

Large-Scale Multidimensional Data Visualization: A Web Service for Data Mining

**Gintautas Dzemyda, Virginijus Marcinkevičius,
Viktor Medvedev**

gintautas.dzemyda@mii.vu.lt



Vilnius University
Institute of Mathematics and Informatics,
Lithuania





Overview

Interaction between humans and machines is one of the areas in computer science that has evolved a lot the last years.

Here we present an approach and architecture of

Web service-based data mining

oriented to

the multidimensional data visualization.

We combine the well-known visualization methods with modern computing possibilities including Web-based architectures and parallel computing.





Visualization Problem

Real data of natural and social sciences are often
high-dimensional

So, it is very difficult to understand these data and extract patterns.

One way for such understanding is to make visual insight into the analyzed data set.



Visualization Problem

Visualization of multidimensional data is a complicated problem followed by extensive researches because it allows to the investigator

- to observe data clusters
- to estimate the inter-nearness between the multidimensional points
- to make proper decisions

Let us have m multidimensional (n -dimensional) vectors

$$X_1, X_2, \dots, X_m \in R^n \quad X_i = (x_{i1}, x_{i2}, \dots, x_{in}), i = \overline{1, m}$$

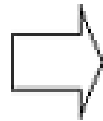
The problem is to get a projection of this set of vectors on the visually perceived low dimensional space R^2 or R^3 .

Denote projections on the plane by $Y_i = (y_{i1}, y_{i2}), i = \overline{1, m}$

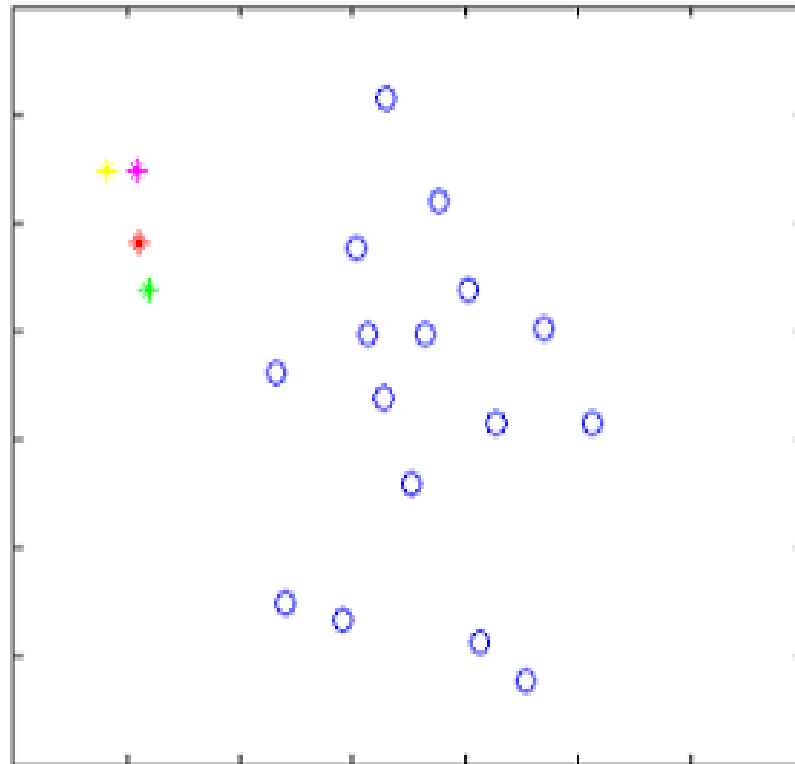


The human being can comprehend visual information more quickly than textual one

0,57	0,11	0,47	0,54	0,84	0,53
0,65	0,14	0,53	0,55	0,89	0,54
0,81	0,13	0,49	0,52	0,92	0,57
0,44	0,16	0,45	0,42	0,99	0,45
0,55	0,11	0,53	0,52	0,92	0,55
0,44	0,16	0,43	0,41	0,98	0,43
0,49	0,12	0,56	0,46	0,82	0,48
0,41	0,17	0,42	0,43	0,98	0,42
0,54	0,12	0,59	0,46	0,85	0,48
0,89	0,16	0,63	0,56	0,91	0,49
0,86	0,16	0,51	0,48	0,87	0,55
0,70	0,13	0,52	0,48	0,81	0,52
0,65	0,14	0,63	0,52	0,89	0,49
0,59	0,11	0,51	0,57	0,89	0,52
0,53	0,11	0,52	0,57	0,89	0,50
0,52	0,13	0,51	0,81	0,92	0,51
0,58	0,12	0,55	0,61	0,95	0,52
0,45	0,10	0,52	0,53	0,92	0,51
0,42	0,17	0,41	0,42	0,98	0,40
0,53	0,11	0,42	0,57	0,91	0,57



2-dimensional vectors are obtained from 6-dimensional vectors;
they are projected onto the plane





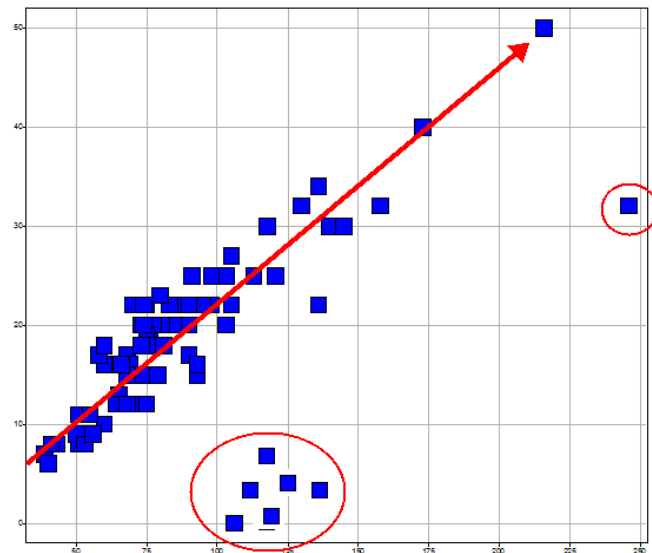
Visualization Problem

The goal of the projection (visualization) methods is to represent the input data items in a lower-dimensional space so that certain properties of the structure of the data set were preserved as faithfully as possible.

$$R^n \rightarrow (\text{projection methods}) \rightarrow R^2$$

R^n

	A	B	C	D	E	F	G	H	I	J
1	1,11091	7,04704	2,68999	3,51433	6,24851	-7,70285	3,98986	-15,0384	5,9192	5,87187
2	-5,15074	1,14707	1,49299	2,55285	-2,27442	-14,6506	7,86902	-7,36946	9,63875	7,11384
3	1,91532	-4,58991	-1,90508	4,07452	-11,9971	-7,92136	9,69161	0,222057	10,8834	5,87123
4	0,020048	0,630172	1,08796	2,26984	-4,79242	-8,15183	4,29621	-0,46226	7,42289	6,15558
5	-0,89494	8,11653	-6,95531	6,61701	-8,21592	-11,2322	7,56654	-8,13229	10,6122	5,83966
6	-0,10323	-10,4471	-3,17563	5,31002	0,14175	-9,40686	9,52215	1,64615	-1,88292	6,64164
7	1,72396	0,97497	5,85485	5,75813	1,33302	-4,35716	3,95199	-1,21654	4,97955	4,7983
8	4,52856	11,1303	-0,52264	3,58449	-5,35348	-9,28854	9,6102	-3,04153	5,72042	5,28749
9	-0,83436	1,05809	7,03569	2,13677	-4,65407	-8,54866	6,28365	-2,71864	17,3228	5,42663
10	2,58882	8,18563	4,12082	5,86436	-3,93221	-0,99236	6,98705	-2,16161	3,14953	5,16401
11	-1,00736	0,836051	-0,27273	5,757	-8,00482	-2,68832	4,08388	2,19967	4,11293	6,58779
12	-2,78432	-0,36631	-5,02407	2,12856	-5,31326	-8,53505	4,3084	-4,39105	5,59385	6,73681
13	6,34916	4,67636	3,14195	2,93509	-4,95993	-11,4985	4,62661	-7,7016	7,09844	7,54618
14	2,35358	9,6925	0,786942	5,20636	-1,26773	-11,0521	5,5869	-0,09746	10,3834	6,49727
15	-3,13531	-3,0484	3,19325	3,82446	1,29874	-6,93801	4,42466	-4,00488	15,2039	4,52903
16	0,192948	9,22675	1,49804	3,56301	-3,13276	-7,37238	11,2558	-1,01955	6,3389	5,51749
17	-6,14694	11,5993	5,03532	3,68279	2,66215	-12,0582	8,4349	-3,82024	-4,39669	8,02664
18	1,66801	-1,53136	2,95147	0,441002	-6,85443	-6,11669	7,83389	-8,50664	12,1942	6,17958
19	-2,42808	-2,35968	-0,9866	2,95973	-0,71374	-6,25115	7,01922	-4,05128	0,133749	7,84149
20	-3,23398	0,739948	1,61084	3,85228	-9,73604	-3,64236	11,4746	-11,9365	9,5091	5,65706
21	-0,02485	3,72569	-5,69587	1,96643	-8,95218	0,753423	8,5288	-6,39854	1,8114	7,20848
22	-1,7316	0,14579	-8,76263	0,305276	-0,84276	-13,4019	7,17862	-2,53552	1,32054	5,16366
23	-0,26349	2,65934	-0,38402	4,20786	-9,42951	-3,06817	6,16654	-7,08867	-3,23498	6,06282
24	0,935001	7,62704	-5,89728	0,043108	-8,97537	-17,8392	5,97077	-13,8924	6,5347	5,41031
25	-1,13148	-3,6351	3,4953	4,03109	11,0707	2,57909	6,92097	-4,71983	5,97176	5,41212
26	-0,42632	10,1015	-4,25269	2,9947	-8,99443	-5,9449	3,81592	-0,83499	17,2262	3,8634
27	1,67647	4,57368	2,22492	2,66839	-7,65192	-8,25385	6,32712	-2,59647	7,04368	6,75276
28	0,42202	5,57122	-3,12425	4,19927	-16,3641	-0,344	9,59166	-7,06135	11,1317	1,69255





Example of Multidimensional Data (Breast Cancer Data)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	C
X_1	5	1	1	1	2	1	3	1	1	b
X_2	5	4	4	5	7	10	3	2	1	b
X_3	3	1	1	1	2	2	3	1	1	b
X_4	6	8	8	1	3	4	3	7	1	b
X_5	4	1	1	3	2	1	3	1	1	b
X_6	1	1	1	1	2	10	3	1	1	b
X_7	2	1	2	1	2	1	3	1	1	b
X_8	2	1	1	1	2	1	1	1	5	b
X_9	4	2	1	1	2	1	2	1	1	b
...
X_{460}	8	10	10	8	7	10	9	7	1	m
X_{461}	5	3	3	3	2	3	4	4	1	m
X_{462}	8	7	5	10	7	9	5	5	4	m
X_{463}	7	4	6	4	6	1	4	3	1	m
X_{464}	10	7	7	6	4	10	4	1	2	m
X_{465}	7	3	2	10	5	10	5	4	4	m
X_{466}	10	5	5	3	6	7	7	10	1	m
X_{467}	5	2	3	4	2	7	3	6	1	m
...

*University of Wisconsin, Clinical
Sciences Center*

x_1 – clump thickness,
 x_2 – uniformity of cell size,
 x_3 – uniformity of cell shape,
 x_4 – marginal adhesion,
 x_5 – single epithelial cell size,
 x_6 – bare nuclei,
 x_7 – bland chromatin,
 x_8 – normal nucleoli,
 x_9 – mitoses,
C – class (**b**enign, **m**alignant)



Dimension Reduction Methods

There exist a lot of methods that can be used for reducing the dimensionality of data, and, particularly, for visualizing the n -dimensional vectors.

- Traditional methods
 - Multidimensional scaling
 - Sammon's projection
 - Principal components
 - Direct methods (Chernoff faces, Andrew's curves, star...)
 - Others
- Neural networks
 - Self-organizing map (SOM)
 - Feed-forward networks
- Combinations of traditional methods and neural networks
- Manifold learning methods (locally linear embedding (LLE), Laplacian Eigenmaps (LE), Isomap...)



Analysis of the Economic and Social Conditions of Central European Countries

Countries

Parameters

1	Hungary
2	Czech Republic
3	Lithuania
4	Latvia
5	Slovakia
6	Poland
7	Romania
8	Estonia
9	Bulgaria
10	Slovenia

x_1 - the infant mortality rate (deaths / 1000 live births)

x_2 - the Gross Domestic Product (GDP) per capita in US dollars obtained taking into account the purchasing power parity of the national currency but not the exchange rate

x_3 - the percentage of GDP developed in the industry and services (not in the agriculture)

x_4 - the export per capita in thousands of US dollars

x_5 - the number of telephones per capita

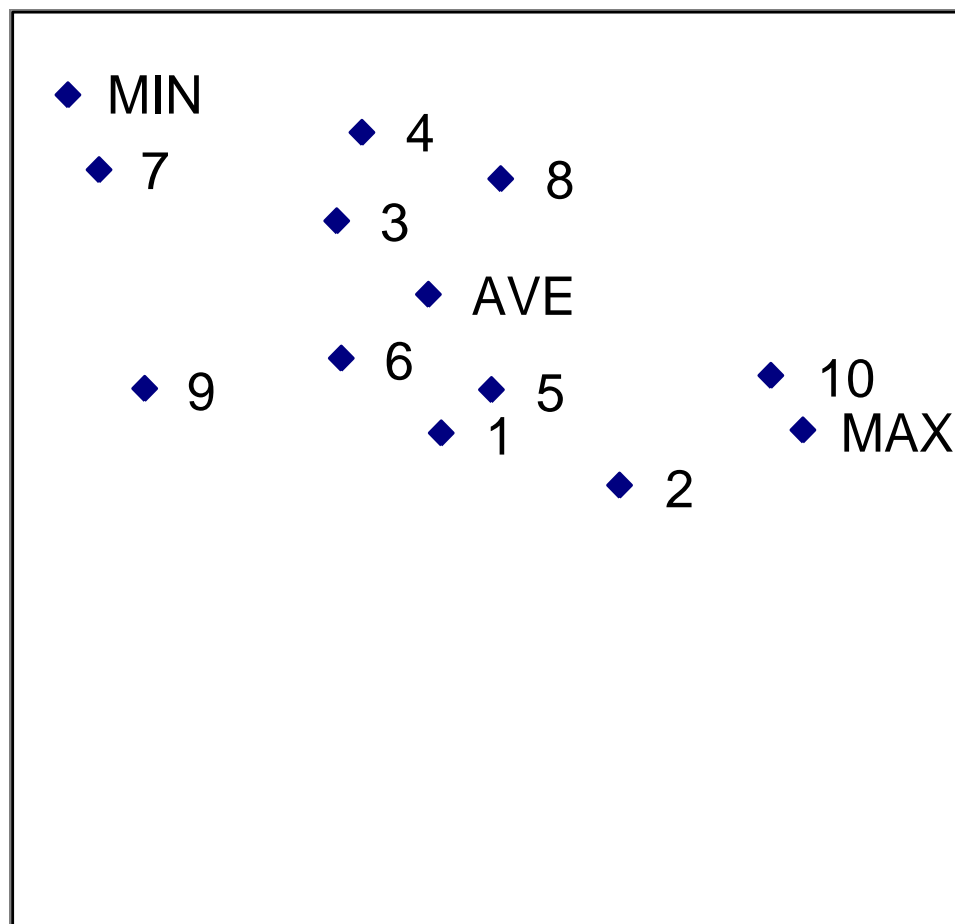
x_6 - the international aid in US dollars per capita



Projection of Countries from the View of Economic and Social Parameters Using MDS

Countries

1	Hungary
2	Czech Republic
3	Lithuania
4	Latvia
5	Slovakia
6	Poland
7	Romania
8	Estonia
9	Bulgaria
10	Slovenia

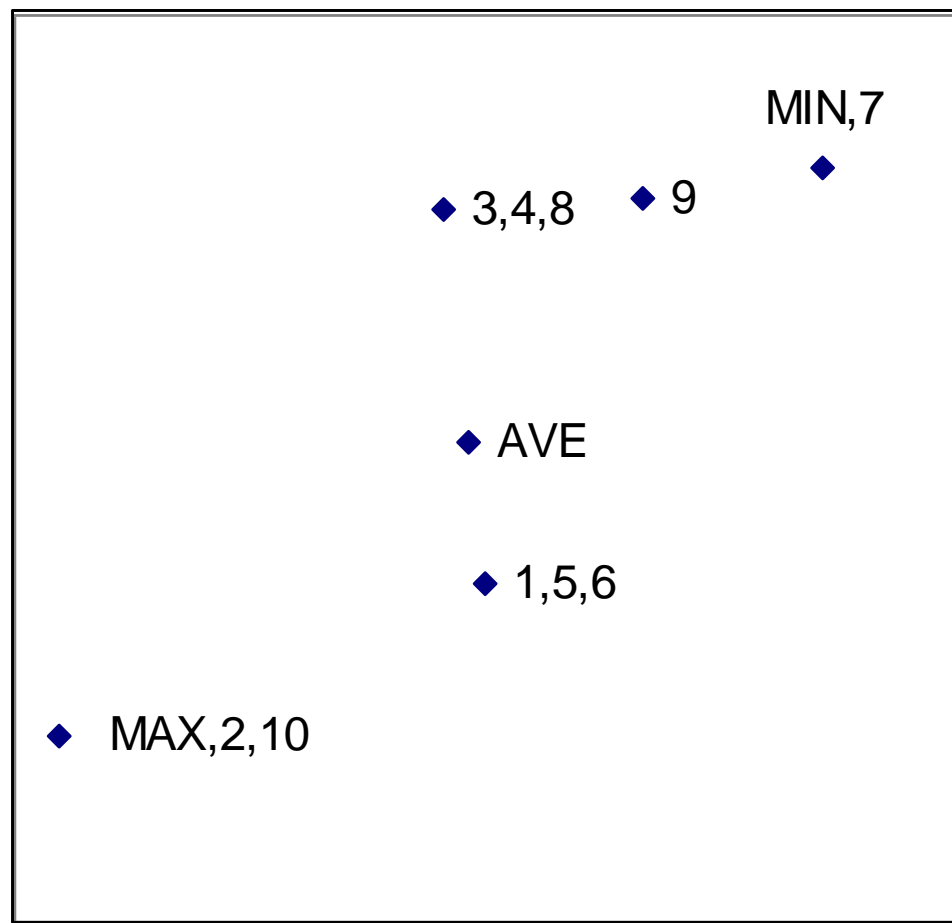




Projection of Countries from the View of Economic and Social Parameters Using SOM+MDS

Countries

1	Hungary
2	Czech Republic
3	Lithuania
4	Latvia
5	Slovakia
6	Poland
7	Romania
8	Estonia
9	Bulgaria
10	Slovenia





Analysis of Physiological Data

The purpose of analysis is to evaluate men's health state and their possibility of going in for sports.

The analysed physiological data set consists of three groups:

- (1) ischemic heart-diseased men (61 items),
- (2) healthy persons (not going in for sports) (110 items),
- (3) sportsmen (161 items).

Non-specific physiological features that are frequently used in clinical medicine and that describe the human functional state are as follows:

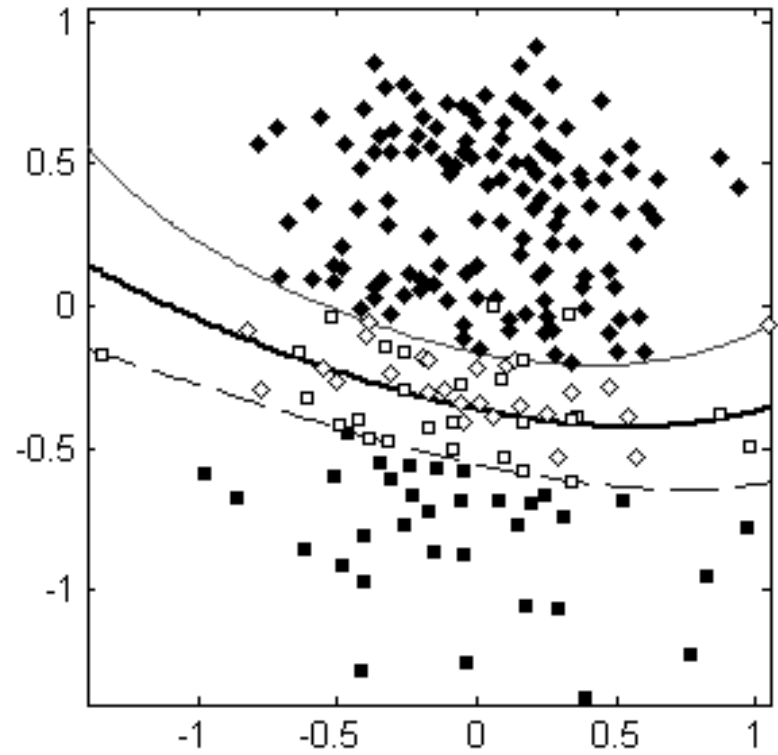
- heart rate (HR),
- interval in the electrocardiogram from point J to the end T of the wave (JT interval),
- systolic blood pressure (SBP),
- diastolic blood pressure (DBP),
- the ratios between some parameters $(SBP-DBP)/SBP$, JT/RR ($RR=60/HR$).



Integrating Classification Results into Visualization

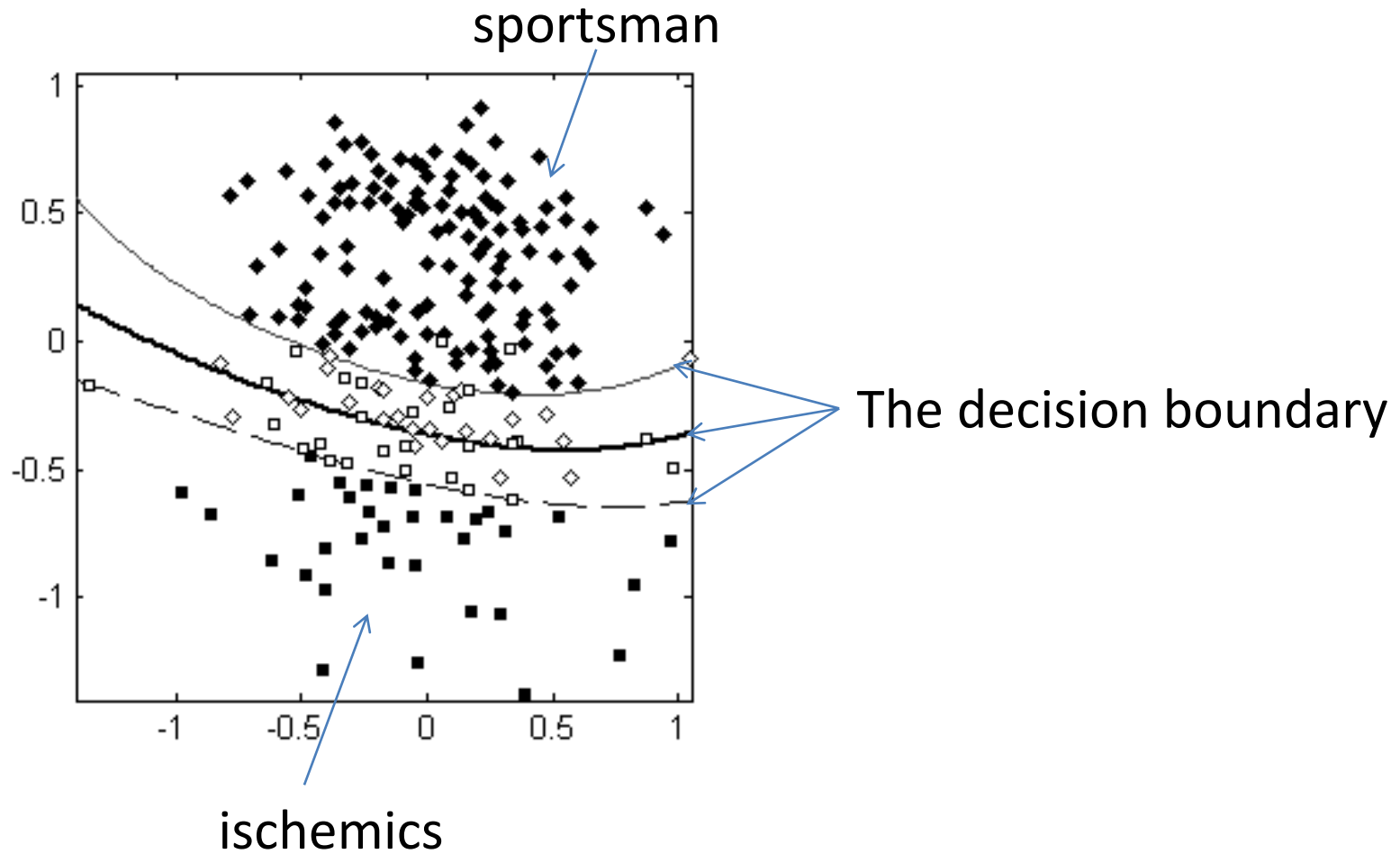
The projections of 17-dimensional data, the support vectors and the decision surfaces are presented:

- the points, corresponding to ischemics, are marked by filled squares;
- the points, corresponding to sportsmen, are marked by filled rhombi;
- the support vectors are marked by unfilled squares or rhombi (total 53);
- the bold line marks the decision surface,
- the light solid line marks the decision boundary of sportsmen,
- the dashed line marks the decision boundary of ischemics.



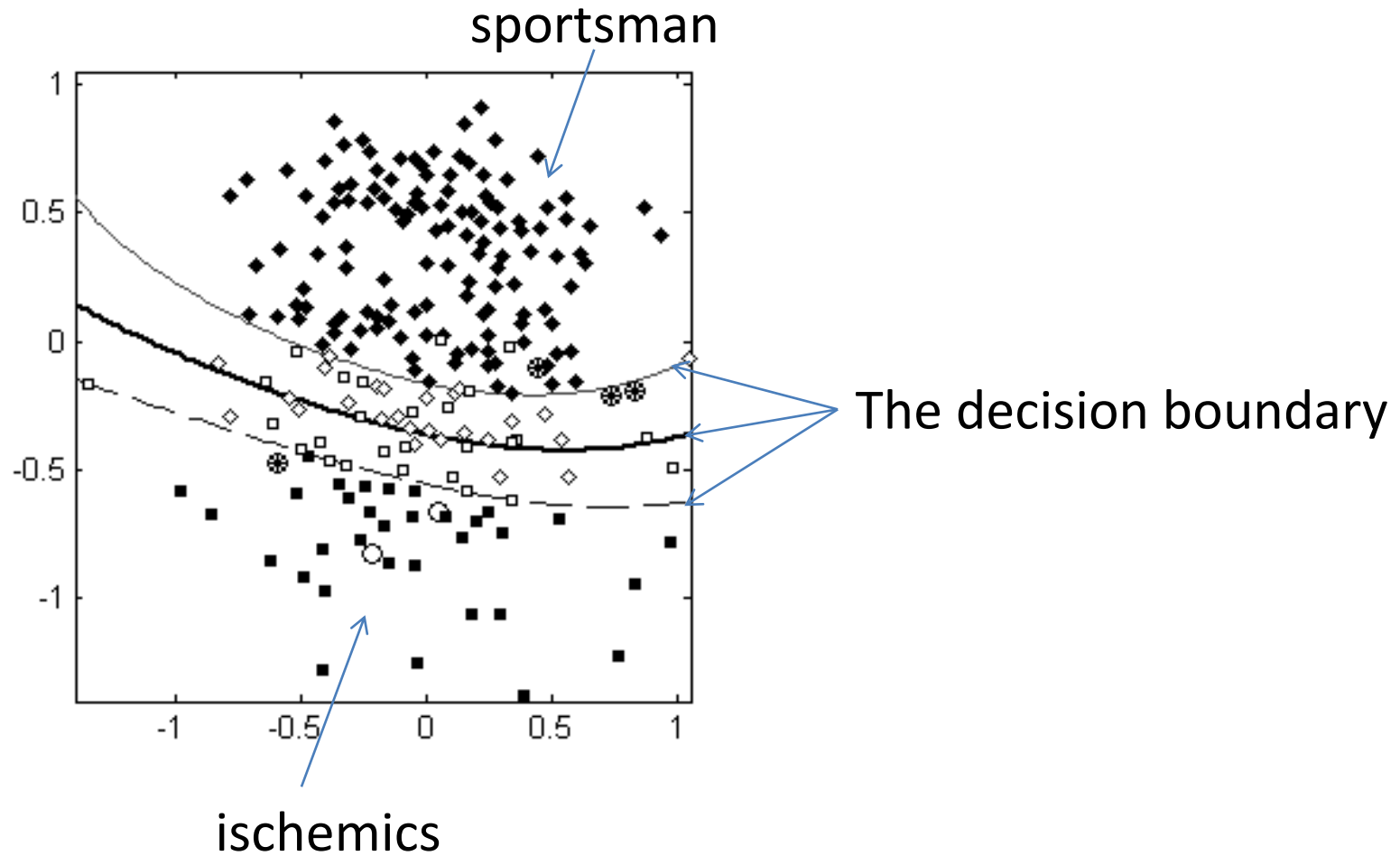


The Decision Boundary





Mapping of New Patients





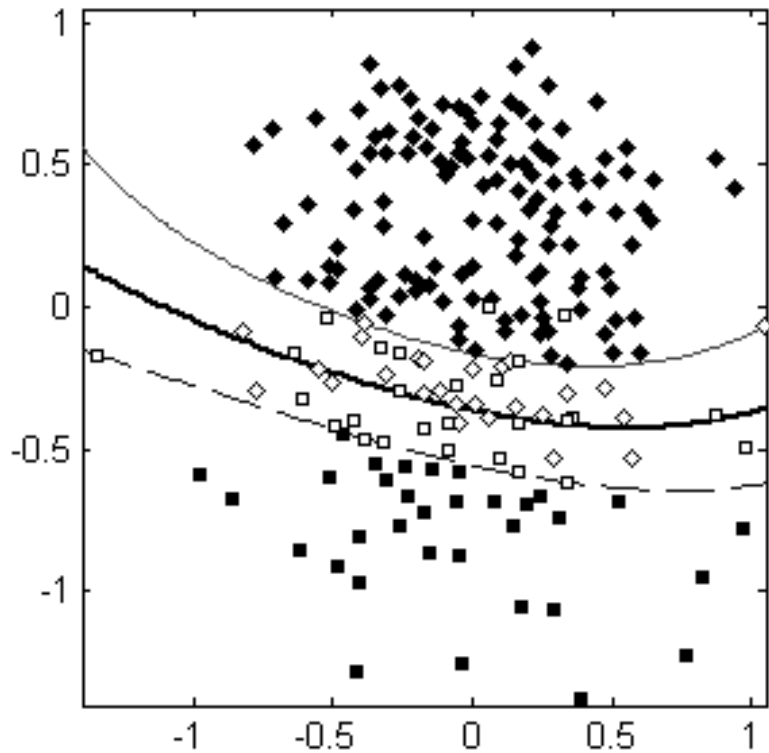
Visualization for the Self-Observation

- Physiological data analysis



Visualization for the Self-Observation

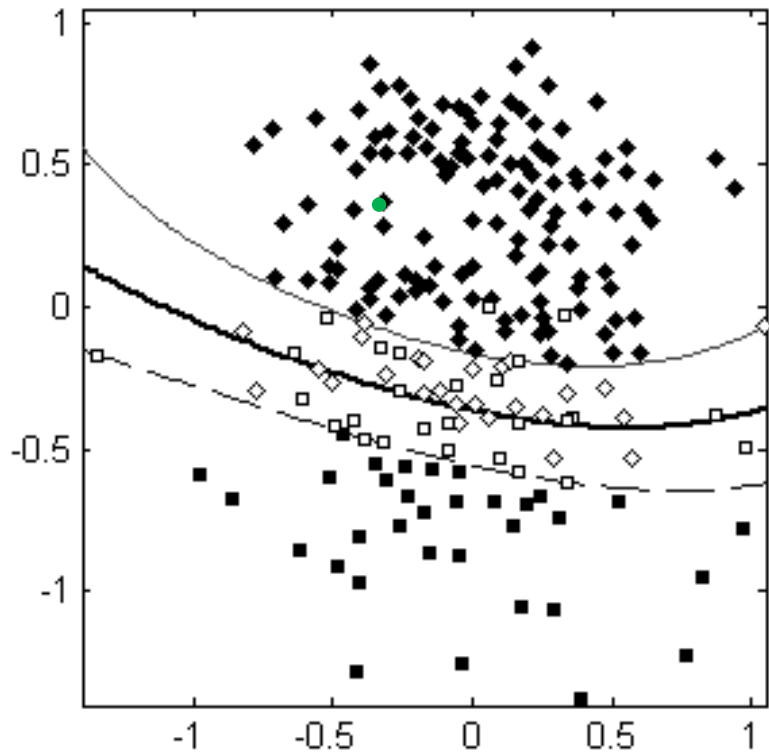
- Physiological data analysis





Visualization for the Self-Observation

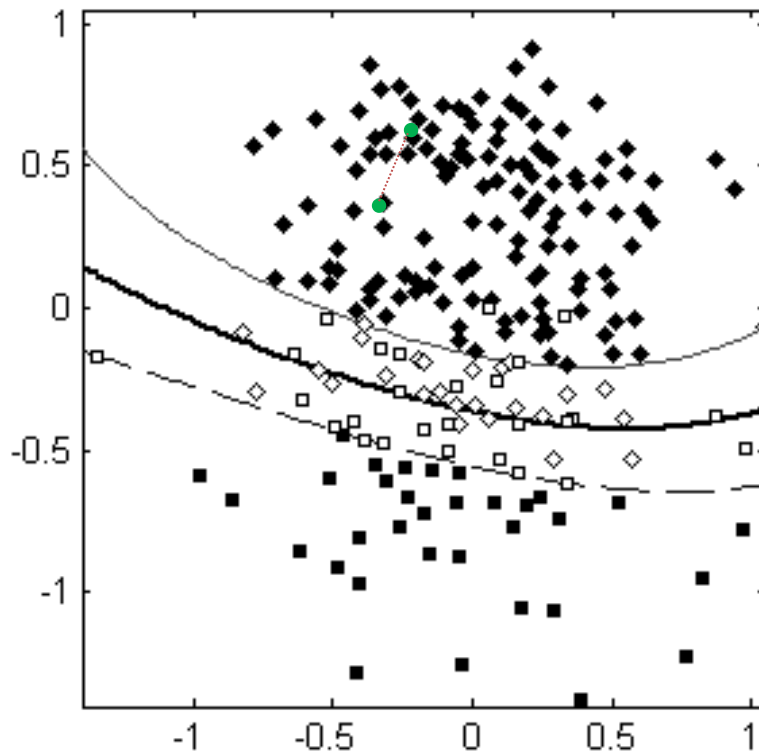
- Physiological data analysis





Visualization for the Self-Observation

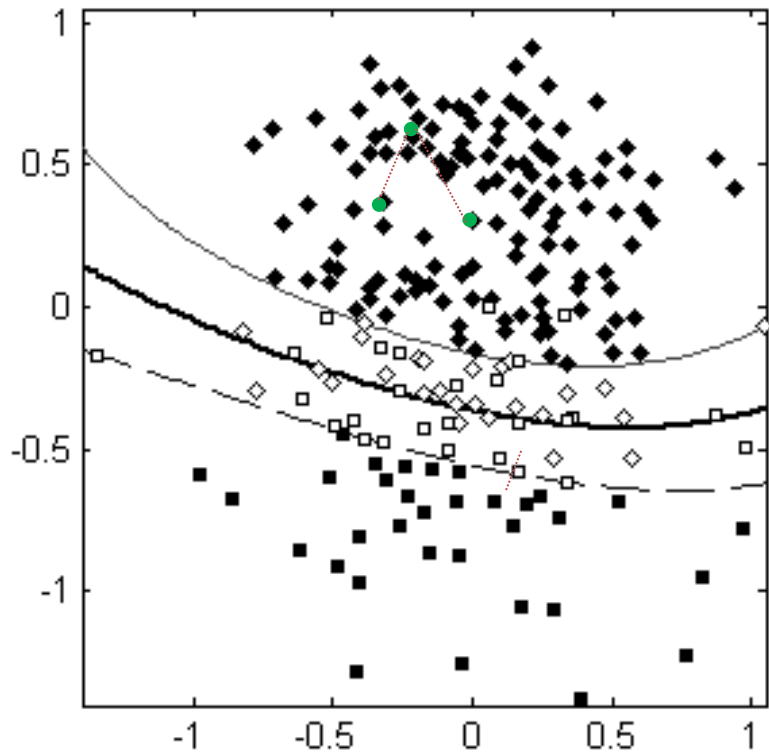
- Physiological data analysis





Visualization for the Self-Observation

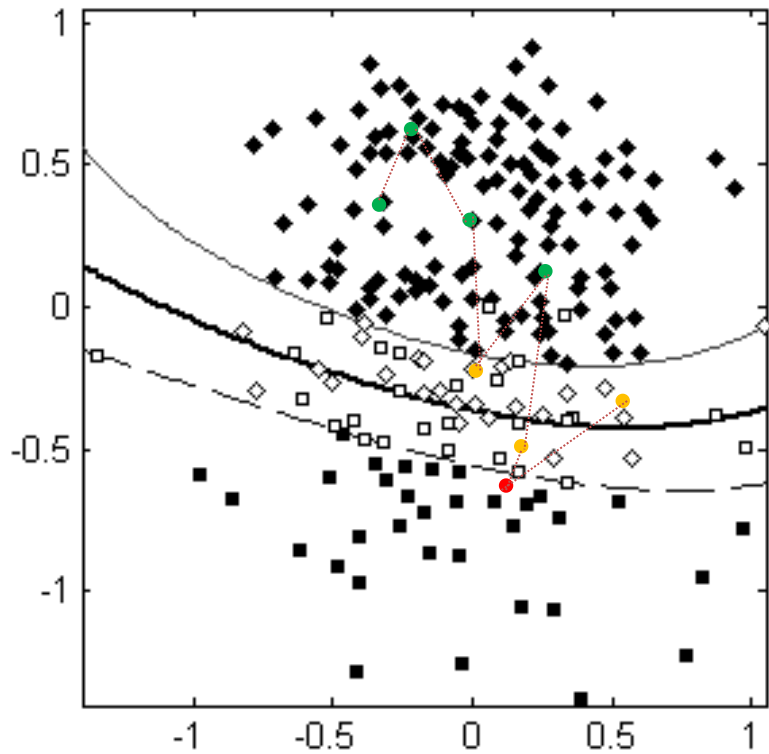
- Physiological data analysis





Visualization for the Self-Observation

- Physiological data analysis





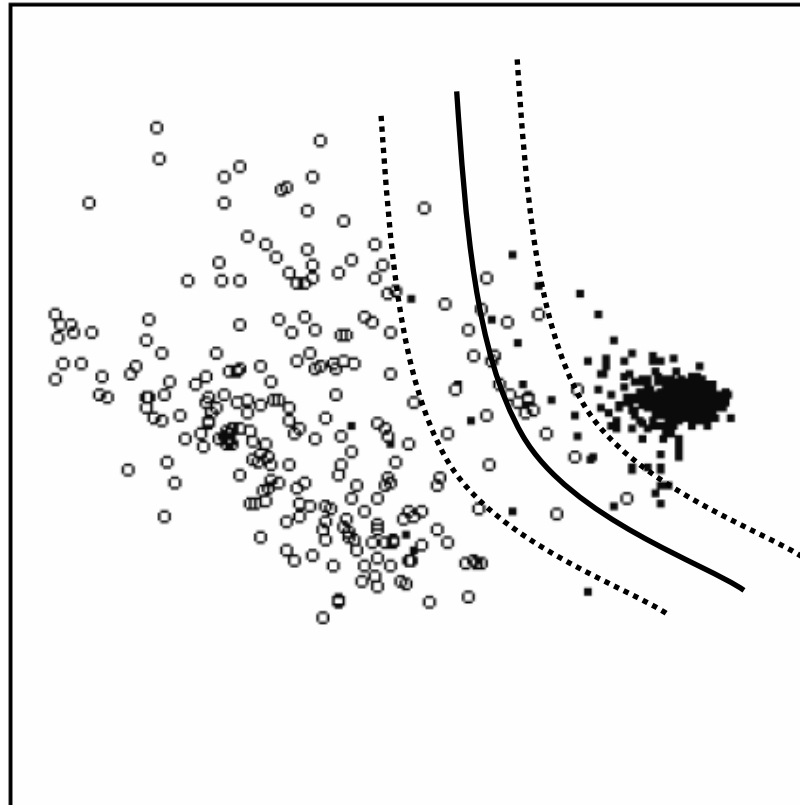
Visualization for Early Diagnosis

- Breast cancer data analysis



Visualization for Early Diagnosis

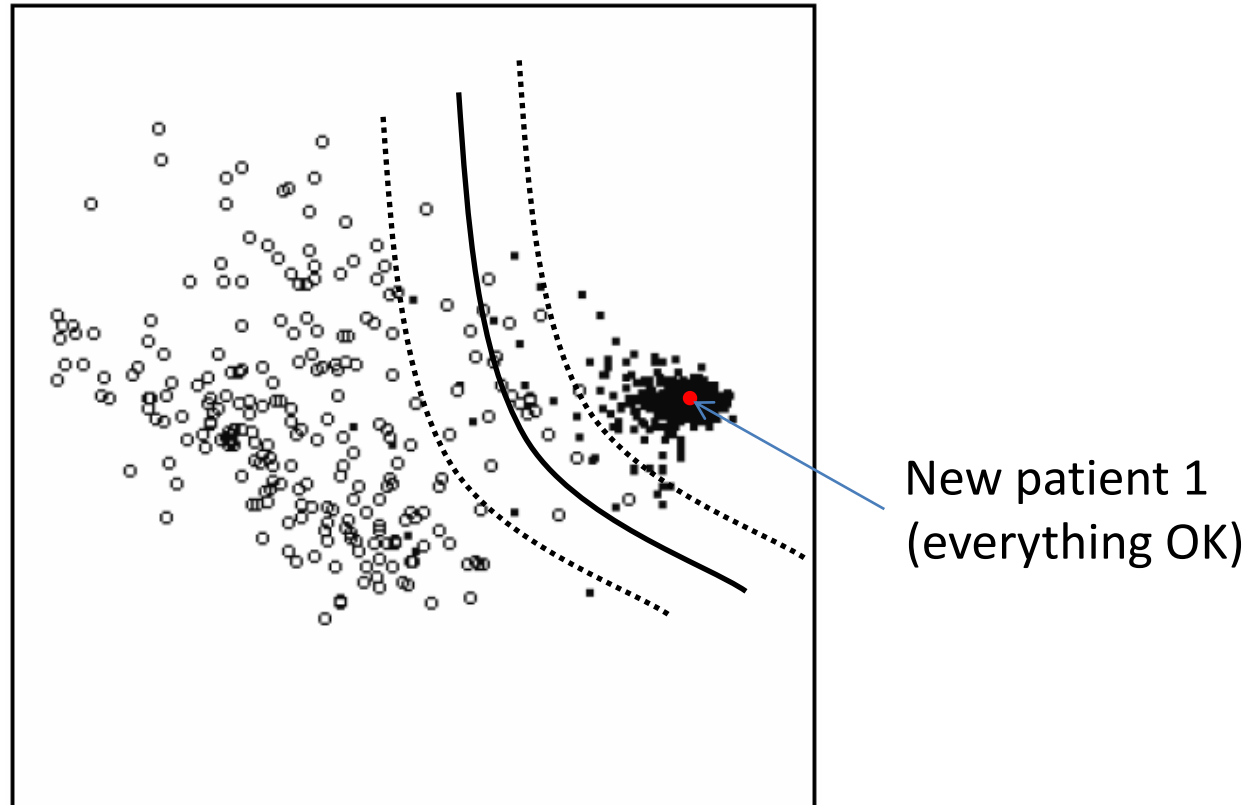
- Breast cancer data analysis





Visualization for Early Diagnosis

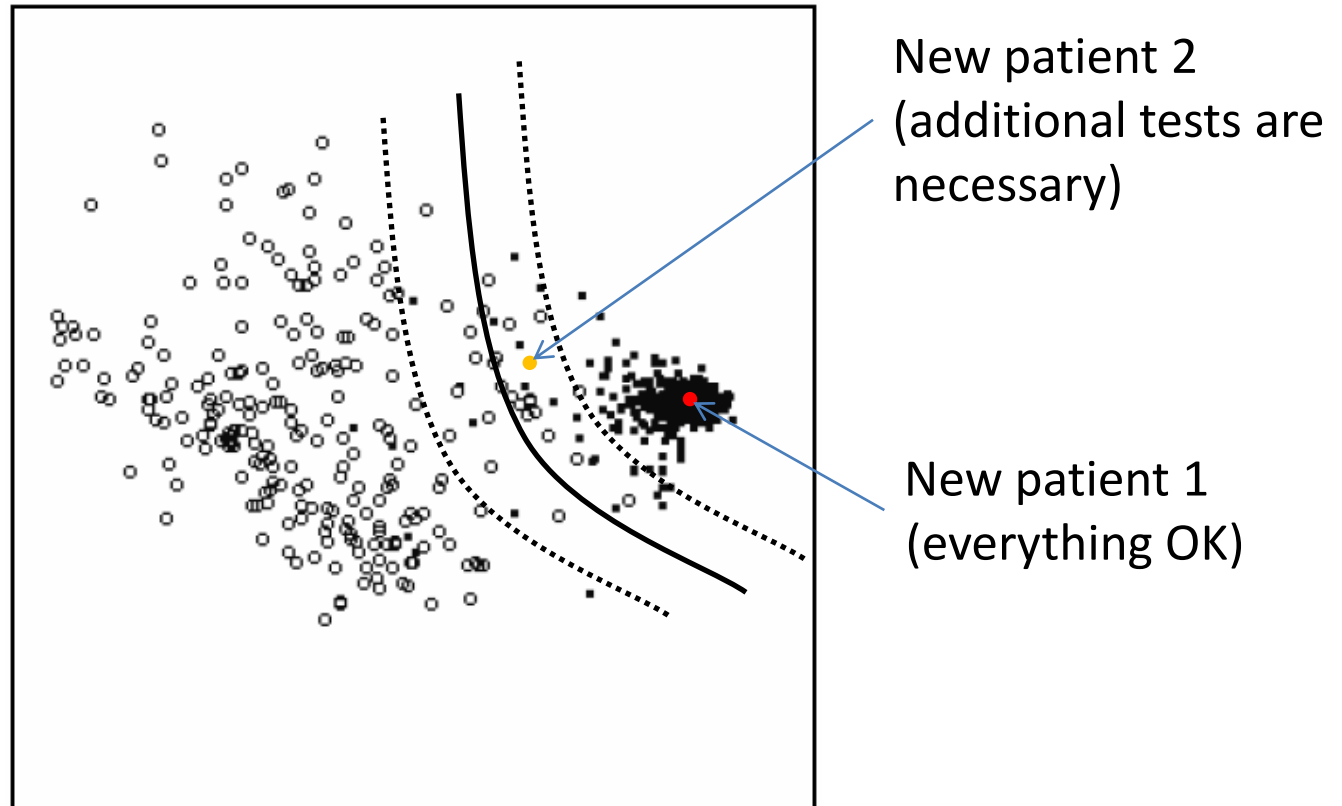
- Breast cancer data analysis





Visualization for Early Diagnosis

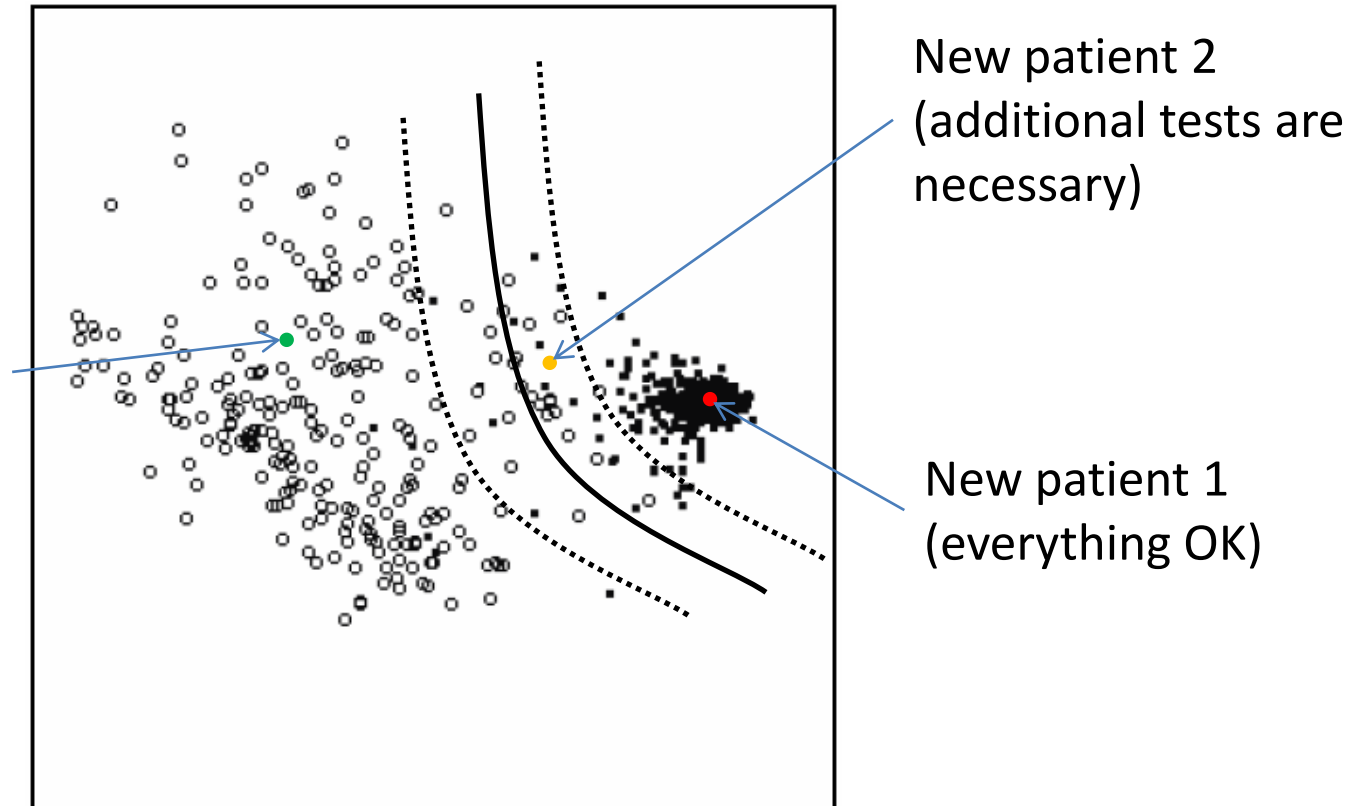
- Breast cancer data analysis





Visualization for Early Diagnosis

- Breast cancer data analysis





WEB Service Based Visualization

The World Wide Web is the ideal platform to implement a **service for visualization** and to **make this service available to customers**.

The proposed service simplifies the usage of visualization methods that are often very sophisticated.



WEB Service Based Visualization

We propose a realization of the service that receives a (large-scale) multidimensional dataset and as a result produces a visualization of the dataset. It also supports different configuration parameters of the data mining methods.

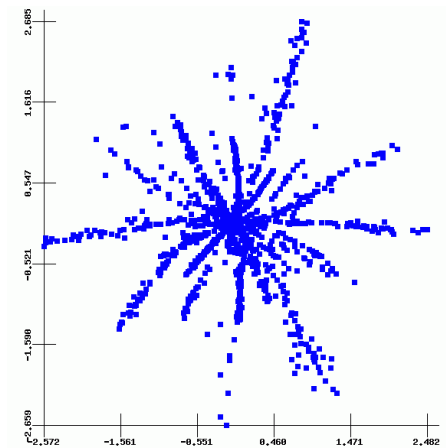
1. WWW

2. Data set

3. Visualization methods

4. Visualization results

	A	B	C	D	E	F	G	H	I	J
1	1.11091	7.04704	2.68999	3.51433	6.24851	-7.70285	3.98986	-15.0384	5.9192	5.87187
2	-5.15074	1.14707	1.49299	2.55285	-2.27442	-14.6506	7.86902	-7.36946	9.63875	7.11384
3	1.91532	-4.58991	-1.90508	4.07452	-11.9971	-7.92136	9.69161	0.222057	10.8834	5.87123
4	0.020048	0.630172	1.08796	2.26984	-4.79242	-8.15183	4.29621	-0.46226	7.42289	6.15558
5	-0.89494	8.11653	-6.95531	6.61701	-8.21592	-11.2322	7.56654	-8.13229	10.6122	5.83966
6	-0.10323	-10.4471	-3.17563	5.31002	0.14175	-9.40686	9.52215	1.64615	-1.88292	6.64164
7	1.72396	0.97497	5.85485	5.75813	1.33302	-4.35716	3.95199	-1.21654	4.97955	4.7993
8	4.52856	11.1303	-0.52264	3.84449	-5.35348	-9.28554	9.6102	-3.04153	5.72042	5.28749
9	-0.83436	1.05809	7.03569	2.13677	-4.65407	-8.54866	6.28365	-2.71864	17.3228	5.42663
10	2.58882	8.18563	4.12082	5.86436	-3.93221	-0.99236	6.98705	-2.16161	3.14953	5.16401
11	-1.00736	0.836051	-0.27273	5.757	-8.00482	-2.68832	4.08388	2.19967	4.11293	6.58779
12	-2.78432	-0.36631	-5.02407	2.12856	-5.31326	-8.53505	4.3084	-4.39105	5.59385	6.73681
13	6.34916	4.67636	3.14195	2.93509	-4.95993	-11.4985	4.62661	-7.7016	7.09844	7.54618
14	2.35358	9.6925	0.786942	5.20636	-1.26773	-11.0521	5.5869	-0.09746	10.3834	6.49727
15	-3.13531	-3.0484	3.19325	3.82446	1.29874	-6.93801	4.42466	-4.00488	15.2039	4.52903
16	0.192948	9.22675	1.49804	3.56301	-3.13276	-7.37238	11.2558	-1.01955	6.3389	5.51749
17	-6.14694	11.5993	5.03532	3.68279	2.66215	-12.0582	8.4349	-3.82024	-4.39669	8.02664
18	1.66801	-1.53136	2.95147	0.441002	-6.85443	-6.11669	7.83389	-8.50664	12.1942	6.17958
19	-2.42808	-2.35968	-0.9866	2.95973	-0.71374	-6.25115	7.01922	-4.05128	0.133749	7.84149
20	-3.23398	0.739948	1.61084	3.85228	-9.73604	-3.64236	11.4746	-11.9365	9.5091	5.65706
21	-0.02485	3.72569	-5.69587	1.96643	-8.95218	0.753423	8.5288	-6.39854	1.8114	7.20848
22	-1.7316	0.14579	-8.76263	0.305276	-0.84276	-13.4019	7.17862	-2.53552	1.32054	5.16366
23	-0.26349	2.65934	-0.38402	4.20786	-9.42951	-3.06817	6.16654	-7.08867	-3.23498	6.06282
24	0.935001	7.62704	-5.89728	0.043108	-8.97537	-17.8392	5.97077	-13.8924	6.5347	5.41031
25	-1.13148	-3.6351	3.4953	4.03109	11.0707	2.57909	6.92097	-4.71983	5.97176	5.41212
26	-0.42632	10.1015	-4.25269	2.9947	-8.99443	-5.9449	3.81592	-0.83499	17.2262	3.8634
27	1.67647	4.57368	2.22492	2.66839	-7.65192	-8.25385	6.32712	-2.59647	7.04368	6.75276
28	0.42202	5.57122	-3.12425	4.19927	-16.3641	-0.344	9.59166	-7.06135	11.1317	1.69255





WEB Service Based Visualization

The Web service for multidimensional data visualization provides a web-based access to several visual data mining methods of different nature and complexity that, in general, allows a visual discovery of patterns and their interpretation in multidimensional data.

The developed software tool allows users to analyze and visualize large-scale multidimensional data sets on the Internet, regardless of time or location, as well as to optimize the parameters of visualization algorithms for better perception of the multidimensional data.



WEB Service Based Visualization

By integrating new powerful technologies into multidimensional data visualization systems, we can get higher performance results with additional functionalities. The basic idea behind Web services is that a specific functionality of software running on one machine of an enterprise is accessible to another machine running at another enterprise using specific protocols over the Internet.

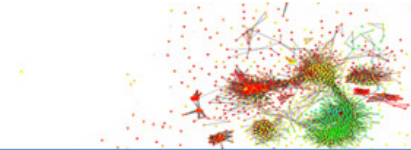
Providing seamless access to systems functionality without downloading the software is the main concept behind Web services.

We are not concerned with other services that might be used in a larger application, but focus simply on a service providing visualization functionality.



http://cluster.mii.lt/visualization

VU MII
Web Service for Data Mining



Home Queue Visualization Results Logout

Number of processors:

Number of iterations:

Visualization method:

Set of basis points:

Computing time:

Datasets:

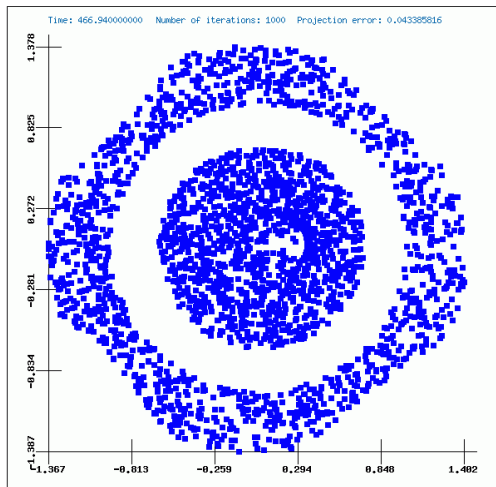
Maximal number of cycles:

Experiments:

newexp090 (2011-10-19 16:23:00) Method: rmds; Data: [1100x100]
newexp089 (2011-10-19 16:23:38) Method: dma; Data: [1100x100]
newexp088 (2011-10-18 19:03:58) Method: mds; Data: [1100x100]
newexp087 (2011-10-13 11:24:38) Method: rmds; Data: [1100x100]
newexp086 (2011-10-13 09:54:24) Method: dma; Data: [1100x100]

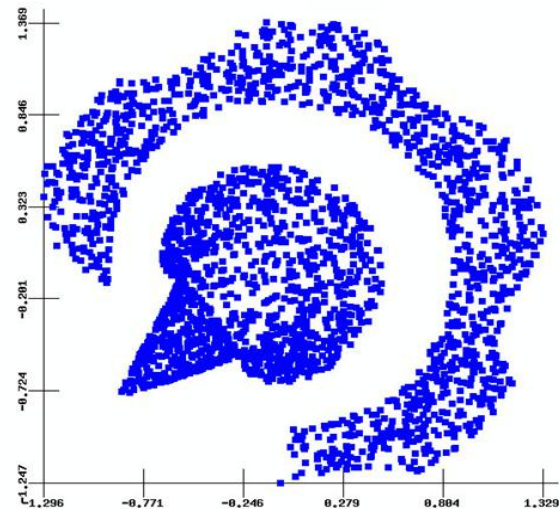
or

Home Queue Visualization Results Logout



Experiments:

-
- newexp000 (2009-06-18 18:29:23)
Method: mds; Number of points: 2573
 - o viz008 (Number of iterations: 100)
 - o viz007 (Number of iterations: 1000)
 - o viz006 (Number of iterations: 1000)
 - o viz_0 <=
 - o viz005 (Number of iterations: 1000)





Advantages

- The proposed Web service can be accessible from any location with internet connectivity and can be used almost on any platform.
- Most of the computational work is performed on the server, with user interaction done on the client.
- The developed software tool allows users to analyze and visualize large-scale multidimensional datasets through the internet, without regard for time or location.



Advantages

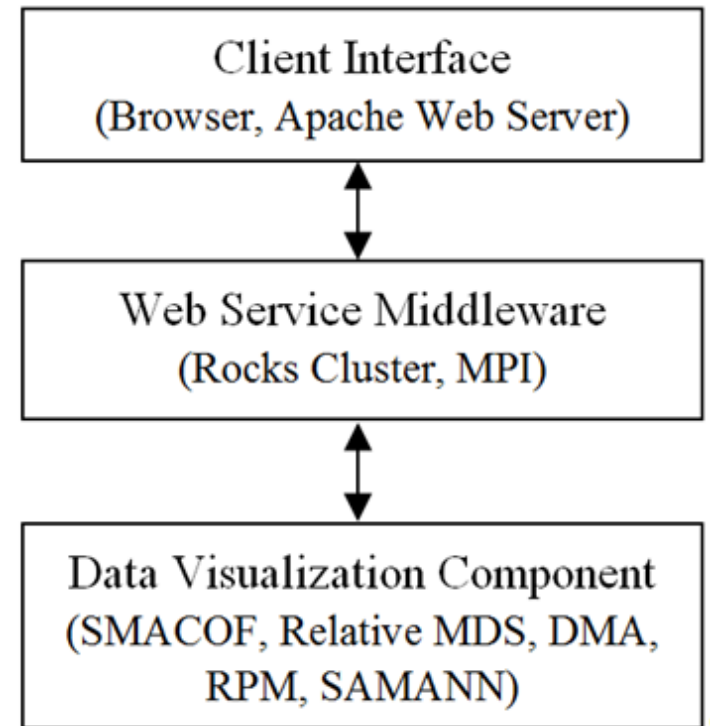
- The proposed service simplifies the usage of the visualization methods and makes them wide-accessible: Multidimensional Scaling (MDS), Relative MDS, Diagonal Majorization algorithm, SAMANN, Relational perspective map.
- For the large-scale multidimensional data visualization a high-performance parallel cluster has been used in our realization. It combines the powers of Web services and parallel computing in a single infrastructure.



Architecture

The proposed Web service architecture for the multidimensional data visualization is a three-layer model.

The Client Interface and Data Visualization Components layers are the main parts of the system. Client's responsibility is sending a data, which must be accepted, processed and returned from the visualization service.





Architecture

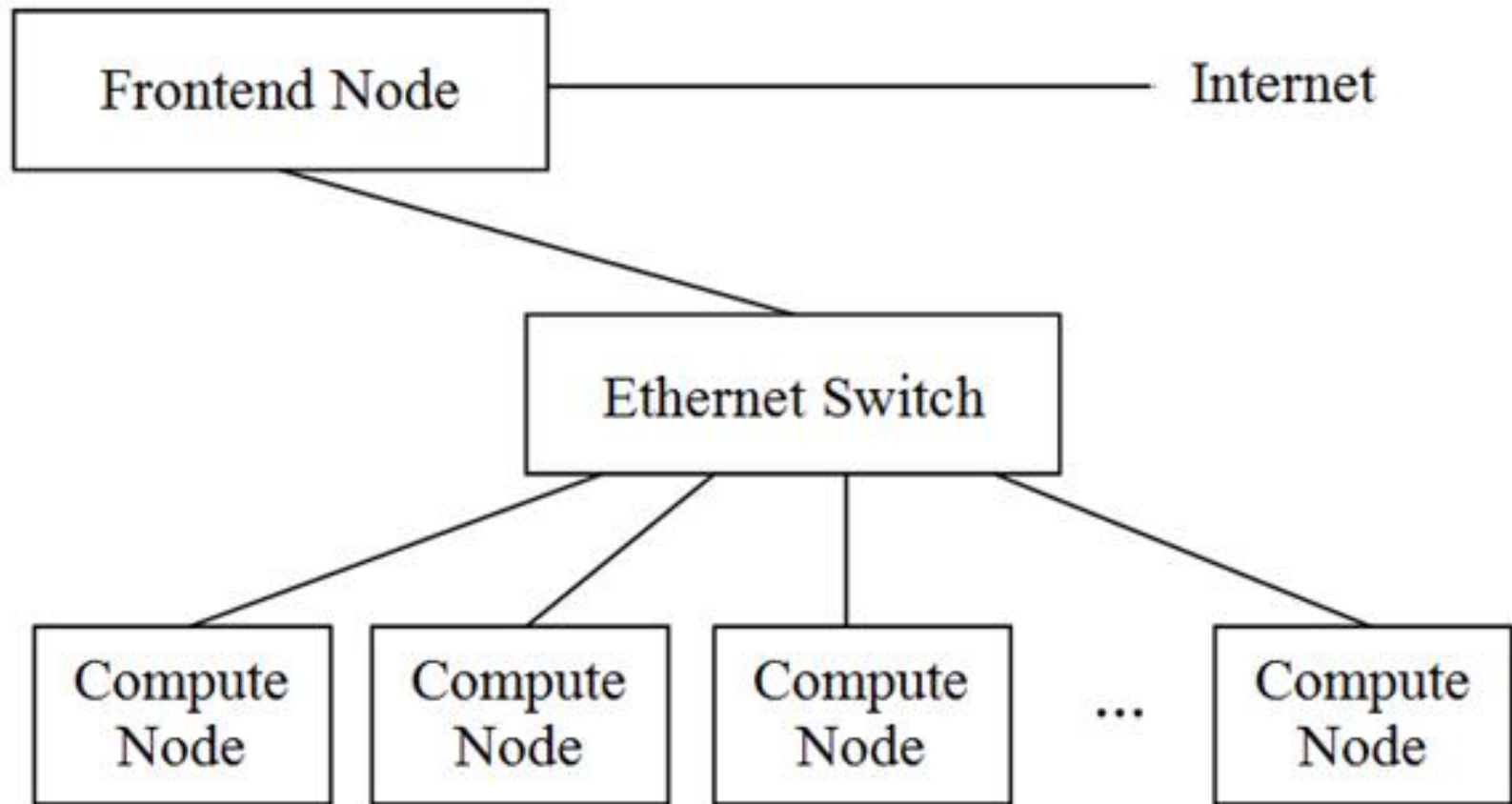
Distributed parallel computing is becoming the de facto architecture for managing the performance of computationally intensive, long-running programs.

In our case it is possible to run parallel visualization components that communicate through MPI (Message Passing Interface).

We suggest to use the design and implementation of the MPI middleware that connects the web service interface and the parallel software tool running on a computational cluster.



Architecture





Client Interface

In the Client Interface, it is possible to choose such parameters:

- Number of processors
- Maximum number of iterations
- Method for a multidimensional data visualization (MDS, SMACOF algorithm, Relative MDS, DMA, RPM, SAMANN)
- Strategies of forming and initializing the set of basis points (on the line, random, maximal dispersion, principal component analysis)
- Maximal computing time
- Upload the client's dataset for visualization
- Maximal number of visualization cycles



Client Interface

Home Queue Visualization Results Logout

Number of processors

Number of iterations

Visualization method

Set of basis points

Number of basis points

Basis point selection method

Computing time

Datasets

Maximal number of cycles

or

Experiments:

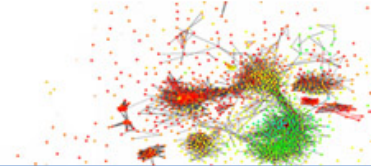
- newexp090 (2011-10-15 17:37:54)
Method: rmds; Data: [1100x100]
- => viz001 - (Number of iterations: 1000, Init Method: Variance, Base: 100, Base init: Random) <=
- viz000 - (Number of iterations: 1000, Init Method: Variance, Base: 100, Base init: Random)
- newexp089 (2011-10-13 11:45:38)
Method: dma; Data: [1100x100]
- newexp088 (2011-10-18 19:03:58)
Method: rmds; Data: [1100x100]



http://cluster.mii.lt/visualization

VU MII

Web Service for Data Mining



Home Queue Visualization Results Logout

Number of processors

Number of iterations

Visualization method

Set of basis points

Number of basis points

Basis point selection method

Computing time

Datasets

Maximal number of cycles

Experiments:

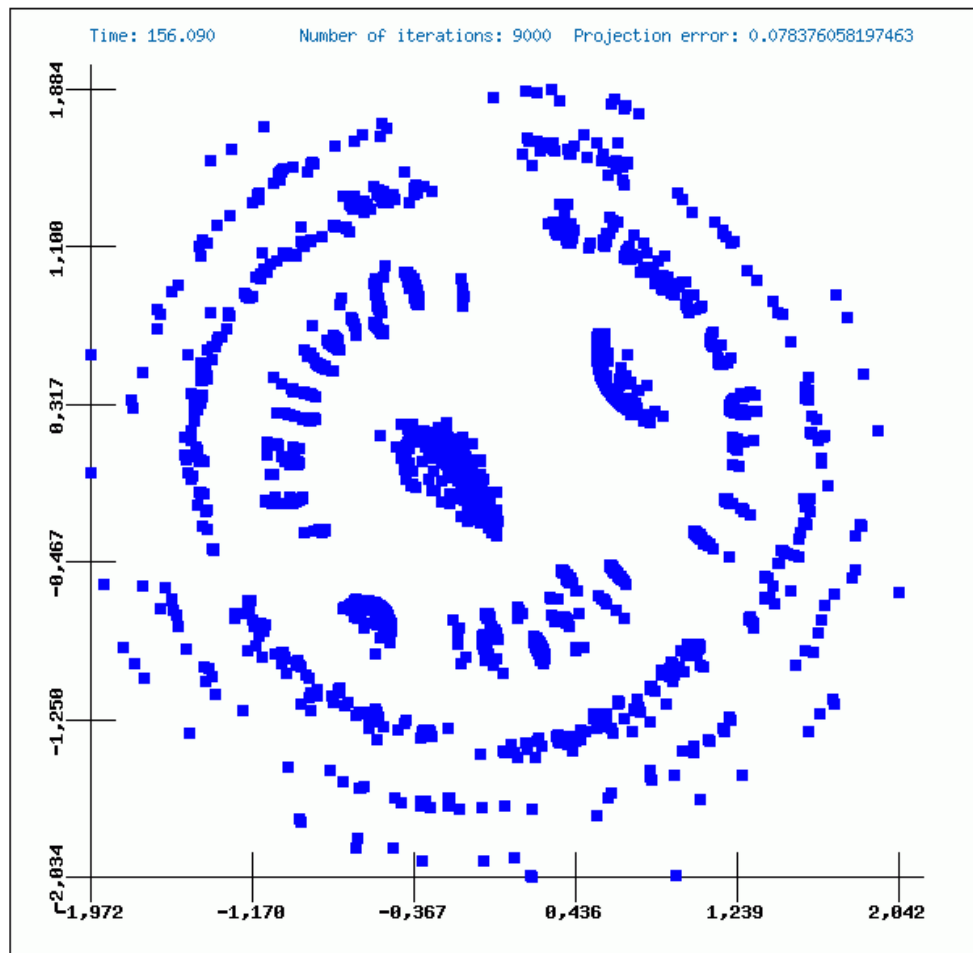
- newexp090 (2011-10-19 16:23:00)
Method: rmds; Data: [1100x100]
- viz002 - (Number of iterations: **1000**, Init Method: **Variance**, Base: **100**, Base init: **Random**)
- => viz001 - (Number of iterations: **1000**, Init Method: **Variance**, Base: **100**, Base init: **Random**) <=
- viz000 - (Number of iterations: **1000**, Init Method: **Variance**, Base: **100**, Base init: **Random**)
- newexp089 (2011-10-19 16:22:51)
Method: dma; Data: [1100x100]

or

Your job 22820 ("experiment.sh") has been submitted

Home Queue Visualization Results Logout

job-ID	prior	name	user	state	submit/start at	queue	slots	ja-task-ID
22820	0.60500	experiment	MdsService	r	10/19/2011 16:23:37	all.q@compute-0-7.local	16	
22821	0.50500	experiment	MdsService	qw	10/19/2011 16:23:27		8	
22822	0.50500	experiment	MdsService	qw	10/19/2011 16:23:29		8	
22823	0.50500	experiment	MdsService	qw	10/19/2011 16:23:30		8	
22824	0.50500	experiment	MdsService	qw	10/19/2011 16:23:31		8	
22825	0.50500	experiment	MdsService	qw	10/19/2011 16:23:33		8	
22826	0.50500	experiment	MdsService	qw	10/19/2011 16:23:34		8	
22827	0.50500	experiment	MdsService	qw	10/19/2011 16:23:35		8	
22828	0.00000	experiment	MdsService	qw	10/19/2011 16:23:37		8	
22829	0.00000	experiment	MdsService	qw	10/19/2011 16:23:38		8	



[Download](#)

Experiments:

[Find](#)

newexp090 (2011-10-19 16:23:00)
Method: rmds; Data: [1100x100]
newexp089 (2011-10-19 16:23:38)
Method: dma; Data: [1100x100]
newexp088 (2011-10-18 19:03:58)
Method: mds; Data: [1100x100]
newexp087 (2011-10-13 11:24:38)
Method: rmds; Data: [1100x100]
newexp086 (2011-10-13 09:54:24)
Method: dma; Data: [1100x100]
newexp085 (2011-10-13 09:53:36)
Method: dma; Data: [1100x100]
newexp084 (2011-10-13 09:38:02)
Method: mds; Data: [1100x100]
newexp083 (2011-10-13 09:28:57)
Method: rmds; Data: [1100x100]
newexp082 (2011-10-13 09:23:16)
Method: mds; Data: [1100x100]
newexp081 (2011-10-13 09:22:32)
Method: mds; Data: [1100x100]
newexp080 (2011-10-13 09:21:35)
Method: dma; Data: [1100x100]
newexp079 (2011-10-13 09:18:20)
Method: rmds; Data: [1100x100]
newexp078 (2011-10-13 08:40:20)
Method: mds; Data: [4731x100]
newexp077 (2011-10-13 08:24:42)
Method: rmds; Data: [1100x100]
newexp076 (2011-10-13 05:27:26)
Method: rmds; Data: [1100x100]



Visualization Methods

At first, client sends the data to the visualization service (Data Visualization Component). In our case, five methods for the multidimensional data visualization are included:

MDS

Relative MDS

Diagonal Majorization algorithm

SAMANN

Relational perspective map

These methods have been chosen for testing the architecture and approach. In the future, the set of options for visualization may be extended.



Visualization Methods (MDS)

Multidimensional scaling (MDS) is a group of methods that project multidimensional data to a low (usually two) dimensional space and preserve the interpoint distances among data as much as possible.

The goal of projection in the **metric multidimensional scaling (MDS)** is to optimize the projection so that the distances between the items in the lower-dimensional space would be as close to the original distances as possible.

$$X_i = (x_{i1}, x_{i2}, \dots, x_{id}) \in R^n; Y_i = (y_{i1}, y_{i2}, \dots, y_{im}) \in R^d$$

d_{ij} the distance between the vectors X_i and X_j in the feature space R^n

d_{ij}^* the distance between the vectors X_i and X_j in the projected space R^d



Visualization Methods (MDS)

The objective function (stress) to be minimized can be written as

$$E_{MDS} = \sum_{\substack{i,j=1 \\ i < j}}^m w_{ij} (d_{ij}^* - d_{ij}).$$

$$w_{ij} = \frac{1}{\sum_{\substack{k,l=1 \\ k < l}}^m (d_{kl}^*)^2}$$

$$w_{ij} = \frac{1}{d_{ij}^* \sum_{\substack{k,l=1 \\ k < l}}^m (d_{kl}^*)^2}$$

$$w_{ij} = \frac{1}{m d_{ij}^*}$$



Visualization Methods (MDS)

The original **MDS** method is unsuitable for large-scale datasets: it takes much computing time or there is not enough computing memory. Furthermore, it is necessary to recalculate the projection of all data points, when a point has to be mapped.

Various modifications of MDS have been proposed for visualization of large datasets: Steerable Multidimensional Scaling, Incremental MDS, Relative MDS, Diagonal Majorization Algorithm (DMA) and etc.

In the Web application proposed, the metric **Multidimensional Scaling SMACOF algorithm** has been used. The SMACOF Algorithm is one of the best optimization algorithms for this type of minimization problem. This method is simple and powerful, because it guarantees a monotone convergence of the stress function.



Visualization Methods (Relative MDS)

Various methods have been proposed for mapping of new points without recalculating all the previously mapped points. **Relative MDS** and **Diagonal Majorization algorithms** are designated to visualize large-scale multidimensional data.

The MDS algorithm does not offer a possibility to project new points on the existing set of mapped points. To get a mapping that presents the previously mapped points together with the new ones requires a complete re-run of the MDS algorithm on the new and the old data points. The main idea of the **Relative MDS** method (which can be easily used for visualizing new points) is to take a subset of the initial multidimensional data set and then map the basic data set, using the MDS. As a second step, the remaining points of initial data are added to the basis layout using the relative mapping.



Visualization Methods (DMA)

Various types of minimization of the stress function are possible. It is possible to use the Guttman majorization algorithm based on iterative majorization and its modification so called **Diagonal Majorization Algorithm (DMA)**. Guttman majorization algorithm is one of the best optimisation algorithms for this type of minimization problem.

DMA attains slightly worse projection error than Guttman majorization algorithm, but computing it faster. Iterative computations of two-dimensional coordinates are based not on all distances between multidimensional points in the input space. This allows us to significantly speed up the visualization process and to save the computer memory essentially.



Visualization Methods (RPM)

The relational perspective map (RPM) method visualizes multidimensional data onto the closed plane (torus surface) so that the distances between data in the lower-dimensional space would be as close as possible to the original distances.

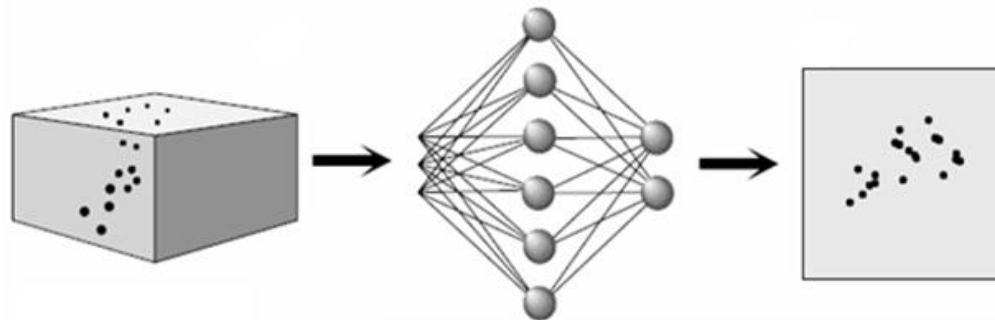
The RPM method also gives the ability to visualize data in a non-overlapping manner so that it reveals small distances better than other known visualization methods.



Visualization Methods (SAMANN)

The combination and integrated use of data visualization methods of a different nature are under a rapid development. The MDS got some attention from neural network researchers.

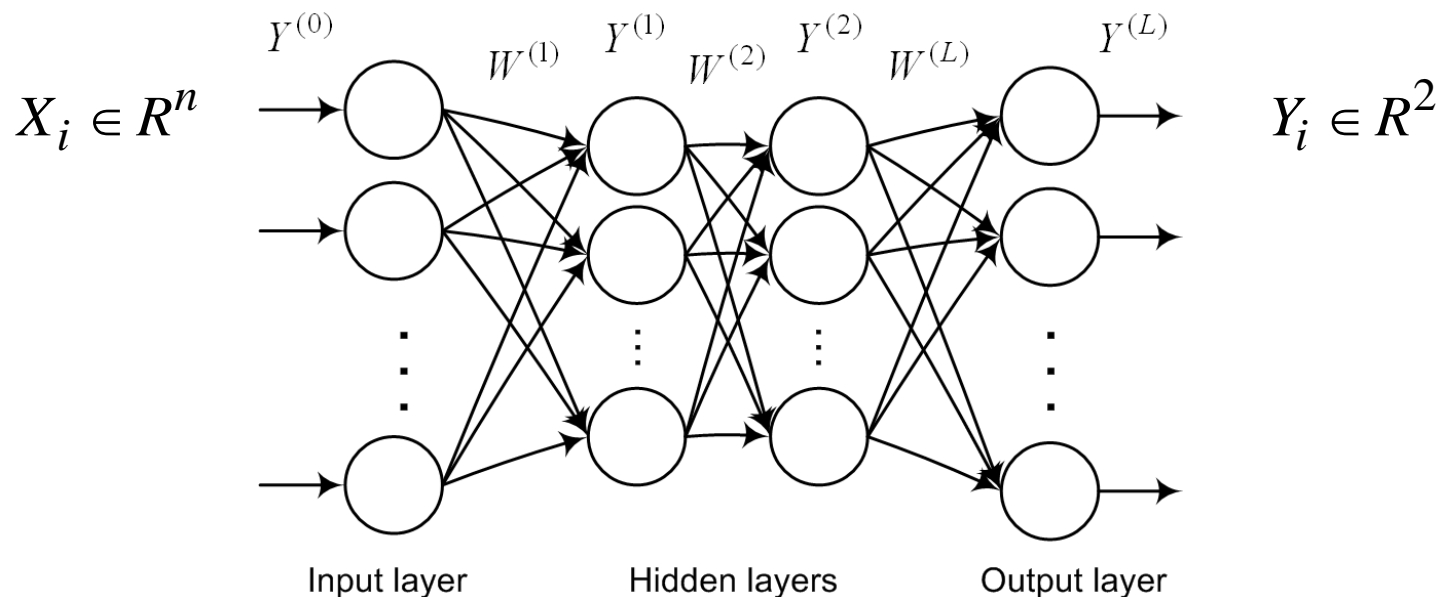
A specific backpropagation-like learning rule (**SAMANN**) has been developed to allow a normal feed-forward artificial neural network to learn Sammon's mapping in an unsupervised way. The network is able to project new multidimensional points after training.





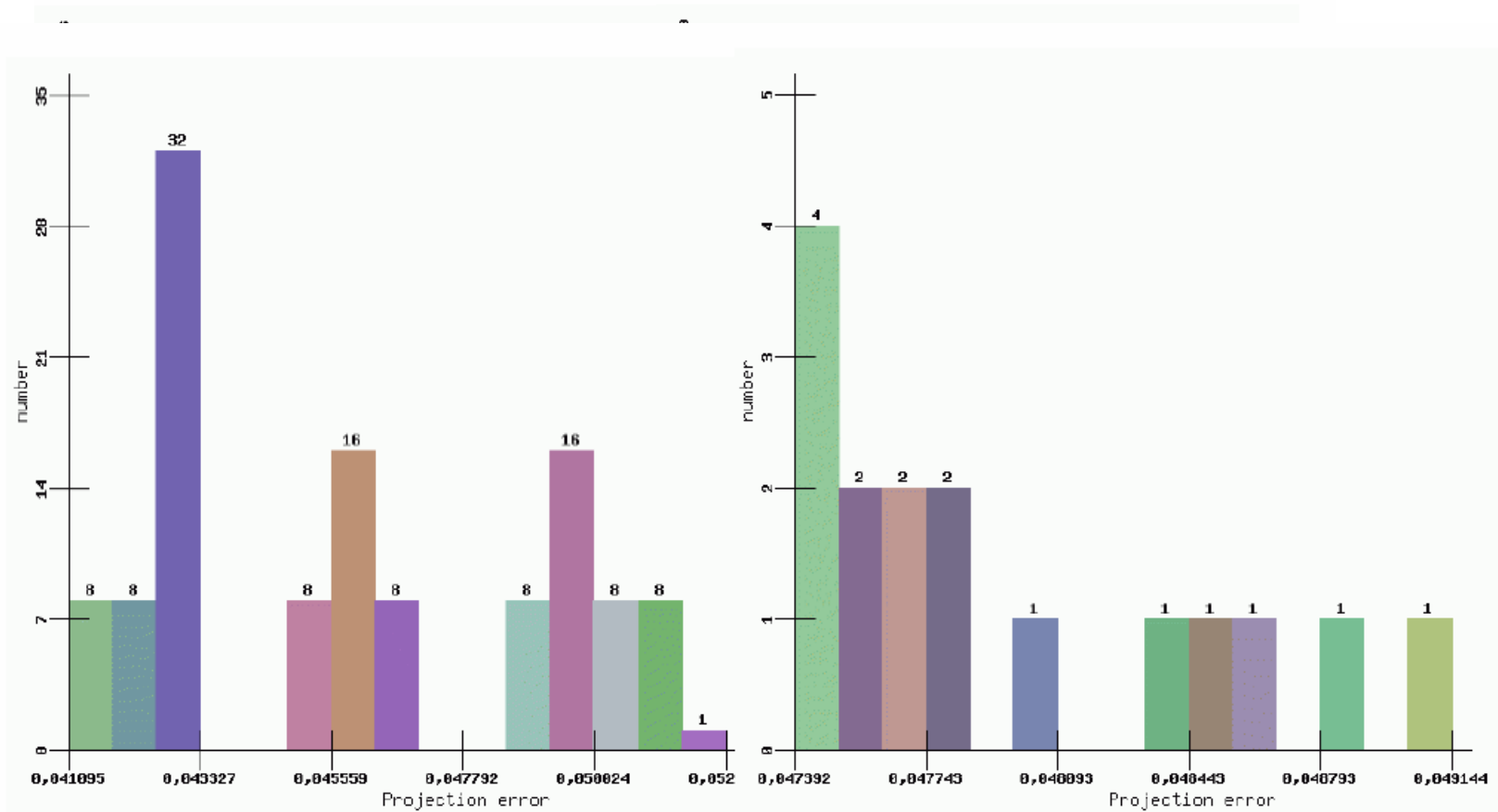
Visualization Methods (SAMANN)

The architecture of the SAMANN network is a multilayer perceptron where the number of input vectors is set to be the input space dimension, n , and the number of output vectors is specified as the projected space dimension, d .





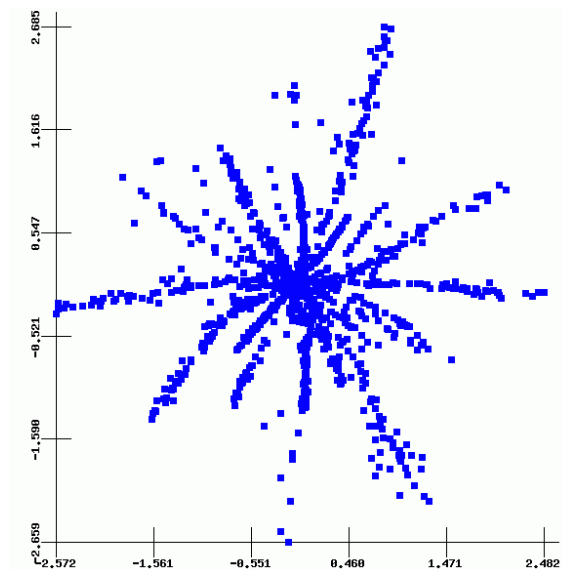
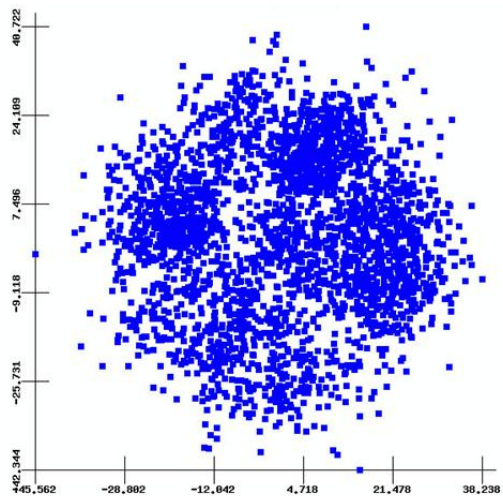
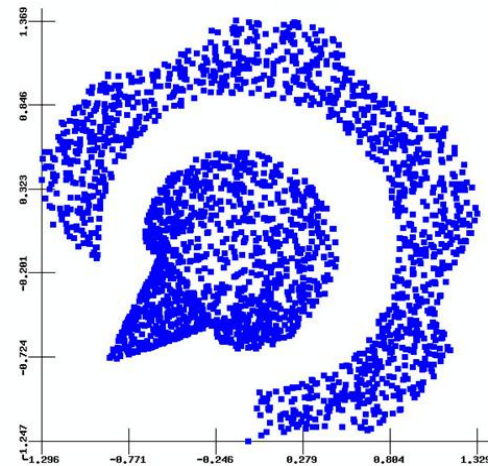
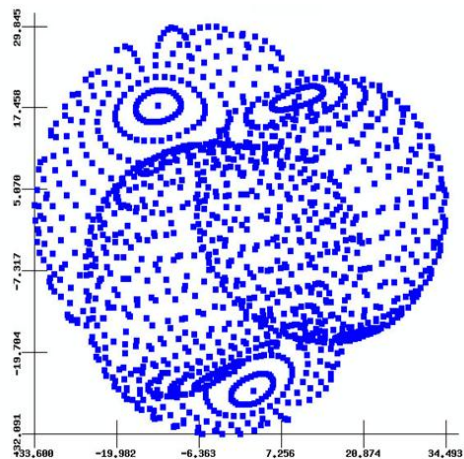
Statistical Information





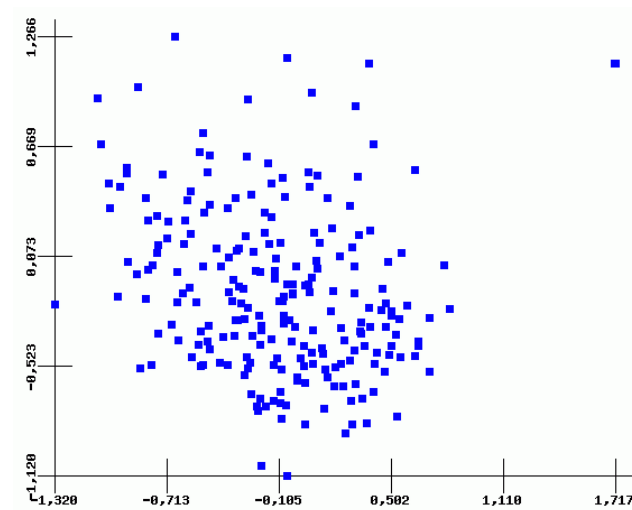
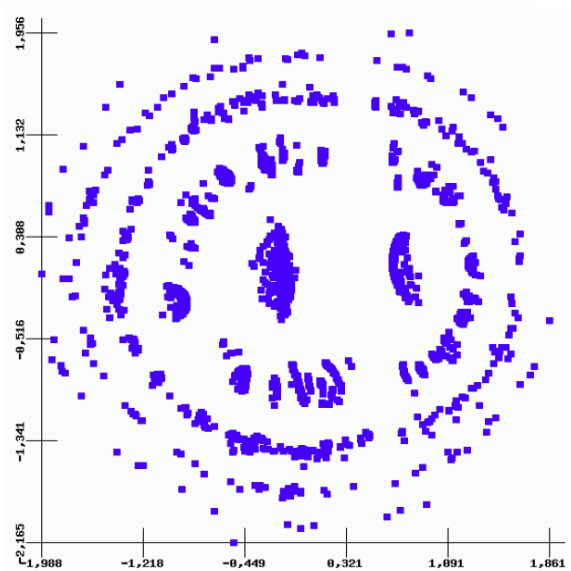
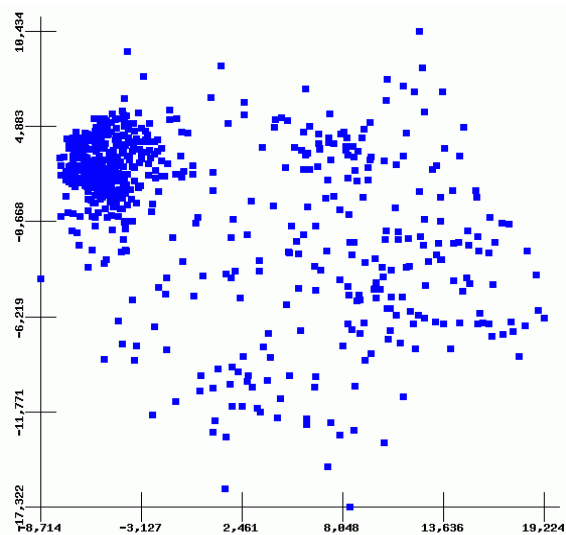
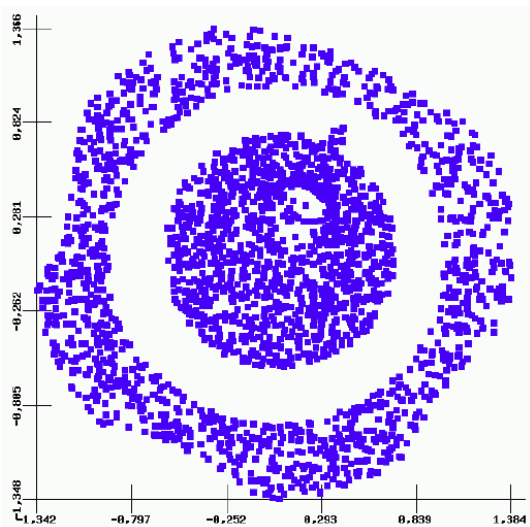
<http://cluster.mii.lt/visualization>

Visualization Results





Visualization Results





Conclusions

An approach and architecture have been proposed for visualization of large-scale multidimensional data, using Web service technologies. This should extend the practical application of multidimensional data analysis and, particularly, visualization techniques.

The proposed service simplifies the usage of visualization methods that are often very sophisticated and include a lot of the know-how of their developers. Five methods for the multidimensional data visualization are included: MDS (SMACOF algorithm), Relative MDS, DMA, RPM and SAMANN. These methods have been chosen for testing the architecture and approach.

In future, the set of options for visualization should be extended.



Conclusions

The main advantage of the proposed approach is that it stimulates the visual data mining and pattern recognition in large-scale multidimensional datasets

Depending on the data set and visualization methods the computations may take the sufficiently large amount of time. The advantage of the service is that the user may not wait for the visualization results online. When computations are completed, the user can download the results at any time he wants.



THANK YOU

<http://cluster.mii.lt/visualization/>
user: gintas/ pass: gintas