owards Visualizing Clusters and asses for Real Valued High mensional Data Sets

Kamalakar Karlapalem

(with Soujanya Vadapalli, Shraddha Agrawal, Nahil Jain, Mounica Maddela, Pallav Tinna)

Indian Institute of Technology, Gandhinagar

International Institute of Information Technology, Hyderabad

kkamal@iitgn.ac.in

Eight International Conference on Contemporary Computing (IC3), August 20-22, 2015

High Dimensional Data Visualization

- D = $\{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional
- d > 3
- n – large
- All real valued
- Need to
  - imagine
  - validate
  - analyze

Motivation

- Seeing helps understanding
- Large data – cannot see completely!
- Dimensions a bigger problem – 4-d and higher
  - Validate classification and clustering results
- Need visualization approaches that
  - provide insight
  - are within canvas
  - can be accurate and/or approximate (metaphor)
  - are like scatter plots
  - can efficiently handle large data and higher dimensions

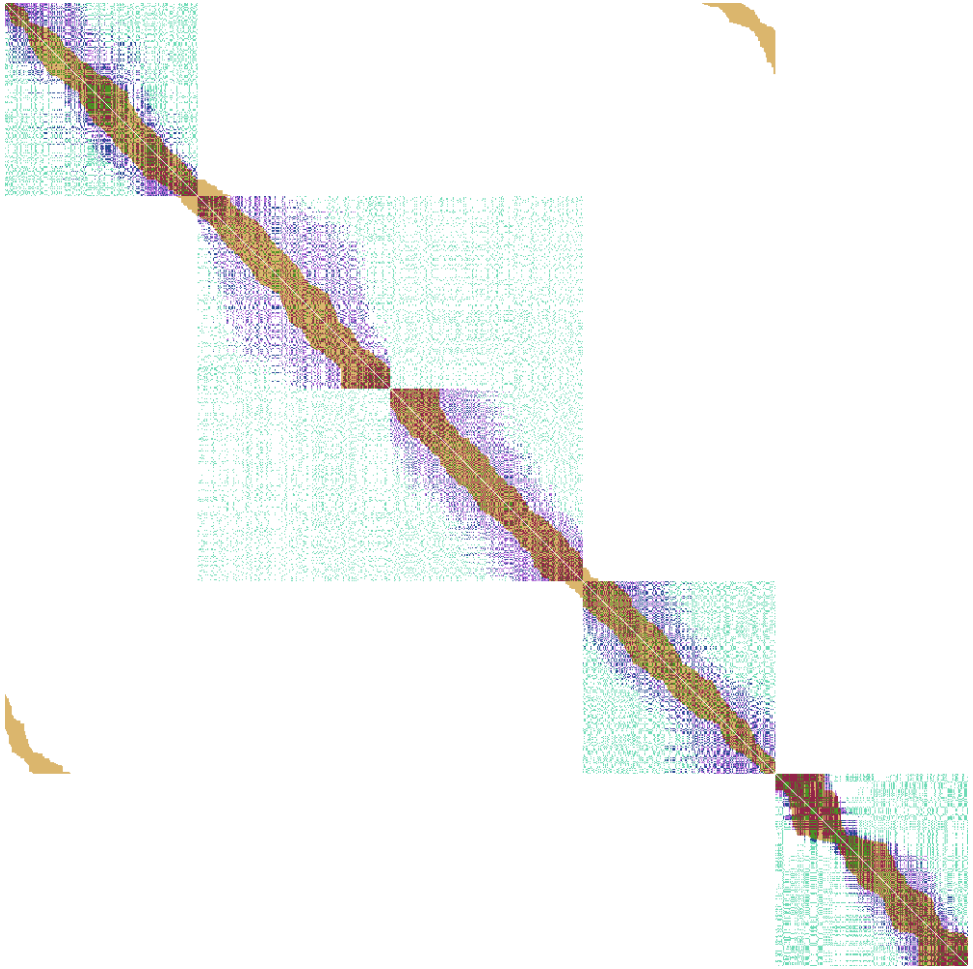
Applications – Some Requirements

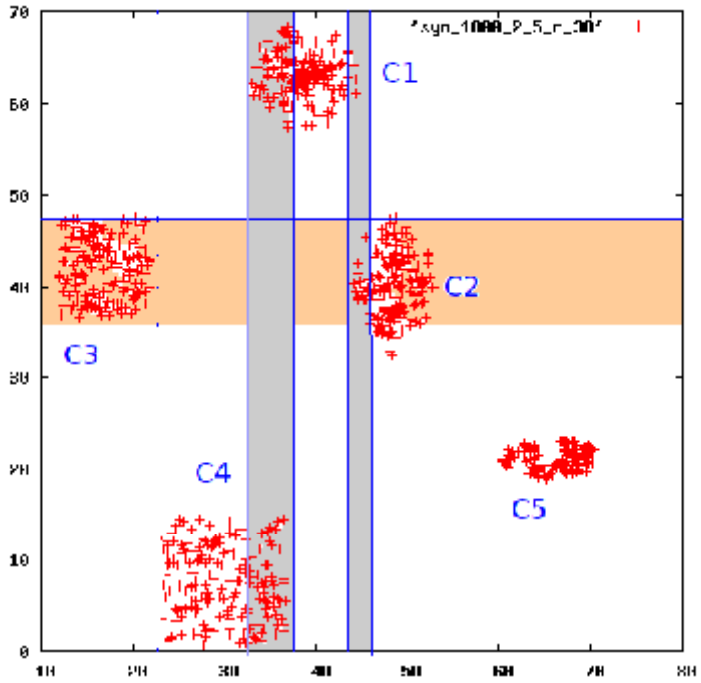
- Across all Subspaces proximity of points
- Shape and size of clusters
- Spread of data across the canvas
- Data Sets
  - Sports
  - Real Estate
  - Spatial-temporal
  - Earthquake
  - Potentially, any real valued data set

Some Problems

- Can we find how clusters in high dimensional data overlap across various subspaces?
  - HEIDI
- Can we visually determine size and shape of a data cluster and explore data set visually?
  - BEADS & PEARLS
- Can we present high dimensional data as a scatter plot?
  - CROVHD
- Useful for
  - Understanding and interpreting data
  - Clustering
  - Classification
  - Image pattern based index

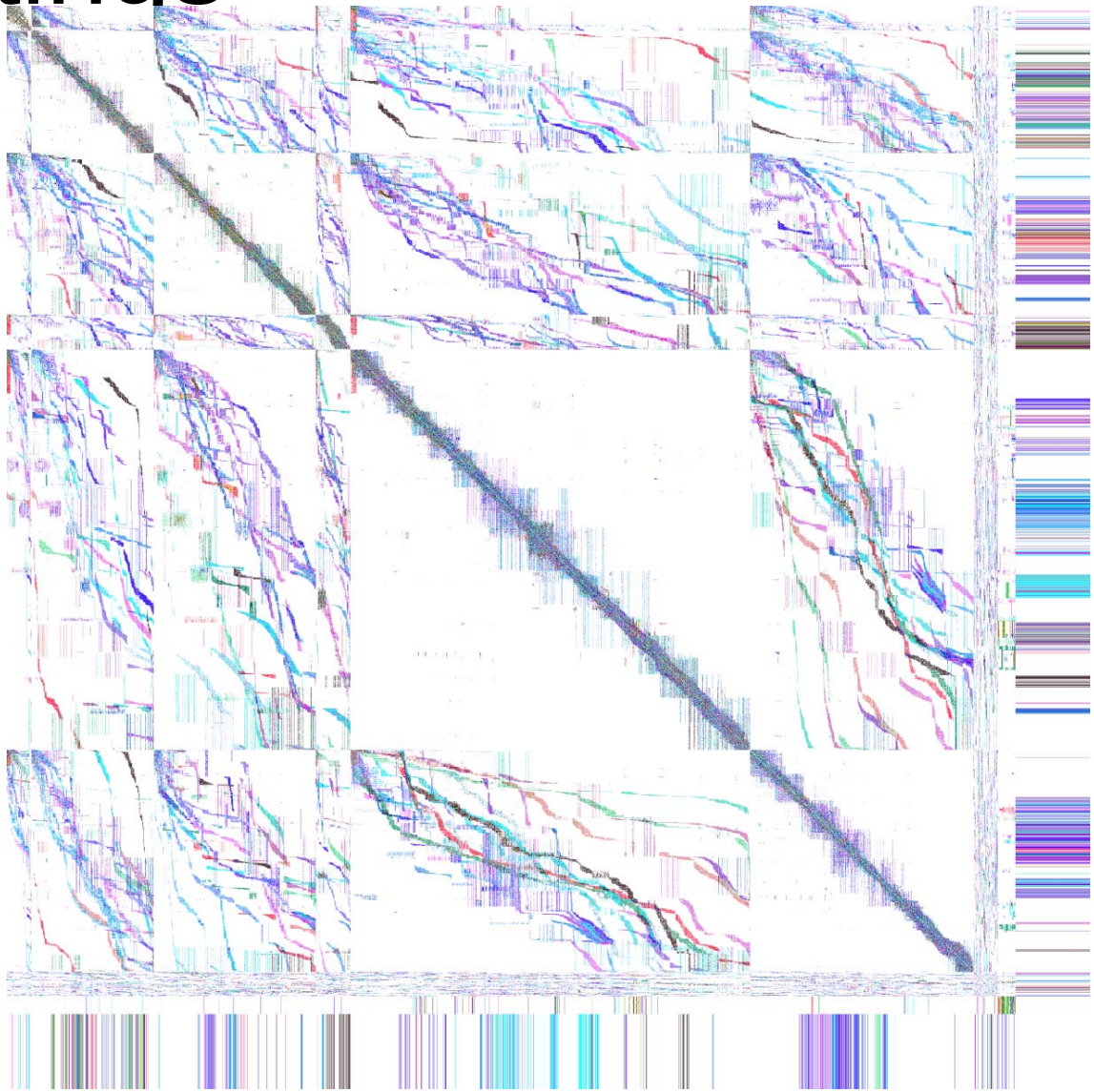
HEIDI - Examples





X – brown; Y – skyblue; {X,Y} - violet

HEIDI Real-estate Property Listings

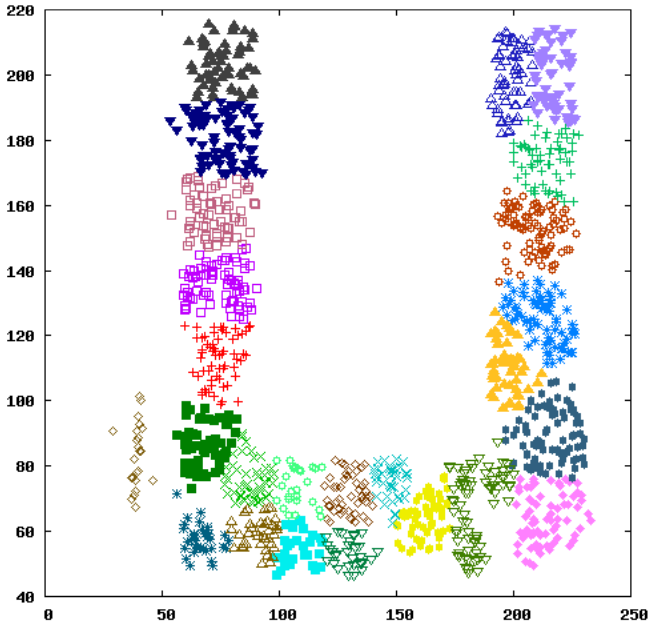


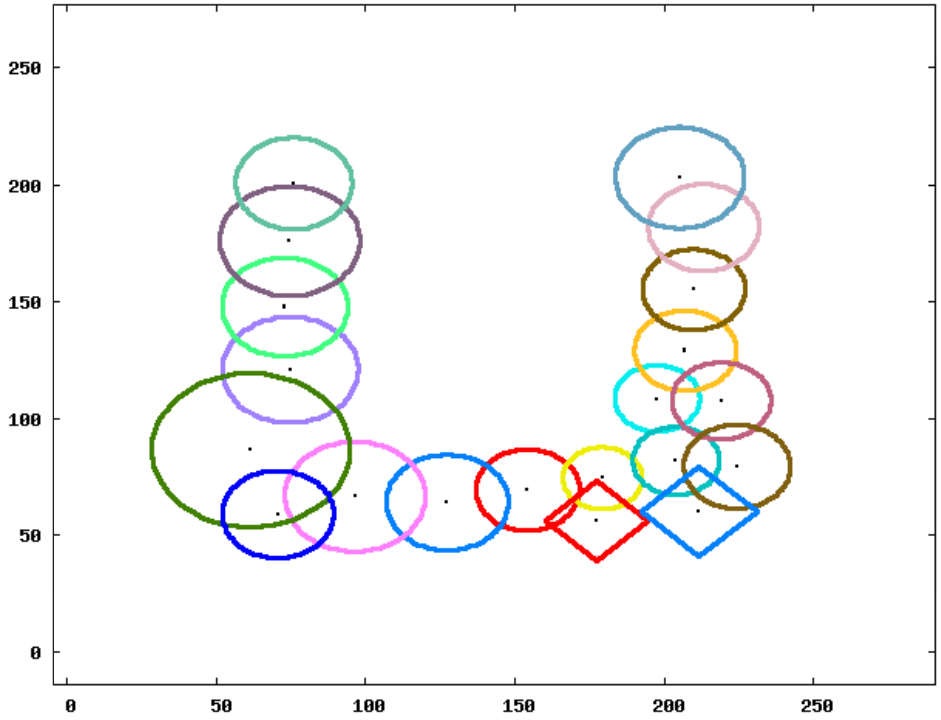
HEIDI – Nearest Neighbors

- D = $\{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional
- Construct a n×n matrix where
  - Element (i,j) is a bit vector
  - Bit **p** of bit vector
    - is set to 1, if x_j is in k nearest neighbor set of x_i,
    - otherwise it is set to 0
    - For the **p^{th}** subspace of the data
  - Length of bit vector is 2^d-1
- Visualize bit-vectors using RGB combination of colors
- Size of matrix is n×n × [(2^d -1) bits mapped to RGB representation based on image type]

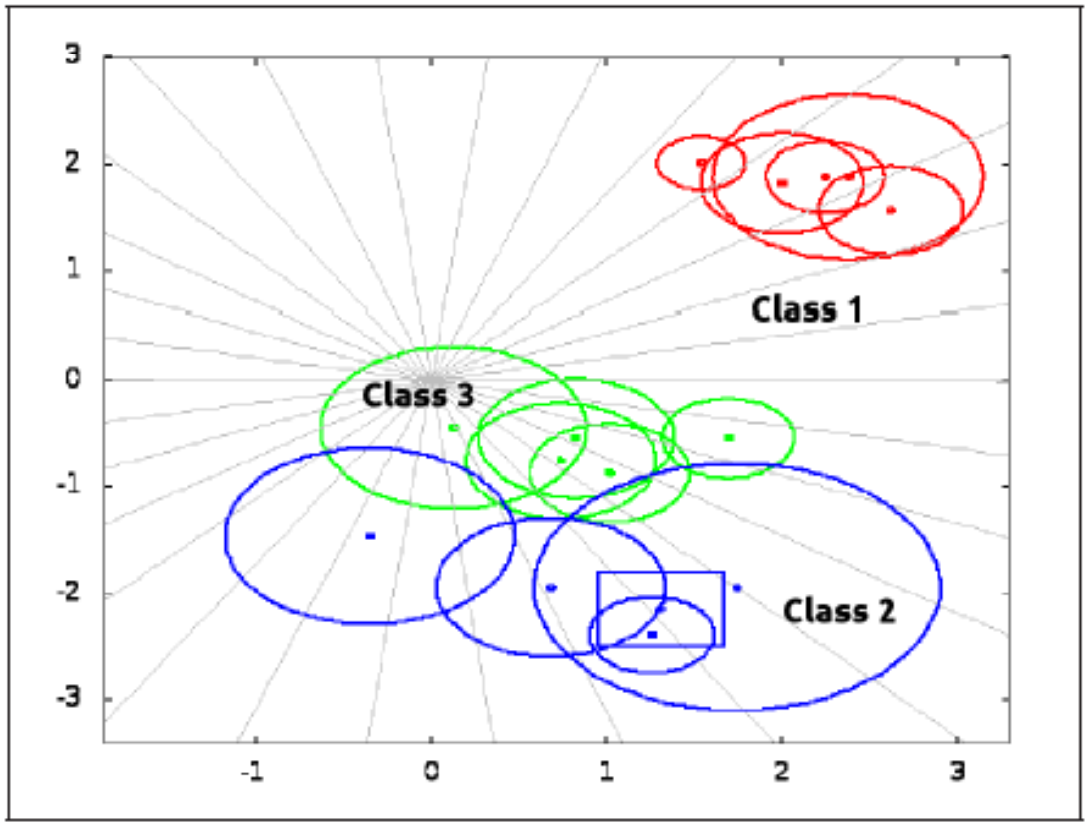
So, what have you got now? – a Heidi Matrix as shown

Beads Example

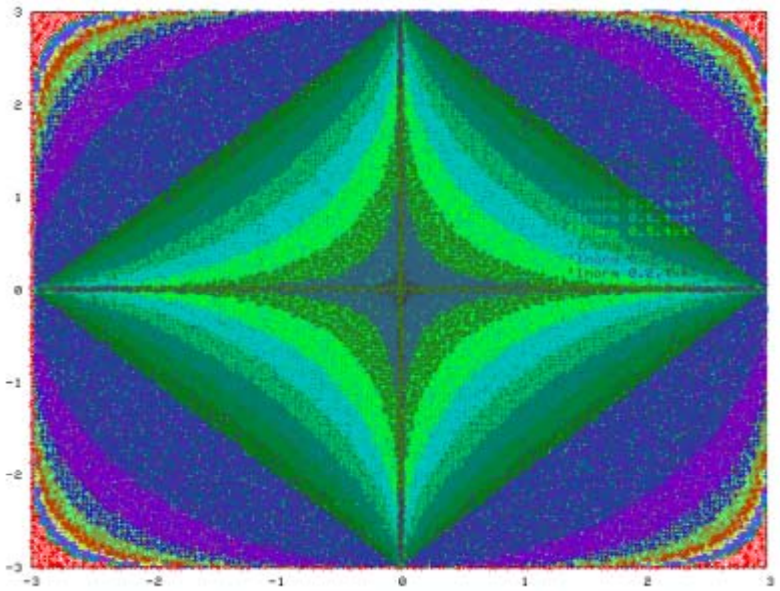


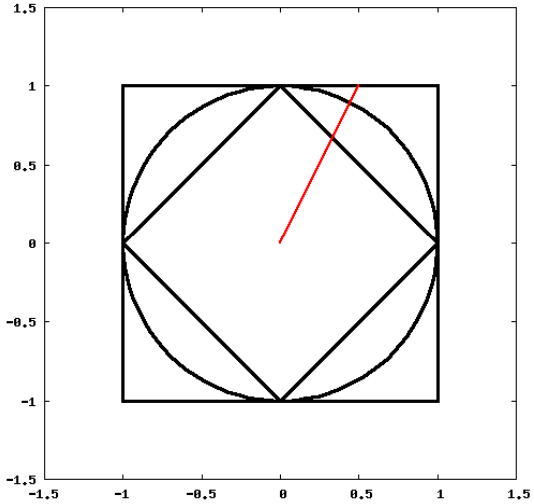


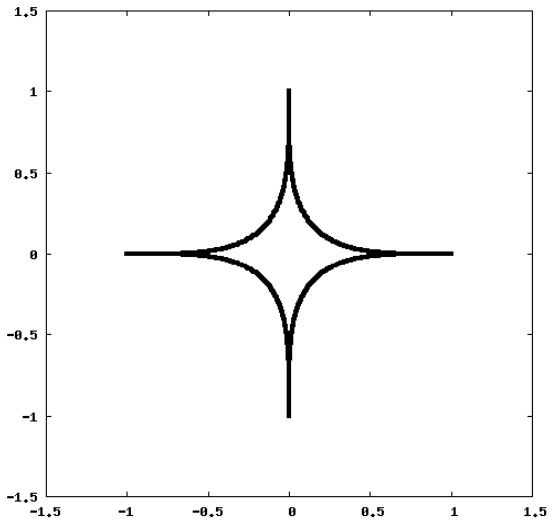
Beads Example – Iris Data Set



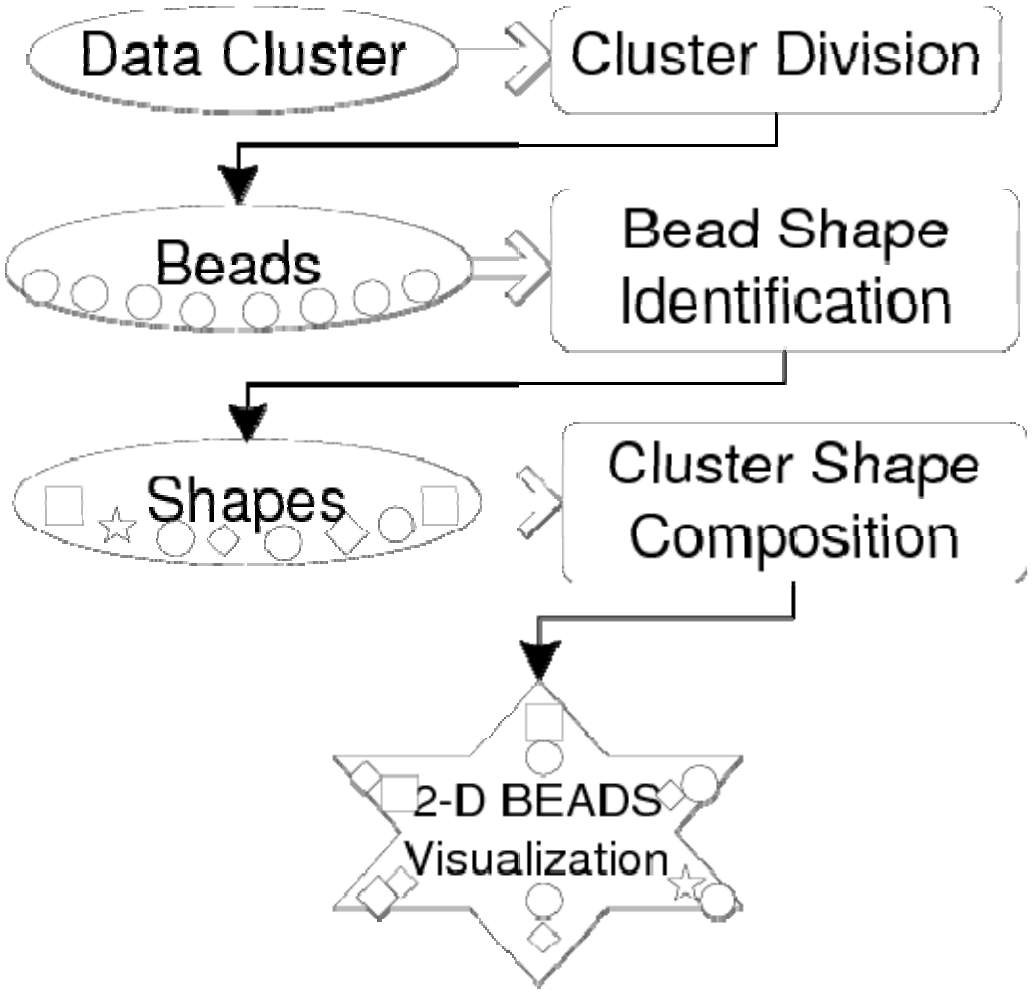
Basis for Beads



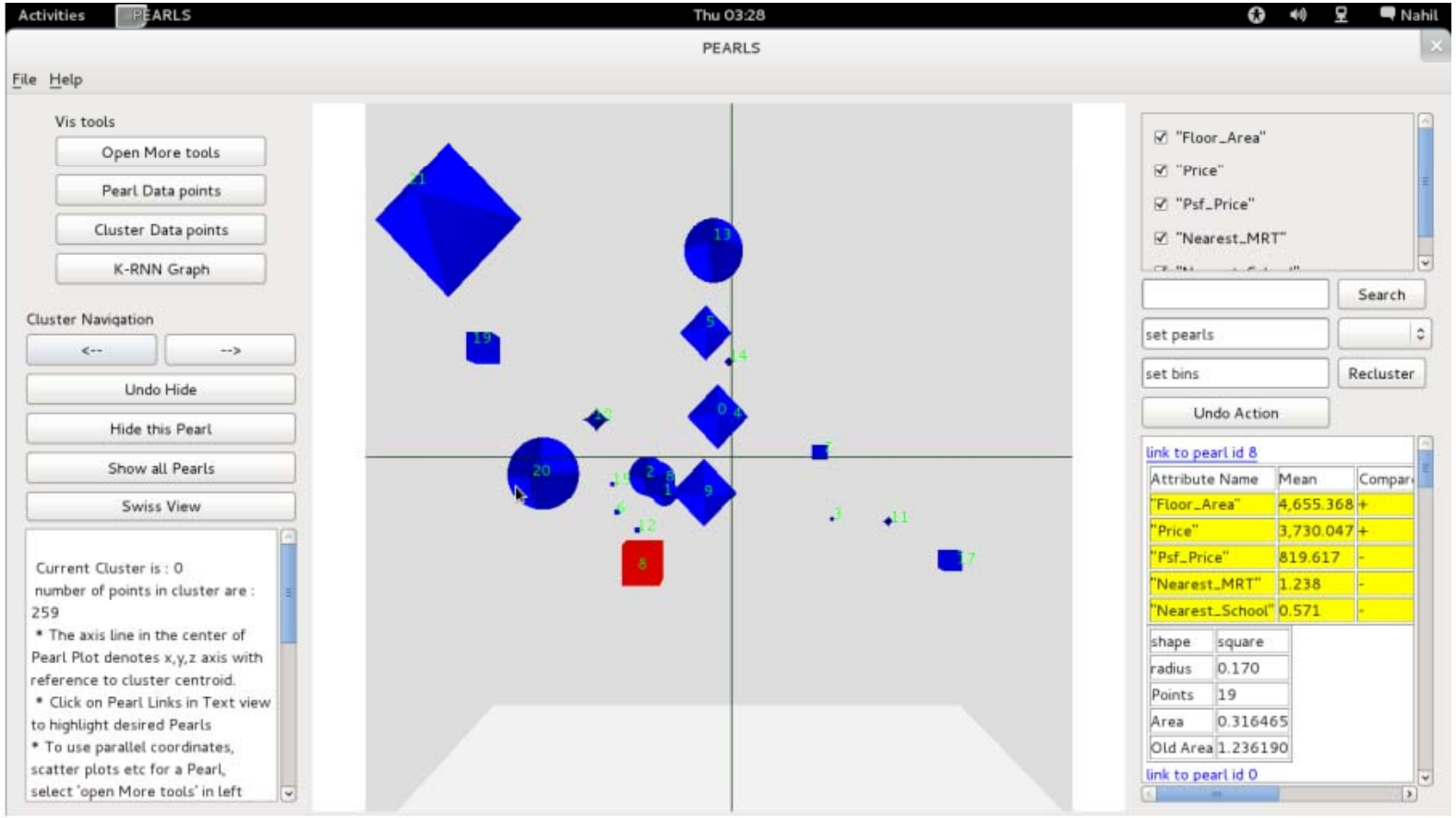




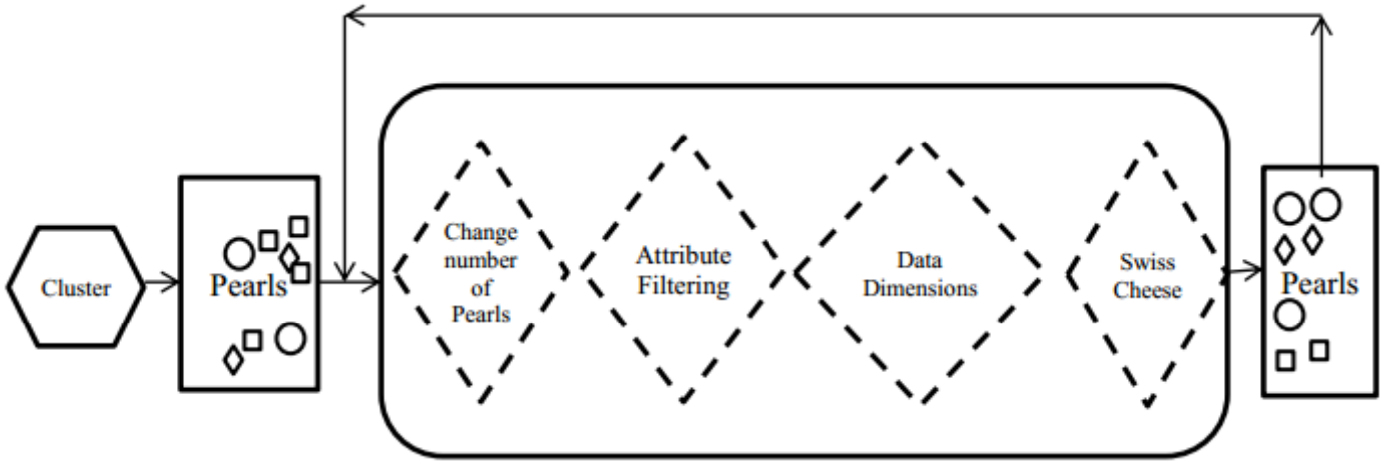
Beads - Approach



PEARLS Visualization



Pearls Visualization System



Visual Explorative Querying

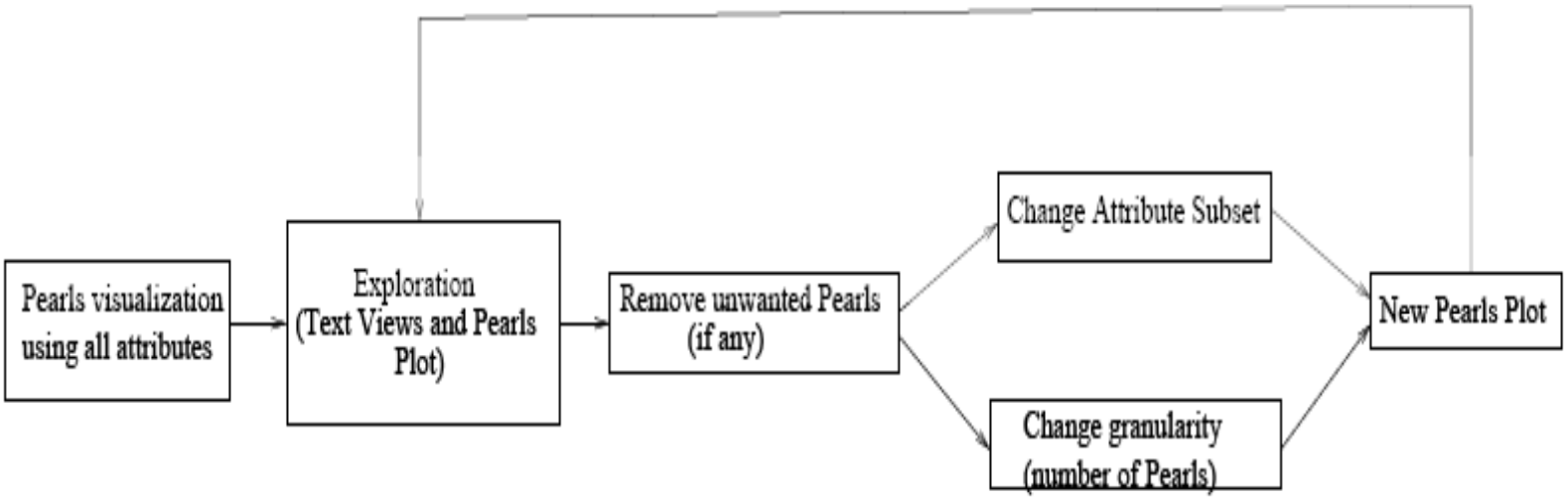
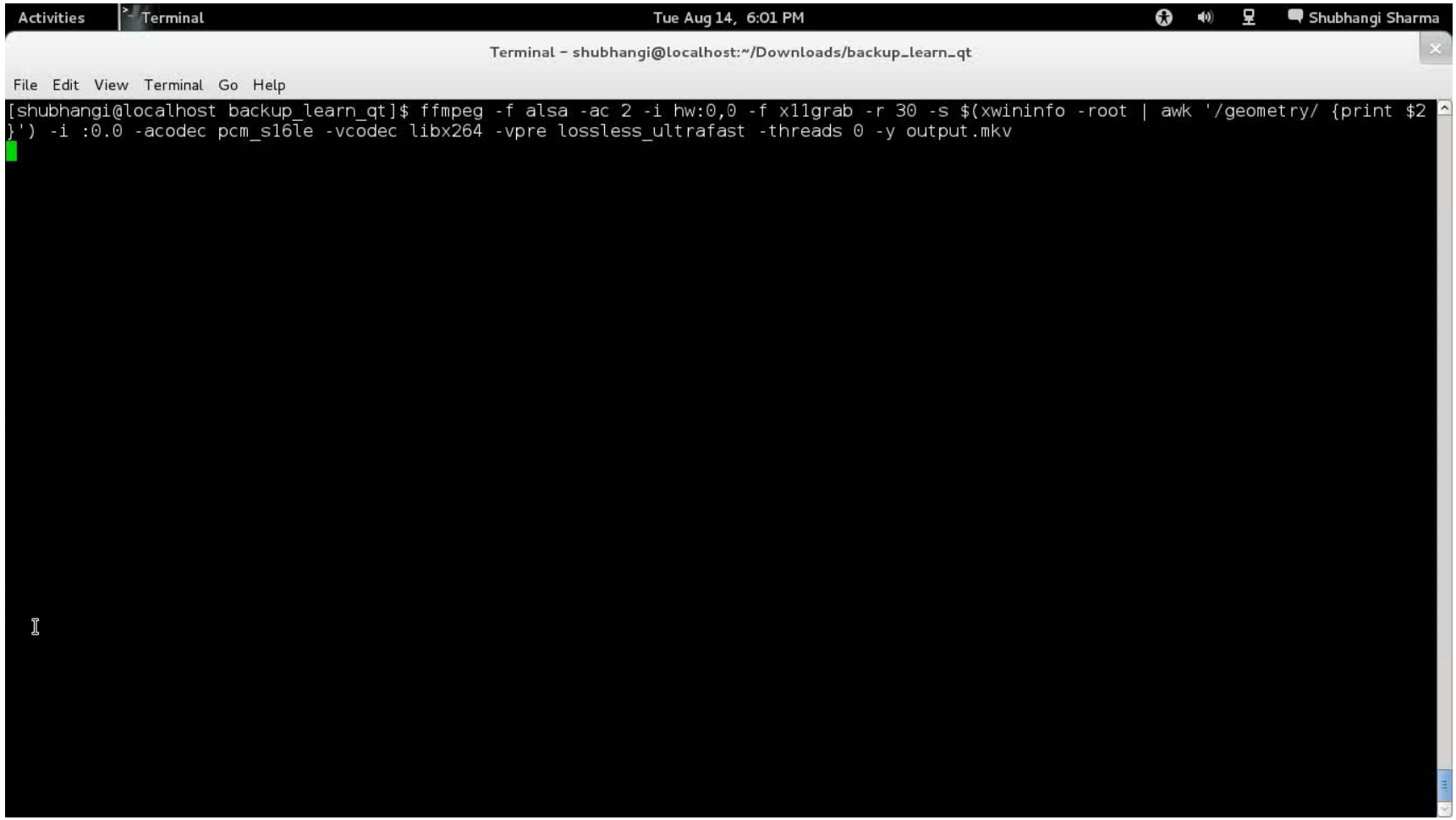
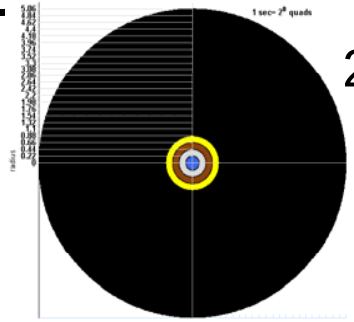


Figure 1: Flowchart

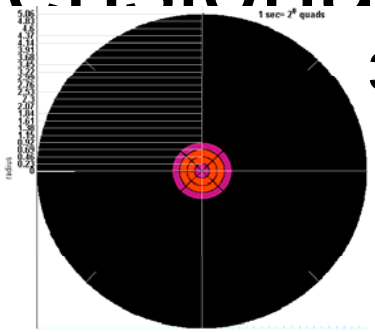
PEARLS Visualization



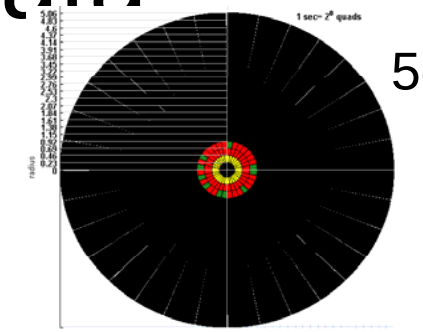
CROVHD – Concentric Rings of Visualization for high dimensional data



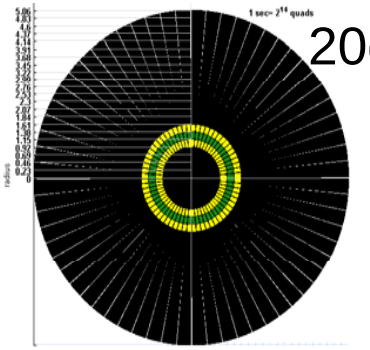
2d



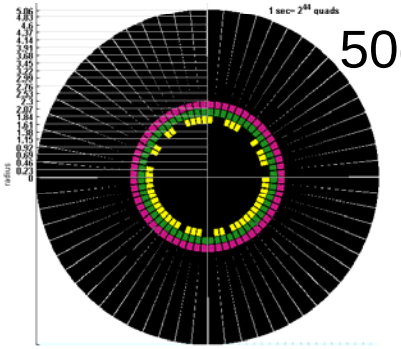
3d



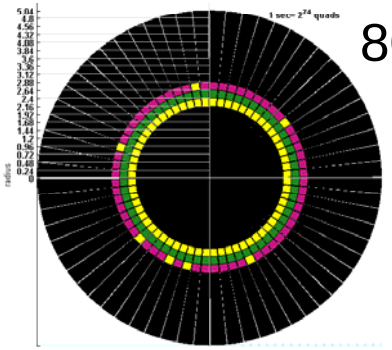
5d



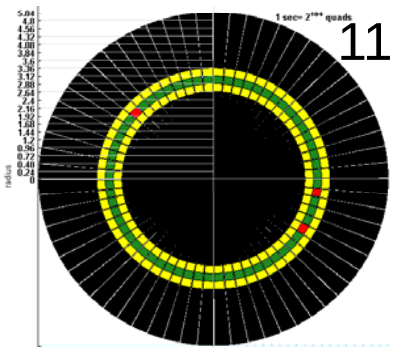
20d



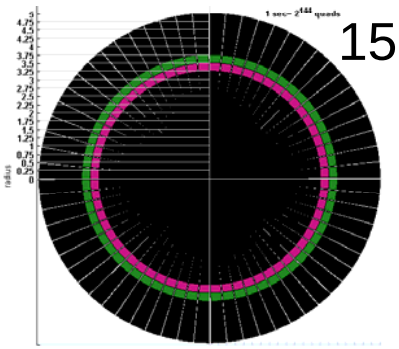
50d



80d



110d



150d

< 100

100 - 450

450 - 700

700 - 1050

1050 - 2000

2000 - 3000

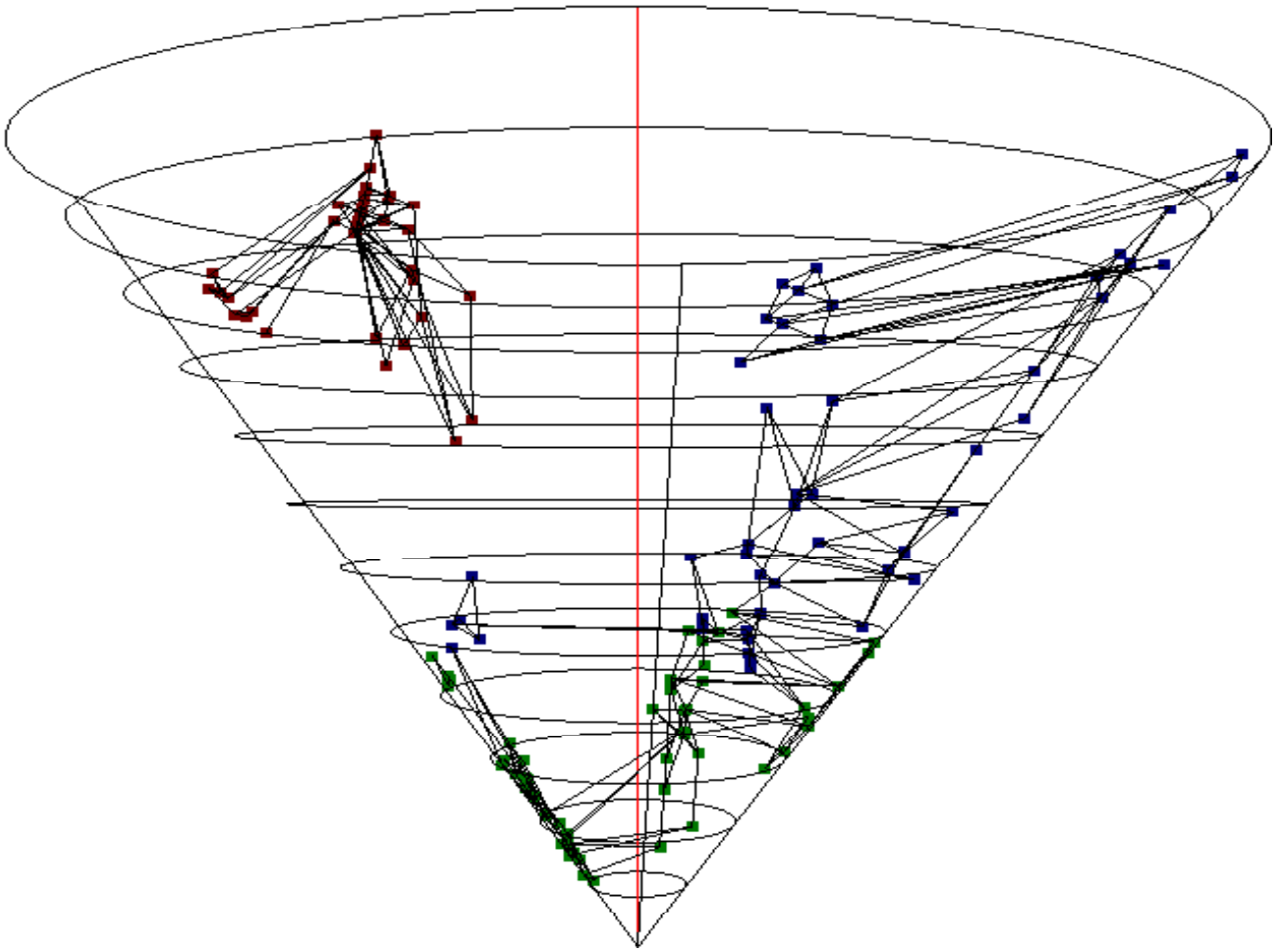
3000 - 4000

4000-7000

7000-10000

> 10000

CROVHD- Example – k-neighbour graph



Nearest Neighbour - Visualization

Related Work

- Parallel Coordinates [Inselberg 1985]
- VISA provides subspace overlap [Assent et al 2007]
- Best fit spheres or ellipsoids at high dimensions [Fitzgibbon, et al 1999, Calafiore 2002]
- Illustrative parallel coordinates [McDonnell & Muellller 2008]
- All 2-d subspaces scatter plots

Summary

- Subspace overlaps in high dimensions - HEIDI
- Different aspects of HEIDI
- Shape and Structure of clusters – BEADS & PEARLS
- High Dimensional Scatter Plots – CROVHD

(VAKD 2009, VAST 2009, LDAV 2013)

Open Problems

- Ordering of points in Heidi
- Tight fit of shapes – composition of shapes – extending to 3d shapes
- Exploration with navigation in Beads and Heidi
- Explorative analysis and analytics from CROVHD
- Time and space efficiency
- Integrated visualization tool kit for R^d data

Take away!

- Subtle work
- Fun with visualization
- Vast open areas to work in
- Dashboards for visual analytics
- Domain specific vertical solutions
- Deep mathematical problems – shape fitting – multiple loss-less visuals

Thank you!

Kamal Karlapalem

kkamal@iitgn.ac.in

Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- CROVDH
- Related Work
- Summary
- Open Problems

Heidi – Visual Relationship Matrix

- D = $\{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional
- Construct a n×n matrix where
  - Element (i,j) is a bit vector
  - Semantics of each bit in bit vector can be user specified
  - The matrix is visualized as an image
  - Patterns in image need to be interpreted

Generalization of gray scale visualization of distance matrix

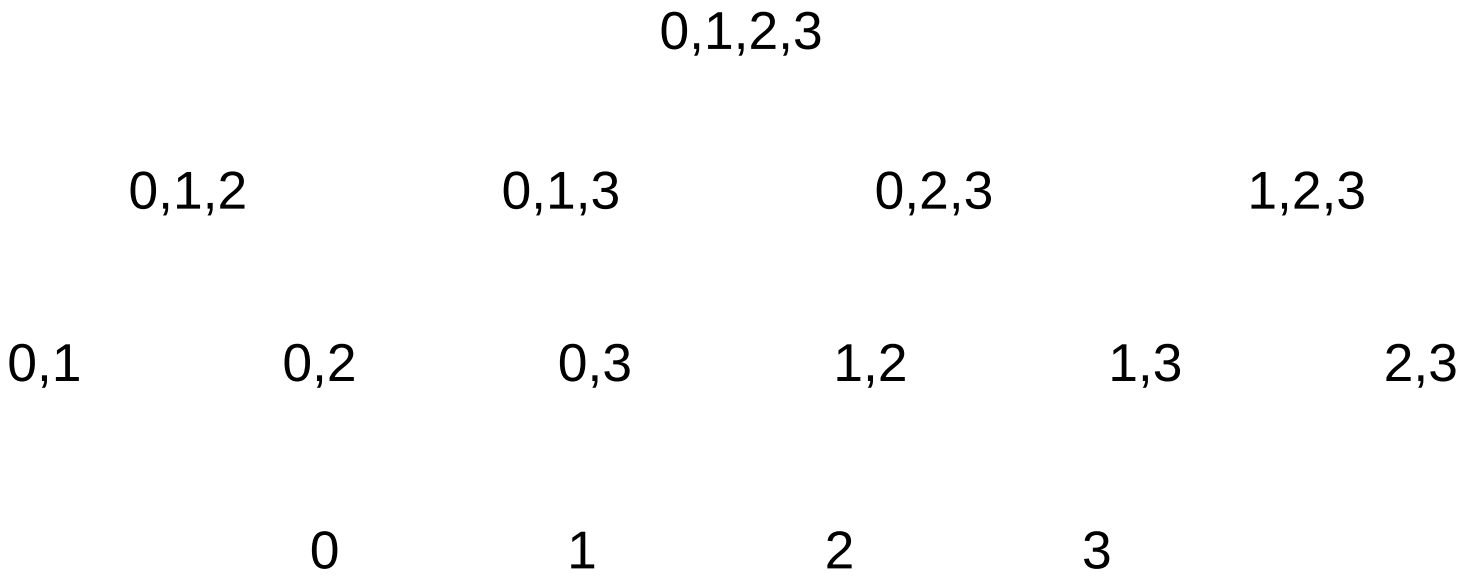
Heidi – specific case – Nearest Neighbors

- D = $\{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional
- Construct a n×n matrix where
  - Element (i,j) is a bit vector
  - Bit **p** of bit vector
    - is set to 1, if x_j is in k nearest neighbor set of x_i,
    - otherwise it is set to 0
    - For the **p^{th}** subspace of the data
  - Length of bit vector is 2^d-1
- Visualize bit-vectors using RGB combination of colors
- Size of matrix is n×n × [(2^d -1) bits mapped to RGB representation based on image type]

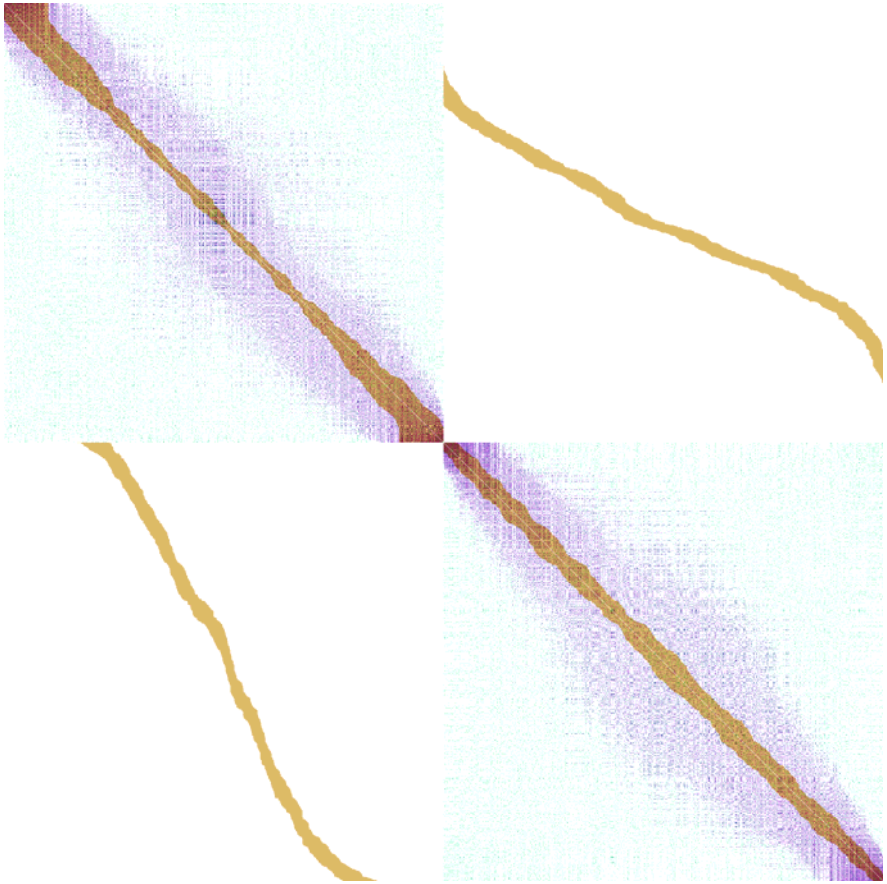
So, what have you got now? – a Heidi Matrix

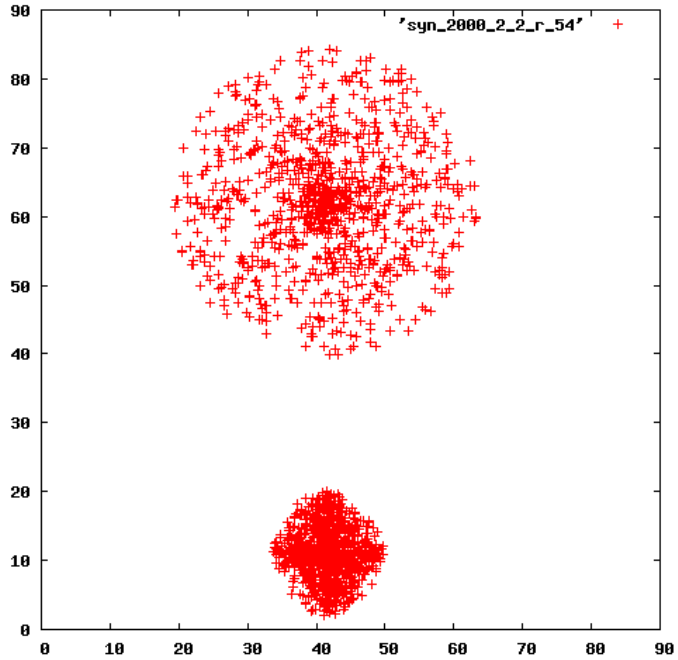
Subspaces

Dimensions – 0, 1, 2, 3;
Number of subspaces = 2^4 = 16;
sets of subspaces = 2^{15}-1



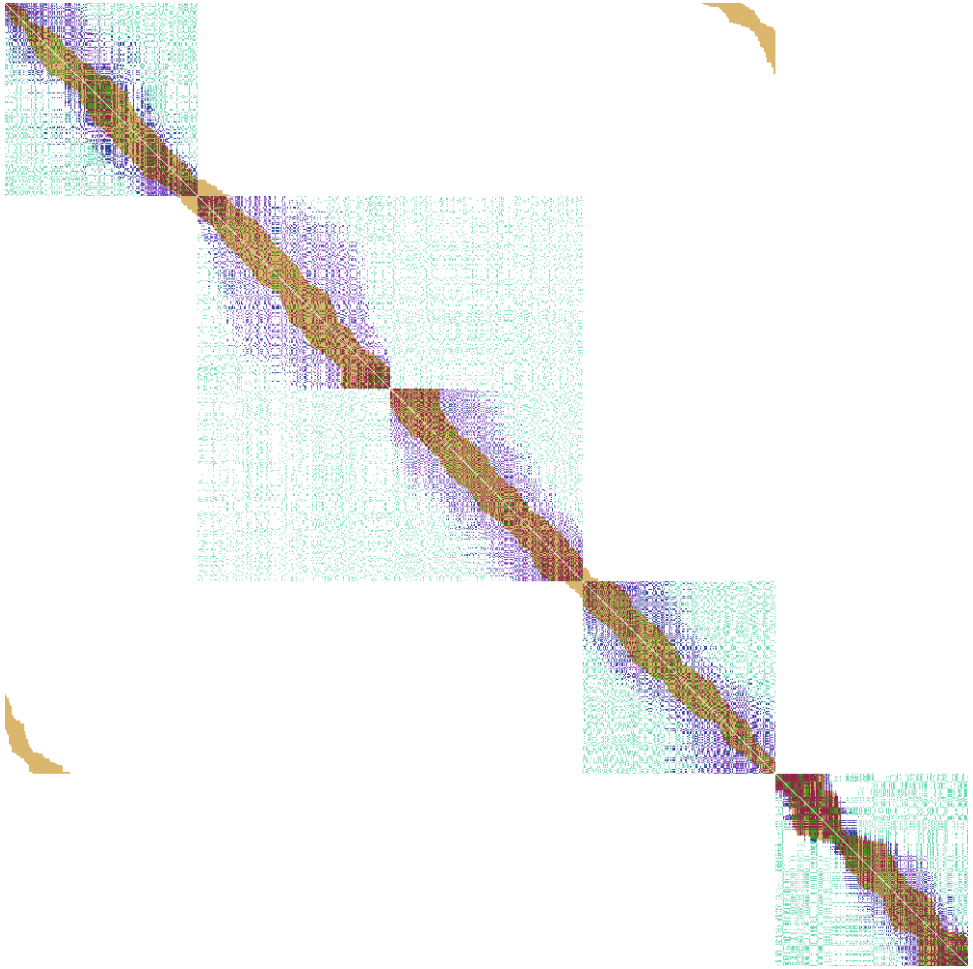
Examples

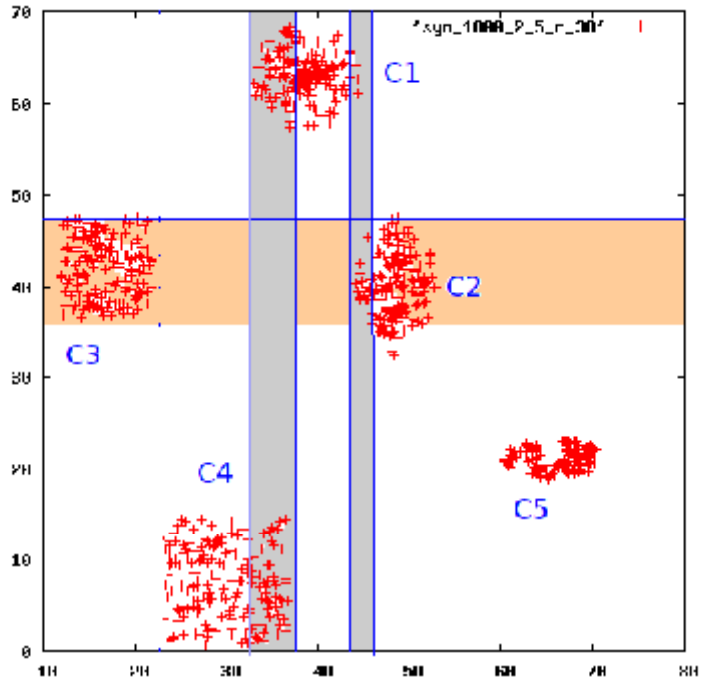




X – brown; Y – skyblue; {X,Y} - violet

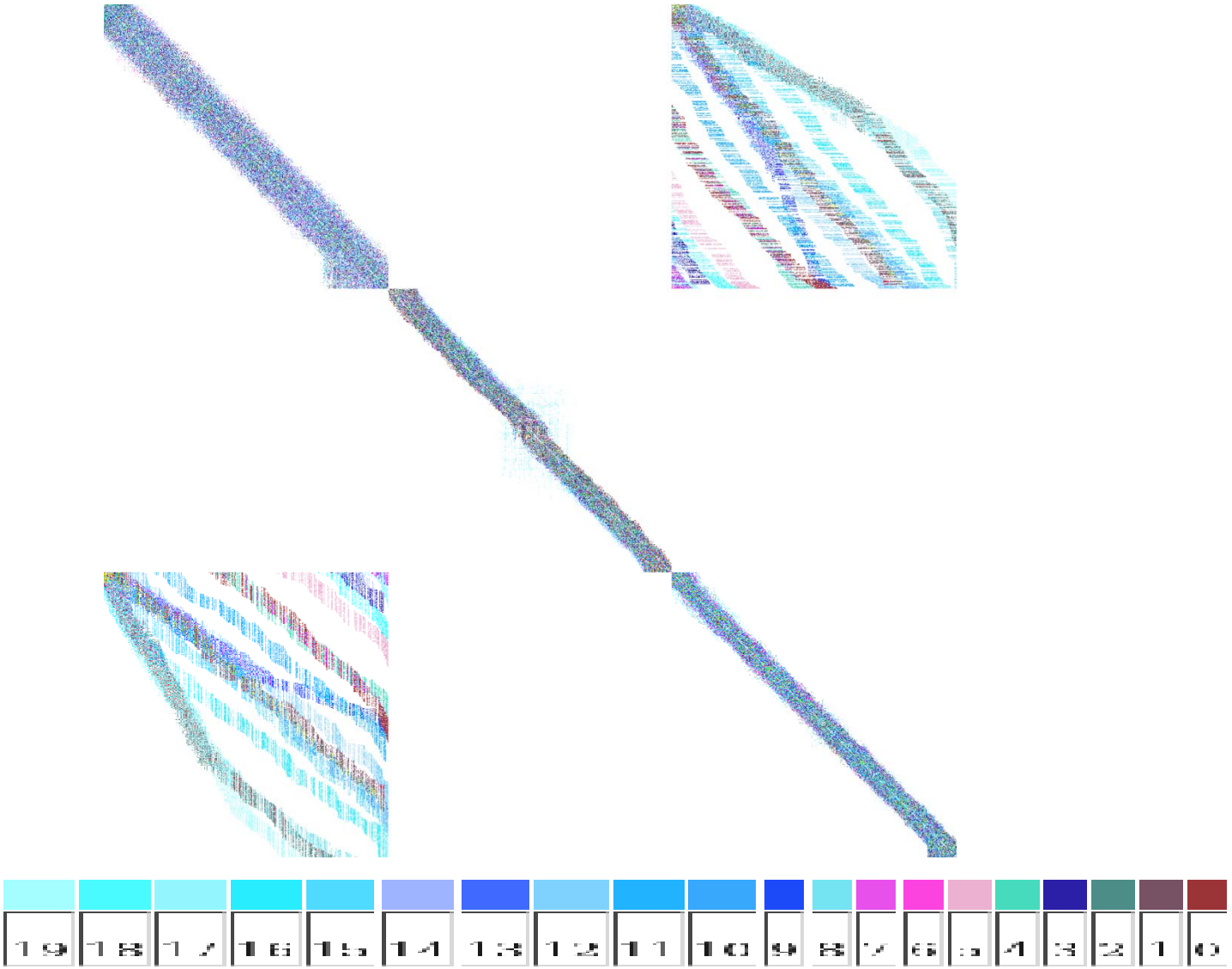
Examples



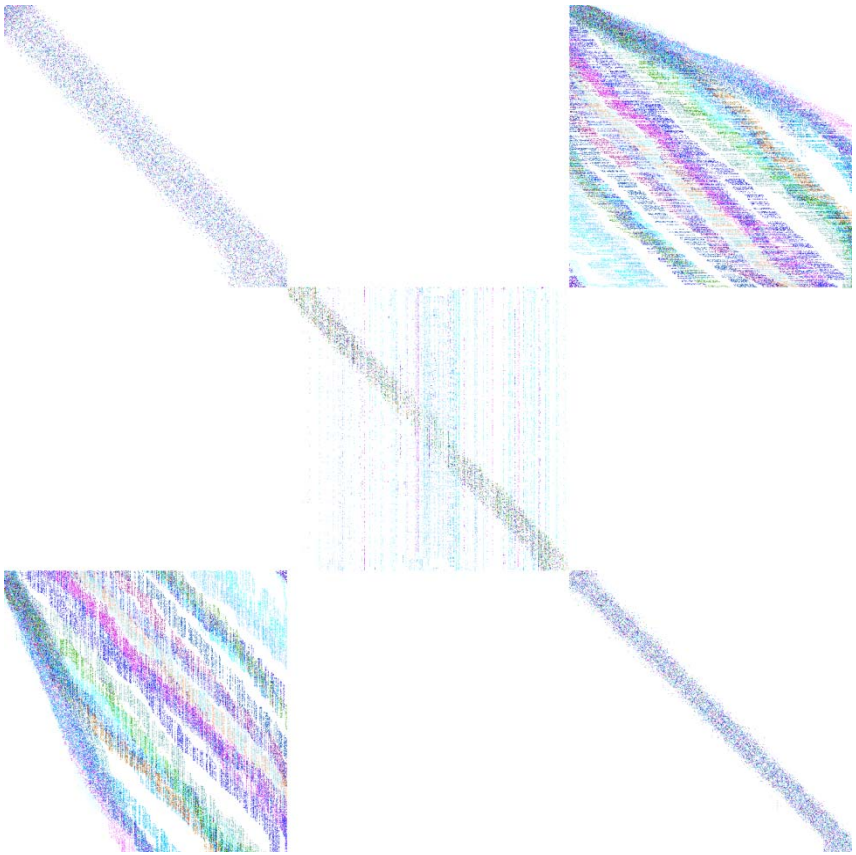


X – brown; Y – skyblue; {X,Y} - violet

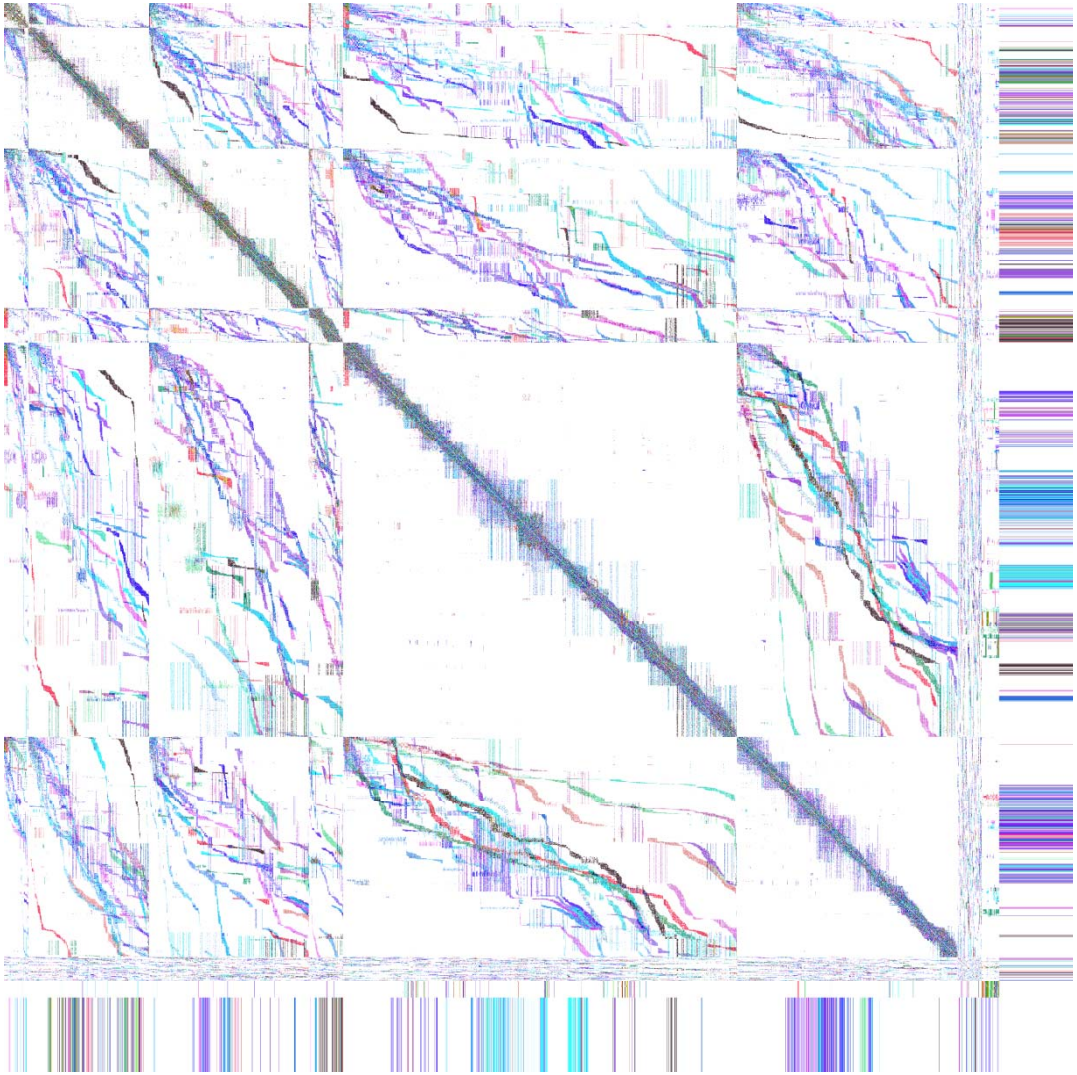
Examples: Composite Heidi – 20d



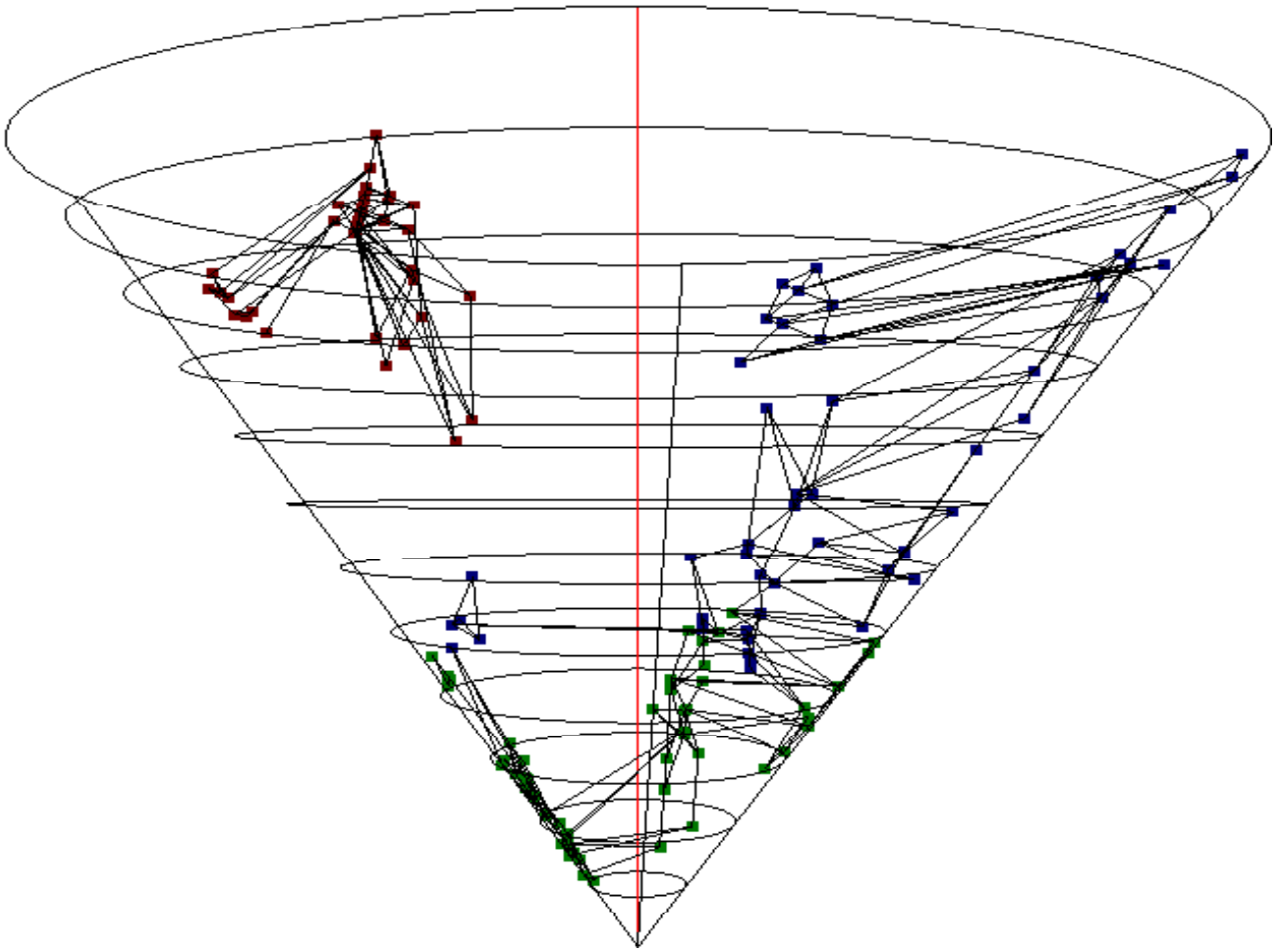
Examples: Composite Heidi=50d



Real-estate Property Listings



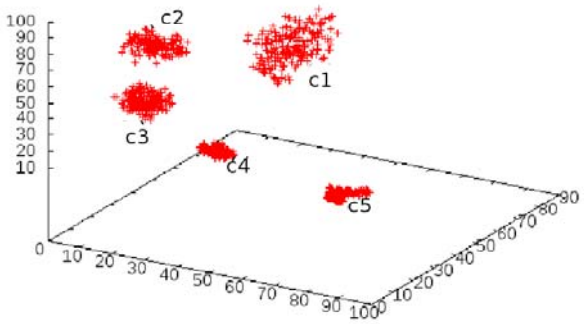
Example – k-neighbour graph

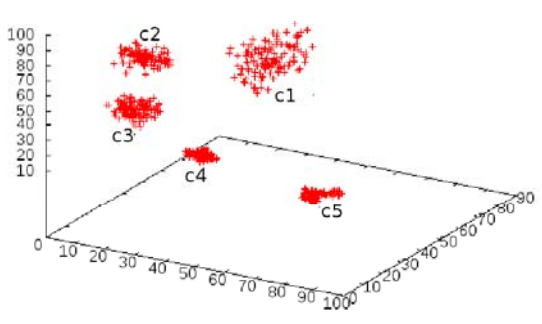


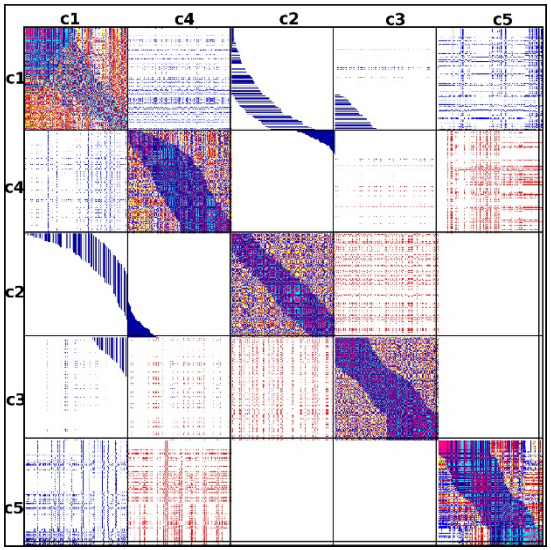
Heidi Matrix - Issues

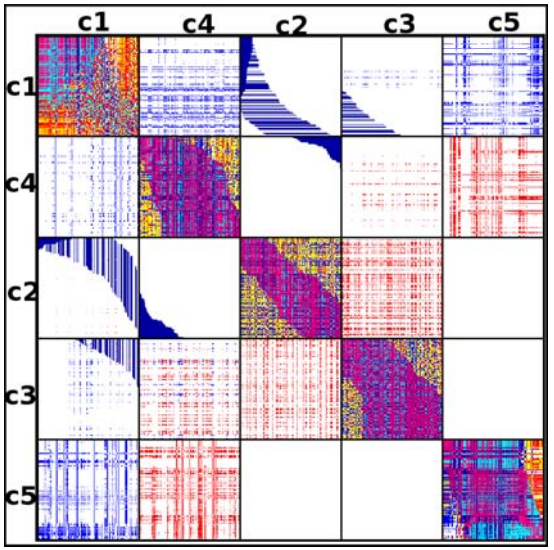
- Ordering of points in a cluster
- Size of the matrix
- Mapping of colors to bit vectors
- Types

Representative Heidi Images



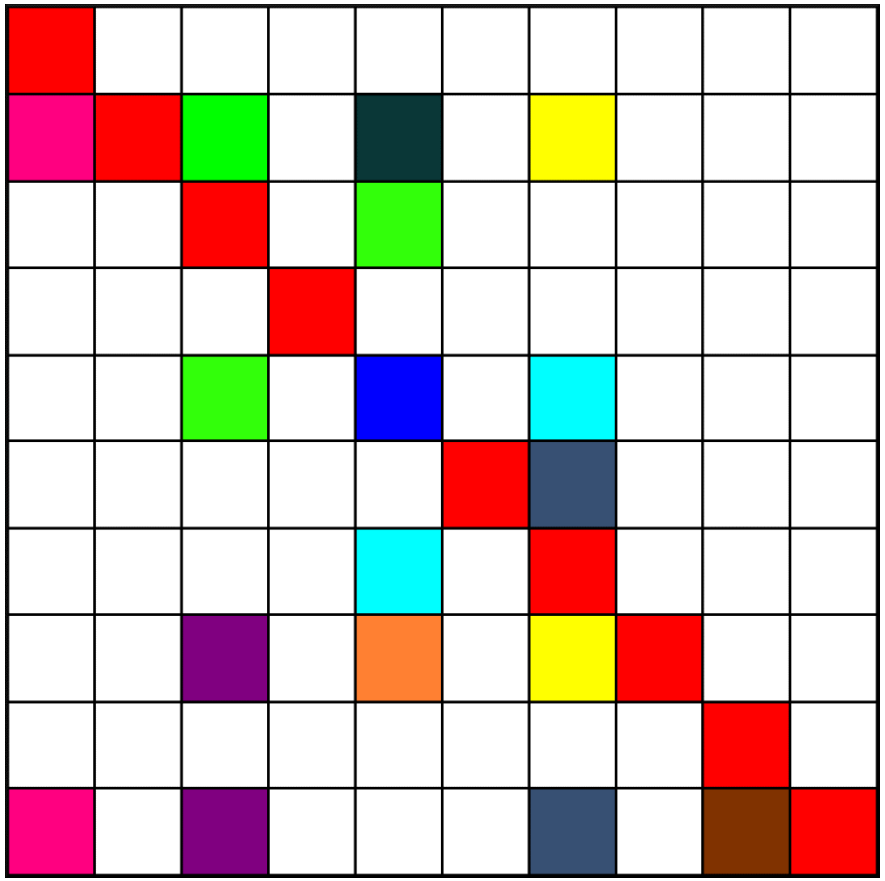






Color								
Set of subspaces	None	{ 2 }	{ 1 }	{ 1 }{ 2 }	{ 0 }	{ 0 }{ 2 }	{ 0 }{ 1 }	{ 0 }{ 1 }{ 2 }

N= 1,00,000 and d=100, prominent subspace



Color	Set of subspaces
	None
	{{ 0 }{ 1 }..{ 99 }}
	{ 1 }
	{{ 0 }{ 1 }..{ 99 }} -{ 1 }
	{{ 0 }{ 1 }..{ 99 }} -{ 0 }
	{ 10 }
	{ 49 }
	{ 65 }
	{ 42 }
	{ 97 }
	{ 51 }
	{ 16 }
	{ 74 }
	{ { 0 }{ 1 }..{ 99 }} - { 2 }
	{ { 0 }{ 1 }{ 2 }..{ 99 }} - { 3 }
	{ 9 }
	{ 38 }
	{ 95 }
	{ 20 }

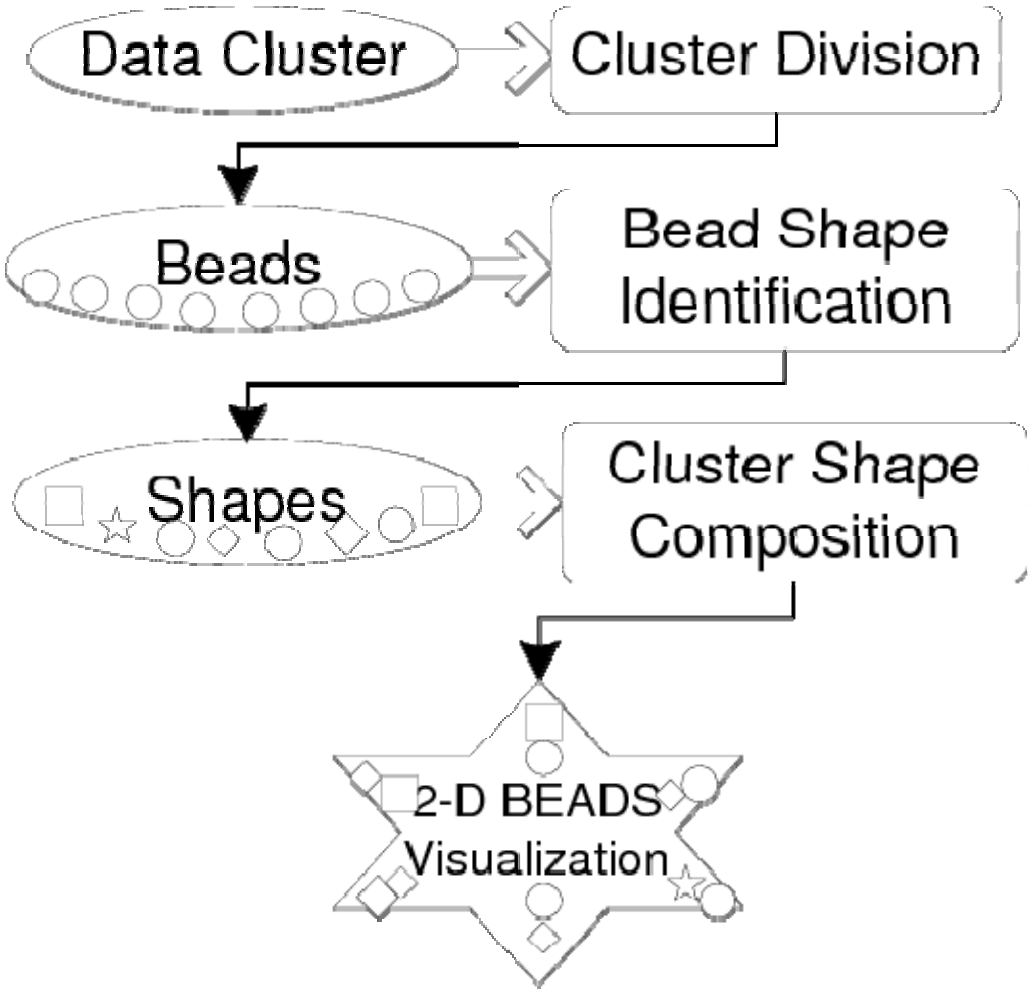
Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- CROVDH
- Related Work
- Summary
- Open Problems

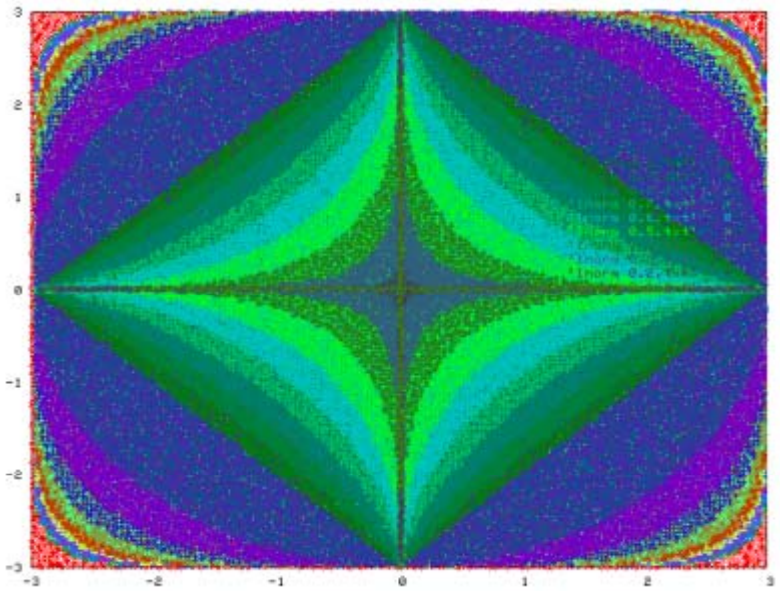
BEADS – Forming a Necklace

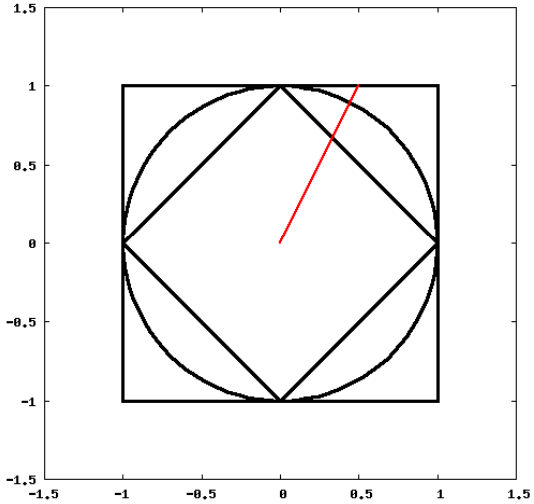
- Given a cluster – that is, a set of points much closer among themselves but well separated from other sets of points
- Need to determine shape and size of the cluster
- Partition points into subsets of points
- Each subset forms a bead
- Beads are mapped to well-specified 2-d shapes
- Beads are placed in canvas to visually represent shape and size of cluster – **a necklace**

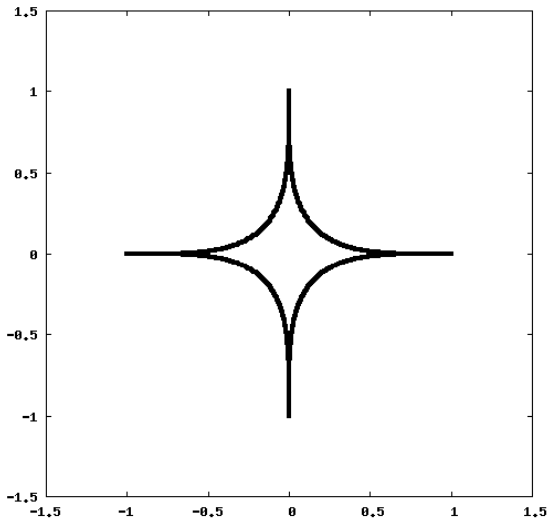
Beads - Approach



Basis for Beads







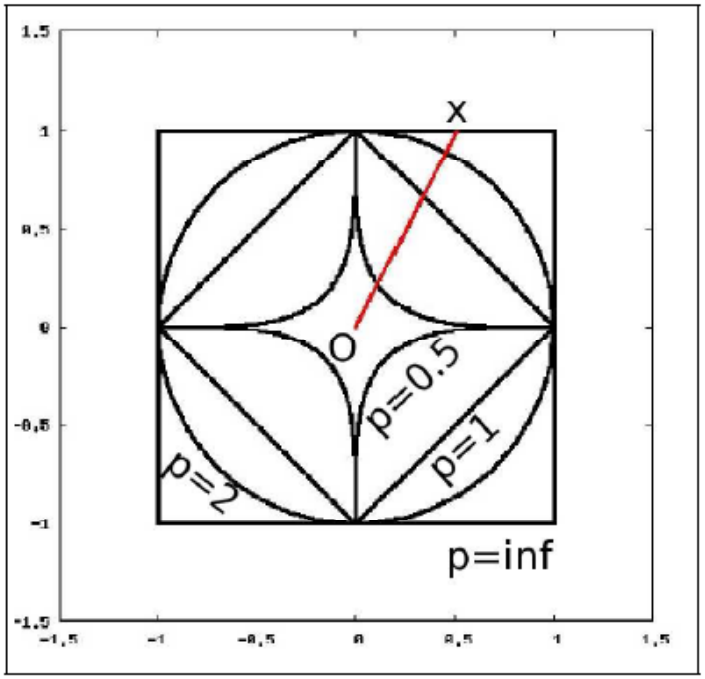
Beads – shape and size

- P = set of distinct *p* values for L_p norm
- Aim: Identify '*p*' and radius 'r_p' that covers the bead tightly

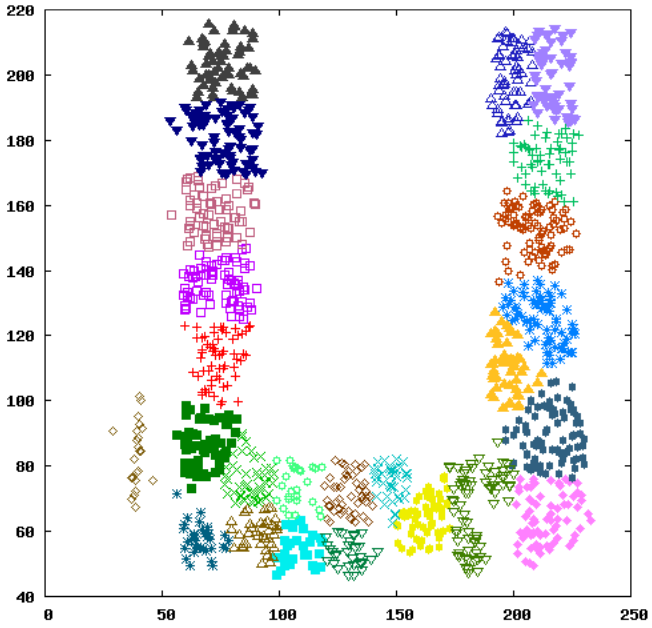
• *Two approaches*

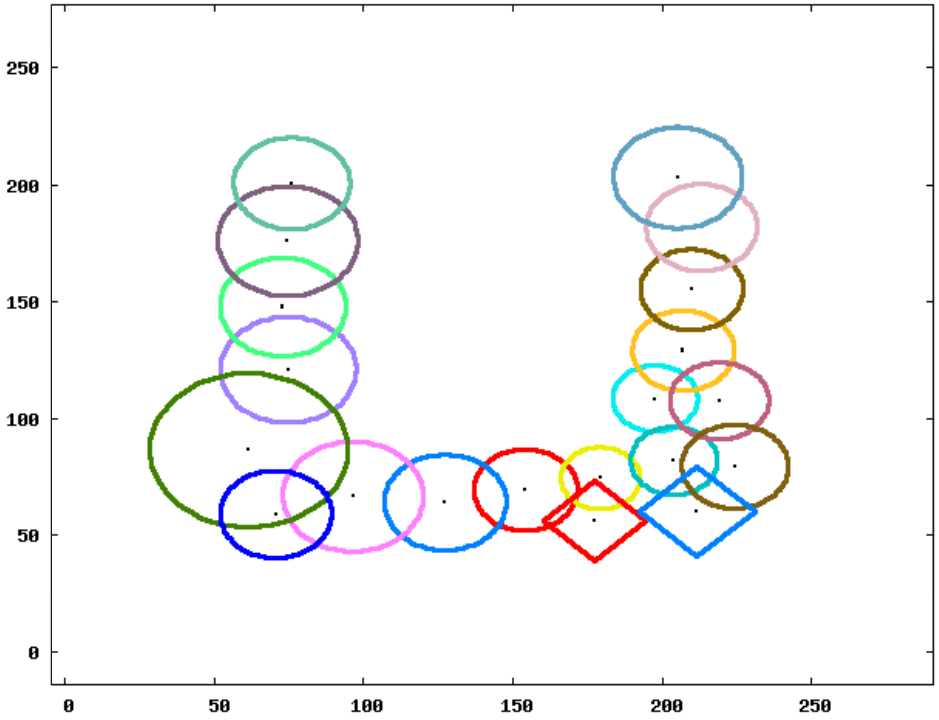
1. *Iterate from p by considering distances between centroid and furthest point using L_p, select the p which has the smallest distance.*
2. *Find the sum of distances among all pairs of points using L_p, and select the p that has smallest sum of distances*

- *The selected p gives the shape.*
- *The size is given by the diameter using the L_p*

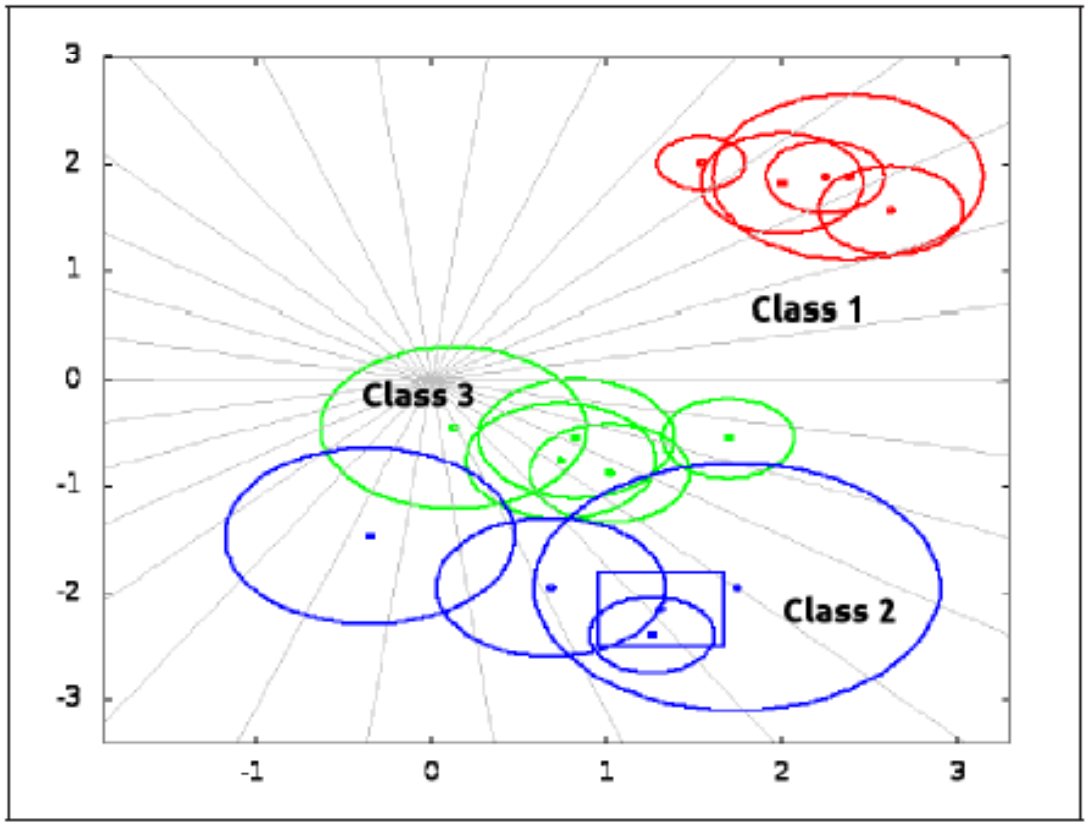


Examples

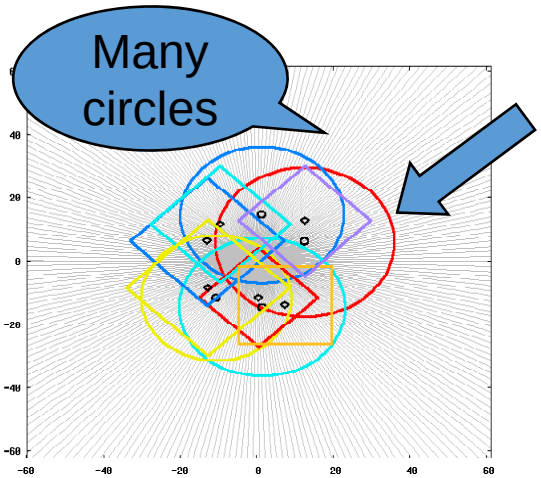

220
200
180
160
140
120
100
80
60
40
0
50
100
150
200
250


250
200
150
100
50
0
0
50
100
150
200
250

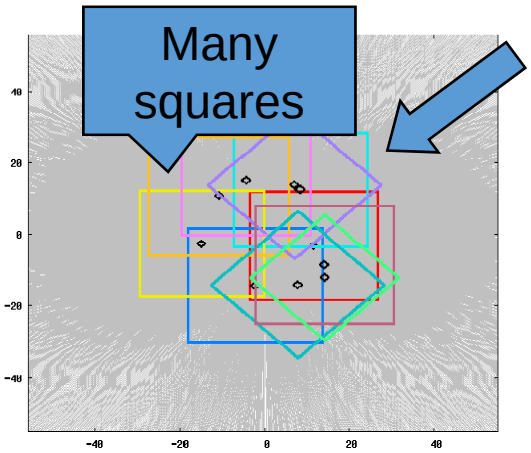
Example – Iris Data Set



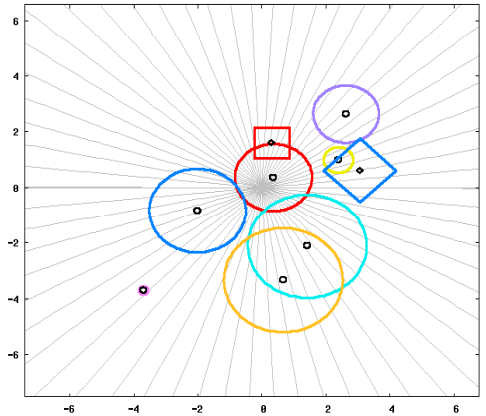
More results



10-D Hyper-sphere



10-D Hyper-cube



5-D NBA Player Data

Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- **Pearls**
- CROVDH
- Related Work
- Summary
- Open Problems

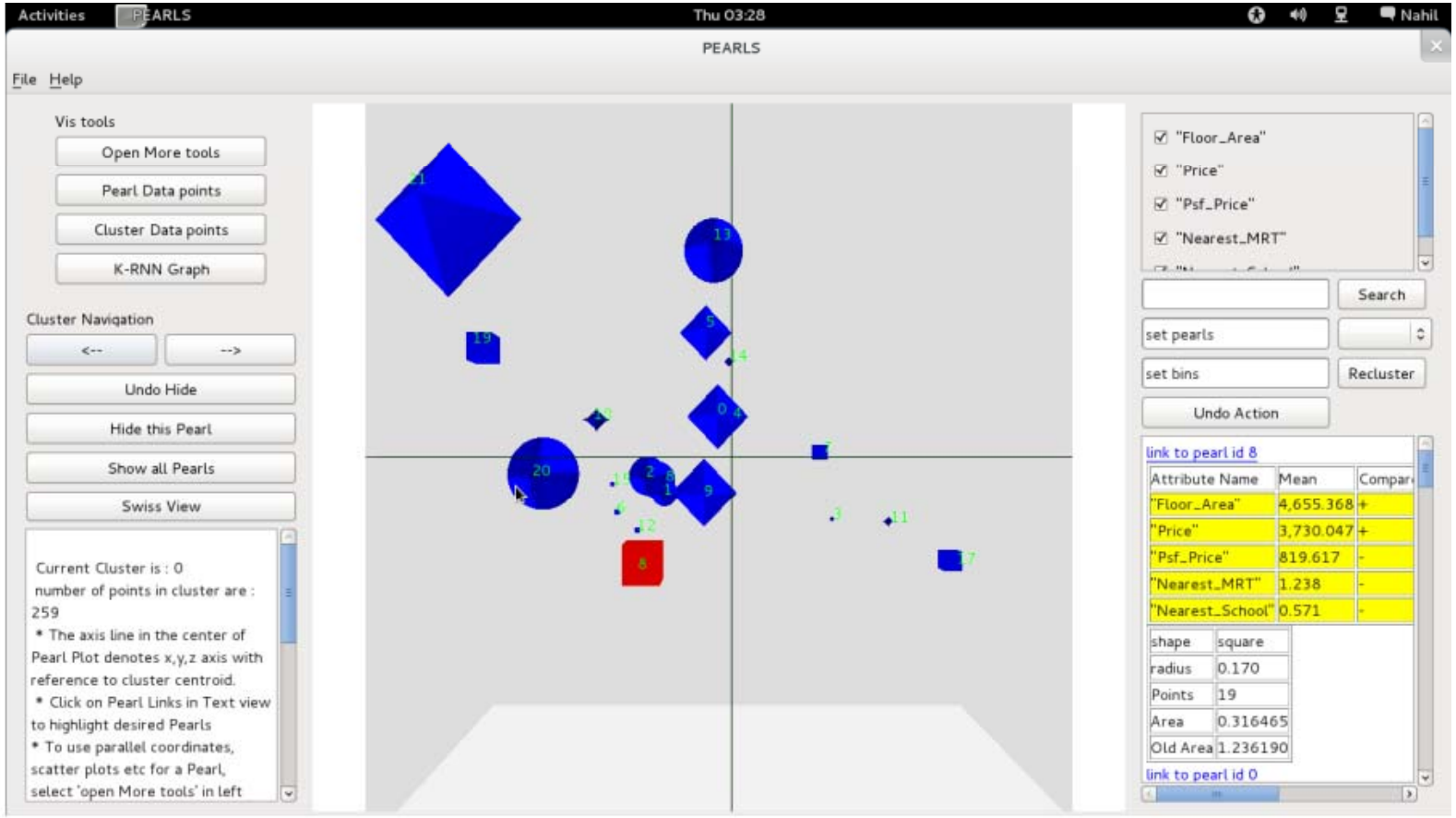
PEARLS

- Basic 3D shapes to visualize high dimensional clusters.
- Level of abstraction between data point & cluster level.
- Interactive techniques make cluster analysis informative and intuitive.
- Techniques for detailed analysis of individual pearls.
- Useful in cluster analysis and concept identification within clusters. (Case Studies)

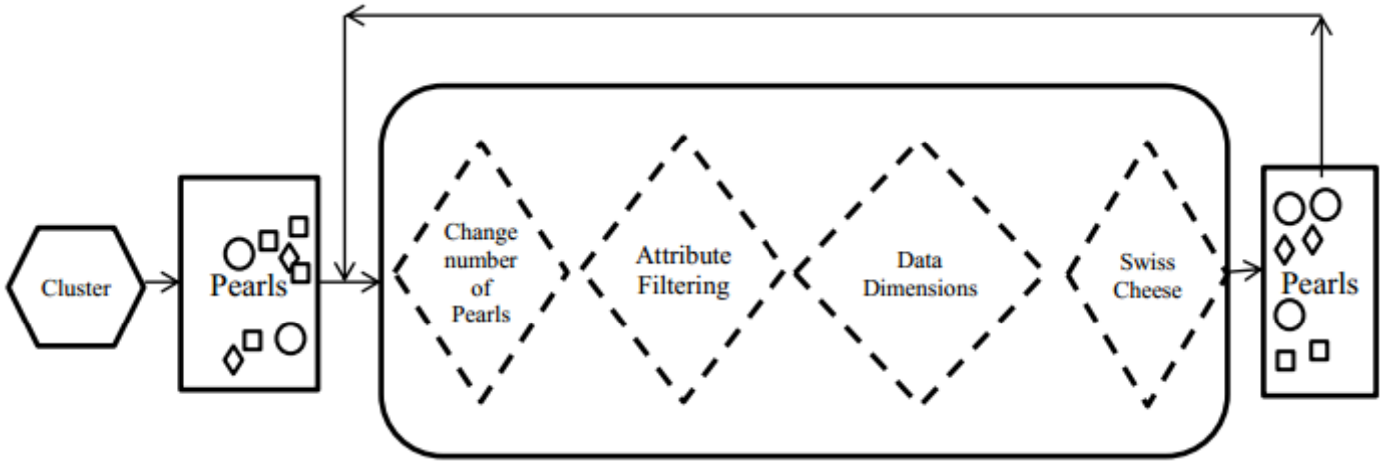
Need for 3-d Pearls

- Overlap in 2 D
  - 3 D gives an extra dimension
  - rotate the camera and view from various angles.
- Position of a bead conveys only its distance from centre and the quadrant.
  - In 3-D, position conveys
    1. distance from cluster centroid
    2. quadrant
    3. value in chosen dimension
- Facilitates data dimension interactive technique due to extra dimension

PEARLS Visualization



Pearls Visualization System



Visual Explorative Querying

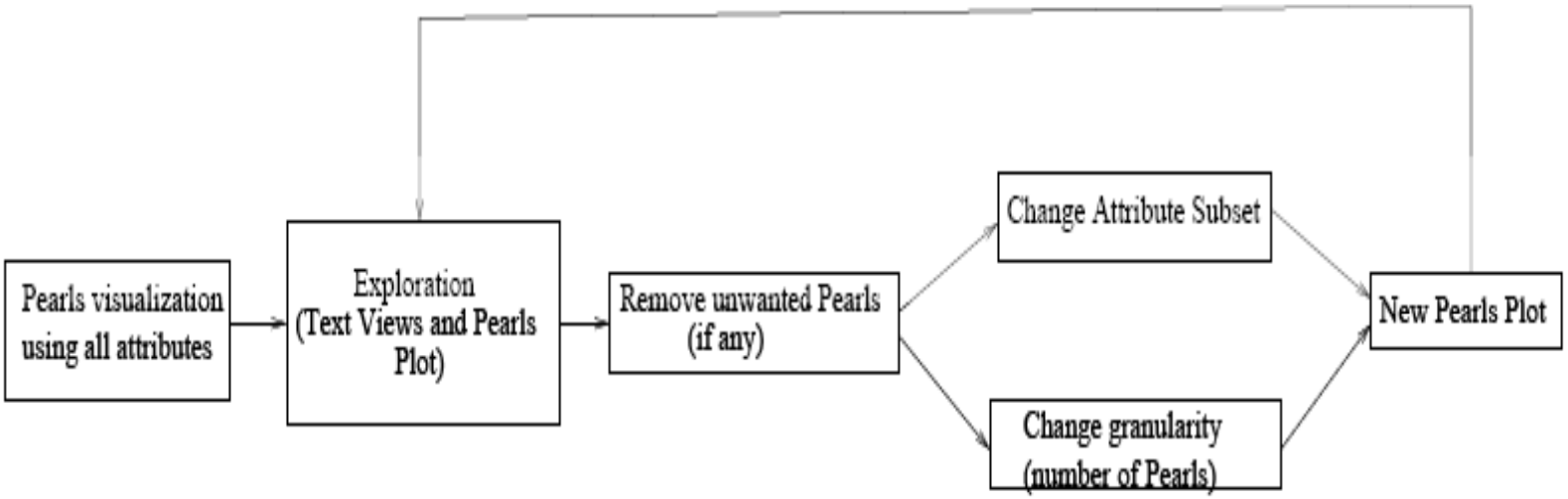


Figure 1: Flowchart

Video of Exploratory Visual Querying

Summary

- PEARLS can be effectively used for visual data analysis
  - exploring cluster as a query result.
  - supports expression of multidimensional queries through interaction and aid of data mining.
  - follow complex lines of inquiry using sequences of simple interactions one can follow complex line of enquiry.
- In a lot of data analysis tasks, it is difficult to specify data points of interest as set of mathematical and Boolean rules.
- It is also difficult to update rules when new interests are found. Moreover, a viewer may not know apriori what they will find interesting.
- PEARLS visualization uses clustering to group points and makes the analysis of dataset easier. This helps in finding interesting data points via exploration.

Summary

- PEARLS does not suffers from drawbacks like
  - inability to plot complete dataset
  - loss of speed and interaction

  number of visual objects(pearls) is << number of data points.
- May suffer from over plotting and decline in legibility when some pearls are overshadowed by larger pearls
  - an effective text based view and ability to rotate the 3-D visualization vertically and horizontally solves this problem.

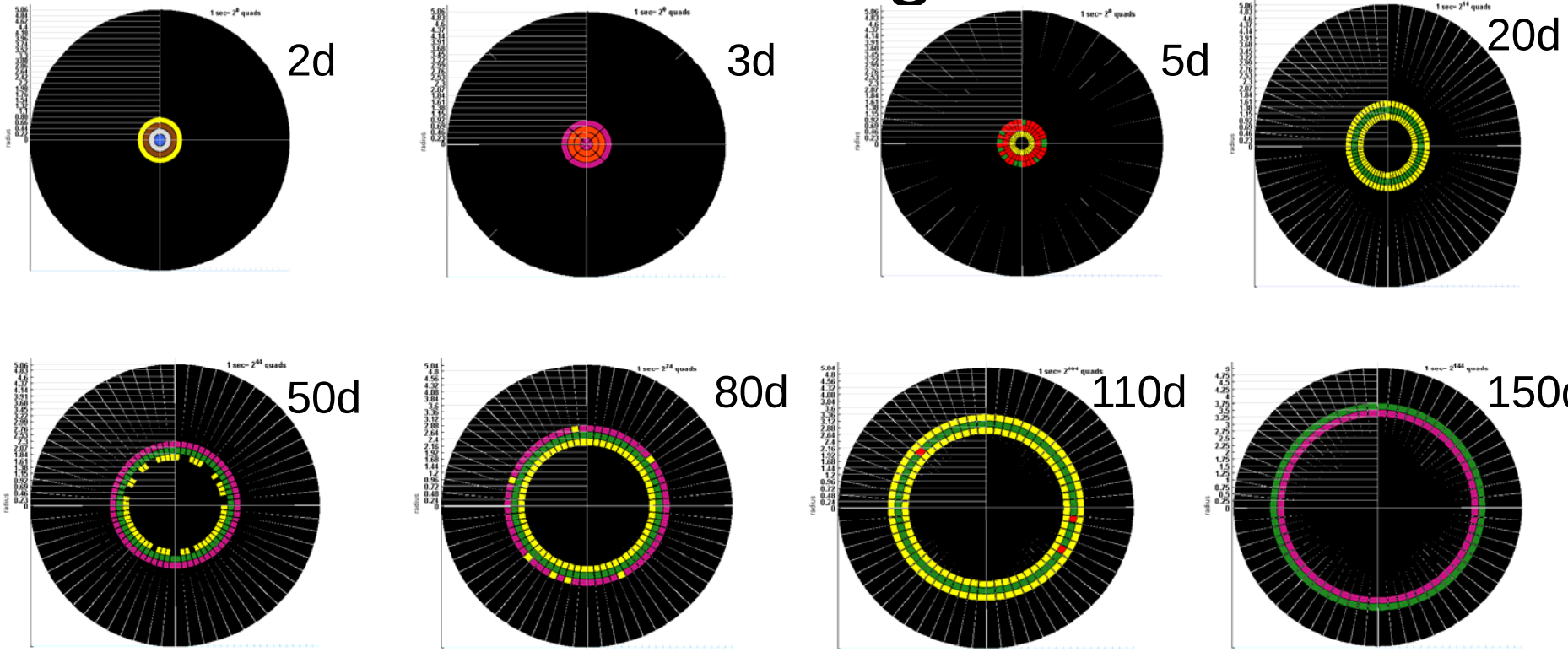
Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- CROVDH
- Related Work
- Summary
- Open Problems

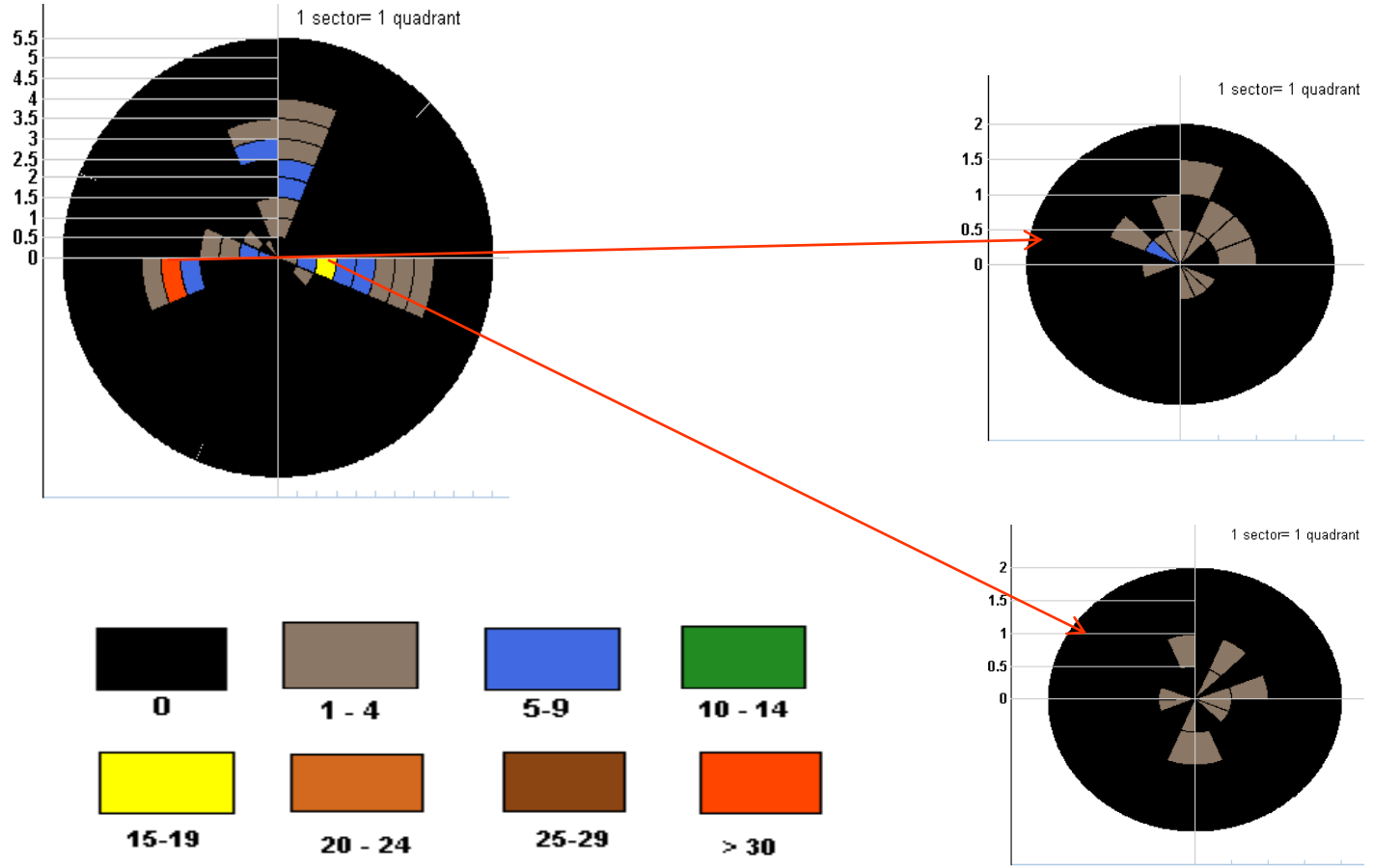
CROVDH – Concentric Rings of Visualization for high dimensional data

- Given a data set $x_1, x_2, \ldots, x_n$ d-dimensional data
- Determine a scatter plot visualization
- Spilt the 2-d space into 2^d quadrants
- Map each x_i to (r, θ) coordinates
  - R is based on distance from centroid to point
  - θ is based on quadrant and the relative angle within quadrant from some base axis
- Divide regions of 2-d space as concentric circles
- Give region colors based on relative density
- Can also show actual points

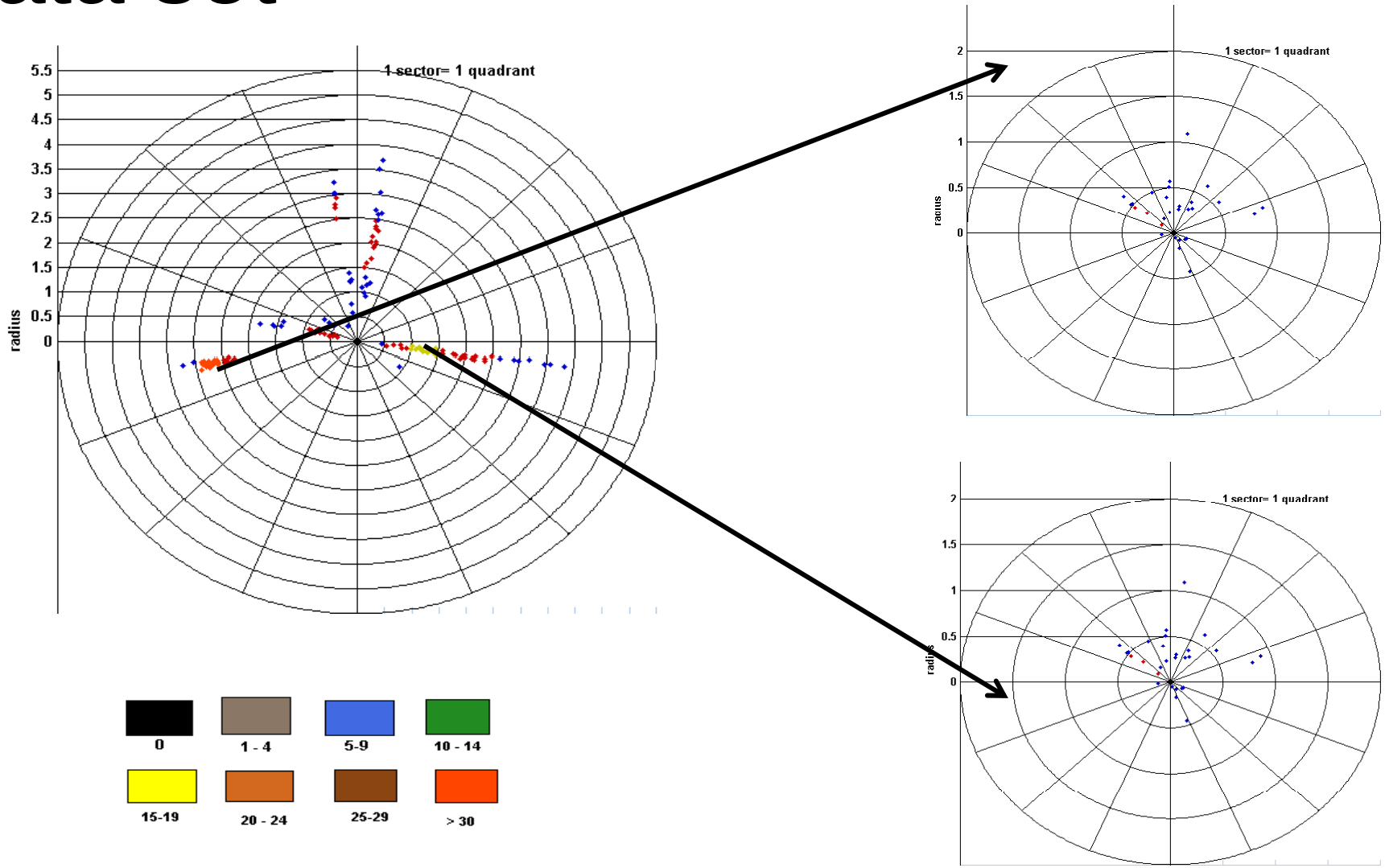
Uniform 100,000 [0,1] points dimensions increasing



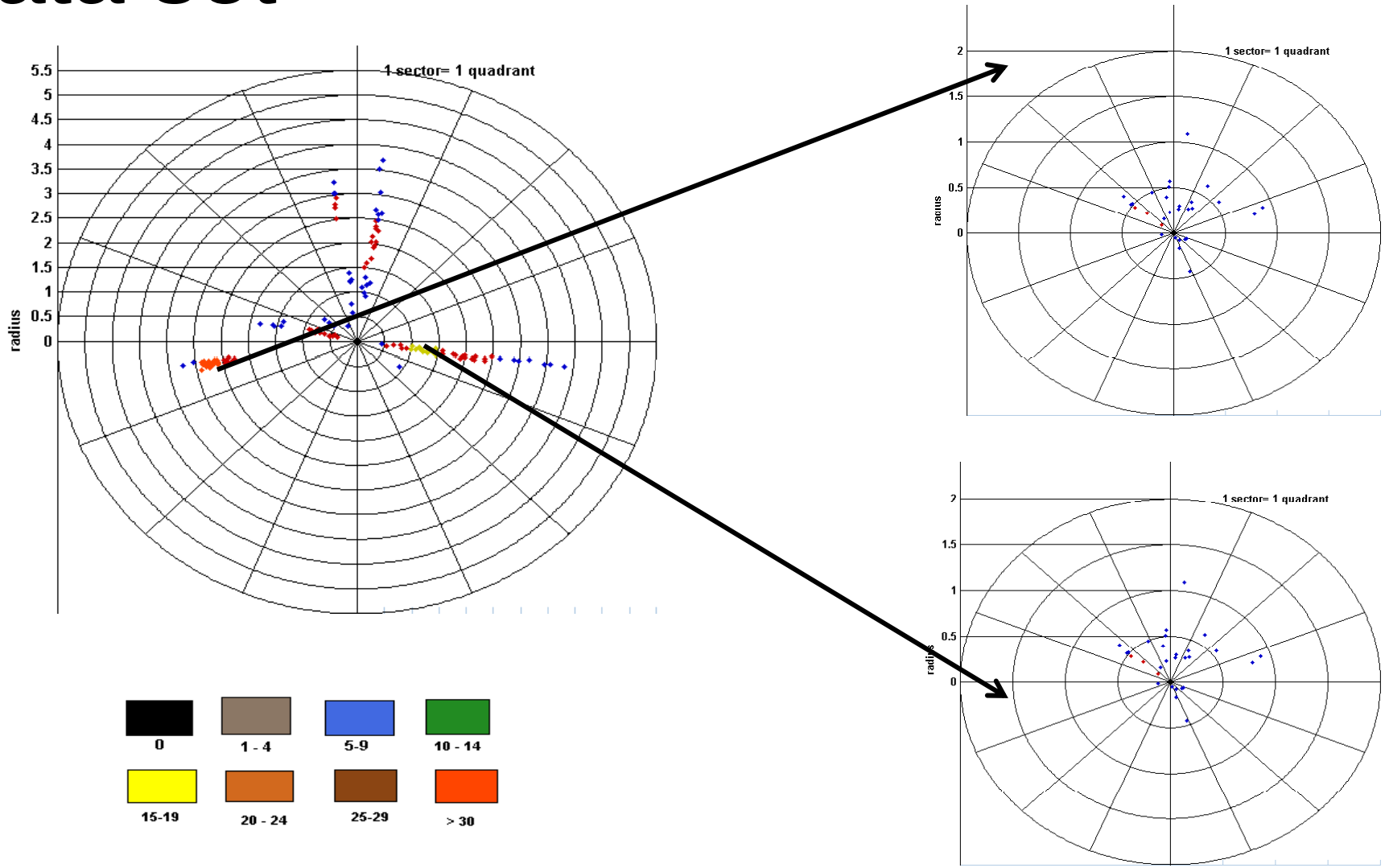
CROVDH Visualization of IRIS data set



CROVDH Visualization of IRIS data set



CROVDH Visualization of IRIS data set



CROVDH

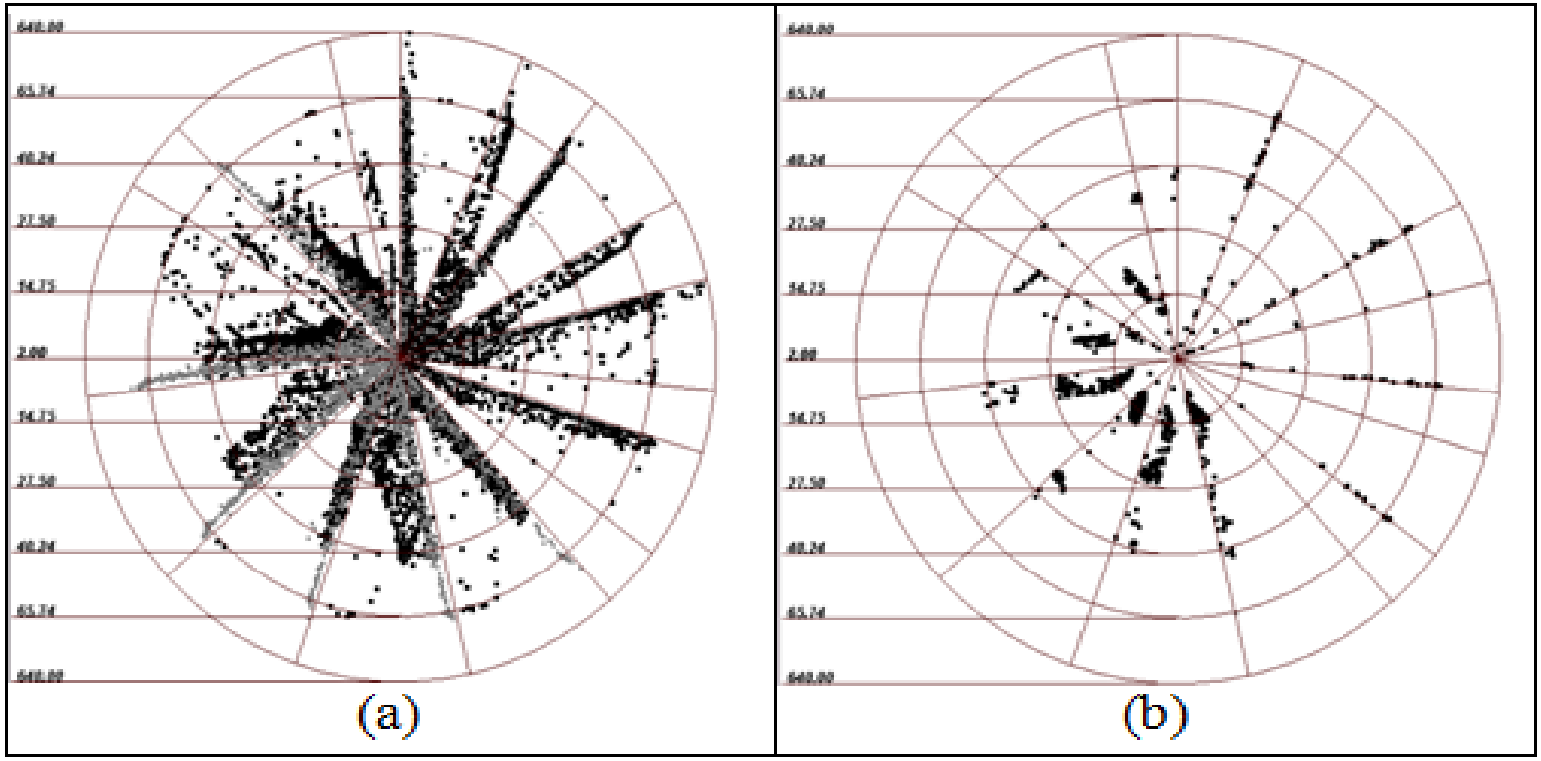


Figure 2 (a) - Scatter plot of 4d synthetic dataset of 10000 instances. The grey boxes represent overlapping points which are plotted in 5(b)

Enhanced CROVDH

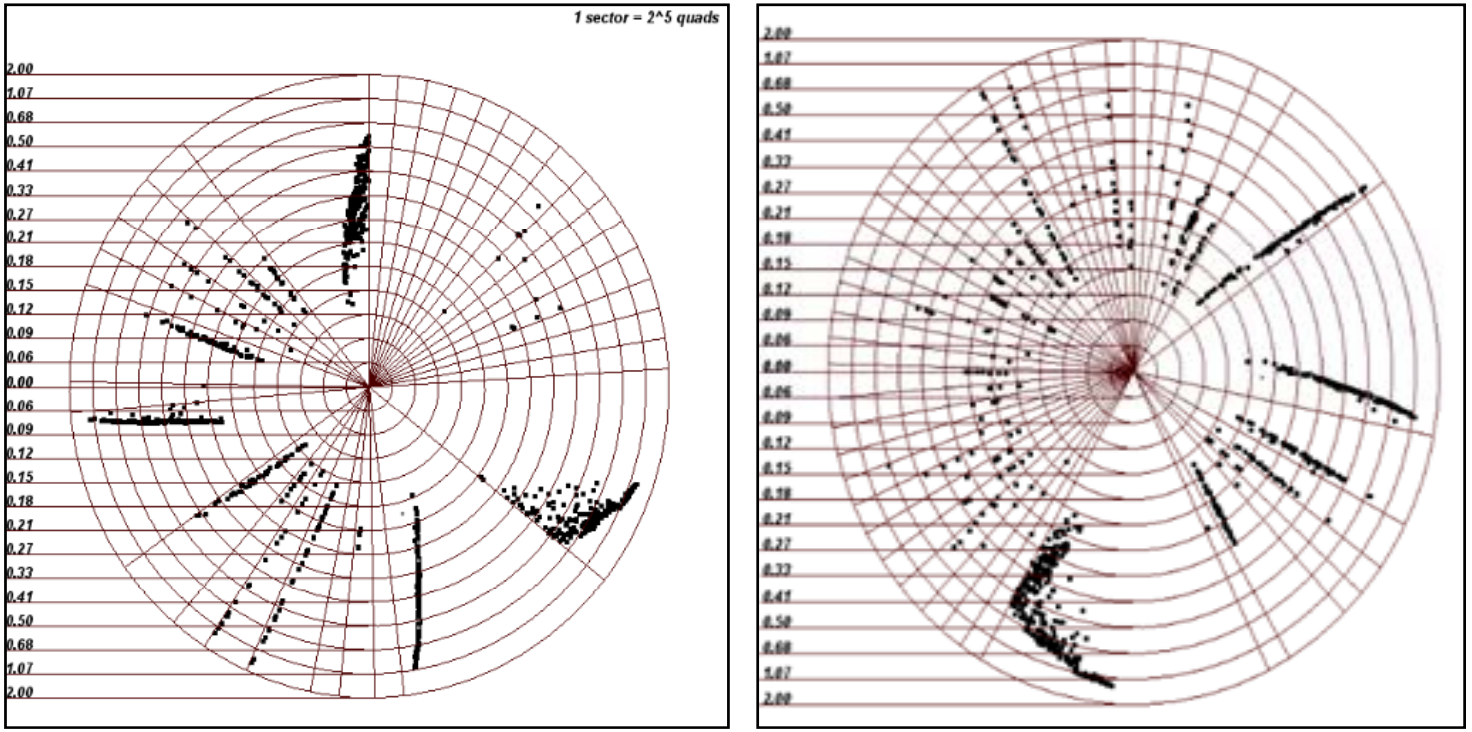
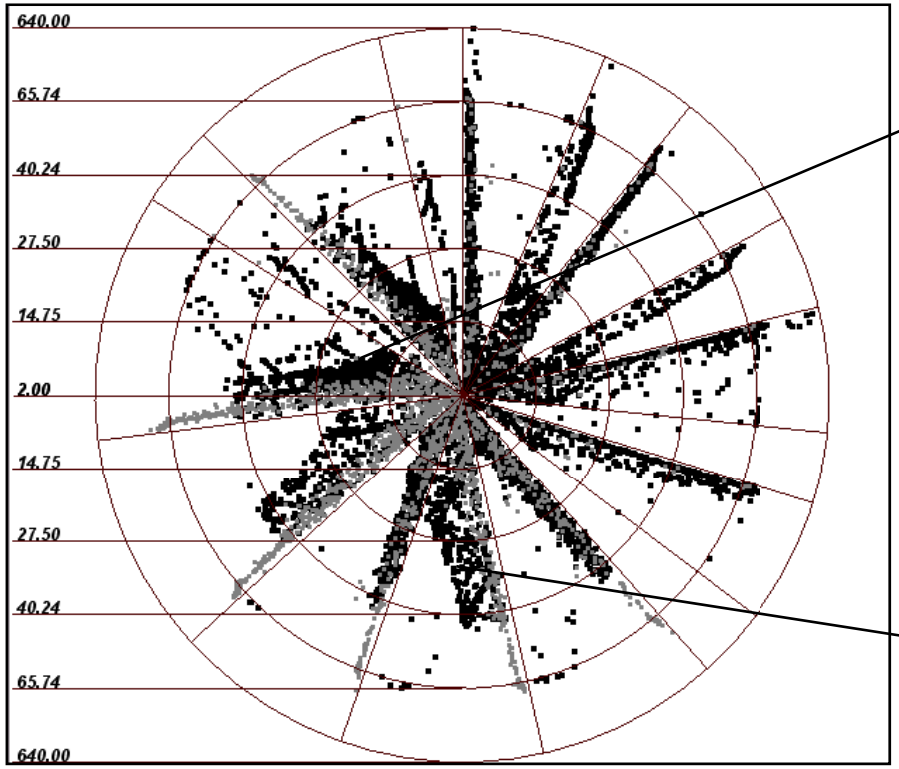


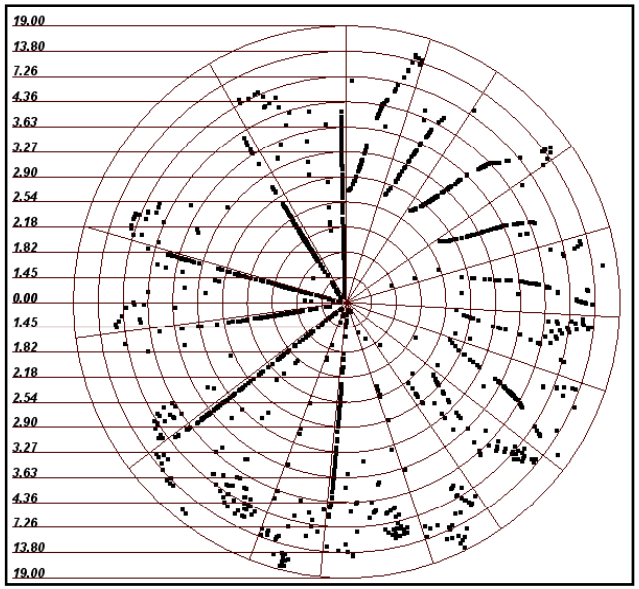
Figure 3(a): Basic CROVHD plot of 10d synthetic dataset with 2000 points. 3(b): Modified CROVHD plot of the same dataset.

Nested View



(a)

(b)



(c)

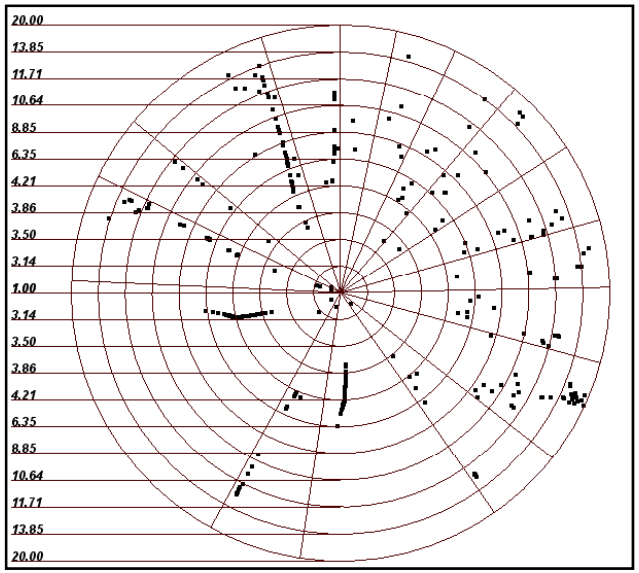


Figure 11: (a) is initial scatter plot of 4d synthetic dataset with 10000 instances. (b) and (c) are scatter plots produced when clicked on respective bins.

Example

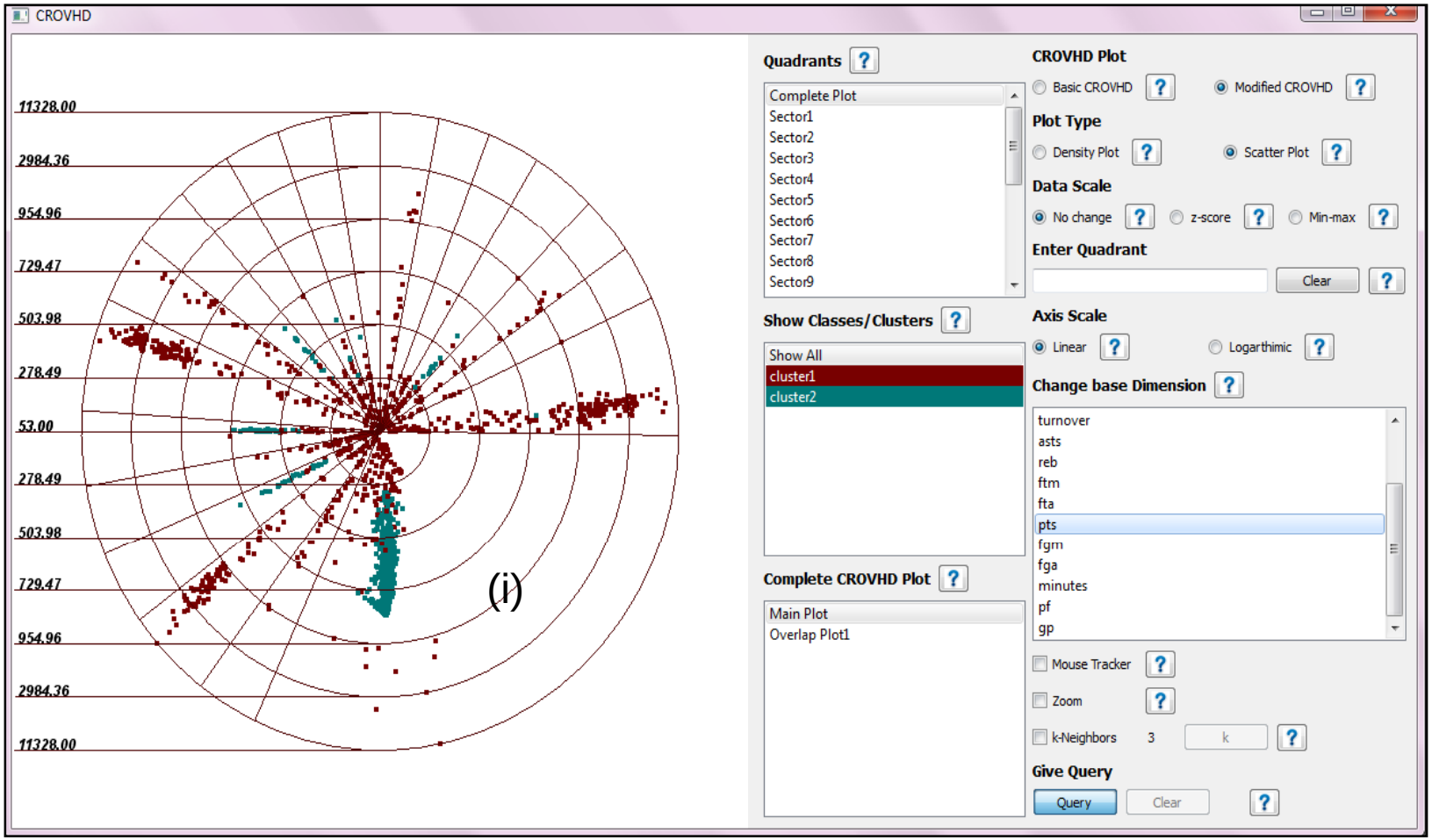


Figure 24: Scatter plot of 15d NBA dataset with 2055 points. This dataset has 2 clusters

Visualizing Nearest Neighbours

- It focuses on representing the data distribution in d-dimensional space on the surface area of a cone.
- 3-d conic visualization explicitly shows neighbours across quadrants, and helps users to comprehend nearest neighbours to perform further analytics.

Example

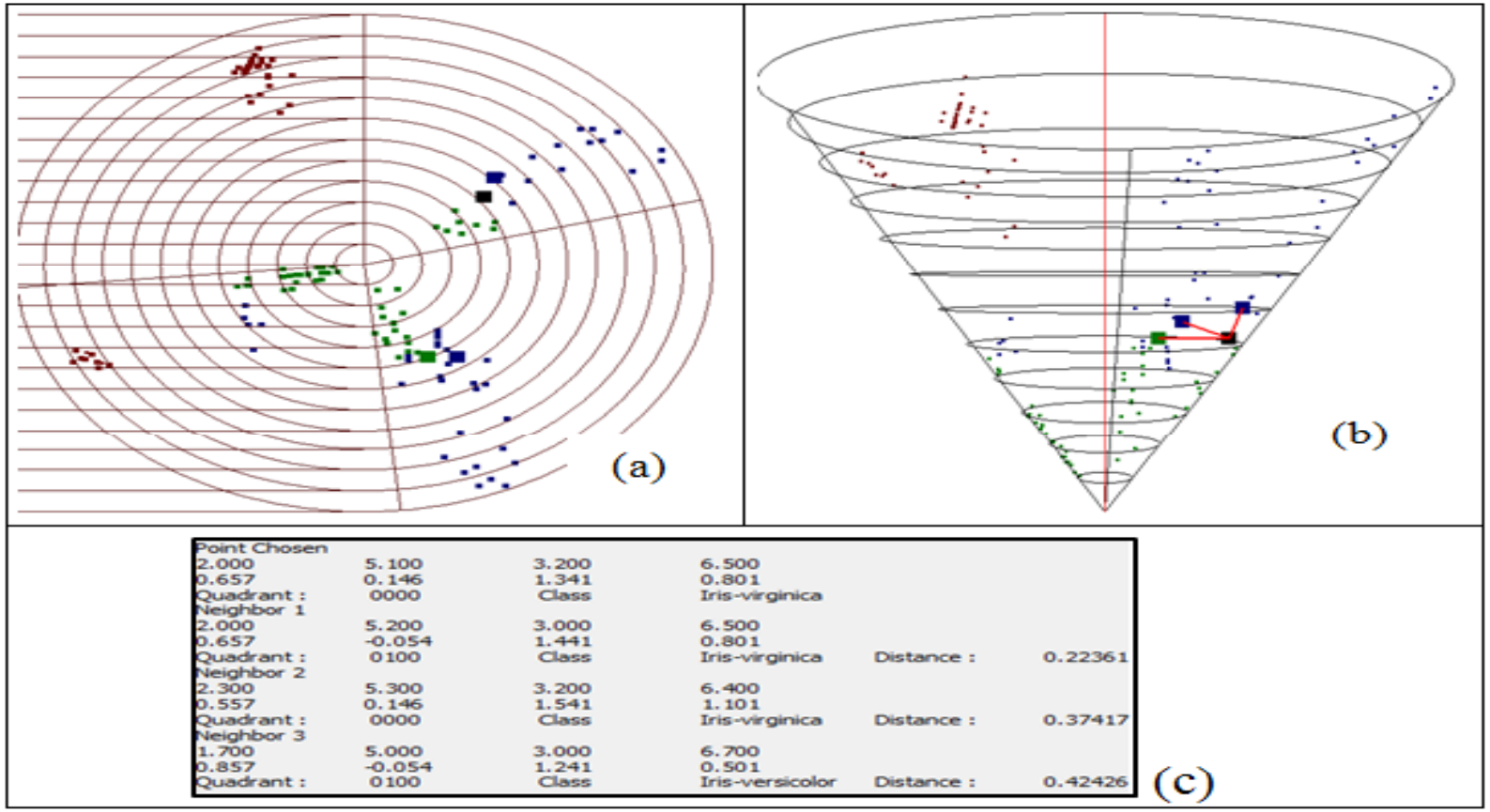
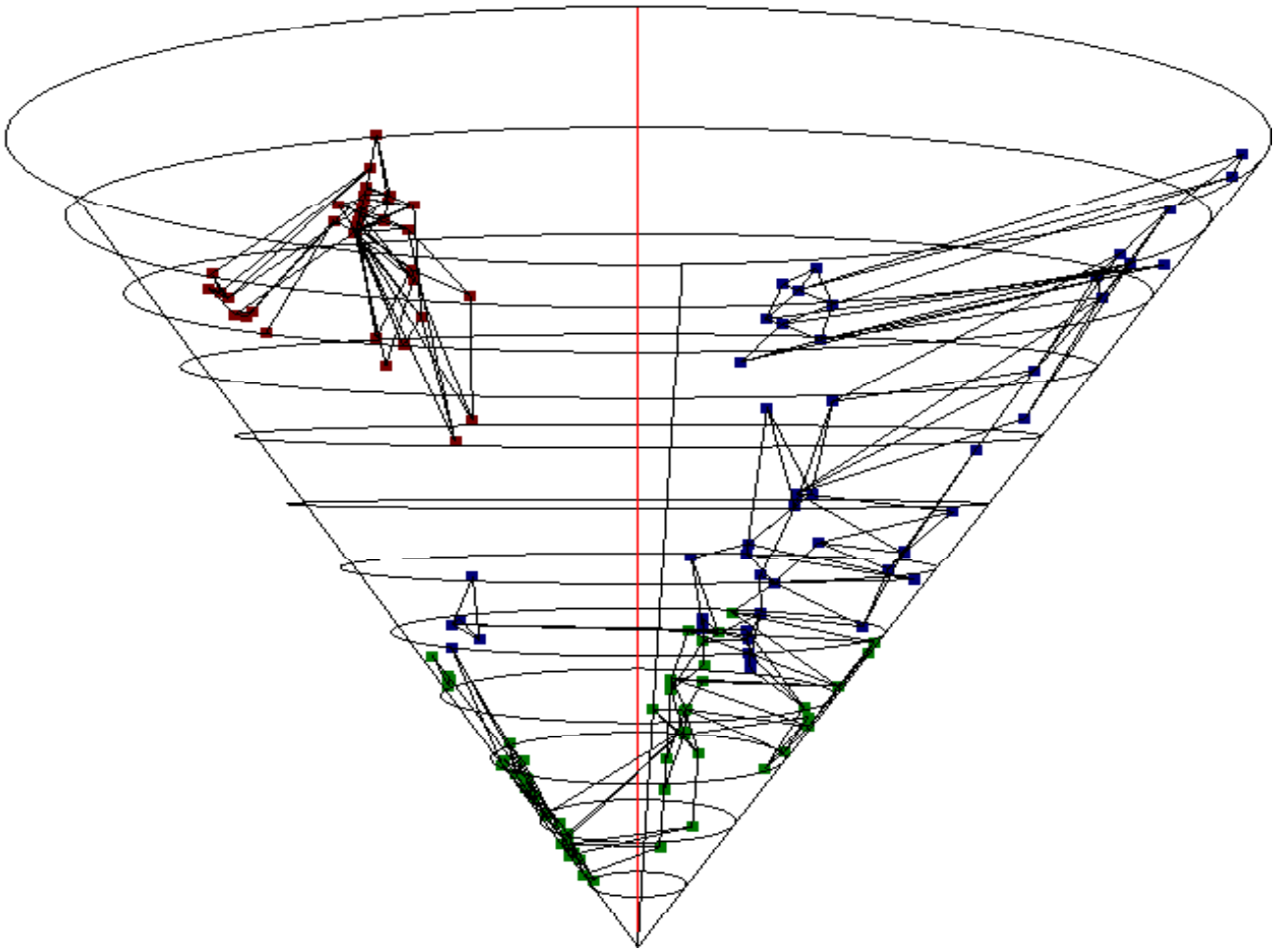


Figure 29 (a) 2D visualization of Iris dataset (4 dimensions and 150 points). (b) 2D visualization is converted into Cone visualization. The selected point (in black) and its neighbours are highlighted. (c) Information about the nearest neighbours.

Example – k-neighbour graph



Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- CROVDH
- Related Work
- Summary
- Open Problems

Related Work

- Parallel Coordinates [Inselberg 1985]
- VISA provides subspace overlap [Assent et al 2007]
- Best fit spheres or ellipsoids at high dimensions [Fitzgibbon, et al 1999, Calafiore 2002]
- Illustrative parallel coordinates [McDonnell & Muellller 2008]
- All 2-d subspaces scatter plots

Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- CROVDH
- Related Work
- Summary
- Open Problems

Summary

- Subspace overlaps in high dimensions - HEIDI
- Different aspects of HEIDI
- Shape and Structure of clusters – BEADS & PEARLS
- High Dimensional Scatter Plots - CROVDH

Outline

- Motivation and Applications
- Problems
- Heidi
- Beads
- CROVDH
- Related Work
- Summary
- Open Problems

Open Problems

- Ordering of points in Heidi
- Tight fit of shapes – composition of shapes – extending to 3d shapes
- Exploration with navigation in Beads and Heidi
- Explorative analysis and analytics from CROVDH
- Time and space efficiency
- Integrated visualization tool kit for R^d data

Take away!

- Subtle work
- Fun with visualization
- Vast open areas to work in
- Dashboards for visual analytics
- Domain specific vertical solutions
- Deep mathematical problems – shape fitting – multiple loss-less visuals