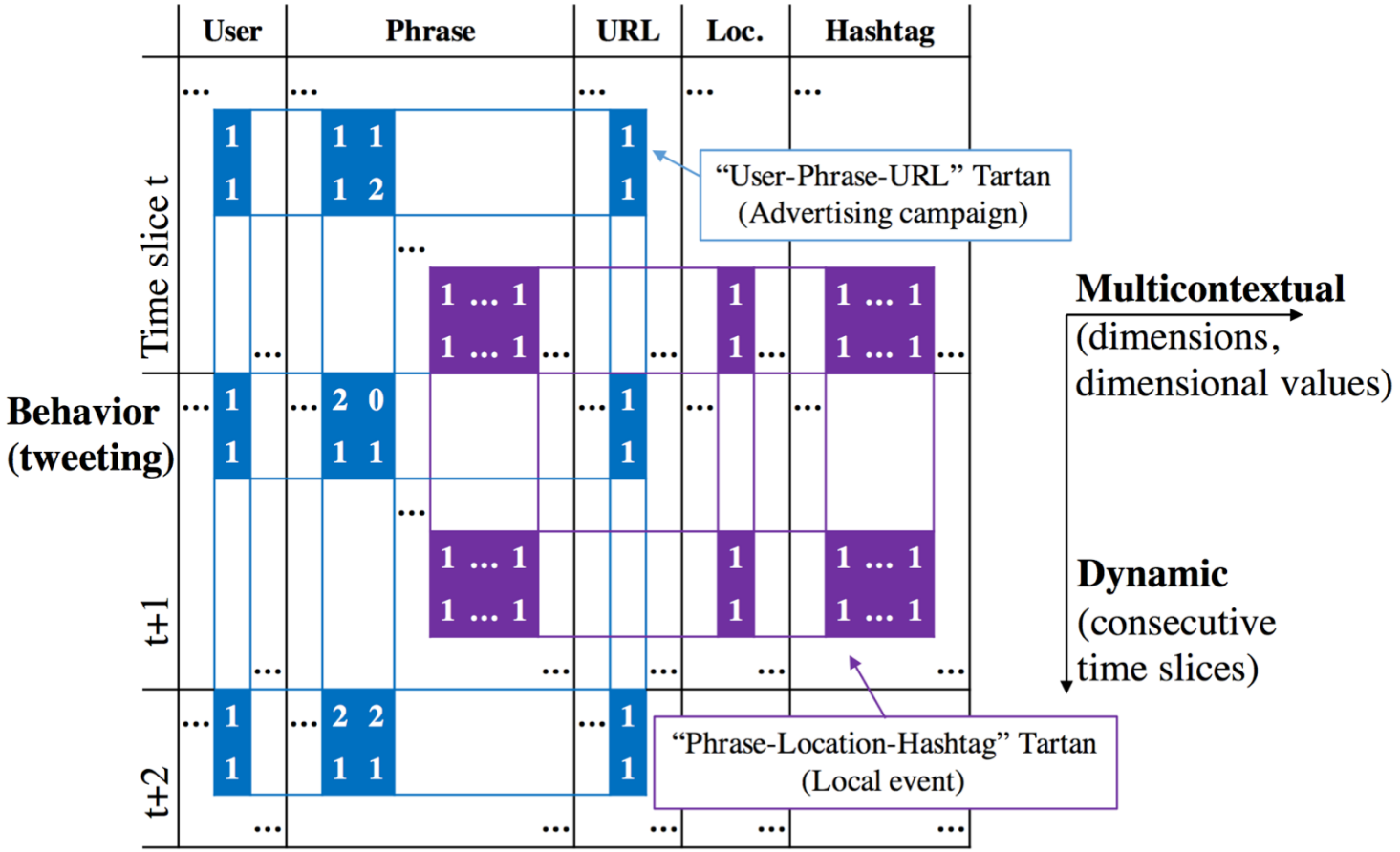


- **Summarization: (dynamic)**
  - **Temporal patterns?**



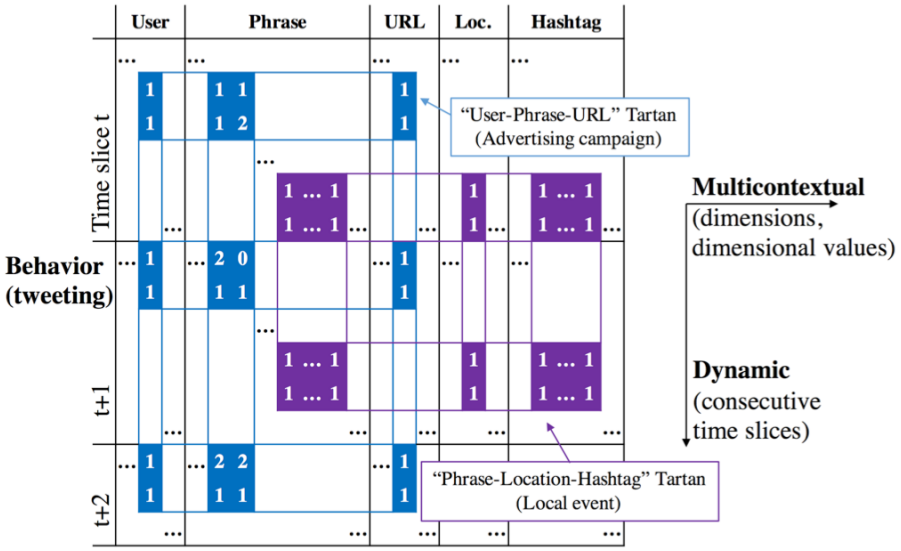
Our Representations for Behavior and Behavioral Summary

- **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

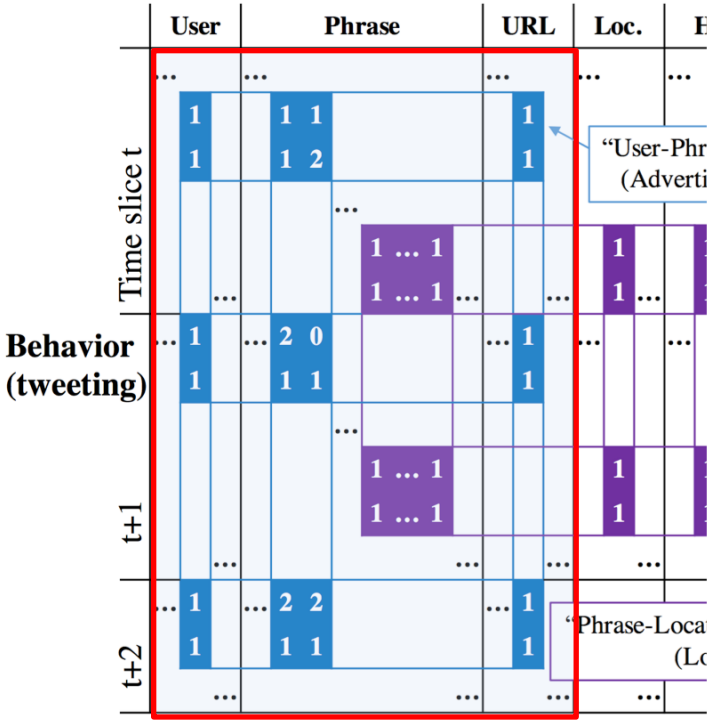
- Employing a lossless encoding scheme
  - The *Minimum Description Length* (MDL) principle
  - Estimating the number of bits that encoding the Tartan can save from merging the meaningful pattern into the encoding of the data

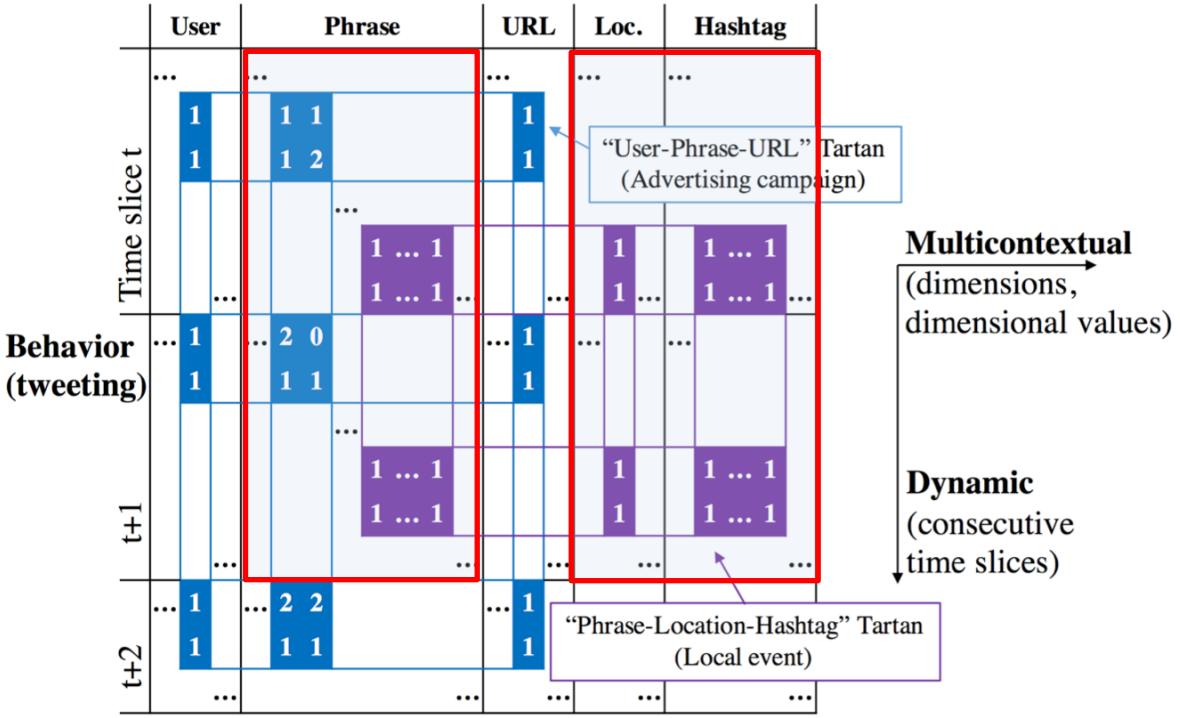
	FSG [18]	GRAPH-CUBE [33]	EVENT-CUBE [29]	MDC [21]	BOW [6]	FEMA [9]	COM2 [2]	CROSS-SPOT [8]	GRAPH-SCOPE [27]	VOG [15]	TIME-CRUNCH [26]	**CATCH-TARTAN**
Principled scoring								√	√	√	√	√
Parameter-free	√				√		√	√	√	√	√	√
Multidimensional		√	√	√	√	√	√	√				√
Multicontextual												√
Timestamp value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

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

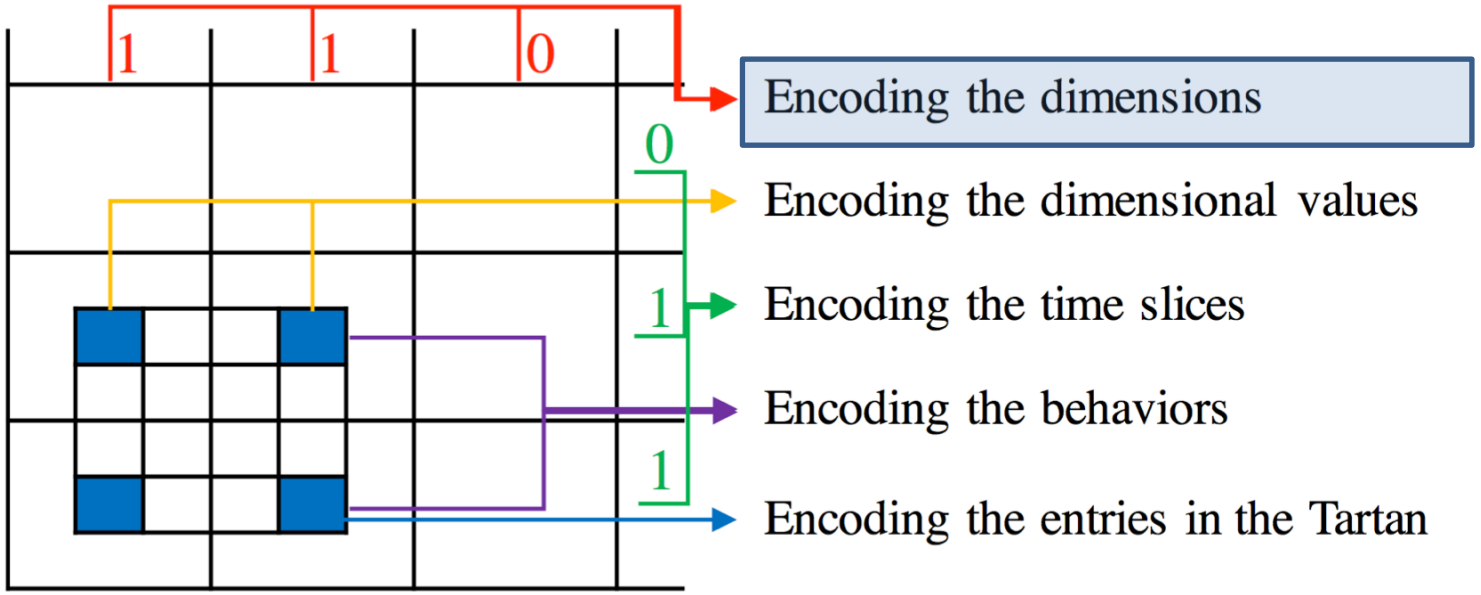
$$\begin{aligned} V &= \left(\sum_{d \in \mathcal{D}} N_d\right)\left(\sum_{t \in \mathcal{T}} E^{(t)}\right). \\ C &= \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \{1,\ldots,E^{(t)}\}, i \in \{1,\ldots,N_d\}} \mathcal{X}_d^{(t)}(b,i). \end{aligned}$$

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

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

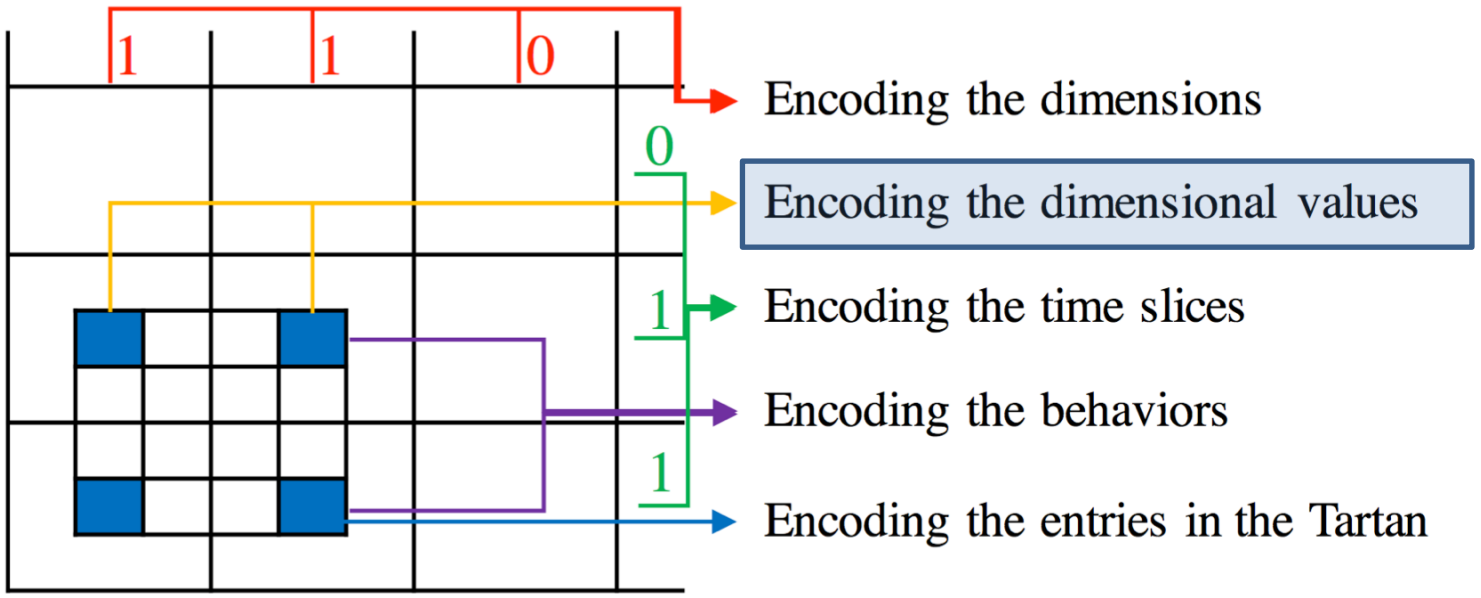
$$L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}) = g(V + C - v - c, C - c);$$

Encoding the Tartan: Dimensions



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

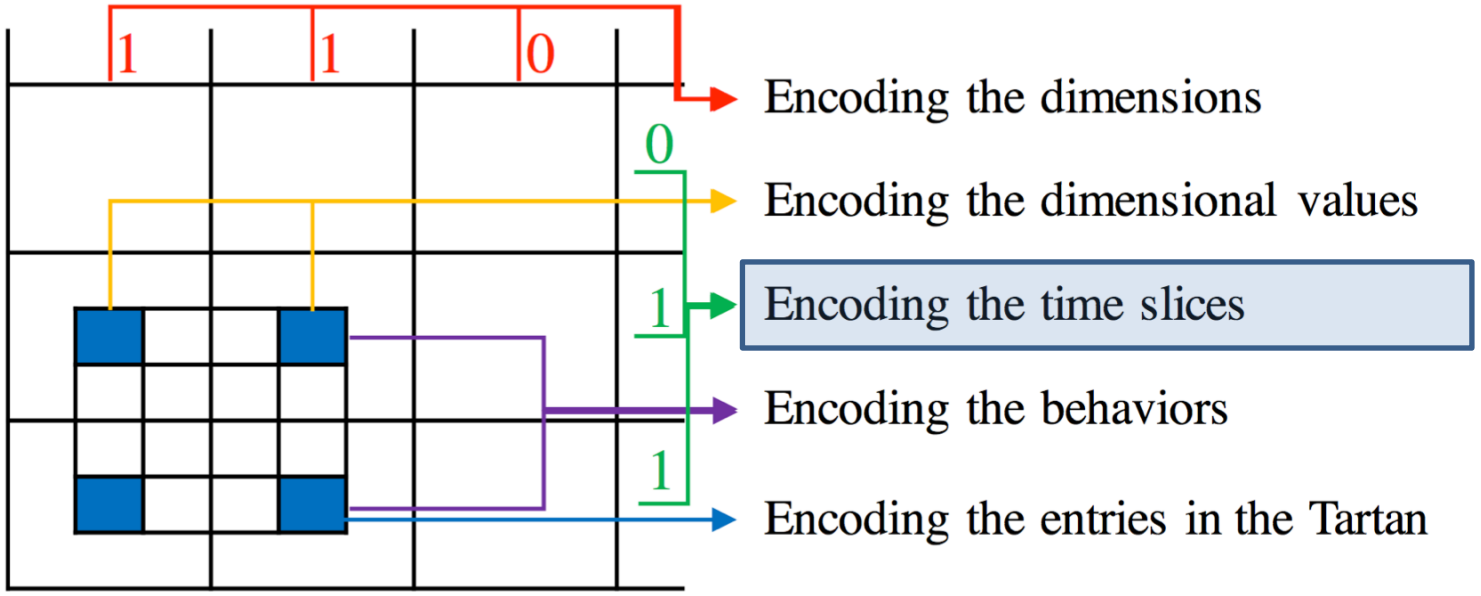
Encoding the Tartan: Dimensional Values



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

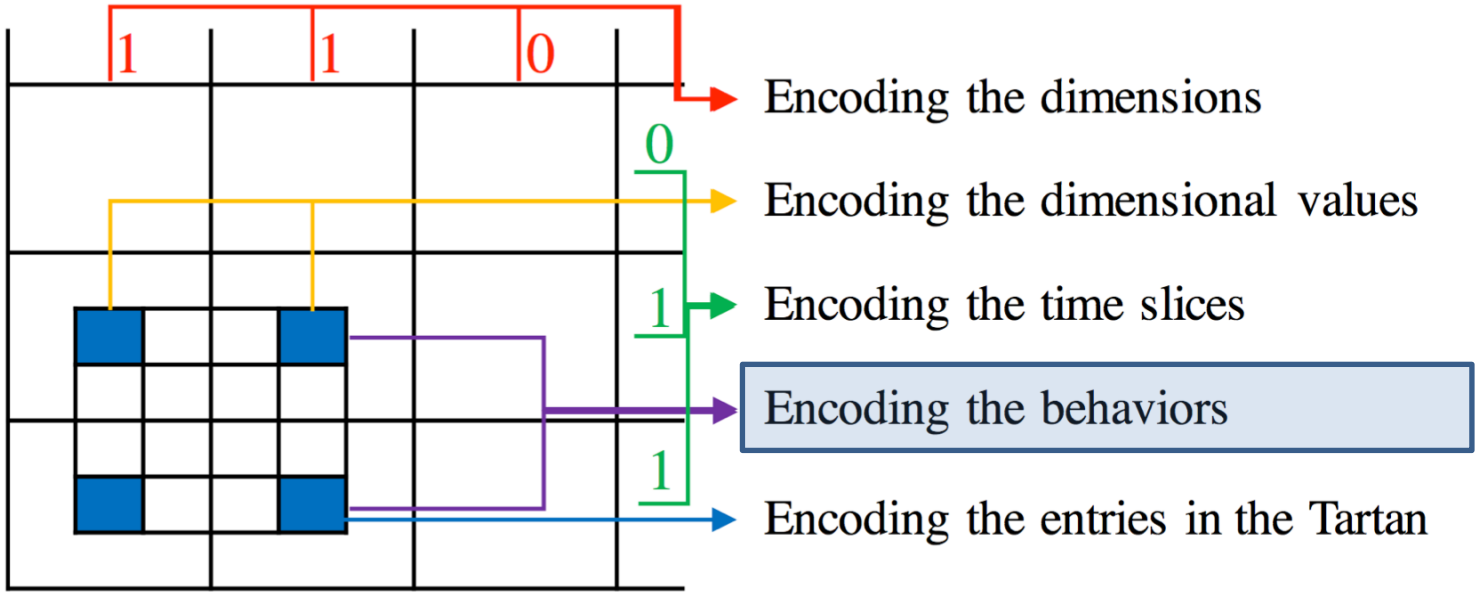
$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding the Tartan: Time Slices



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

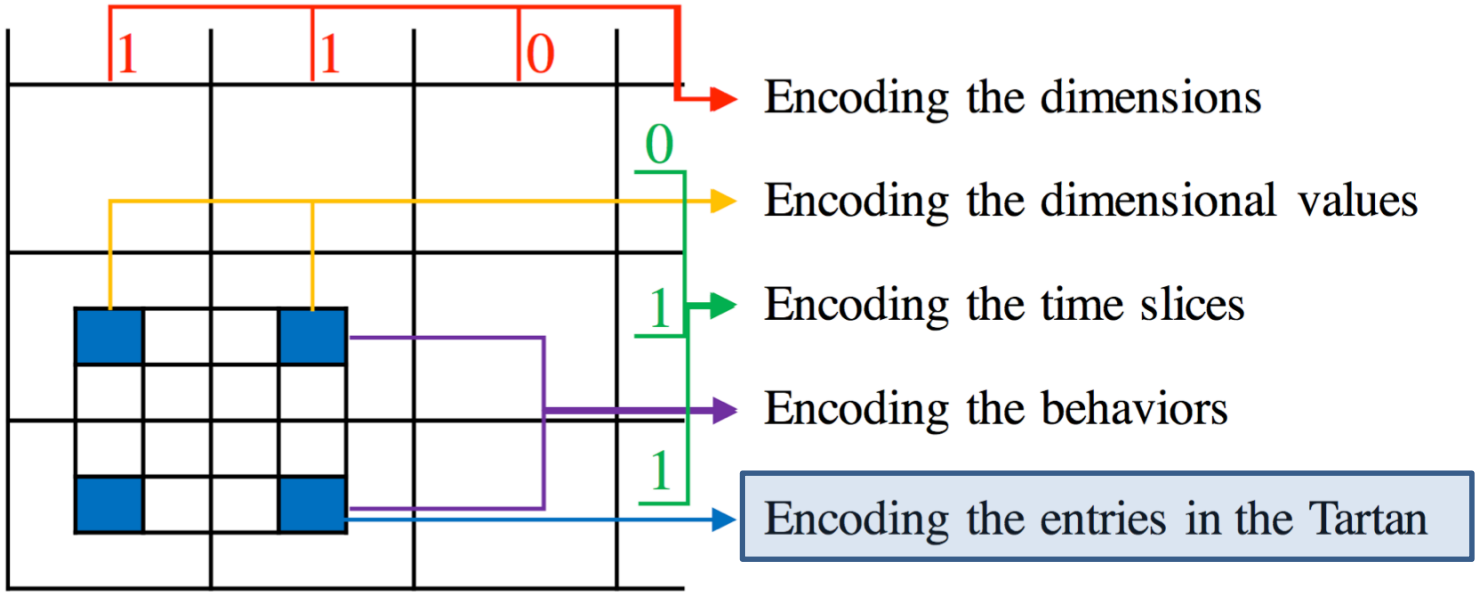
Encoding the Tartan: Behaviors



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

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

Encoding the Tartan: Entries



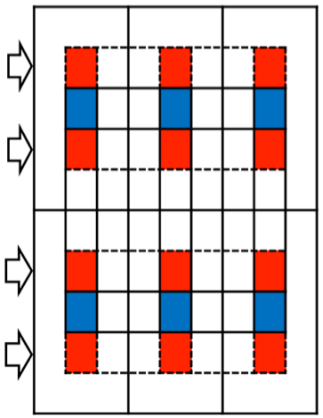
$$v = \left(\sum_{d\in\mathcal{D}} n_d\right)\left(\sum_{t\in\mathcal{T}} e^{(t)}\right).$$

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

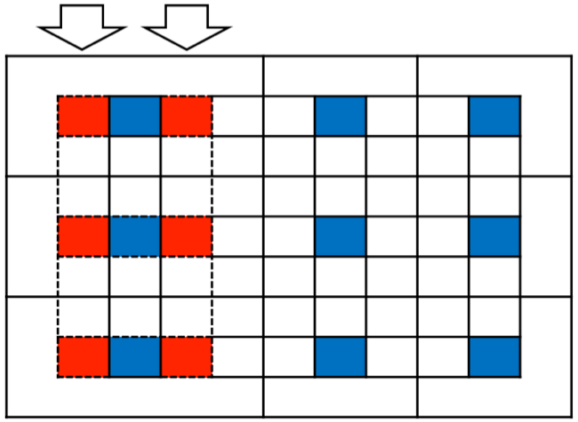
$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c}\log\frac{c}{v+c} + \frac{v}{v+c}\log\frac{v}{v+c}\right).$$

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

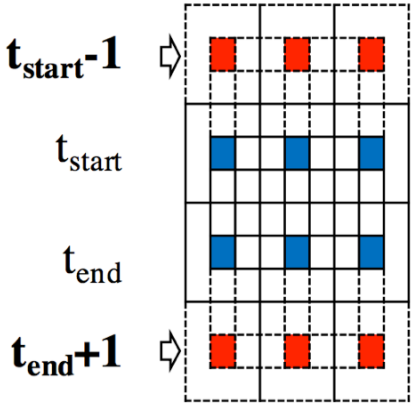
Greedy Search for the Local Minimum



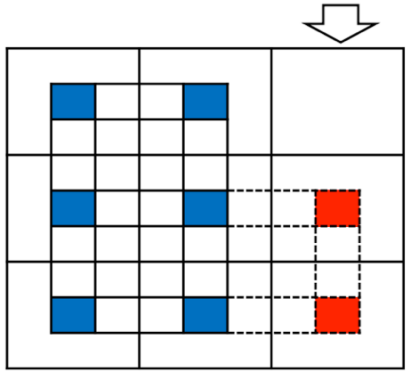
(a) Update the set of behaviors.



(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

16:30

16:30:31 My prediction Ravens 34 Niners 31
16:30:57 Ready for the big game :D, my prediction 24-20 SF #SuperBowl
16:31:14 My prediction for superbowl.. 48.. Jets over Bears 17-13 Mark Sanchez MVP
16:32:24 I predict Baltimore Ravens will win 27 to 24 or 25 or 26. Basically it will be a close game.

"my prediction"

user	phrase	hashtag	URL	3,397 tweets
(3,325)	226	(0)	(0)	

Tartan #1: (1 dim) 16:30-17:30

17:00

17:30

17:30:51 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:01 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:16 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:19 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist

"make your prediction"

user	phrase	RT @user	URL	196 tweets
(196)	4	1	1	

Tartan #2: (3 dims) 17:00-18:00

18:00

18:55:03 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47
18:55:04 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47
18:55:44 RT @Ravens: David Akers is good from 36 yards to make the score 7-3 Ravens. Nice job by the defense to tighten up in the red zone.

"7-3", "1st Qtr"

user	phrase	RT @user	URL	215 tweets
(213)	21	3	(0)	

Tartan #3: (2 dims) 18:30-19:30

18:30

19:00

20:20:01 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6
20:20:02 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs
20:20:04 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6
20:20:05 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs

"halftime show"

user	phrase	RT @user	URL	617 tweets
(617)	11	4	4	

Tartan #4: (3 dims) 20:00-21:00

19:30

20:00

20:20:47 (Manhattan, NY)...and every one of those girls took #ballet #Beyonce #superbowl
20:22:01 (New York, NY) I have the biggest lady boner for Beyonce #BeyonceBowl #DestinyBowl #DestinysChild #SuperBowl
20:24:32 (Manhattan, NY) No one can ever top that performance by Beyonce. EVER. #Beyonce #superbowl #halftimeshow

"beyonce", #beyonce, #superbowl, #DestinysChild

location	phrase	hashtag	URL	166 tweets
2	55	17	(0)	

Tartan #5: (3 dims) 20:00-21:00

20:30

21:00

21:44:42 Ahora si pff #49ers 23-28 #Ravens
21:44:44 Baltimore #Ravens 28-23 San Francisco #49ers
21:44:50 FG Akers #49ers 23-28 #Ravens 3Q 3:10 #SuperBolwXLVII #SuperBowl #NFL

"28-23", #49ers, #Ravens

user	phrase	hashtag	URL	653 tweets
(650)	69	11	(0)	

Tartan #6: (2 dims) 21:00-22:00

21:30

22:00

22:42:27 Congratulations Ravens!!!!
22:42:43 Congratulations Ray Lewis and the Ravens.
22:42:43 Game over.! Ravens won ray got his retirement ring now all y'all boys and girls go to sleep .!
22:42:52 "@LetThatBoyTweet: Game over. Ravens win the Super Bowl."

"congratulations", "game over"

user	phrase	hashtag	URL	1,950 tweets
(1942)	248	(0)	(0)	

Tartan #7: (1 dim) 22:00-23:30

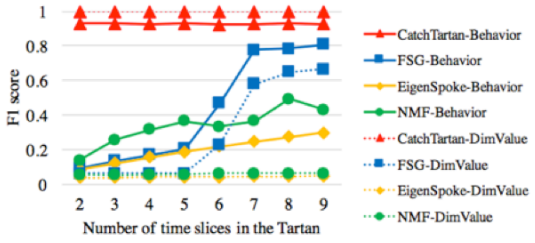
Quantitative Analysis: Accuracy and Efficiency in Synthetic Experiments

- Tartan distribution

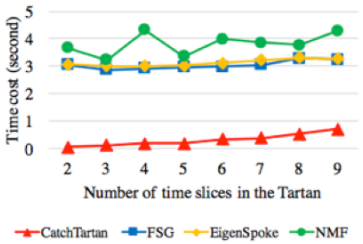
1. $T^{\mathcal{A}} \in [2,\ 9]$, the number of consecutive time slices in the Tartan $\mathcal{A}$, 4 as default;
2. $e^{(t)} \in [100,\ 2,000]$, the number of behaviors in the time slice, $1,000$ as default;
3. $D^{\mathcal{A}} \in [2,\ 9]$, the number of dimensions in $\mathcal{A}$, 3 as default;
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

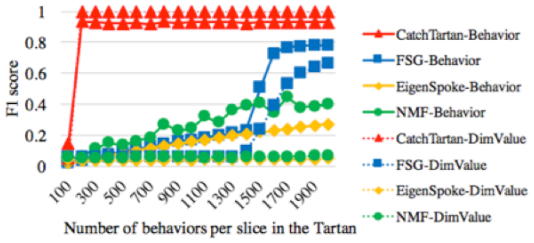
6. $T \in [5, 30]$, the total number of time slices in the dataset, 10 as default;
7. $E^{(t)} \in [1,000,\ 10,000]$, the number of behaviors per time slice in the dataset, $5,000$ as default;
8. $N_d \in [1,000,\ 2,000]$, the number of values per dimension in the data, $1,000$ as default.



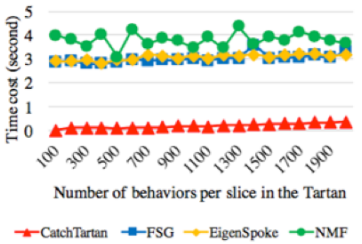
(a) F1 score vs $T^{\mathcal{A}}$.



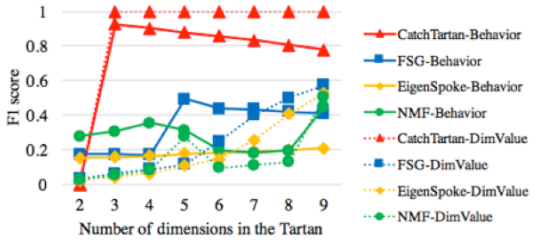
(b) Time cost vs $T^{\mathcal{A}}$.



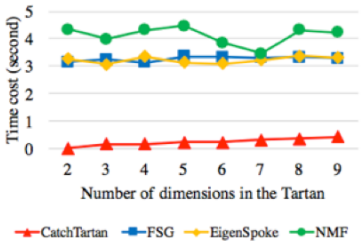
(c) F1 score vs $e^{(t)}$.



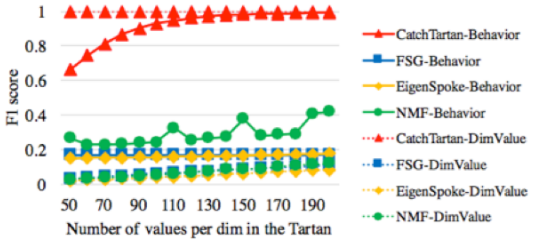
(d) Time cost vs $e^{(t)}$.



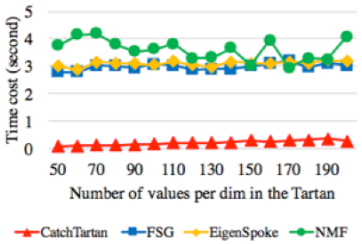
(e) F1 score vs $D^{\mathcal{A}}$.



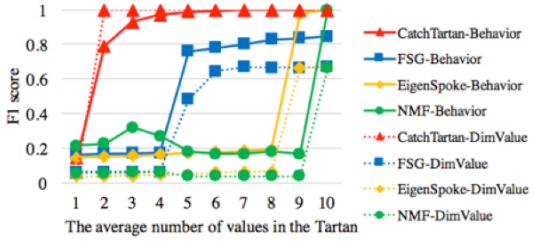
(f) Time cost vs $D^{\mathcal{A}}$.



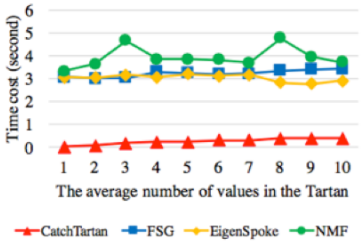
(g) F1 score vs n_d.



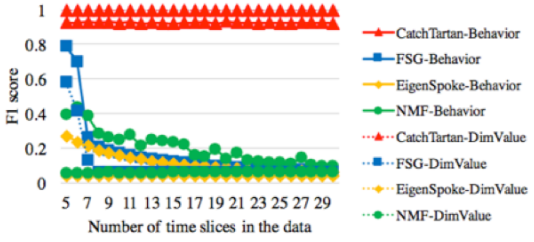
(h) Time cost vs n_d.



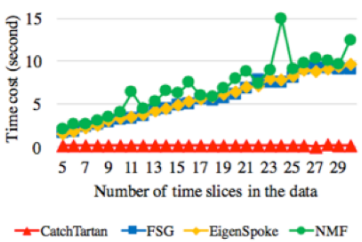
(i) F1 score vs ρ.



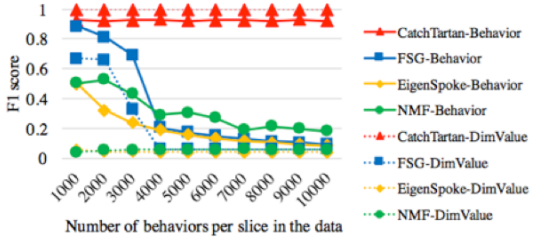
(j) Time cost vs ρ.



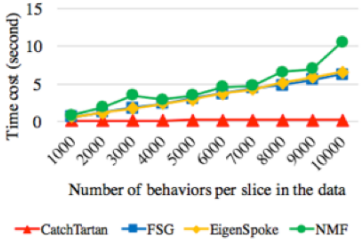
(k) F1 score vs T.



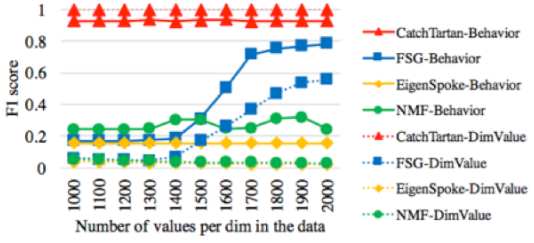
(l) Time cost vs T.



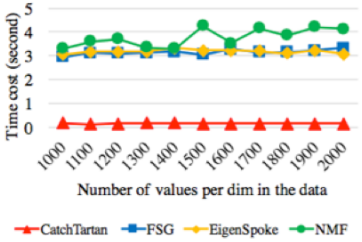
(m) F1 score vs $E^{(t)}$.



(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