# Challenges in High Dimensional Data Visualization

December 21st 2011, COMAD – 2011

**Kamalakar Karlapalem**

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

**Centre for Data Engineering**

International Institute of Information Technology

Hyderabad, India

kamal@iiit.ac.in

# Outline

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

# 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!

# Motivation

» Seeing helps understanding

» Large data – cannot see completely!

» Dimensions a bigger problem – 4-d and higher

- Validate classification and clustering results

# 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

# 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

# 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

# Applications – Some Requirements

» Across all Subspaces proximity of points

» Shape and size of clusters

## Applications – Some Requirements

» Across all Subspaces proximity of points

» Shape and size of clusters

» Spread of data across the canvas

## 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

# Outline

» Motivation and Applications

» Problems

» Heidi

» Beads

» CROVDH

» Related Work

» Summary

» Open Problems

# Some Problems

» Can we find how clusters in high dimensional data overlap across various subspaces?

- HEIDI

# 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?
  - BEADS

# 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?
  - BEADS

» Can we present high dimensional data as a scatter plot?
  - CROVDH

# 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?
  - BEADS

» Can we present high dimensional data as a scatter plot?
  - CROVDH

» Useful for
  - Understanding and interpreting data
  - Clustering
  - Classification
  - Image pattern based index

# Outline

» Motivation and Applications

» Problems

» Heidi

» Beads

» CROVDH

» Related Work

» Summa ry

» Open Problems

# Heidi – Visual Relationship Matrix

» D = {$x_1$, $x_2$, …, $x_n$} n- points, d – dimensional

# Heidi – Visual Relationship Matrix

» $D = \{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional

» Construct a $n \times n$ matrix where

- Element (i,j) is a bit vector

# Heidi – Visual Relationship Matrix

» $D = \{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional

» Construct a $n \times n$ matrix where

- Element (i,j) is a bit vector
- Semantics of each bit in bit vector can be user specified

# Heidi – Visual Relationship Matrix

» $D = \{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional

» Construct a $n \times n$ 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

# Heidi – Visual Relationship Matrix

» $D = \{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional

» Construct a $n \times n$ 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

# Heidi – Visual Relationship Matrix

» D = {$x_1$, $x_2$, …, $x_n$} 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 $n \times n$ 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

# Heidi – specific case – Nearest Neighbors

» D = $\{x_1, x_2, \ldots, x_n\}$ n- points, d – dimensional

» Construct a $n \times n$ 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 $2^d - 1$

# Heidi – specific case – Nearest Neighbors

» D = {$x_1$, $x_2$, …, $x_n$} n- points, d – dimensional
» Construct a $n \times n$ 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 $n \times n \times [(2^d - 1)$ 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**

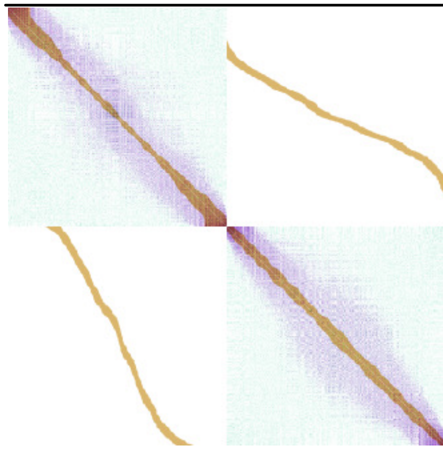
0,1,2,3

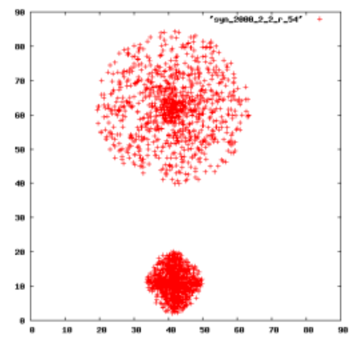
0,1,2 0,1,3 0,2,3 1,2,3

0,1 0,2 0,3 1,2 1,3 2,3

0 1 2 3

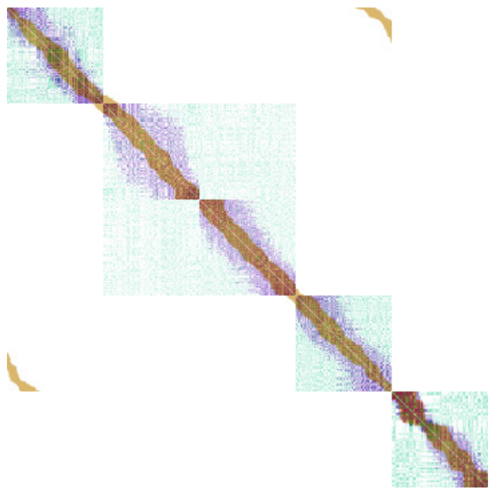
# Examples

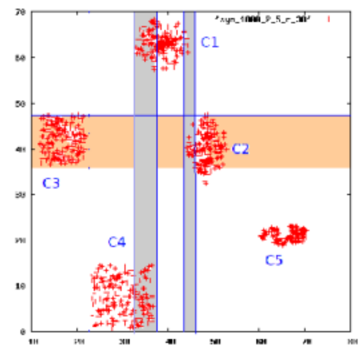





X – brown; Y – skyblue; {X,Y} - violet

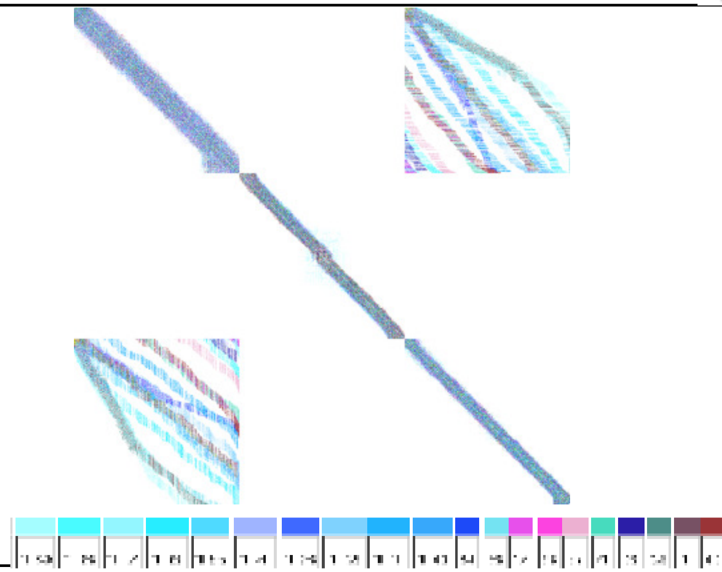
# Examples



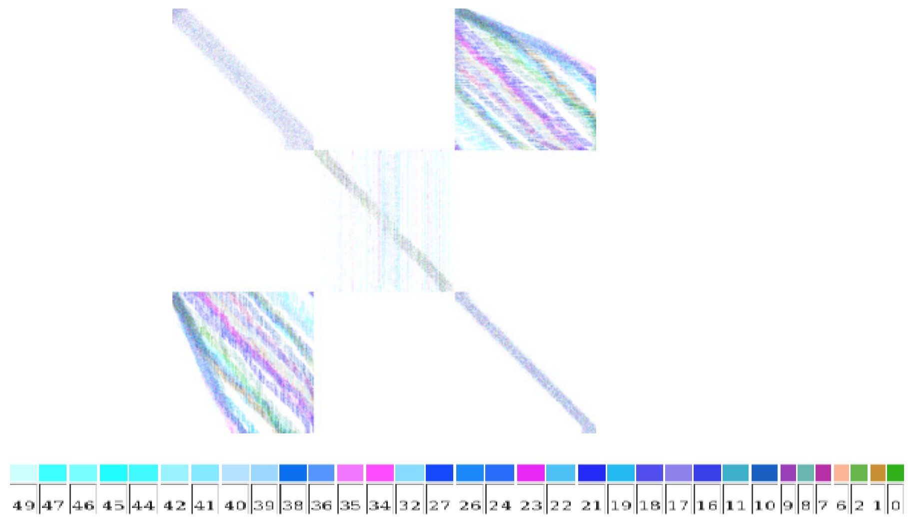



X – brown; Y – skyblue; {X,Y} - violet

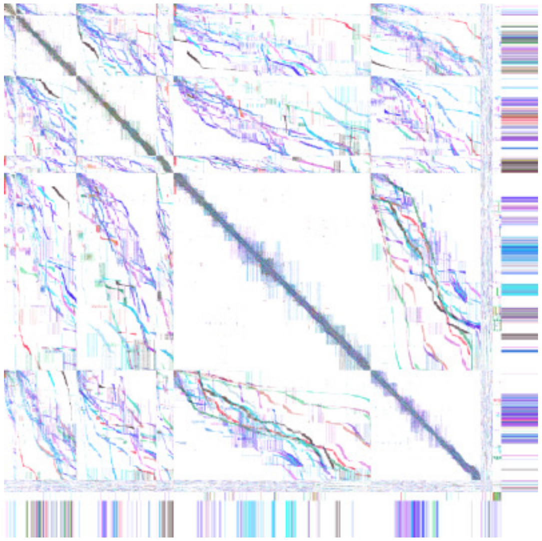
# Examples: Composite Heidi – 20d



# Examples: Composite Heidi=50d



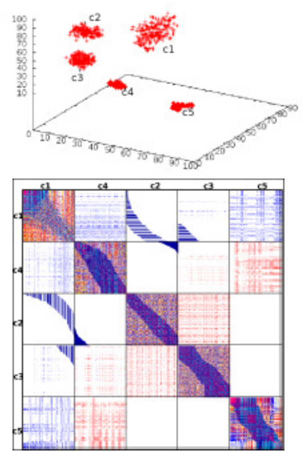
# Real-estate Property Listings

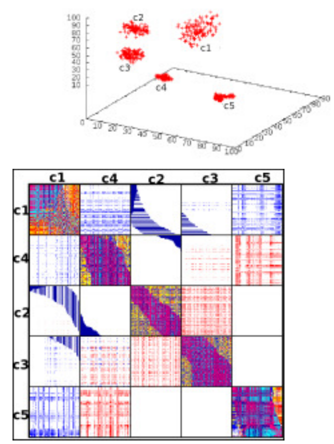


## Heidi Matrix - Issues

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

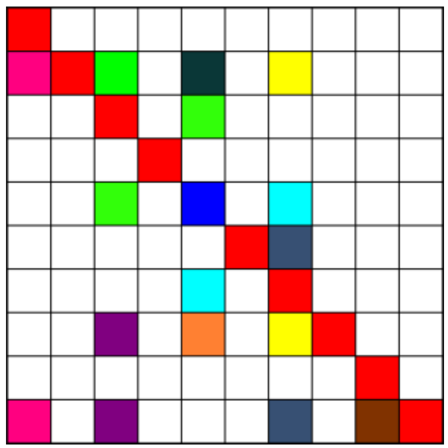
# Representative Heidi Images



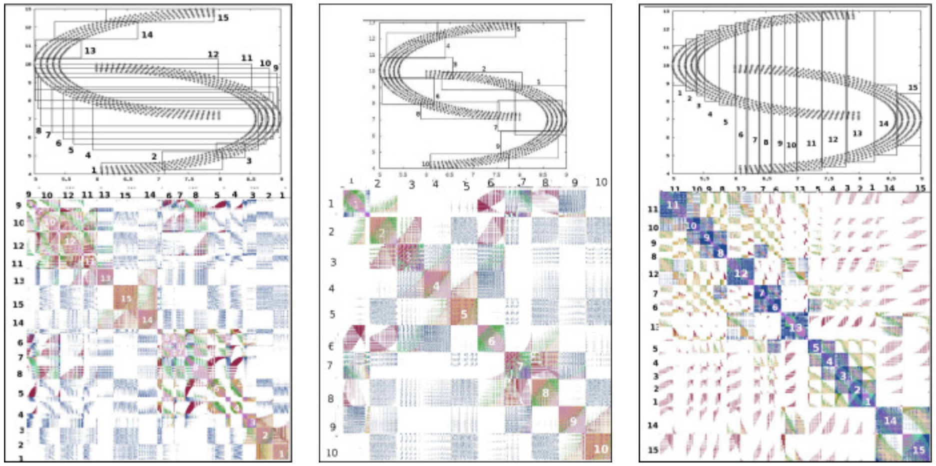

| Color | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Set of subspaces** | None | { 2 } | { 1 } | { 1 }{ 2 } | { 0 } | { 0 }{ 2 } | { 0 }{ 1 } | { 0 }{ 1 }{ 2 } |

# 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 } |

# Application – index visualization



R-tree, R-tree quadratic splitting, R*-tree

# 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

# 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

# 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

# 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 – 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 shapes

# 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

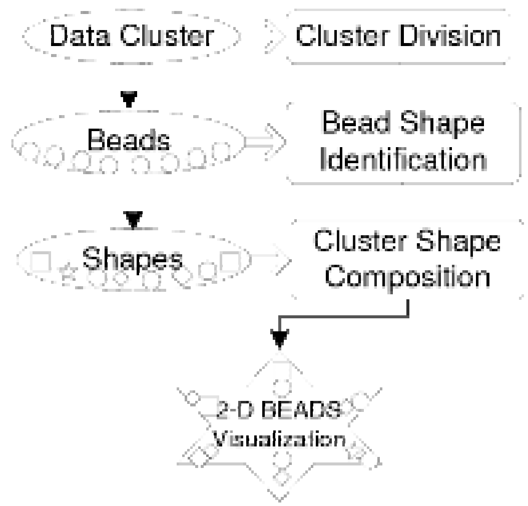
» Partition points into subsets of points

» Each subset forms a bead

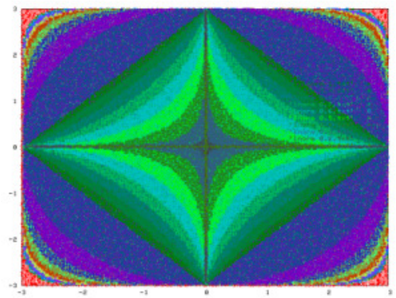
» Beads are mapped to well-specified 2-d shapes

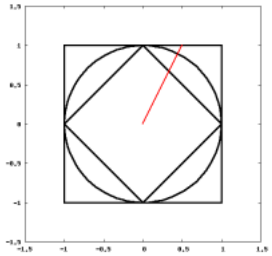
» Beads are placed in canvas to visually represent shape and size of cluster – **a necklace**

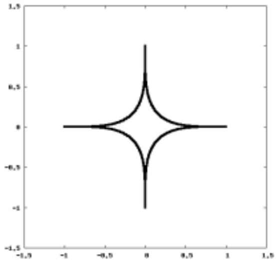
# Beads - Approach



# Basis for Beads





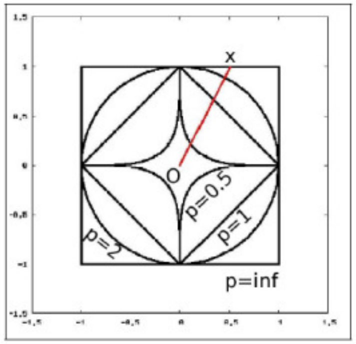


# Beads – shape and size

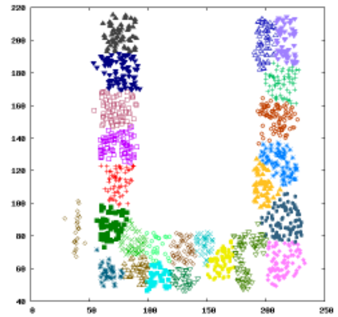
- P = set of distinct *p* values for $L_p$ norm
- Aim: Identify '*p*' and radius '$r_p$' that covers the bead tightly
- *Two approaches*

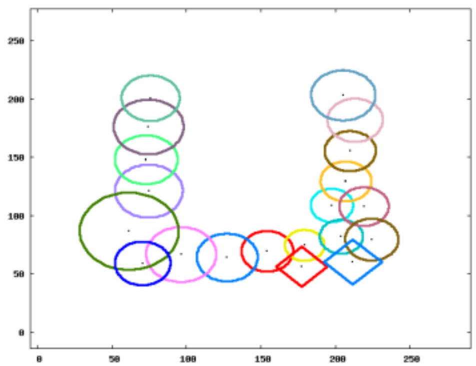
1. *Iterate from p by considering distances between centroid and furthest point using $L_p$, select the p which has the smallest distance.*
2. *Find the sum of distances among all pairs of points using $L_p$, and select the p that has smallest sum of distances*

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

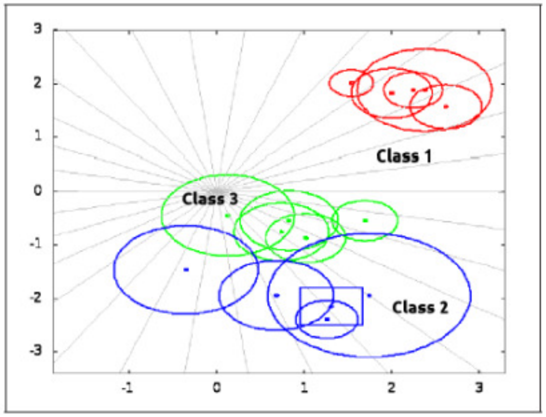


# Examples

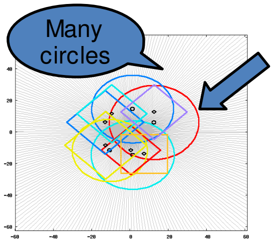




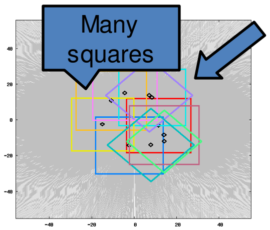
# Example – Iris Data Set



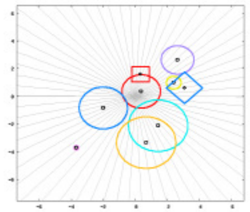
# More results




**10-D Hyper-sphere**




**10-D Hyper-cube**



**5-D NBA Player Data**

# Outline

- » Motivation and Applications
- » Problems
- » Heidi
- » Beads
- » 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

## 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

## 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 $(r, \theta)$ coordinates

- R is based on distance from centroid to point
- $\theta$ is based on quadrant and the relative angle within quadrant from some base axis

# CROVDH – Concentric Rings of Visualization for high dimensional data

» Given a data set $x_1$, $x_2$, …, $x_n$ d-dimensional data

» Determine a scatter plot visualization

» Spilt the 2-d space into $2^d$ quadrants

» Map each $x_i$ to $(r, \theta)$ coordinates

- R is based on distance from centroid to point
- $\theta$ is based on quadrant and the relative angle within quadrant from some base axis

» Divide regions of 2-d space as concentric circles

## 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 $(r, \theta)$ coordinates

- R is based on distance from centroid to point
- $\theta$ is based on quadrant and the relative angle within quadrant from some base axis

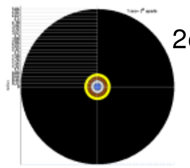
» Divide regions of 2-d space as concentric circles

» Give region colors based on relative density

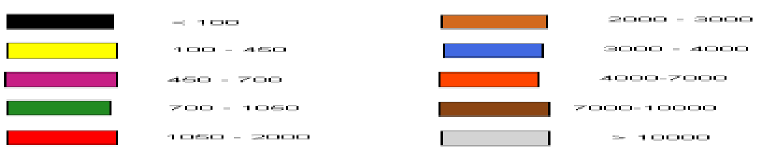
» Can also show actual points

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

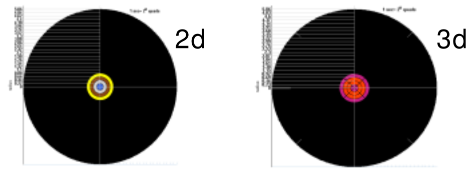
# Uniform 100,000 [0,1] points dimensions increasing

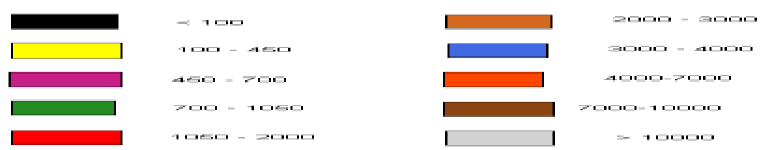


2d

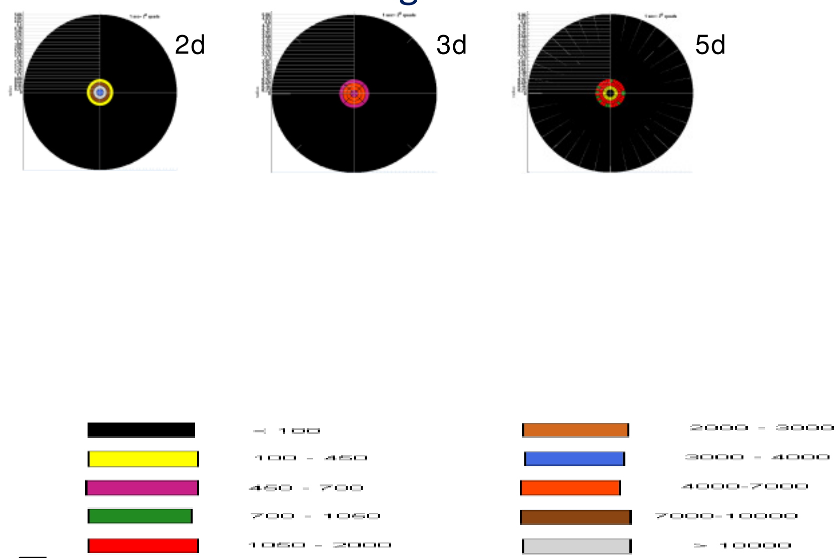


# Uniform 100,000 [0,1] points dimensions increasing

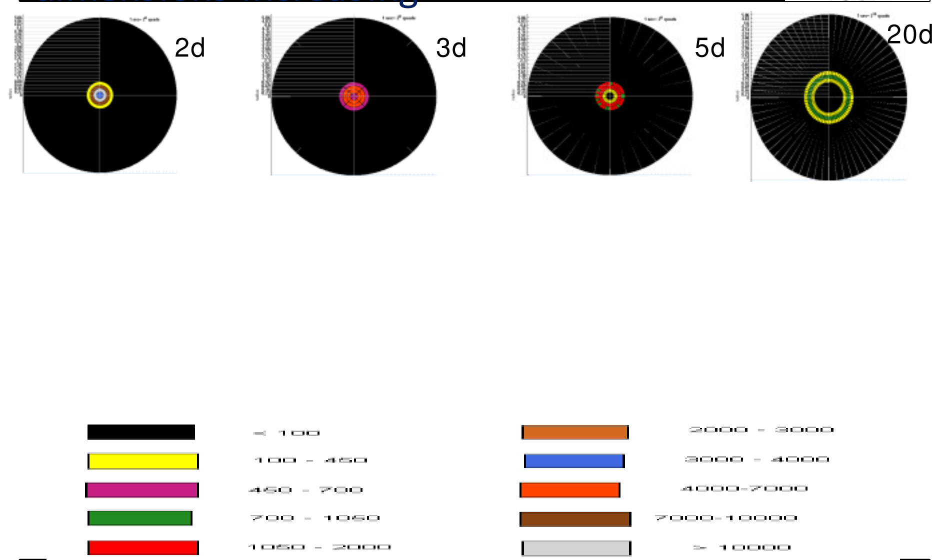


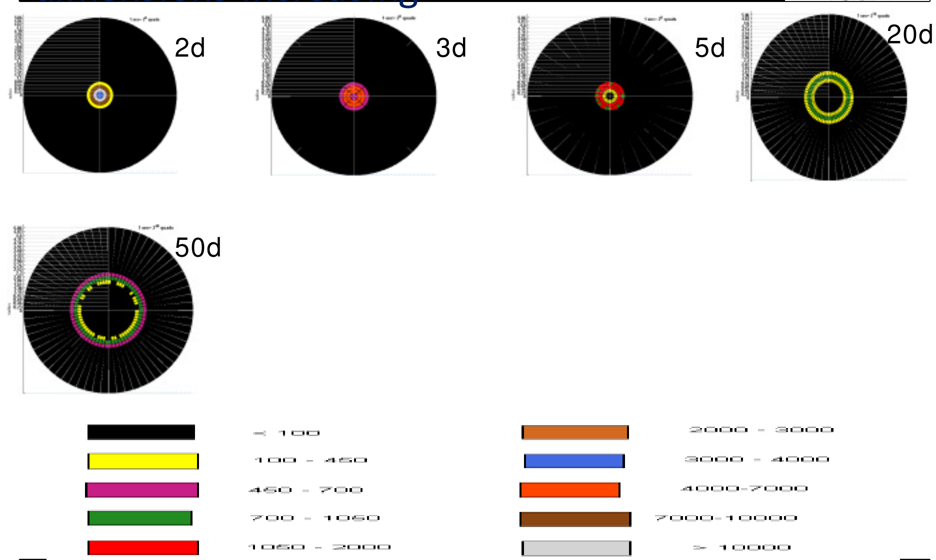
# Uniform 100,000 [0,1] points dimensions increasing



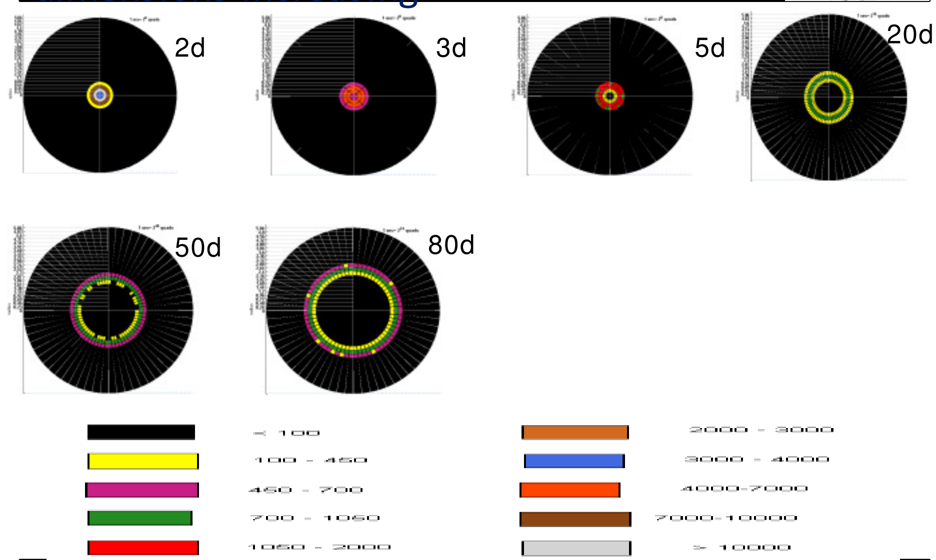
# Uniform 100,000 [0,1] points dimensions increasing



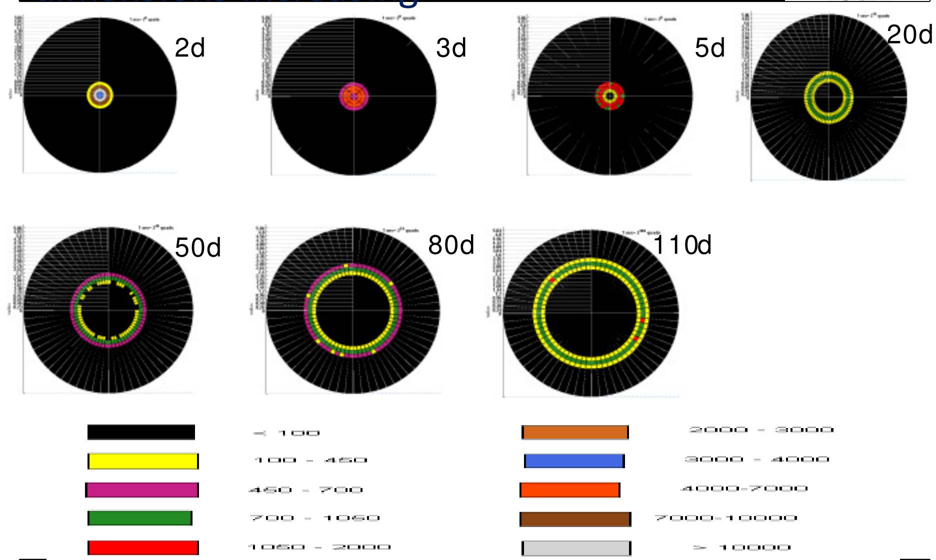
# Uniform 100,000 [0,1] points dimensions increasing



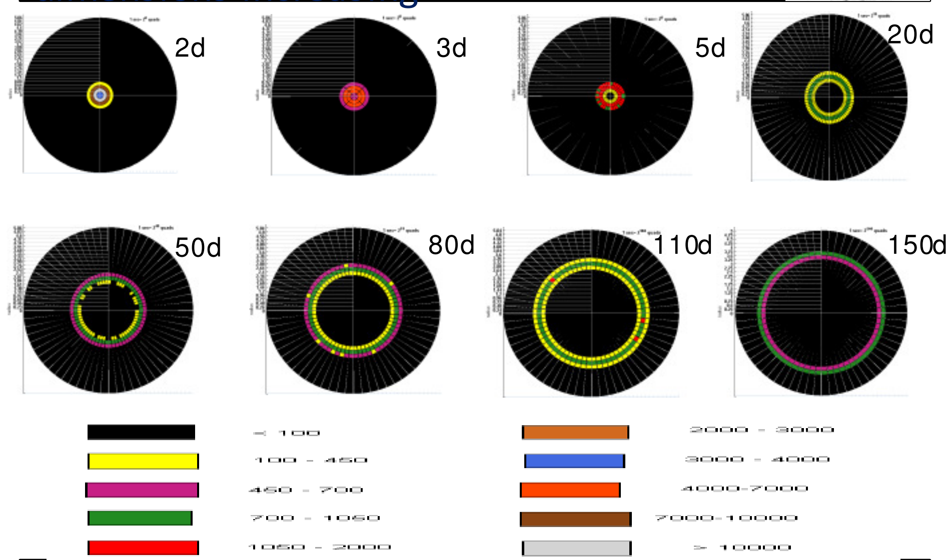
# Uniform 100,000 [0,1] points dimensions increasing



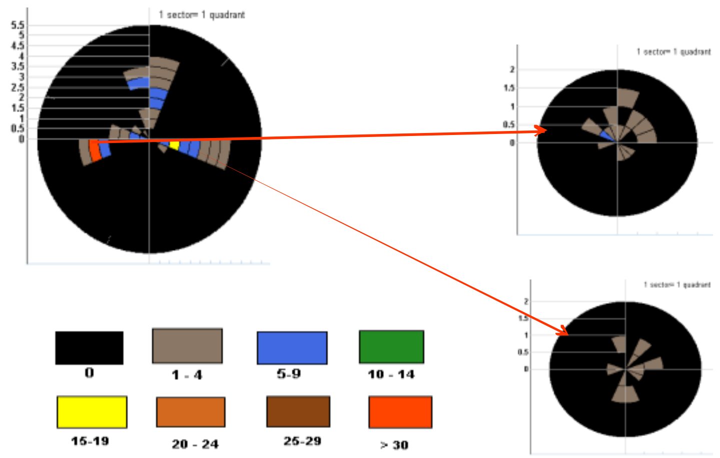
# Uniform 100,000 [0,1] points dimensions increasing



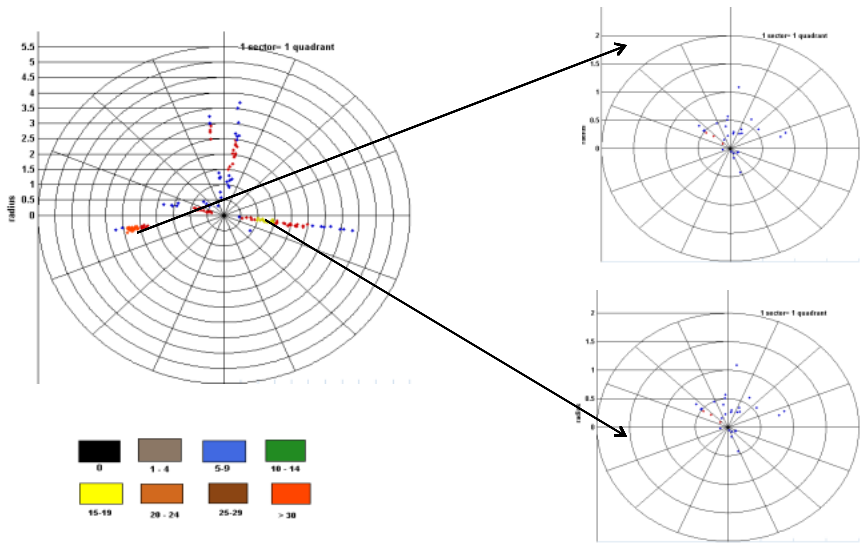
# Uniform 100,000 [0,1] points dimensions increasing



## CROVDH Visualization of IRIS data set



# CROVDH Visualization of IRIS data set



# 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

» Applications of Heidi

» Shape and Structure of clusters – Beads

» 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