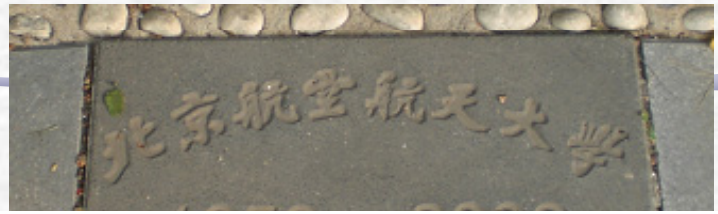
ニセコ町教育長賞

水彩「昔からの夢」　水田百香（札屯田北中３年）

ひたむきな夢の実現に作者の喜びが伝わります。遊び心も描き込んだなかなかの力作です。


北京航空航天大学

HOKKAIDO UNIVERSITY

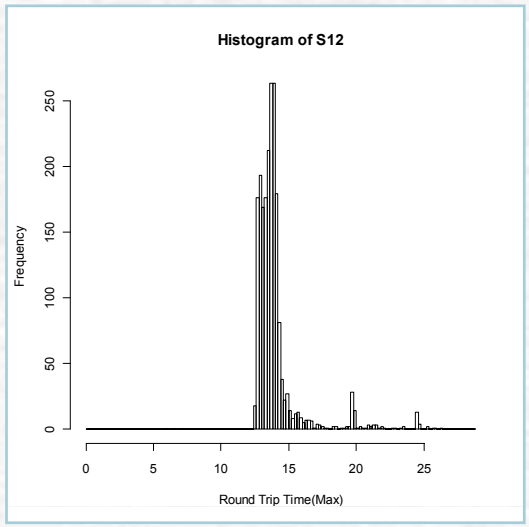
Analysis of Symbolic Data with Functional Data Analysis

Masahiro MIZUTA

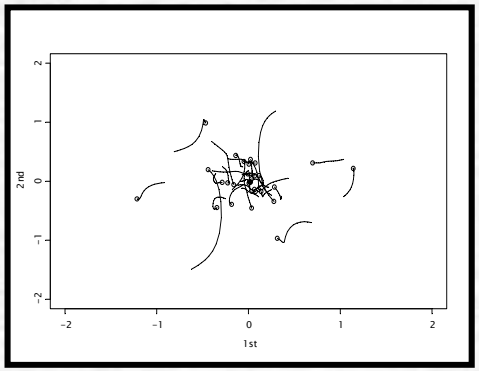
Advanced Data Science Lab.

Hokkaido University, JAPAN

Outline of our study



MDS for distribution valued dissimilarity data

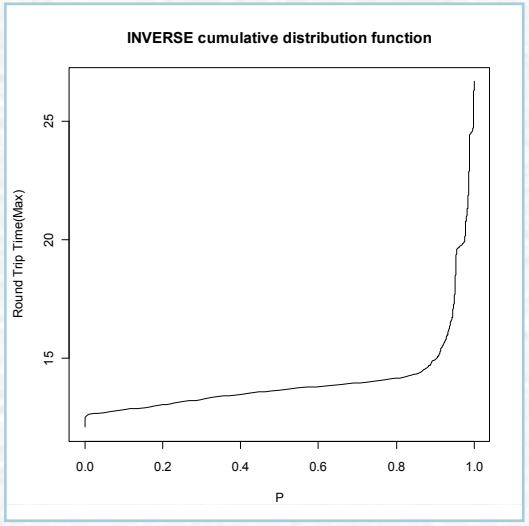


Distribution Valued Dissimilarity Data

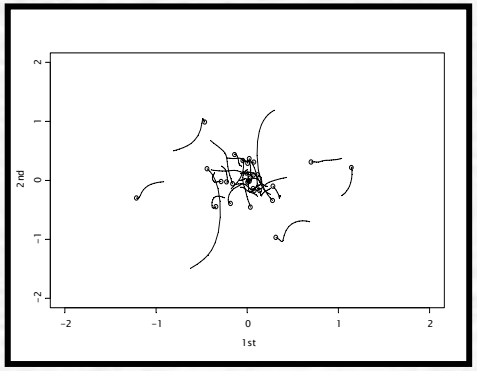
s_{ij}

Symbolic Data Analysis

Outline of our study



Functional MDS



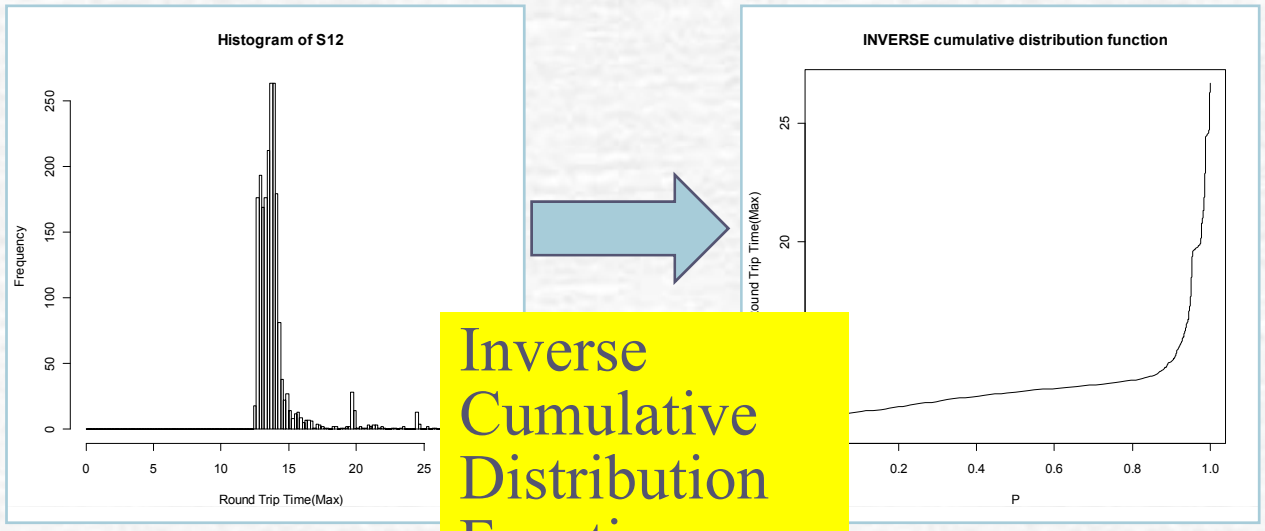
Functional Dissimilarity Data

$s_{ij}(t)$

Functional Data Analysis

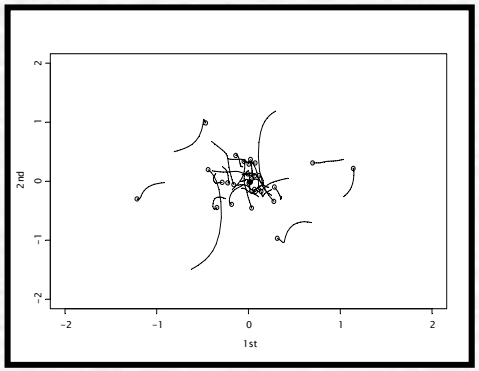
Mizuta, M. (2003): Multidimensional scaling for dissimilarity functions with continuous argument, *Journal of the Japanese Society of Computational Statistics,* Vol.15, No.2, pp.223-227.

Outline of our study



Inverse Cumulative Distribution Function

Functional MDS



Distribution Valued Dissimilarity Data

s_{ij}

Symbolic Data Analysis

Functional Dissimilarity Data

$s_{ij}(P)$

Functional Data Analysis

List of Contents

- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

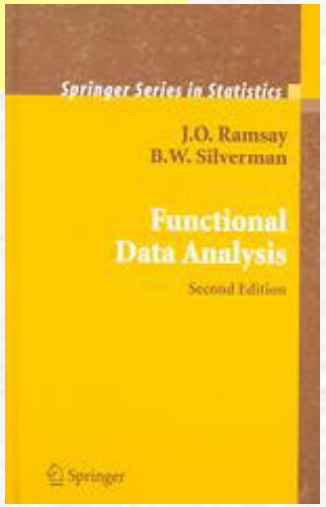
List of Contents

- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

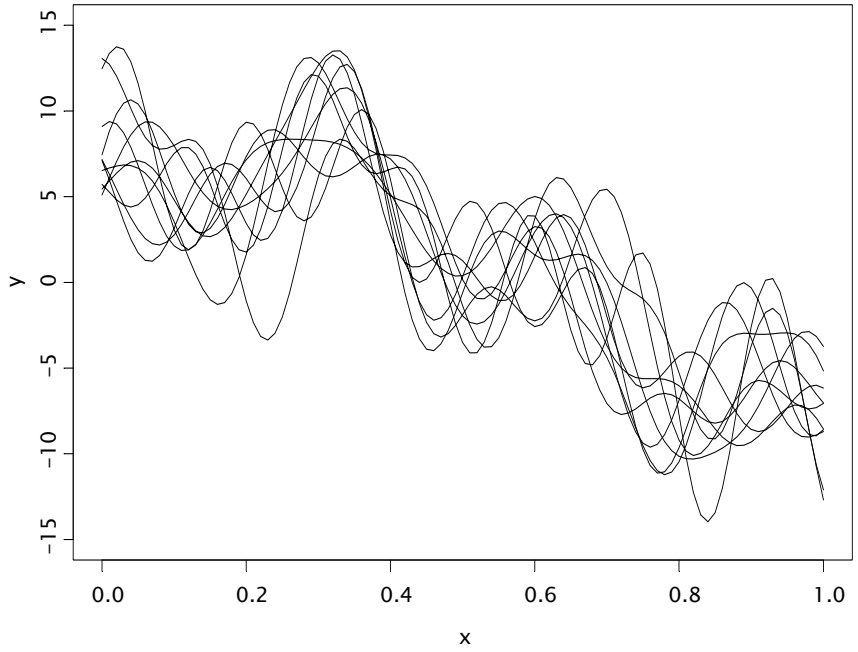
Functional Data Analysis (FDA)

J.O.Ramsay and B.W.Silverman(1982)
When the data are functions, *Psychometrika*

J.O.Ramsay and B.W.Silverman(1997,2005)
Functional Data Analysis, springer



Typical Functional Data



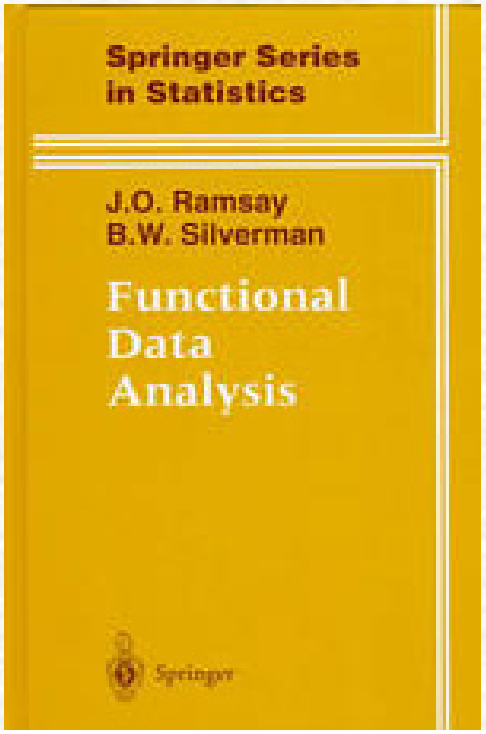
$$t \in [0,1]$$

$$f_i(t) \in \mathrm{R}^1$$

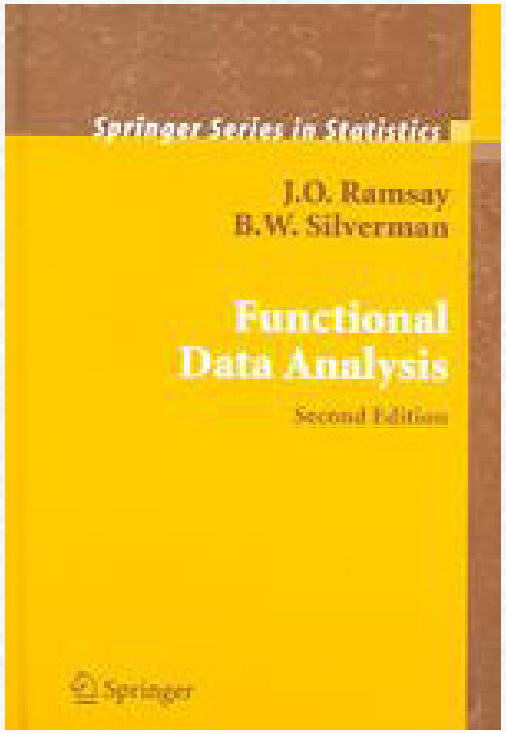
$$(i = 1,2,\cdots,n)$$

Functional Data Analysis (Ramsay and Silverman)

1st Edition (1997)

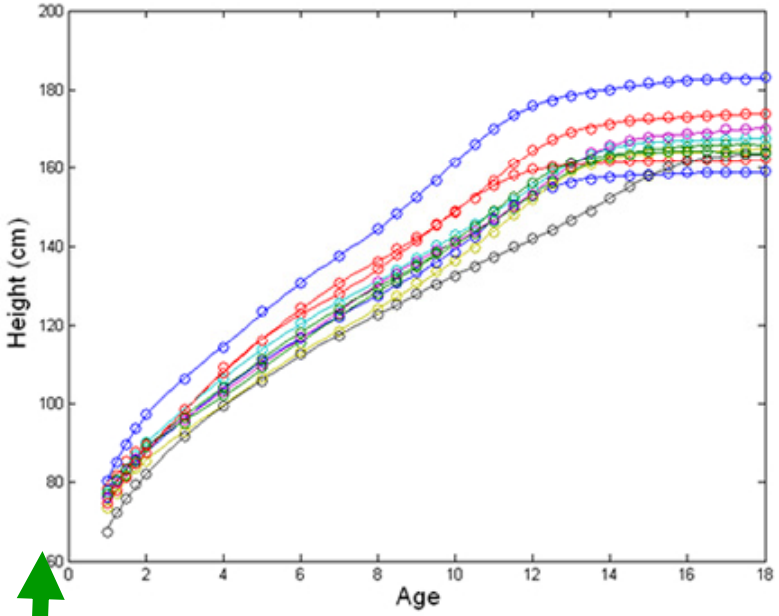


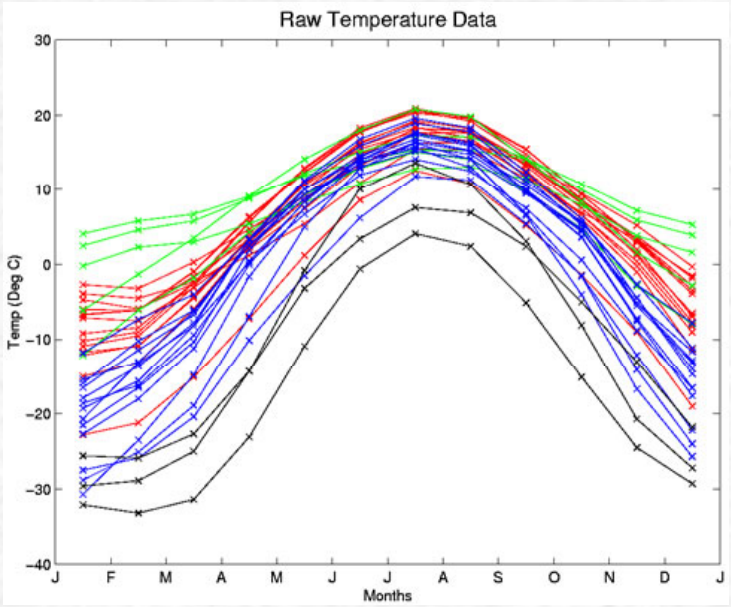
2nd Edition (2005)

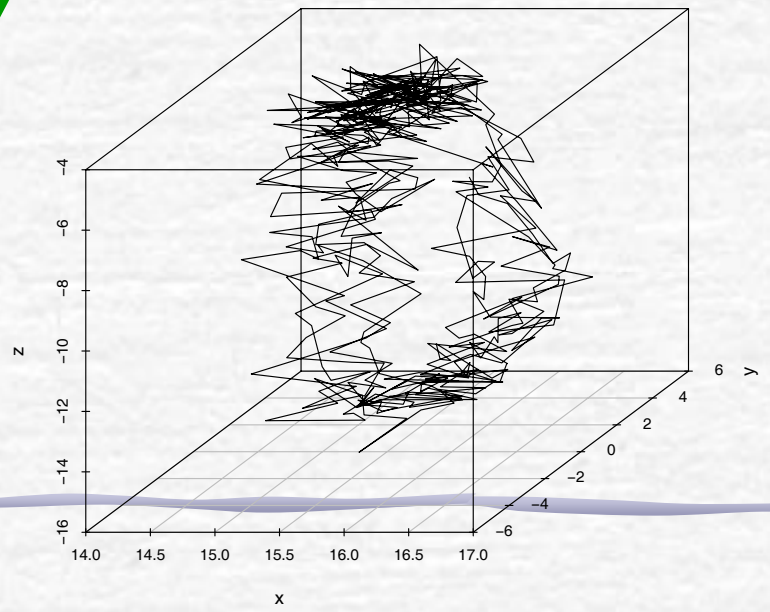


# of Pages	310	426
# of Figures	98	151
Weight	605g	775g
Price	EUR 100.50	EUR 74.50

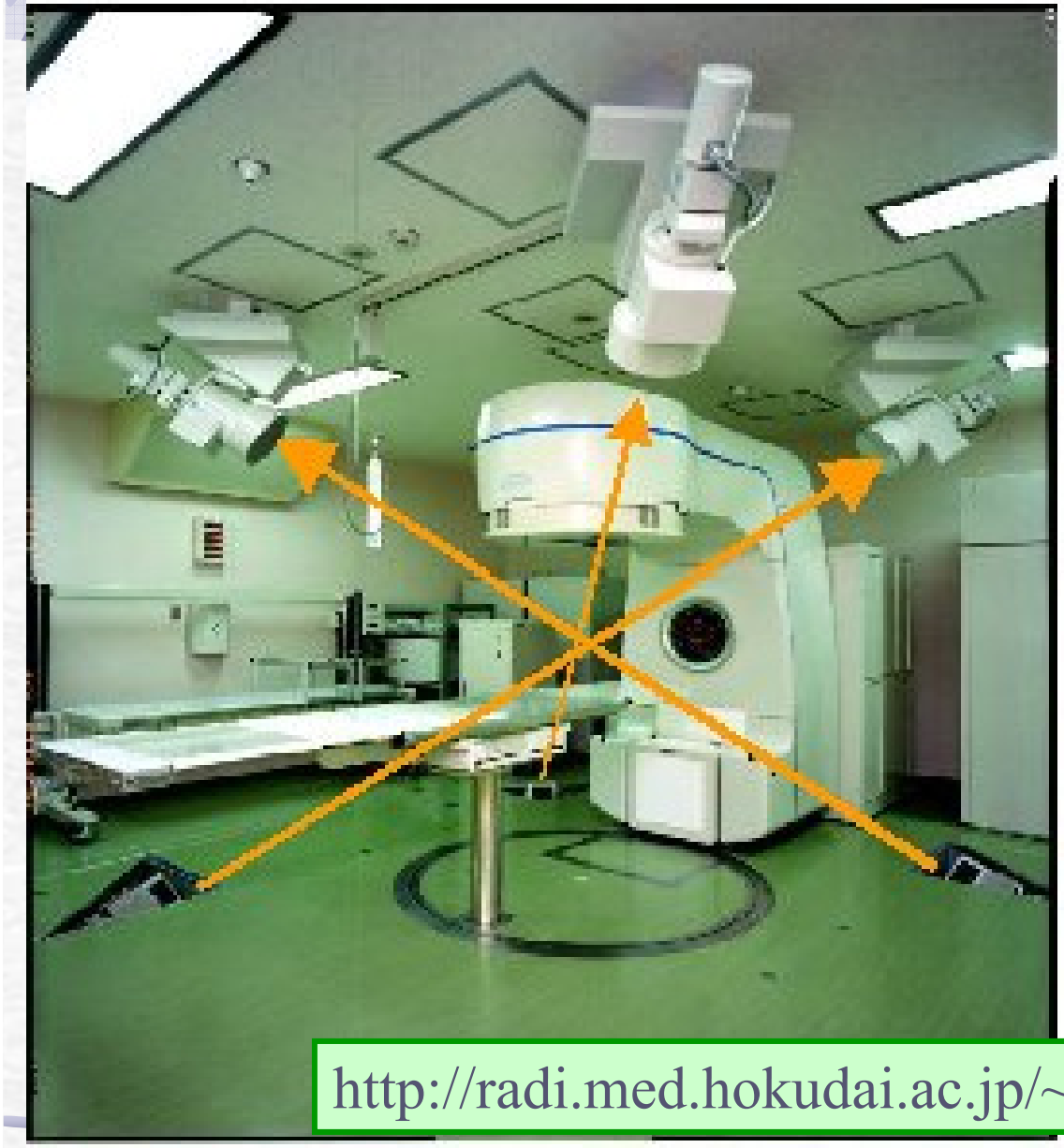
Functional Data

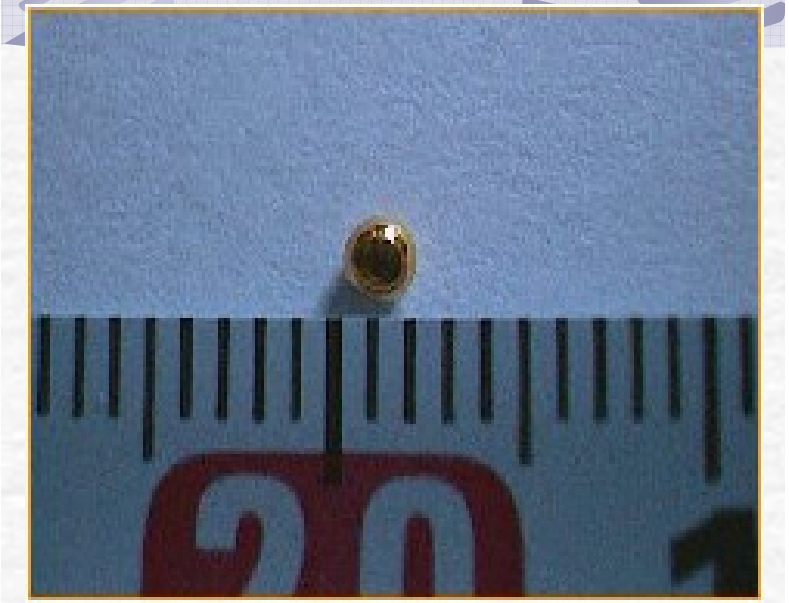






http://ego.psych.mcgill.ca/misc/fda/

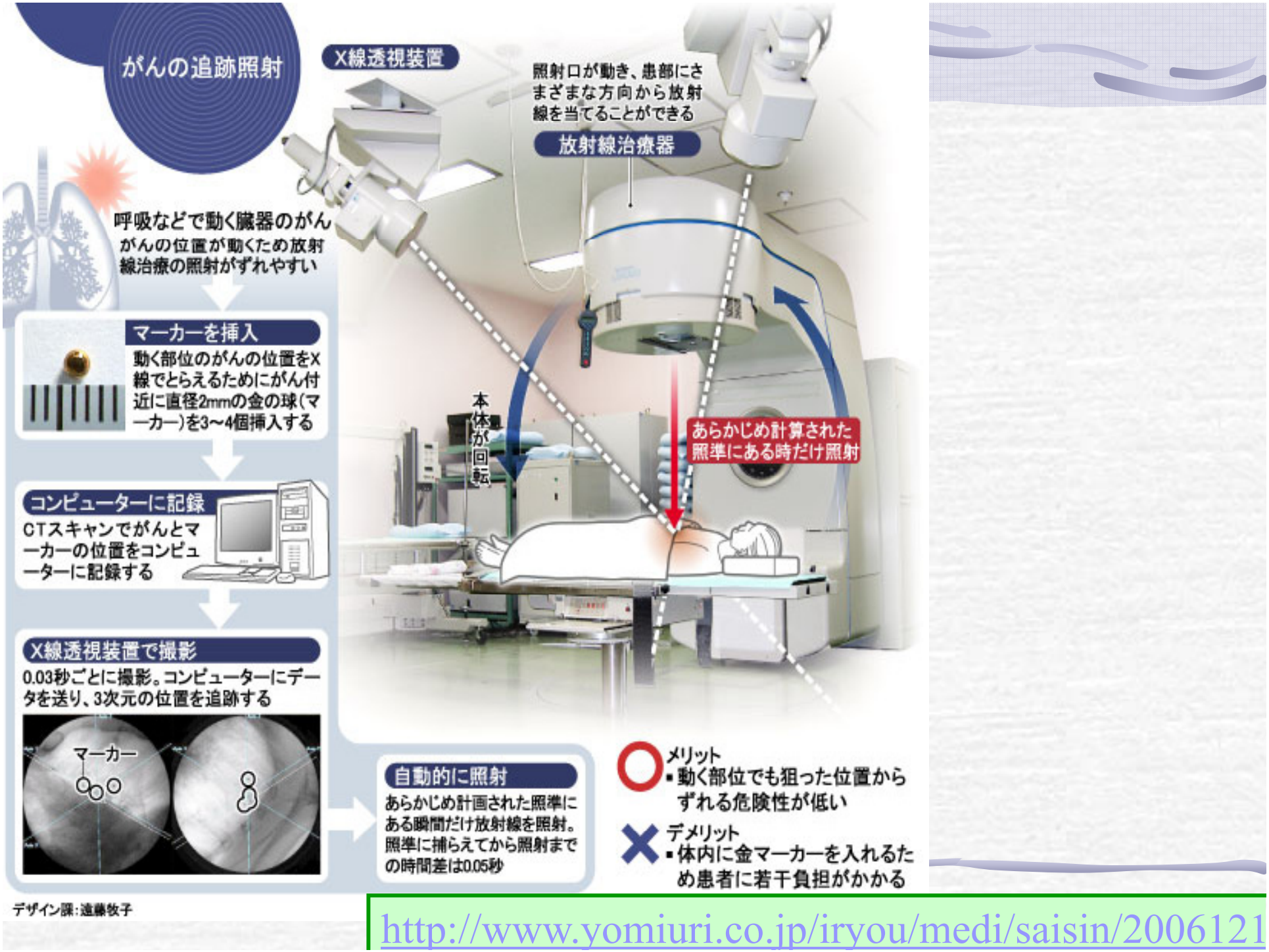




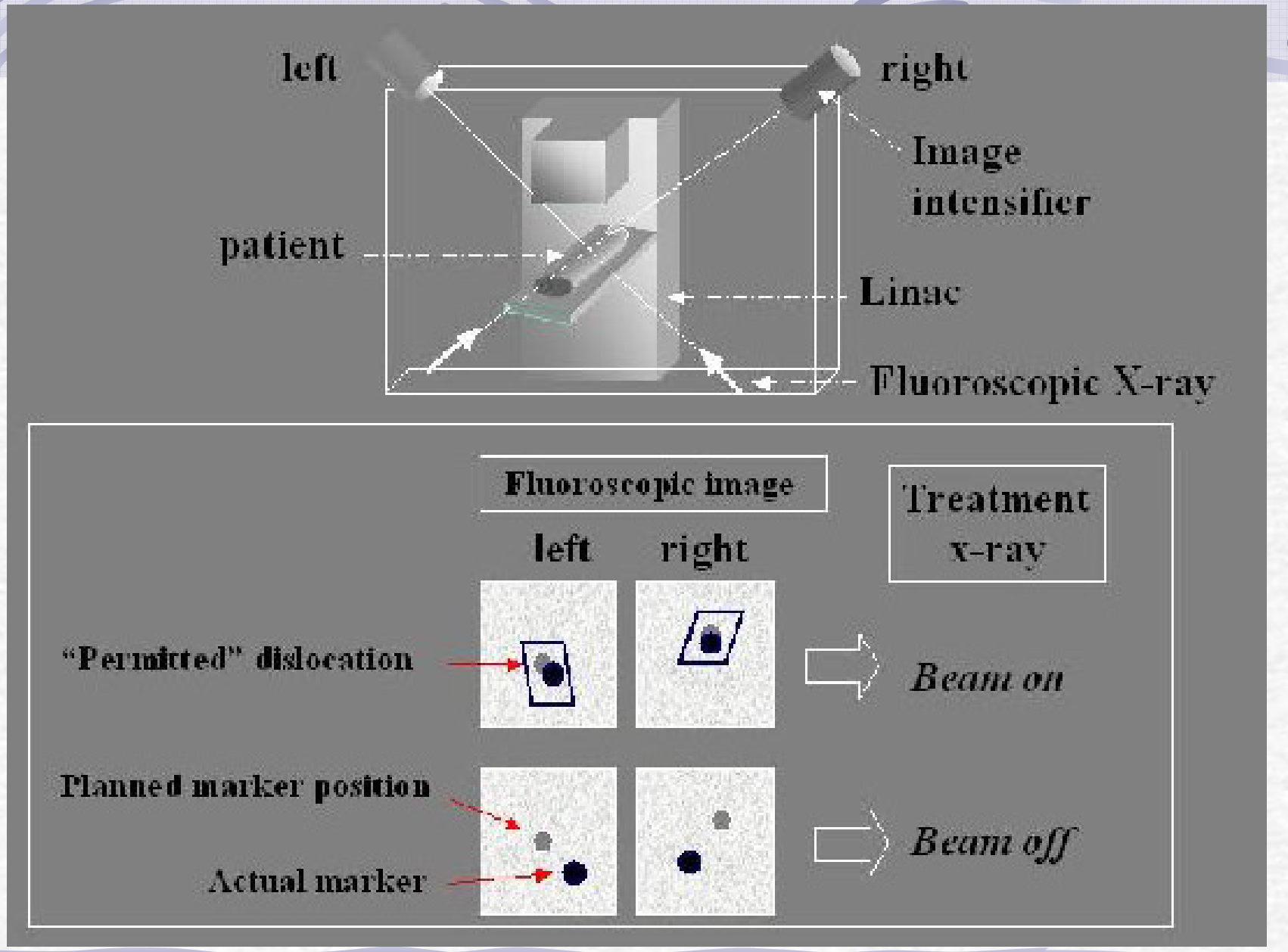
直径: 2mm

http://radi.med.hokudai.ac.jp/~radiotherapy/index.html

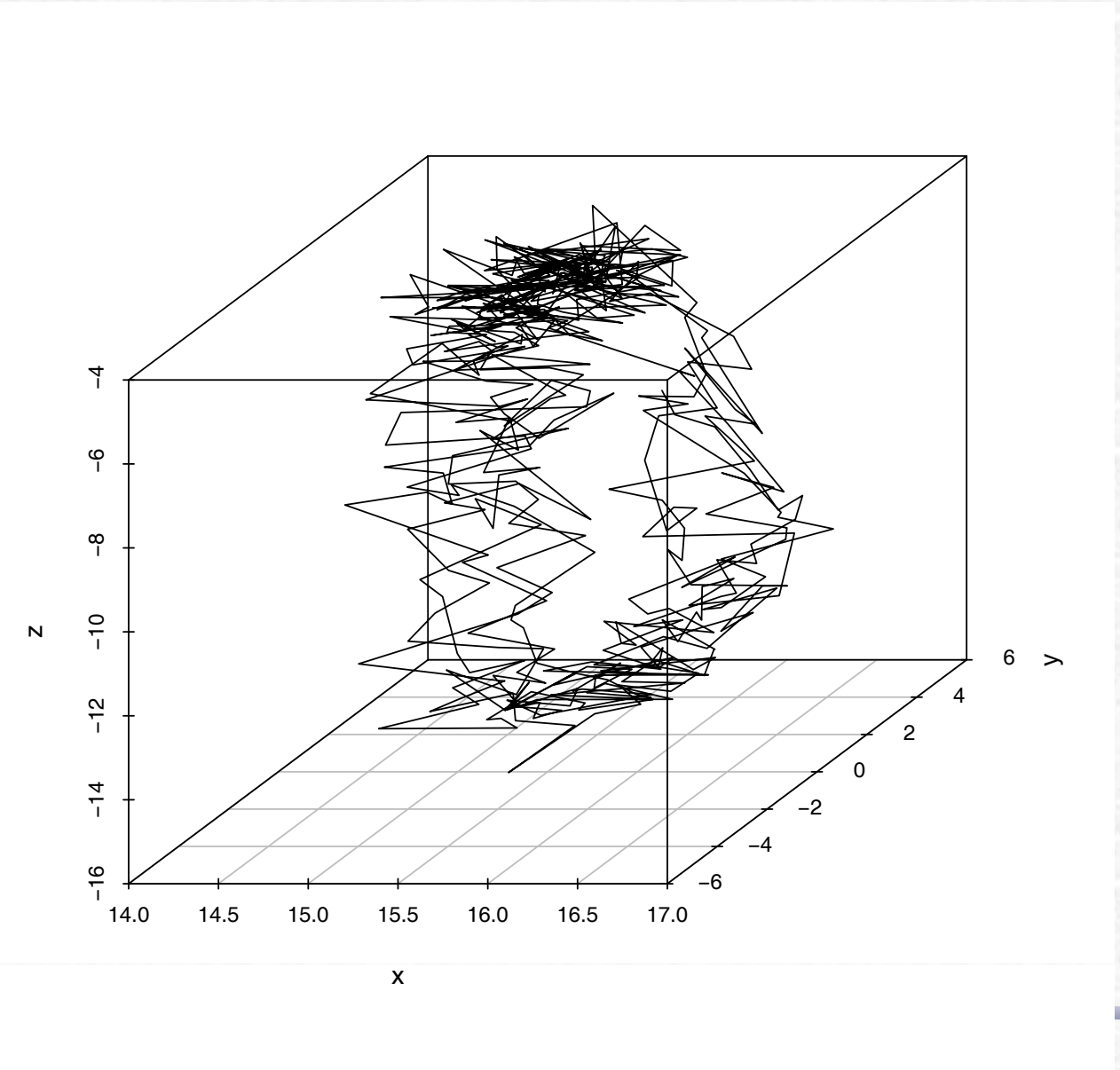
動体追跡照射　Real-Time Tumor-Tracking Radiation Therapy



http://www.yomiuri.co.jp/iryou/medi/saisin/2006121


left
right
Image intensifier
patient
Linac
Fluoroscopic X-ray
Fluoroscopic image
left
right
Treatment x-ray
"Permitted" dislocation
Beam on
Planned marker position
Actual marker
Beam off

Functional Data



When we get Data...

Typical approach of Functional Data Analysis

$$x(t_k) = \{x_i(t_k)\}, \qquad k = 1, 2, \cdots$$

Smoothing

$$x(t) = \{x_i(t)\}, \quad t \in [a, b]$$

Basis Expansion

$$x_i(t) = a_{i1}\phi_1(t) + a_{i2}\phi_2(t) + \cdots + a_{im}\phi_m(t)$$

$$\{(a_{i1}, a_{i2}, \cdots, a_{im})\}$$

List of Contents

- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

What is MDS?

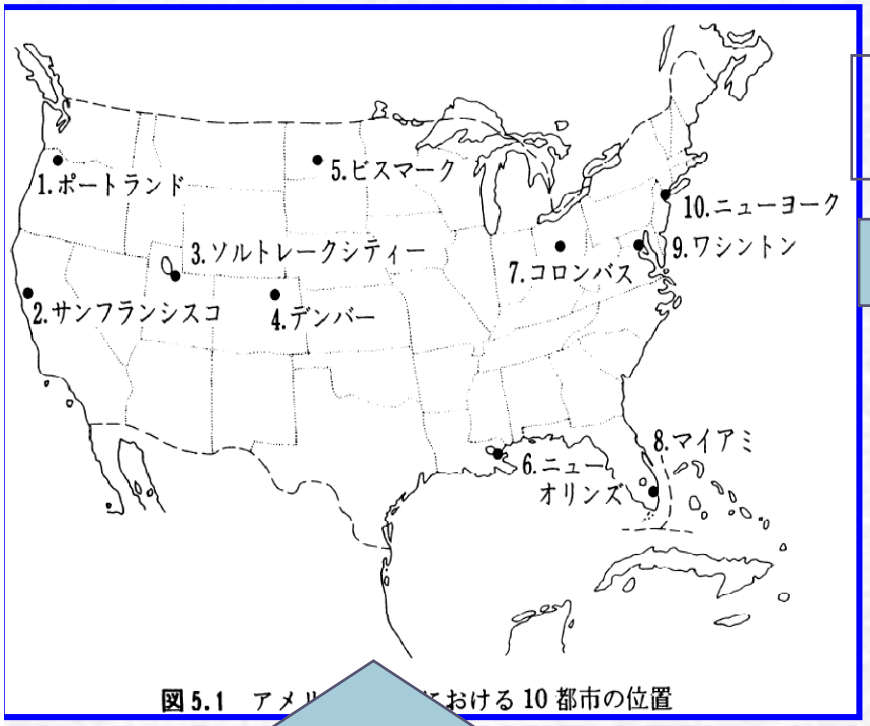


図 5.1　アメリ　　における 10 都市の位置

Euclidean Distances (easy)

表 5.1　アメリカ合衆国の 10 都市の間の距離

		1	2	3	4	5	6	7	8	9	10
1	ポートランド	0.00	8.89	10.16	15.84	16.84	33.01	32.69	43.47	37.78	39.31
2	サンフランシスコ	8.89	0.00	9.64	15.37	20.45	30.75	34.09	41.61	39.26	41.47
3	ソルトレークシティー	10.16	9.64	0.00	6.12	11.20	22.98	24.56	33.61	29.75	31.86
4	デンバー	15.84	15.37	6.12	0.00	8.48	17.20	18.75	27.64	23.91	26.23
5	ビスマーク	16.84	20.45	11.20	8.48	0.00	20.77	16.25	29.52	21.17	22.49
6	ニューオリンズ	33.01	30.75	22.98	17.20	20.77	0.00	13.01	10.86	15.61	19.06
7	コロンバス	32.69	34.09	24.56	18.75	16.25	13.01	0.00	15.99	5.19	7.69
8	マイアミ	43.47	41.61	33.61	27.64	29.52	10.86	15.99	0.00	14.84	17.68
9	ワシントン	37.78	39.26	29.75	23.91	21.17	15.61	5.19	14.84	0.00	3.45
10	ニューヨーク	39.31	41.47	31.86	26.23	22.49	19.06	7.69	17.68	3.45	0.00

（単位　100 km）

MDS

Multidimensional Scaling (MDS)

Data Points $\{\mathbf{x}_i\} \in \mathbf{R}^p$

$$d_{ij} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\| \ (i, j = 1,2,\cdots,n)$$

Distances among Data Points $\{d_{ij}\} \in \mathbf{R}$

Dissimilarities among Objects $\{s_{ij}\} \in \mathbf{R}$

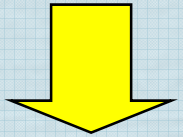
Data Points $\{\mathbf{x}_i\} \in \mathbf{R}^p$

Configuration

Functional MDS

$\{s_{ij}(t)\}$

Dissimilarities among Objects $\{s_{ij}\} \in \mathbf{R}$



$\{x_i(t)\}$

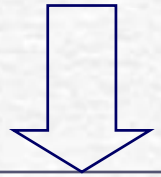
Data Points $\{\mathbf{x}_i\} \in \mathbf{R}^p$

Configuration

Multidimensional Scaling (MDS)

Functional MDS

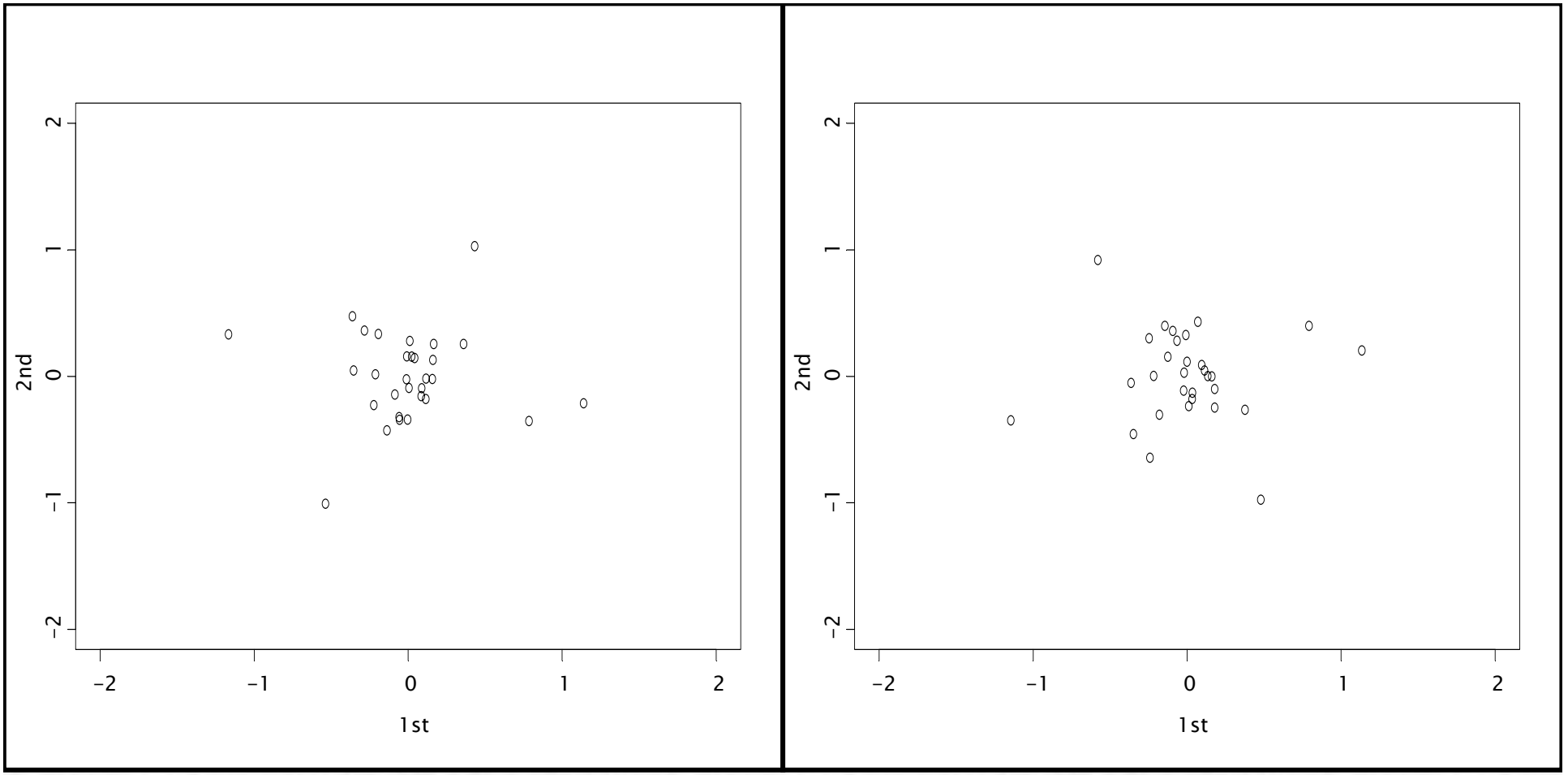
$$S(t_1), S(t_2), \cdots, S(t_m)$$



Conventional MDS

$$\mathrm{X}(t_1), \mathrm{X}(t_2), \cdots, \mathrm{X}(t_m)$$

$$d_{ij}(t) = \left\| \mathbf{x}_i(t) - \mathbf{x}_j(t) \right\| \approx s_{ij}(t)$$



$t = t_0$

$t = t_0 + \delta$

Functional MDS

$$S(t_1), S(t_2), \cdots, S(t_m)$$

$$\mathrm{X}(t_1), \mathrm{X}(t_2), \cdots, \mathrm{X}(t_m)$$

Orthogonal Matrix

$$\mathrm{Q}(t_1)\mathrm{X}(t_1), \quad \mathrm{Q}(t_2)\mathrm{X}(t_2), \cdots, \mathrm{Q}(t_m)\mathrm{X}(t_m)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\left\|\mathbf{x}_i(t_{k-1}) - \mathrm{Q}(t_k)\mathbf{x}_i(t_k)\right\|^2 \rightarrow \min$$

Functional MDS

$$S(t_1), S(t_2), \cdots, S(t_m)$$

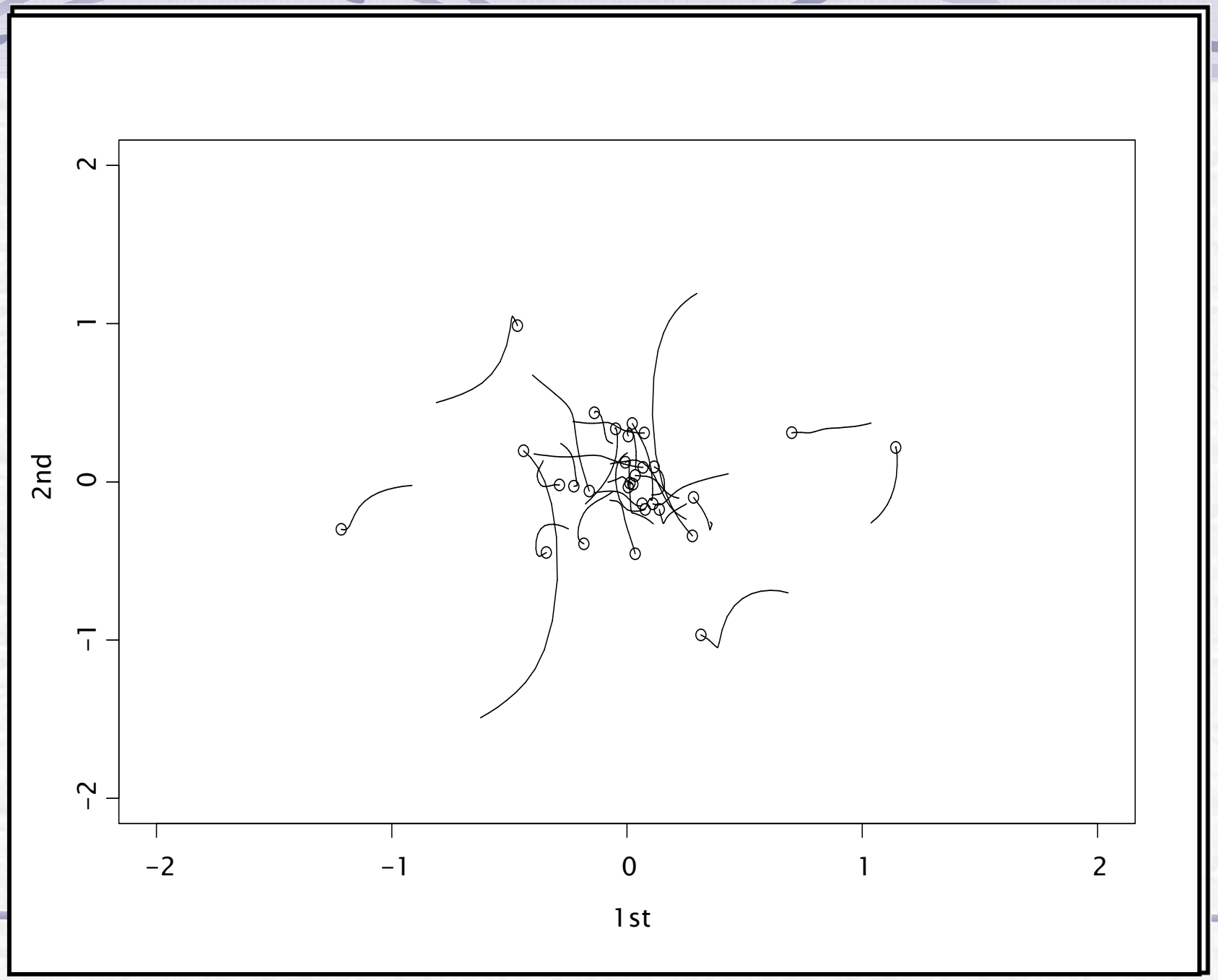
Conventional MDS

$$\mathrm{X}(t_1), \mathrm{X}(t_2), \cdots, \mathrm{X}(t_m)$$

$$\mathrm{Q}(t_1)\mathrm{X}(t_1), \quad \mathrm{Q}(t_2)\mathrm{X}(t_2), \cdots, \mathrm{Q}(t_m)\mathrm{X}(t_m)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\left\|\mathbf{x}_i(t_{k-1}) - \mathrm{Q}(t_k)\mathbf{x}_i(t_k)\right\|^2 \rightarrow \min$$

$$\mathrm{Q}(t_k) \equiv \left(\mathrm{X}(t_k)^{\mathrm{T}}\mathrm{X}(t_{k-1})\mathrm{X}(t_{k-1})^{\mathrm{T}}\mathrm{X}(t_k)\right)^{-1/2}\mathrm{X}(t_{k-1})^{\mathrm{T}}\mathrm{X}(t_k)$$


2nd
2
1
0
−1
−2
−2
−1
0
1
2
1st

Example Data 1

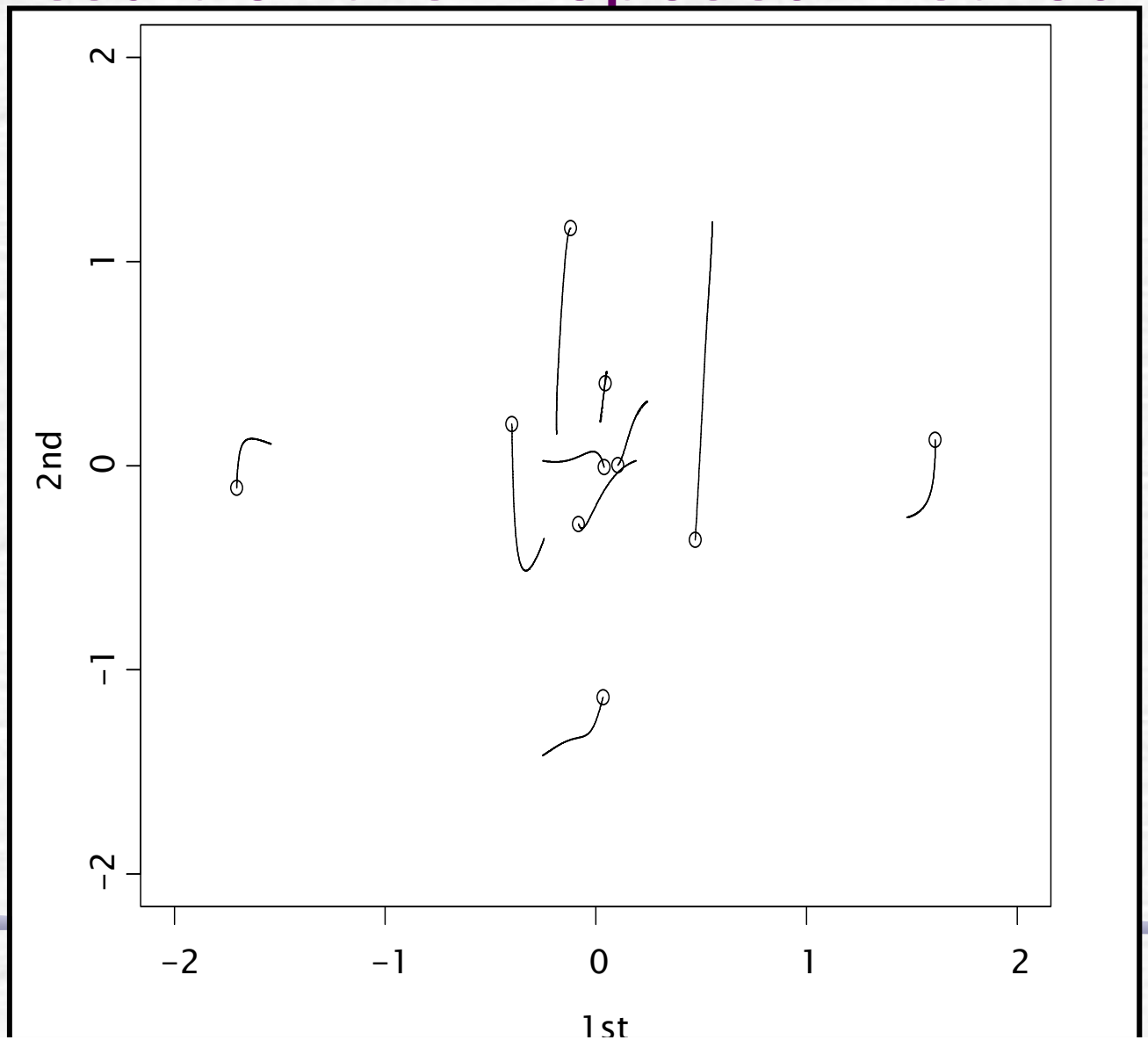
$$s_{ij}(t) = s_{ij} + td_{ij}$$

$$s_{ij} \sim N(const, 1)$$

$$d_{ij} \sim N(const, 0.01^2)$$

Sample Size: 10

Result of the Proposed Method



Random Walk Data

$$s_{ij}(0) = s_{ij}$$

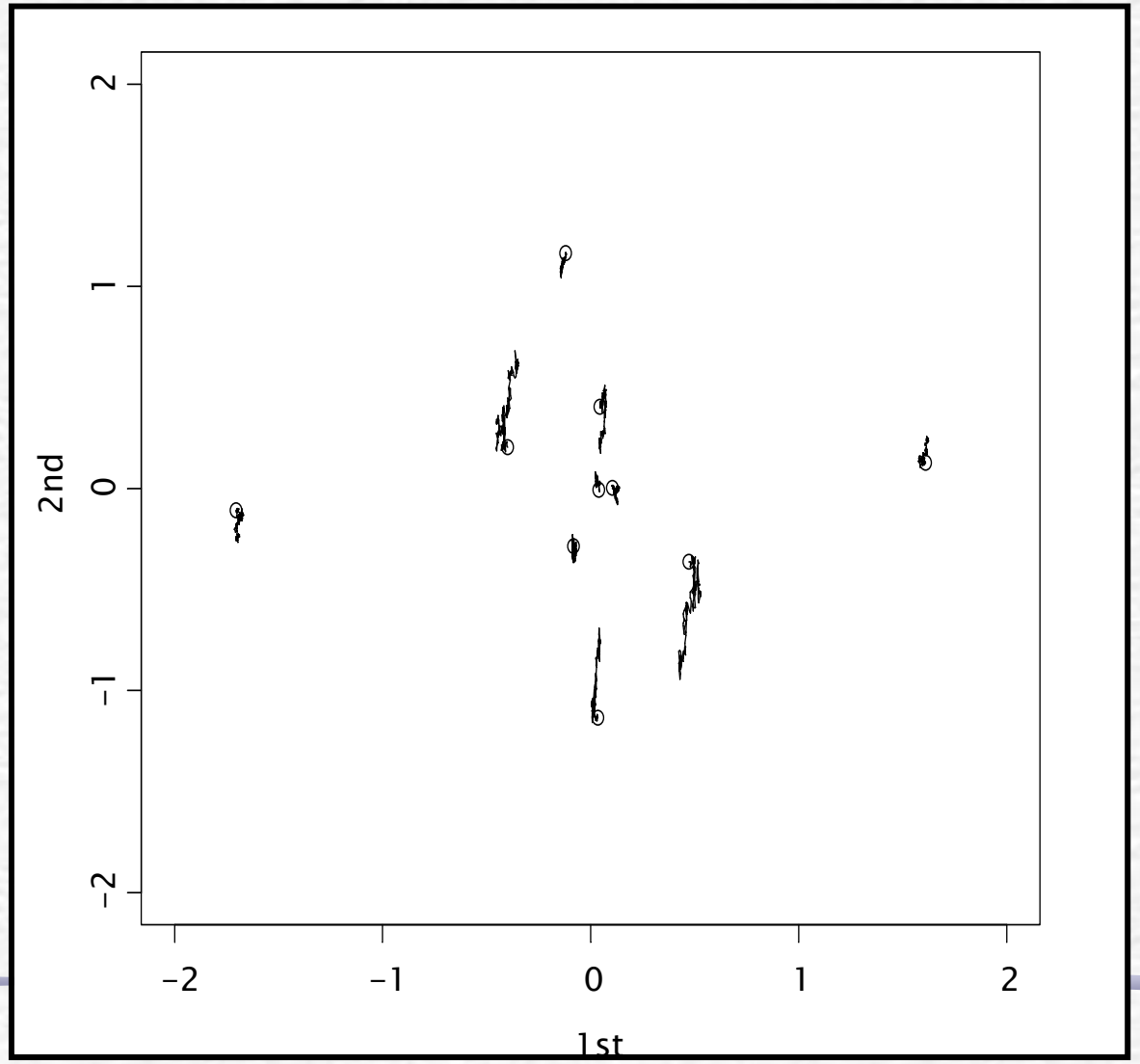
$$s_{ij}(t+1) = s_{ij}(t) + e_{ijt}$$

$$e_{ijt} \sim N(0, \delta^2)$$

Sample Size:10

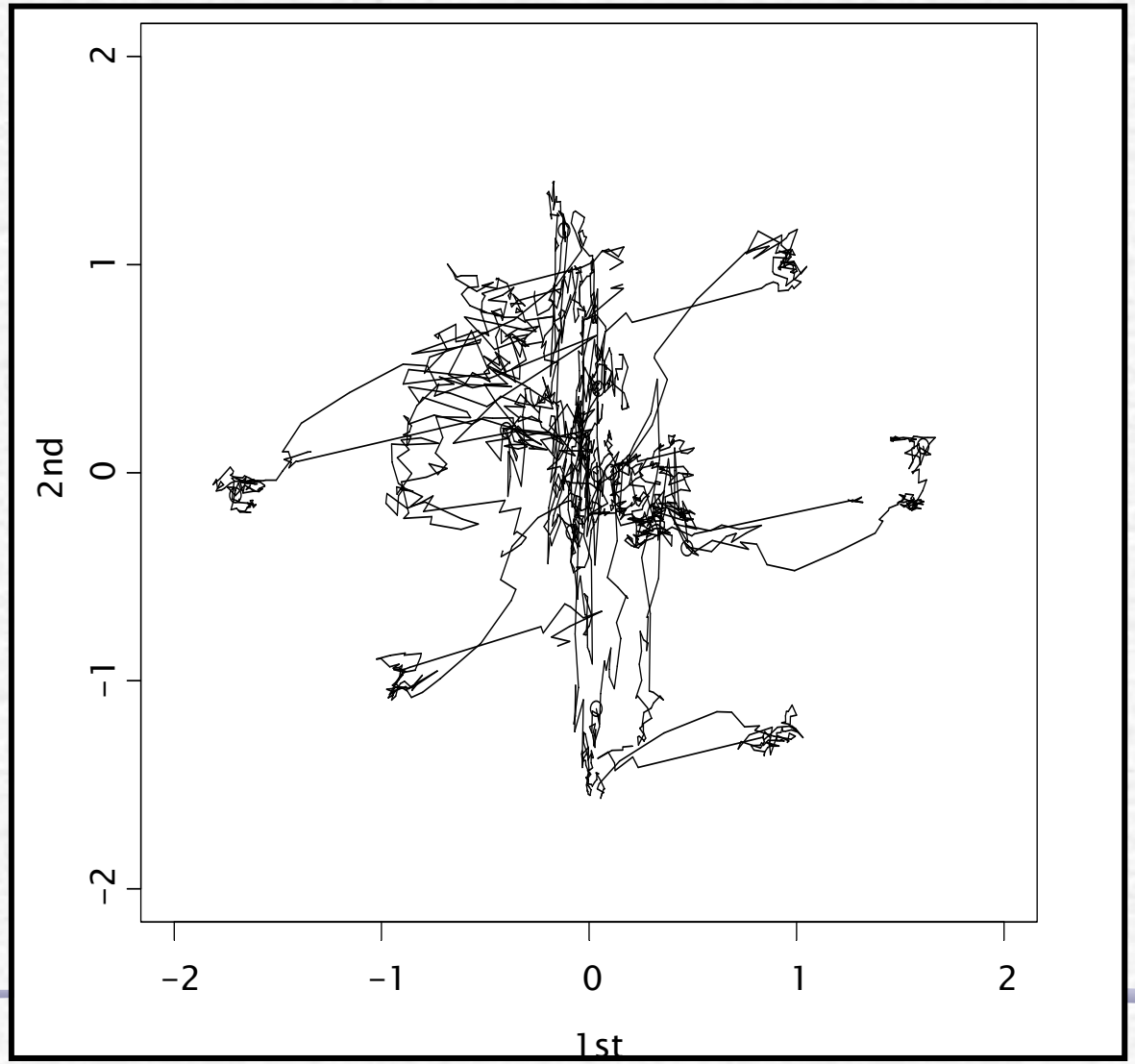
Random Walk Data

δ=0.01



Random Walk Data

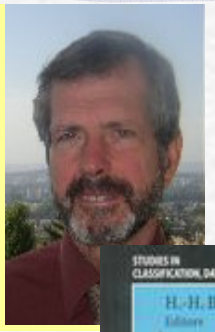
$\delta = 0.1$



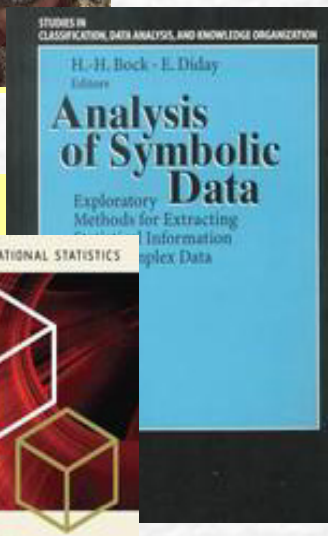
List of Contents

- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

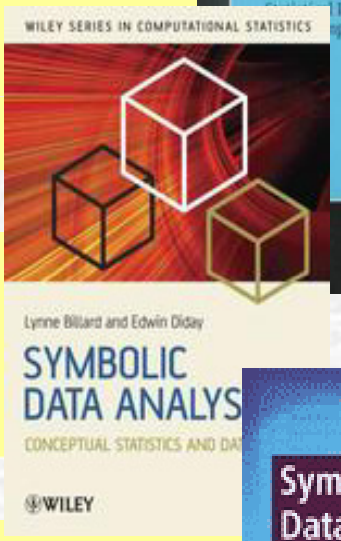
E.Diday (1987)
The symbolic approach in clustering and related methods of data analysis, *IFCS-87*



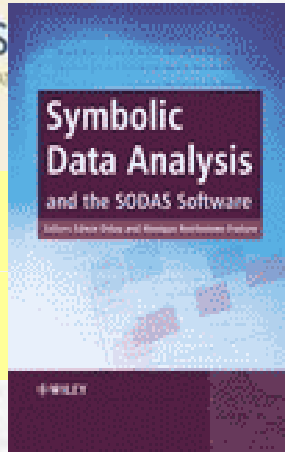
H.Bock and E.Diday eds.(2000)
Analysis of Symbolic Data, springer



L.Billard and E.Diday(2007)
Symbolic Data Analysis, Wiley



E.Diday and M.Noirhomme-Fraiture(2008)
Symbolic Data Analysis and SODAS, Wiley



- 1st level units

  Individuals

- 2nd level units

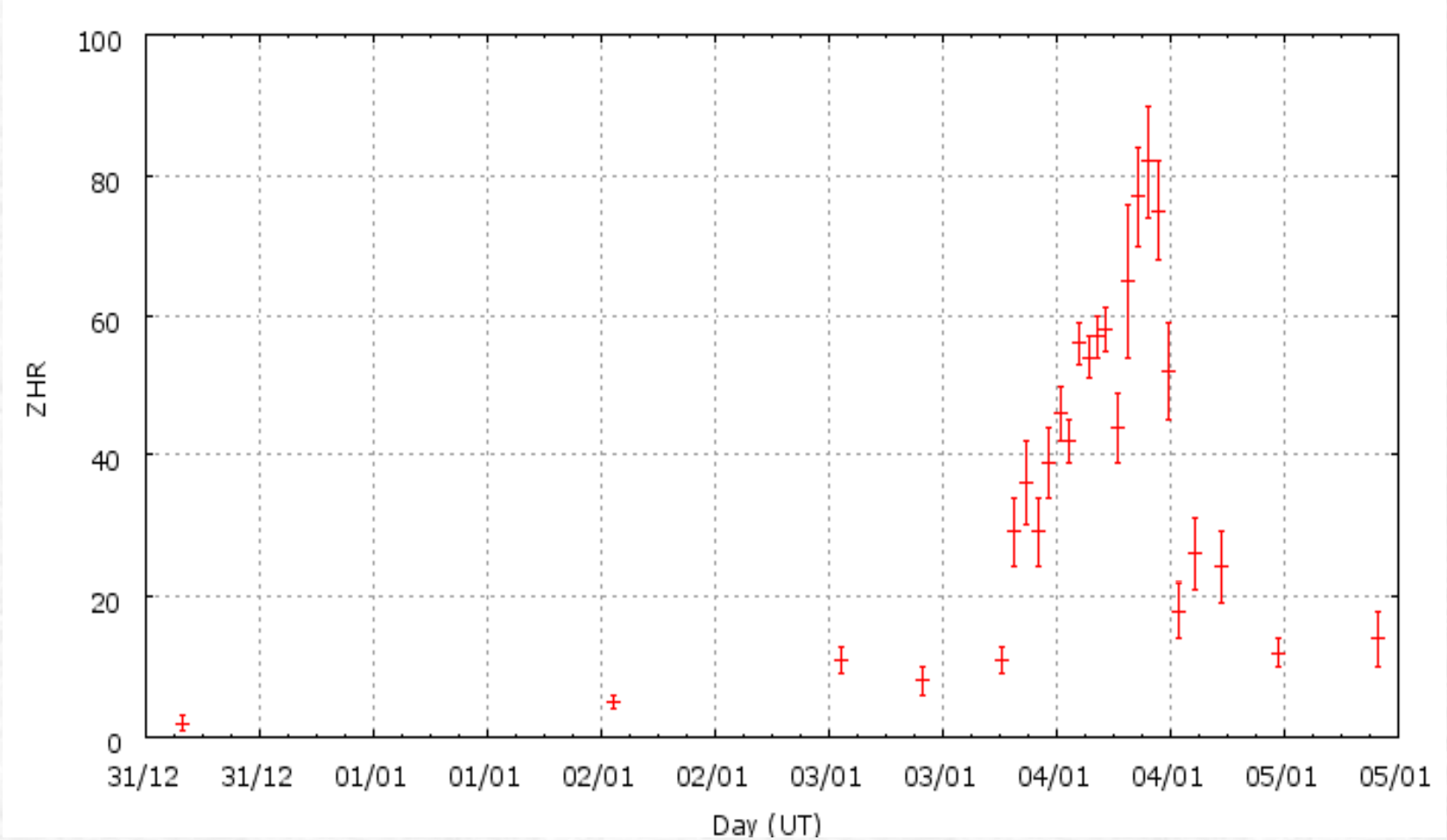
  Classes: a set of individuals (the set of 100 ostriches)

  Categories: with a categorical variables (the category ‘ostrich’ for the variable ‘species’)

  Concept: extent: a set of individuals which satisfy the concept. Intent: a way to find the extent (the set of all ostriches which exist, have existed or will exist)

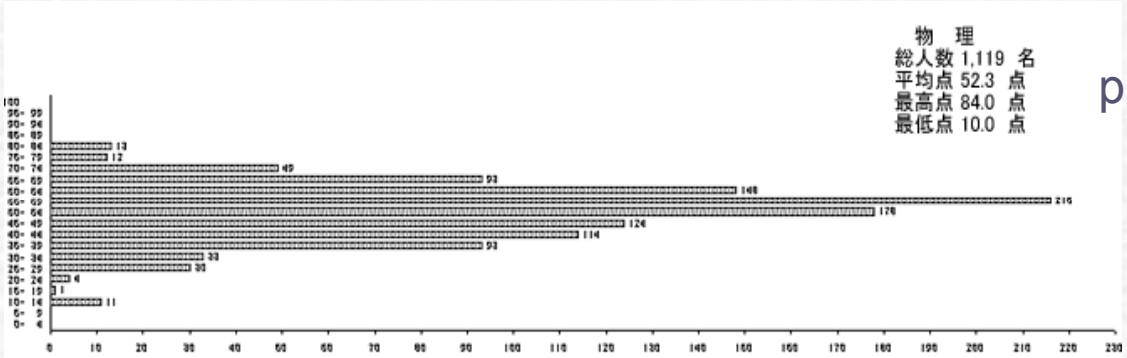
ostrich

Example of Symbolic Data (Interval)

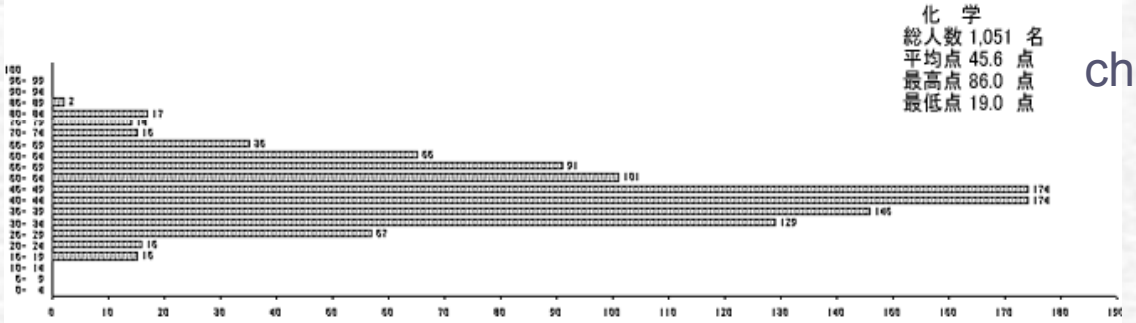


http://www.imo.net/live/quadrantids2008/

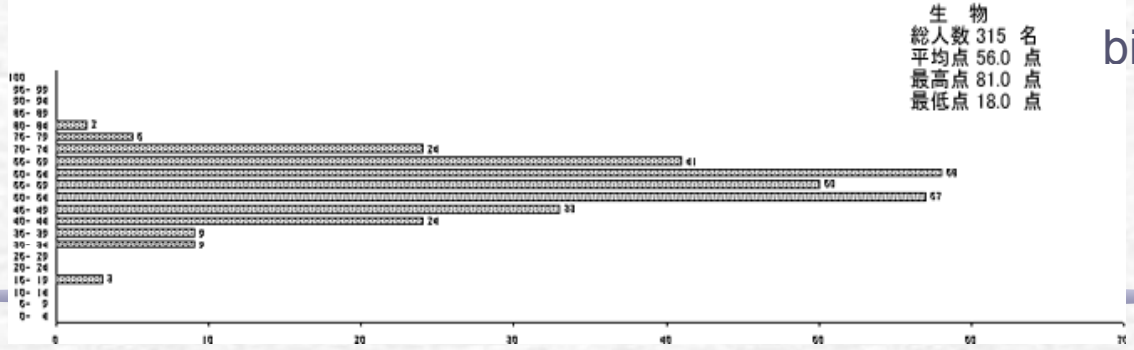
Example of Symbolic Data (distribution)



physics



chemistry

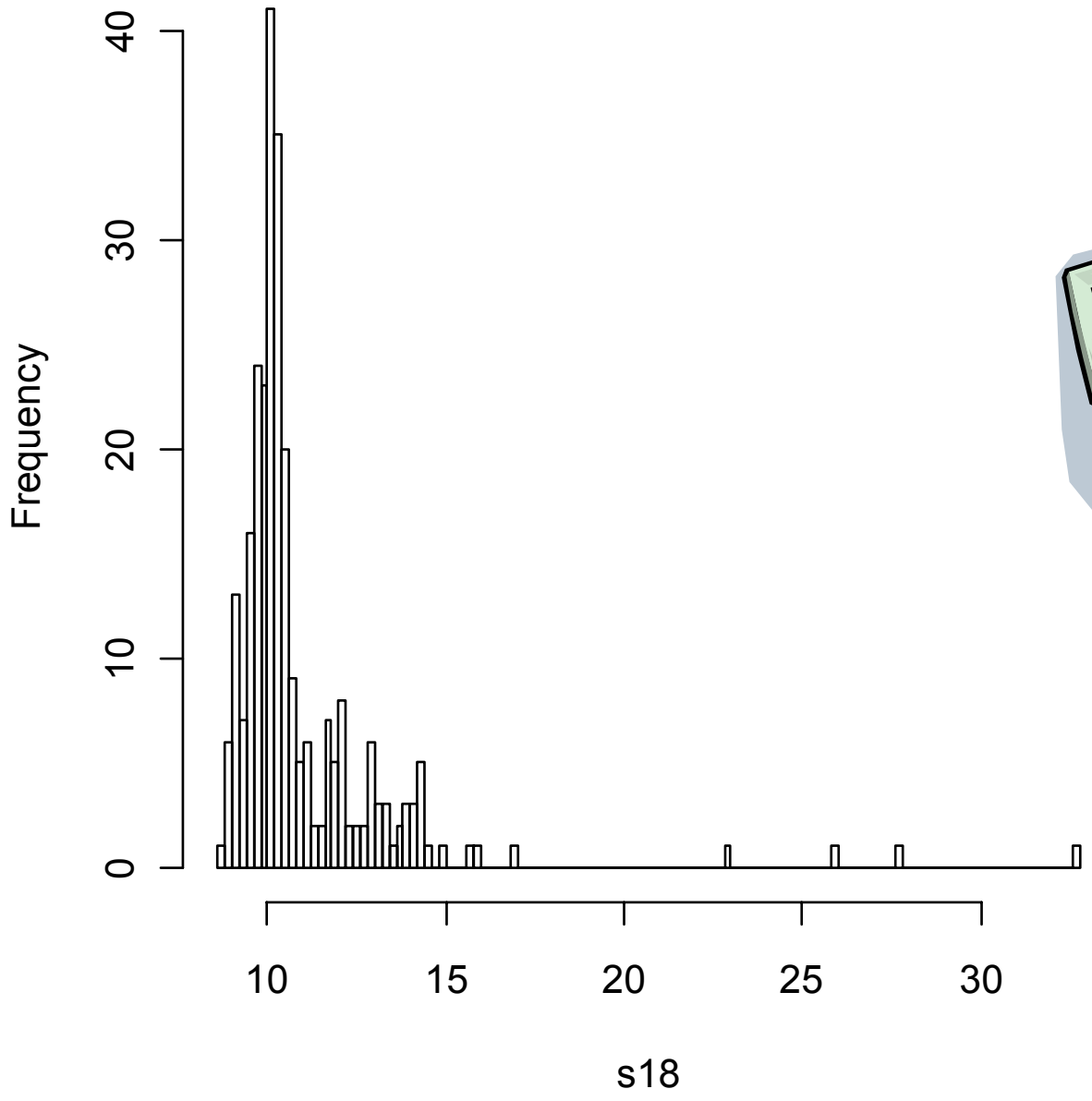


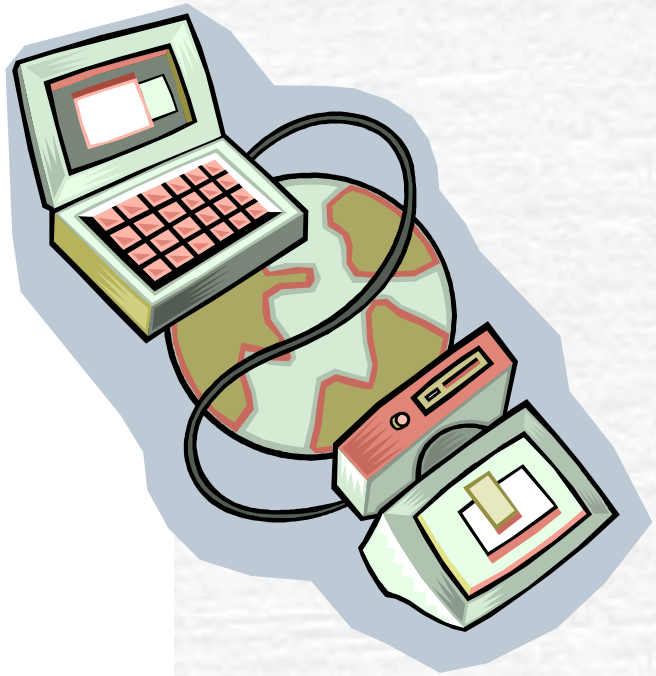
biology

http://www.jasso.go.jp/eju/result_1401_04.html

List of Contents

- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks


Histogram of s18max MTU=1500
Frequency
0
10
20
30
40
10
15
20
25
30
s18



List of Contents

- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

What is MDS?

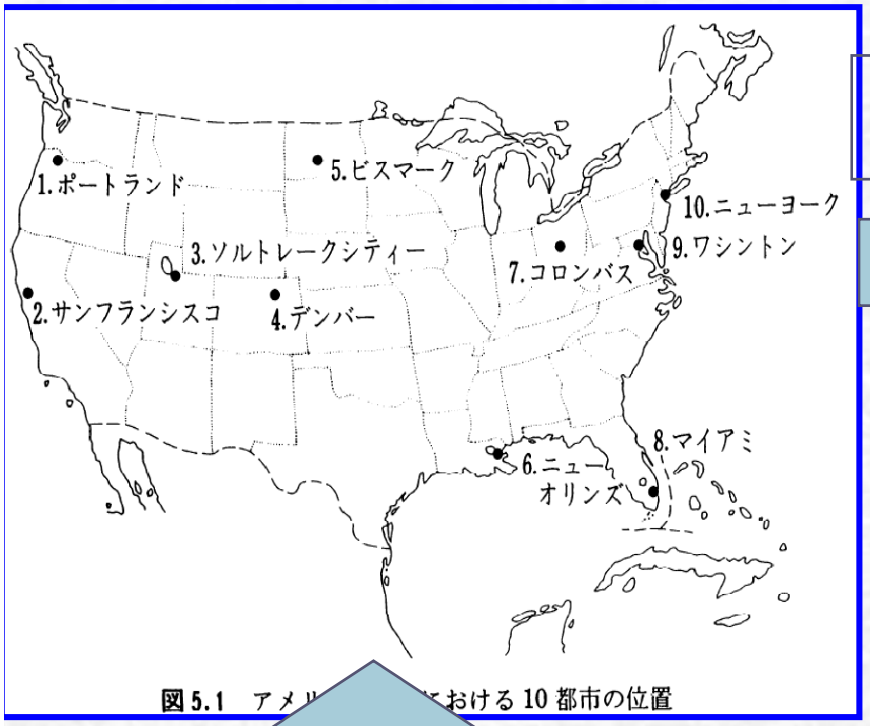


図 5.1　アメリ　　における 10 都市の位置

Euclidean Distances (easy)

表 5.1　アメリカ合衆国の 10 都市の間の距離

		1	2	3	4	5	6	7	8	9	10
1	ポートランド	0.00	8.89	10.16	15.84	16.84	33.01	32.69	43.47	37.78	39.31
2	サンフランシスコ	8.89	0.00	9.64	15.37	20.45	30.75	34.09	41.61	39.26	41.47
3	ソルトレークシティー	10.16	9.64	0.00	6.12	11.20	22.98	24.56	33.61	29.75	31.86
4	デンバー	15.84	15.37	6.12	0.00	8.48	17.20	18.75	27.64	23.91	26.23
5	ビスマーク	16.84	20.45	11.20	8.48	0.00	20.77	16.25	29.52	21.17	22.49
6	ニューオリンズ	33.01	30.75	22.98	17.20	20.77	0.00	13.01	10.86	15.61	19.06
7	コロンバス	32.69	34.09	24.56	18.75	16.25	13.01	0.00	15.99	5.19	7.69
8	マイアミ	43.47	41.61	33.61	27.64	29.52	10.86	15.99	0.00	14.84	17.68
9	ワシントン	37.78	39.26	29.75	23.91	21.17	15.61	5.19	14.84	0.00	3.45
10	ニューヨーク	39.31	41.47	31.86	26.23	22.49	19.06	7.69	17.68	3.45	0.00

（単位　100 km）

MDS

Multidimensional Scaling (MDS)

Data Points $\{\mathbf{x}_i\} \in \mathbf{R}^p$

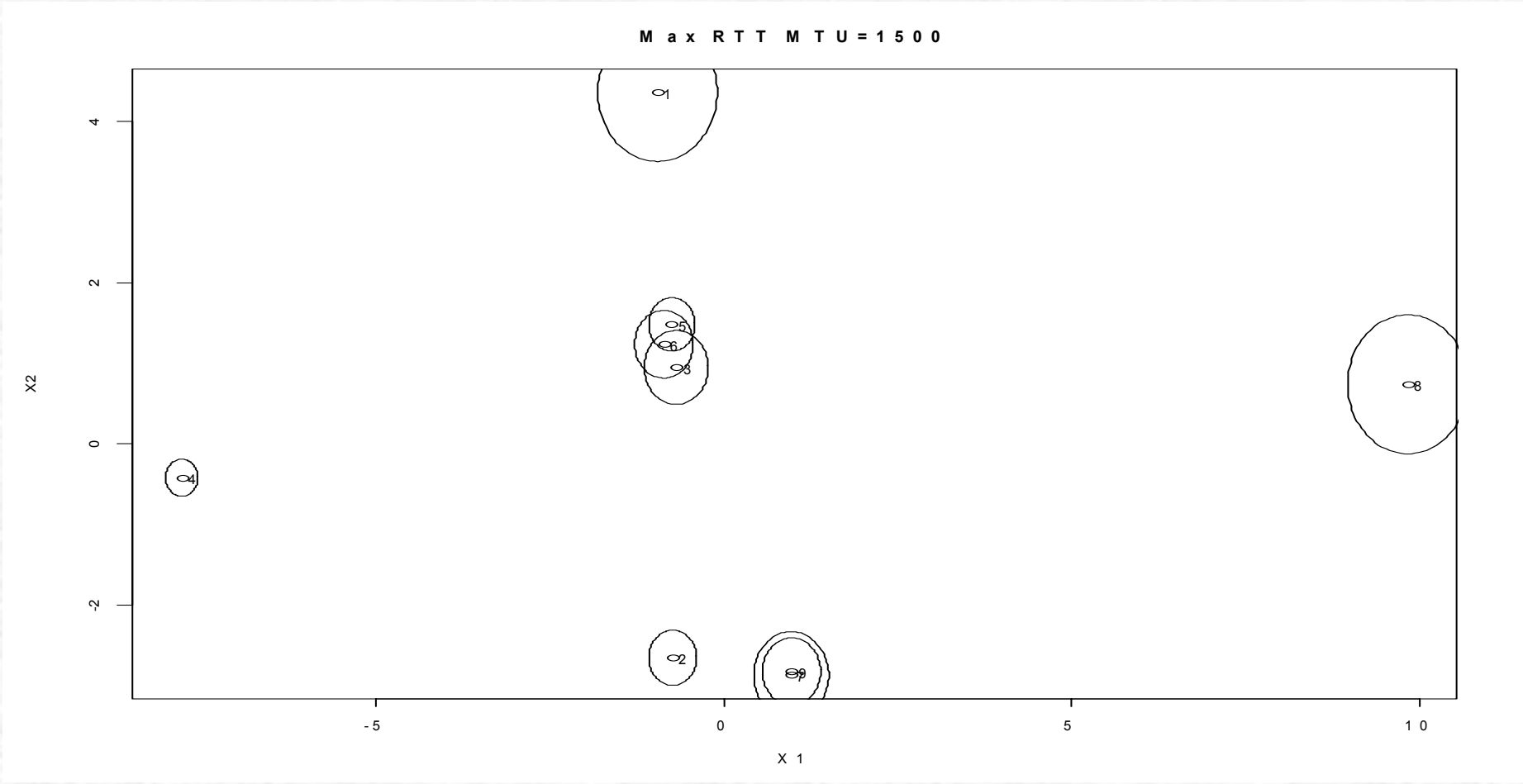
$$d_{ij} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\| \ (i, j = 1,2,\cdots,n)$$

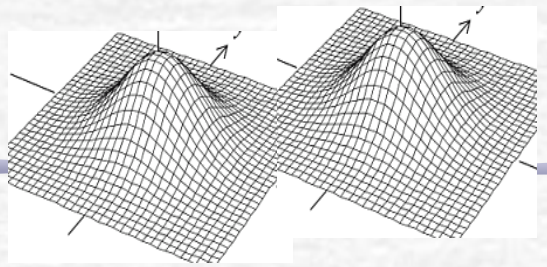
Distances among Data Points $\{d_{ij}\} \in \mathbf{R}$

Dissimilarities among Objects $\{s_{ij}\} \in \mathbf{R}$

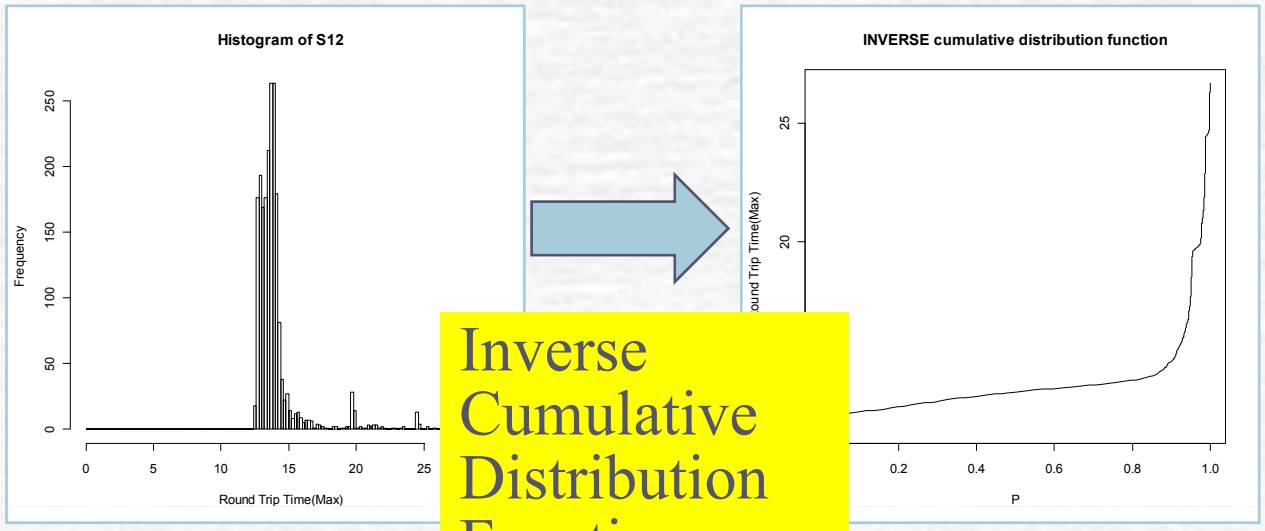
Data Points $\{\mathbf{x}_i\} \in \mathbf{R}^p$

Configuration


Max RTT MTU=1500
X2
X1
4
2
0
-2
-5
0
5
10
1
2
3
4
5
6
7
8
9

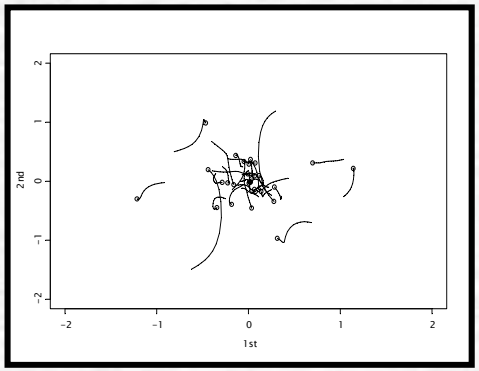


Outline of our study



Inverse Cumulative Distribution Function

Functional MDS



Distribution Valued Dissimilarity Data

s_{ij}

Symbolic Data Analysis

Functional Dissimilarity Data

$s_{ij}(P)$

Functional Data Analysis

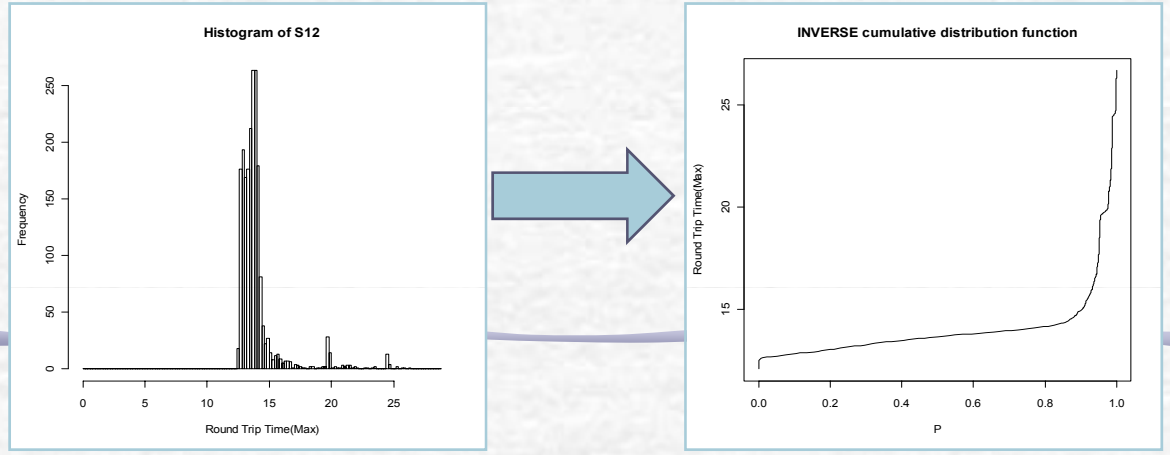
MDS for Distribution Valued Dissimilarity Data

Data:

Dissimilarity $S = \{s_{ij}; i, j = 1, \cdots, n\}$

s_{ij}:Distribution

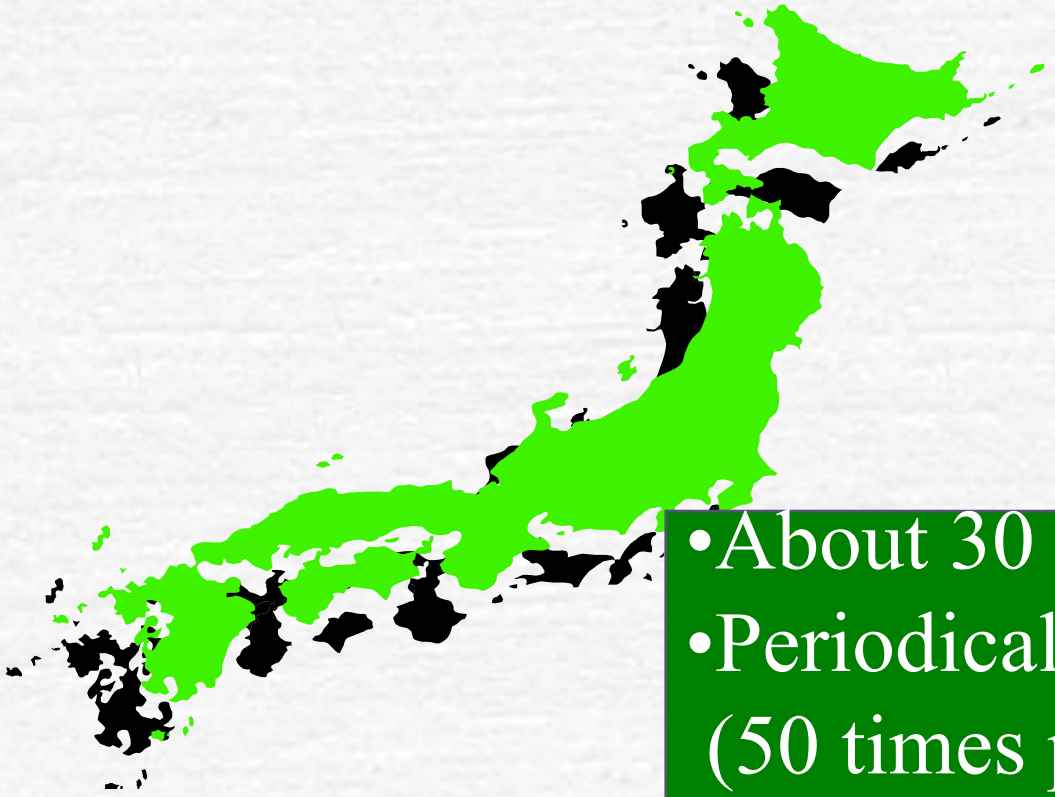
- s_{ij}→Inverse Cumulative Distribution Function
- Apply Functional MDS



List of Contents

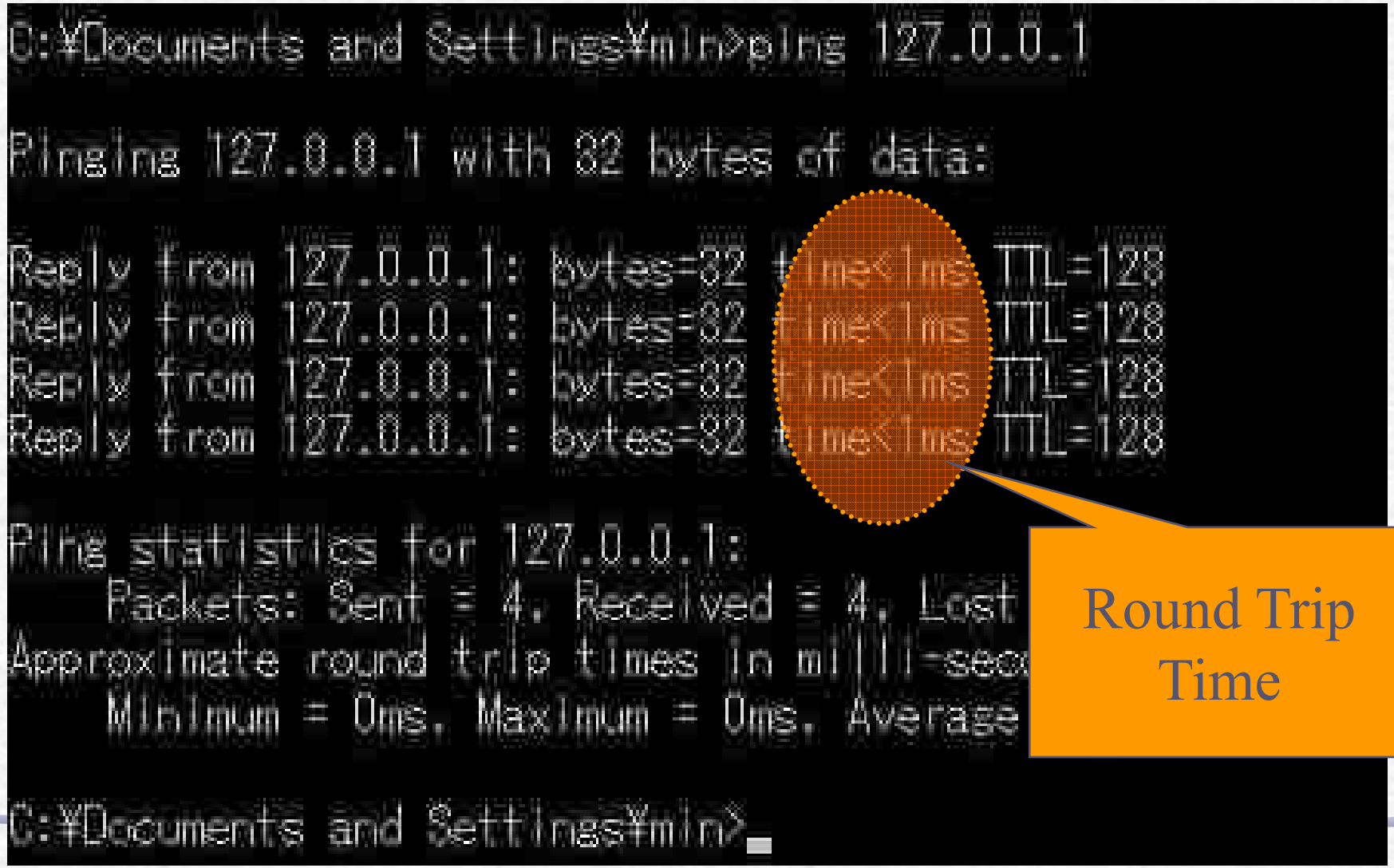
- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

Resilient Project



- About 30 hosts
- Periodical ‘ping’
  (50 times per every 5 min.)
- Store all data to
  Database server

‘ping’ on WindowsXP



Raw data (example)

200511122229060200000000001,200511111922240000001816910,
200410011542000197,200410011542000197,2005-11-1,
00:02:07,1500,0.132000,0.185000,0.155000,0,0.016000
200511122229060200000000002,200511111922240000001816910,
200410011542000197,200410011542000197,2005-11-01,
00:07:08,1500,0.105000,0.199000,0.150000,0,0.019000
200511122229060200000000003,200511111922240000001816910,
200410011542000197,200410011542000197,2005-11-01,
00:12:08,1500,0.107000,0.208000,0.137000,0,0.025000
……

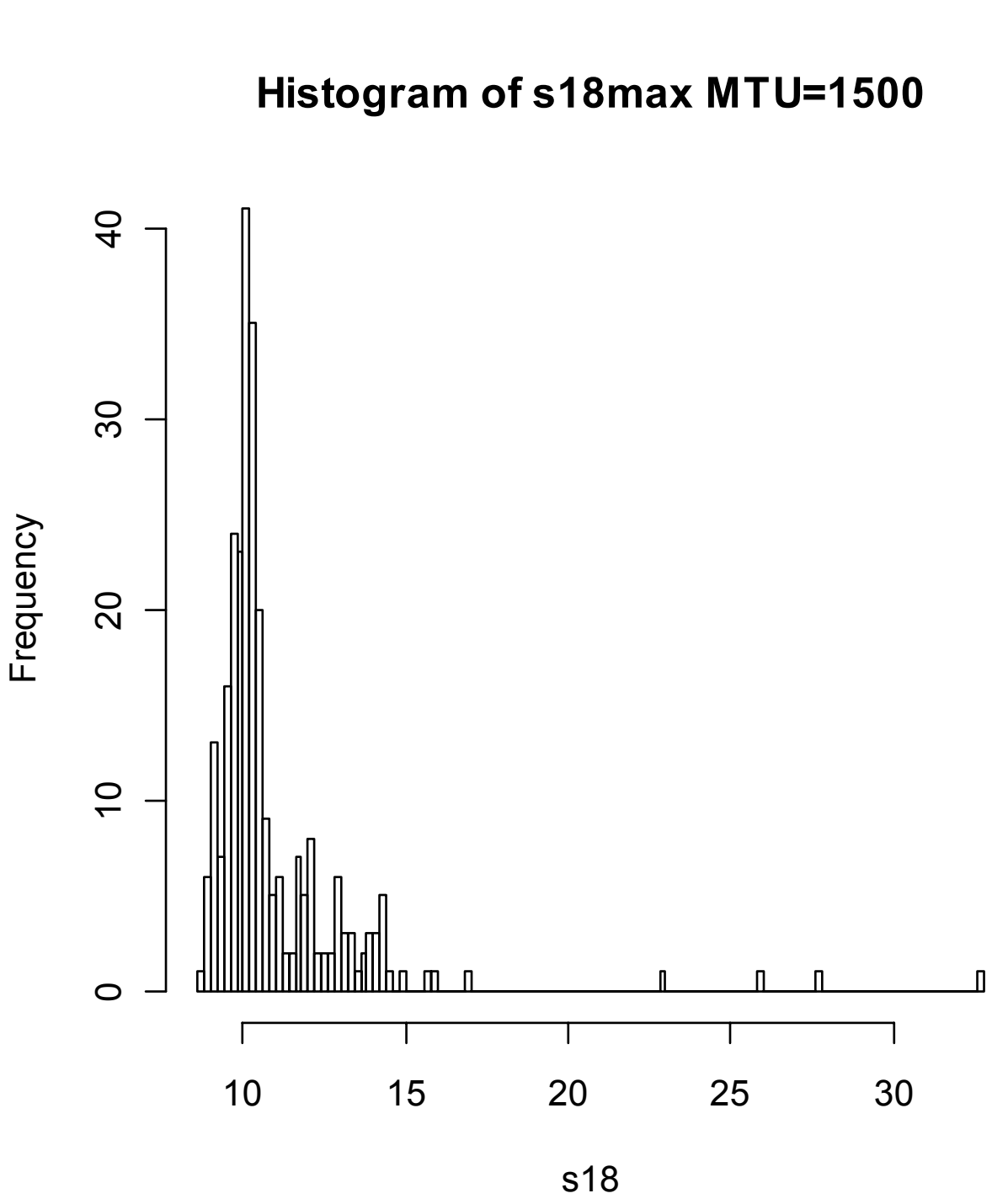
- Source
- Destination
- Start Date, Time

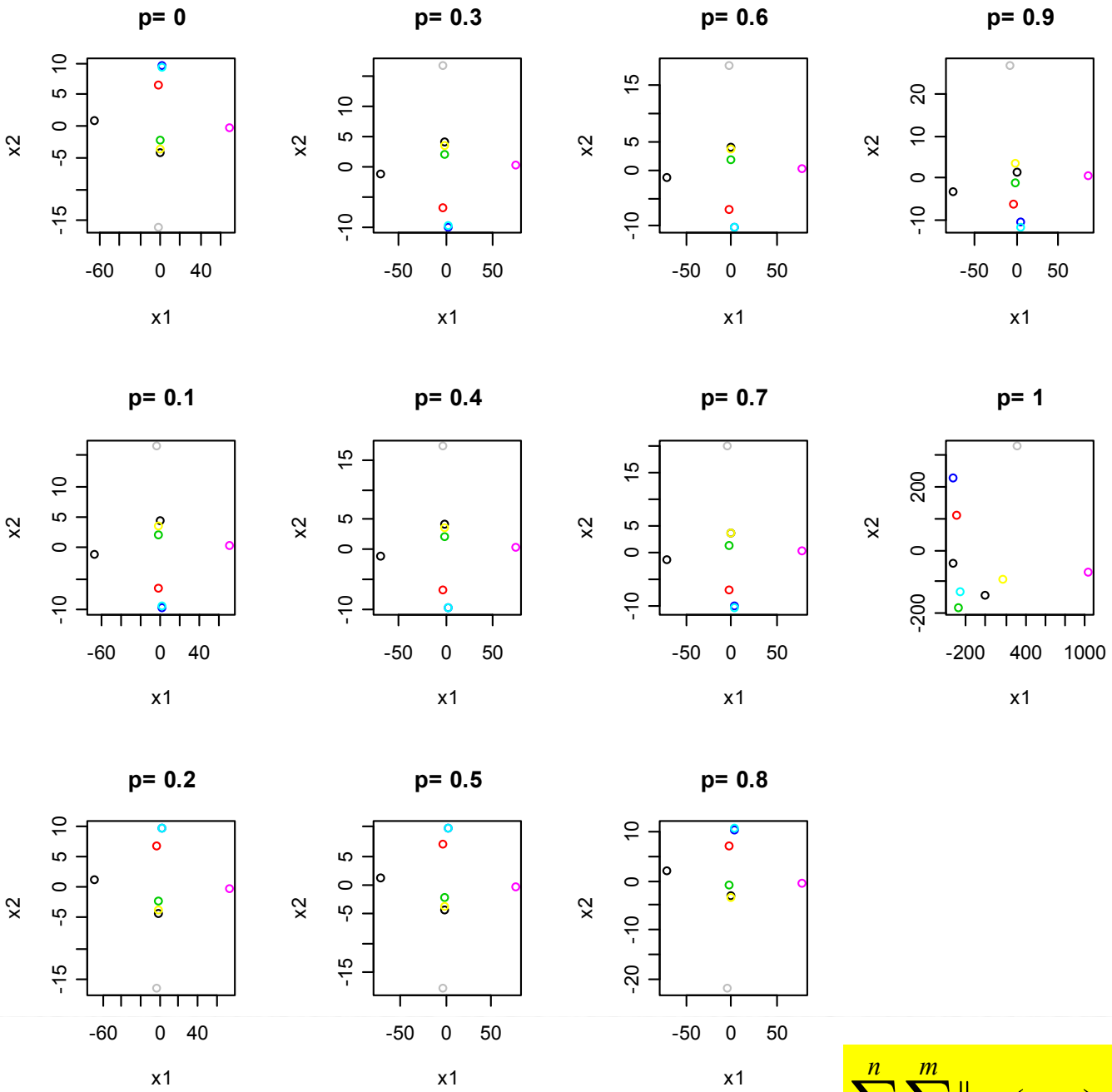
- Minimum RTT
- Ave. RTT
- Maximum RTT
- Loss Ratio

Measured Data

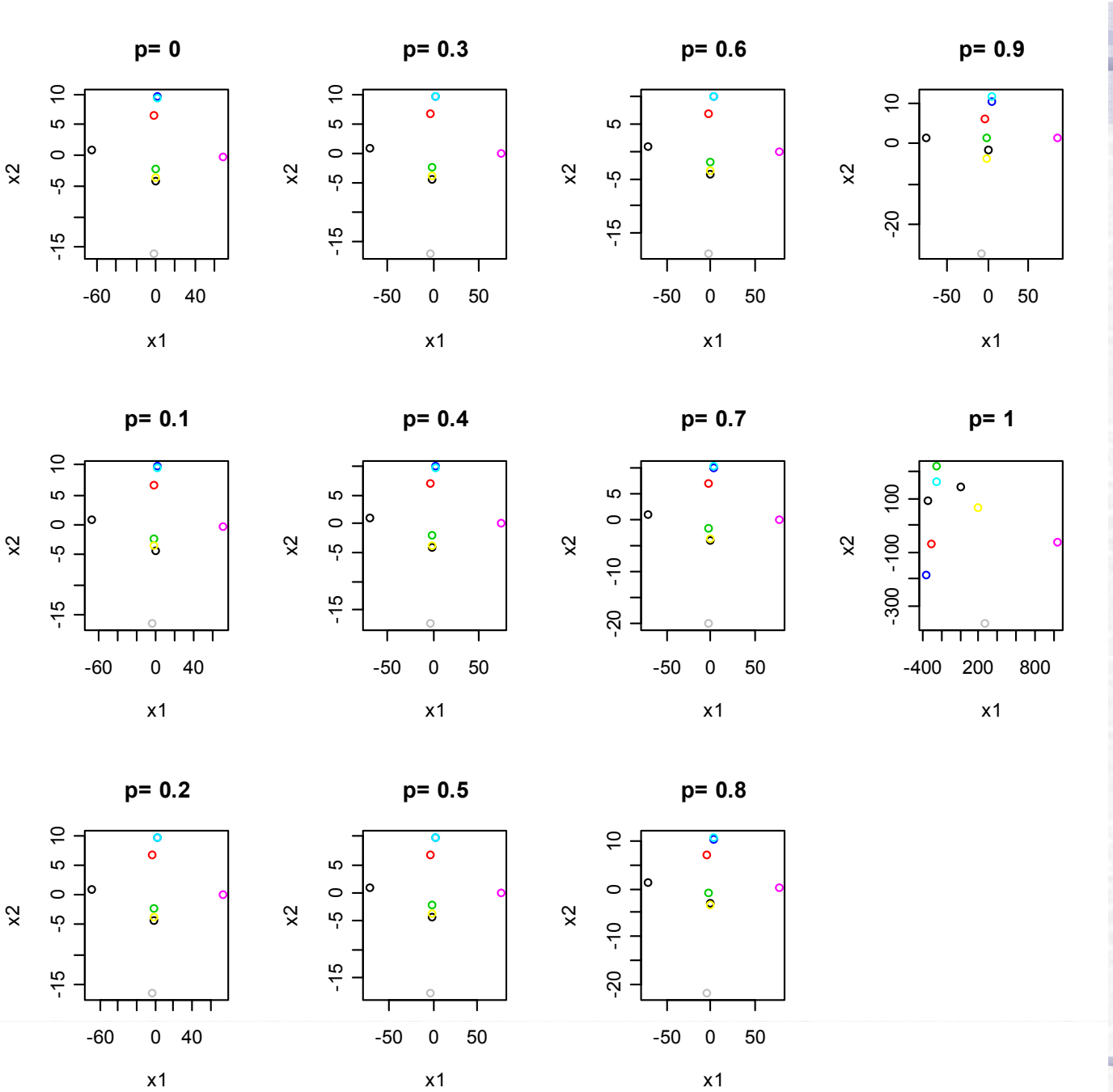
- From Nov.1,2005 to Jan.14,2006
- 50 Ping packets every 5 min.
- Minimum / Average / Maximum Round Trip Time (RTT)
- Number of Hosts: 35 (by IP-Address)

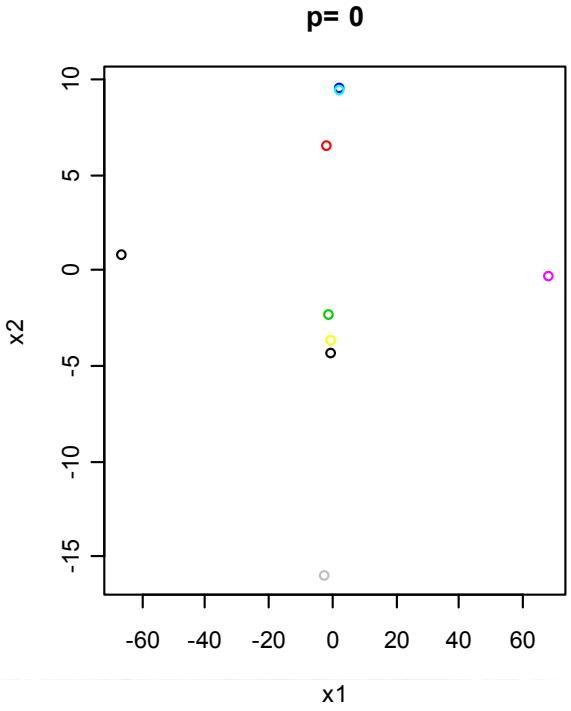
- For example, Jan.10, 2006
- Min. / Aver. / Max. (RTT)
- # of Hosts: 9, 1728 times for any pairs

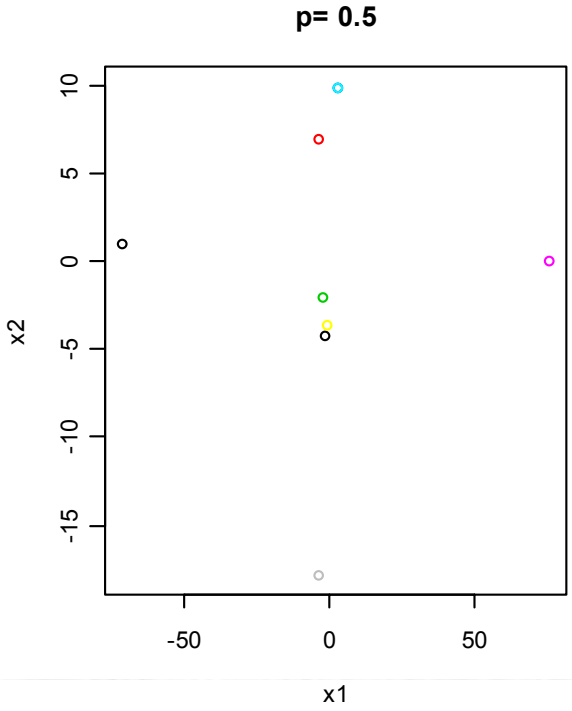

Histogram of s18max MTU=1500
Frequency
0
10
20
30
40
10
15
20
25
30
s18

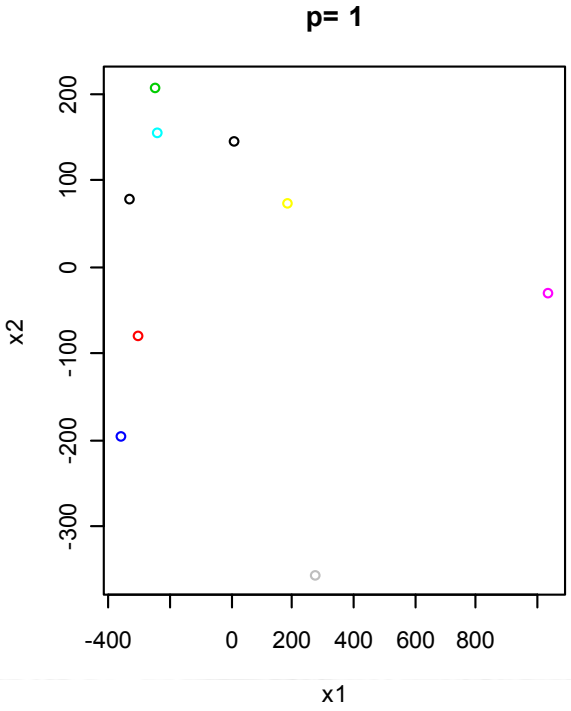


$$\sum_{i=1}^{n}\sum_{k=1}^{m}\left\|\mathbf{x}_i(t_{k-1})-\mathrm{Q}(t_k)\mathbf{x}_i(t_k)\right\|^2 \rightarrow \min$$


p= 0
p= 0.3
p= 0.6
p= 0.9
p= 0.1
p= 0.4
p= 0.7
p= 1
p= 0.2
p= 0.5
p= 0.8
x1
x2


p= 0
x2
10
5
0
-5
-10
-15
-60
-40
-20
0
20
40
60
x1


p= 0.5
x2
10
5
0
-5
-10
-15
-50
0
50
x1


p= 1
x2
200
100
0
-100
-200
-300
-400
0
200
400
600
800
x1

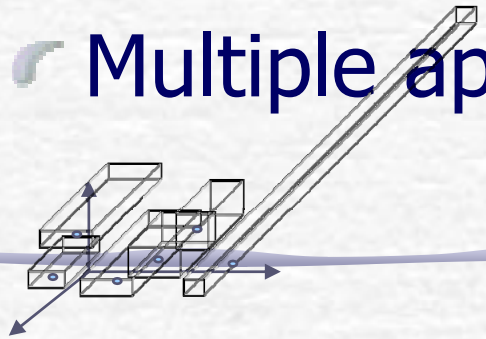
List of Contents

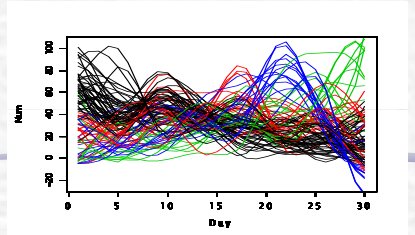
- Functional Data Analysis
- MDS for Functional Dissimilarity Data
- Symbolic Data Analysis
- Distribution Valued Dissimilarity Data
- MDS for Distribution Valued Dissimilarity Data
- Actual Data
- Concluding Remarks

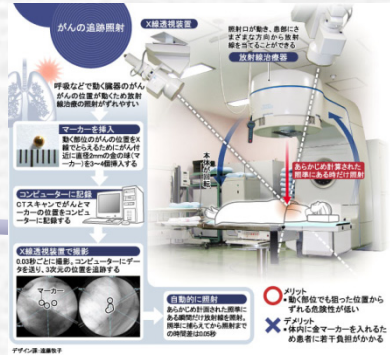
Concluding Remarks

- Various types of data
- Functional Data Analysis, Symbolic Data Analysis
- Distribution-valued data
- MDS,Clustering,…

- Multiple approach to complex data







Various types of Data