A slide show of our lecture note

Intelligent Data Processing

Brest State Technical University
Akira Imada, professor

Last modified on 18 October 2020

In order for data processing
to be
Intelligent

In order for data processing
to be
Intelligent

Collect a set of rules from experts
by human language

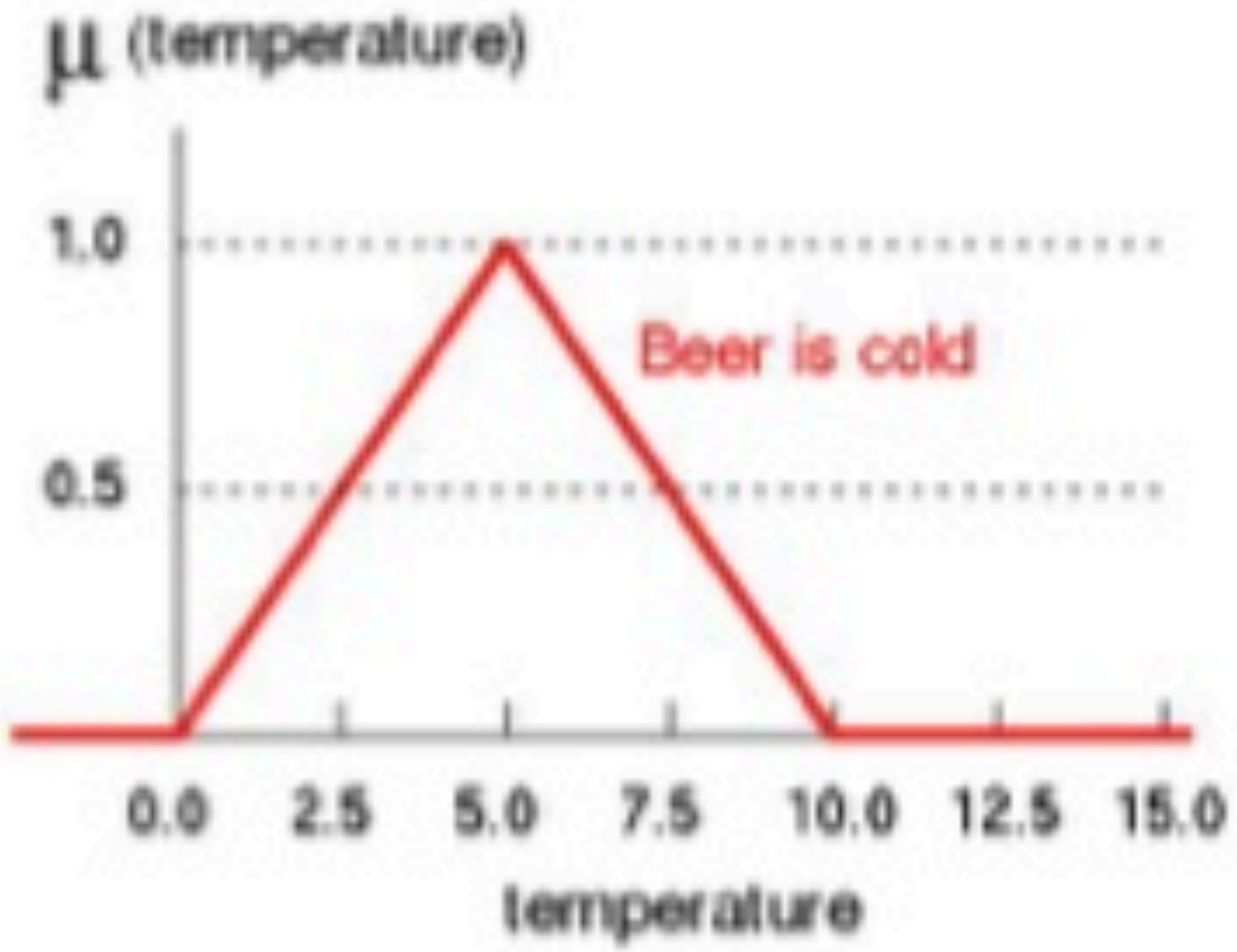
Then translate it to numerical expression
so that machine can understand

=> let's use here Fuzzy logic

I. Fuzzy Basic Arithmetics

Membership Function

In Fuzzy logic the probability of "*how likely A is true*" is called membership value of A and expressed as μ_A. E.g., assuming $A =$ "beer is cold," $\mu_A = 1$ when temperature of beer is 5°C, while $\mu_A = 0.5$ when temperature of beer is 10°C, and $\mu_A = 0$ when temperature of beer is 15°C.



AND and OR

Membership of A AND B and A OR B are given, respectively, as

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$$

and

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$$

IF-THEN

Membership of IF A THEN B has proposed by many but here we use this Larsen's proposal.

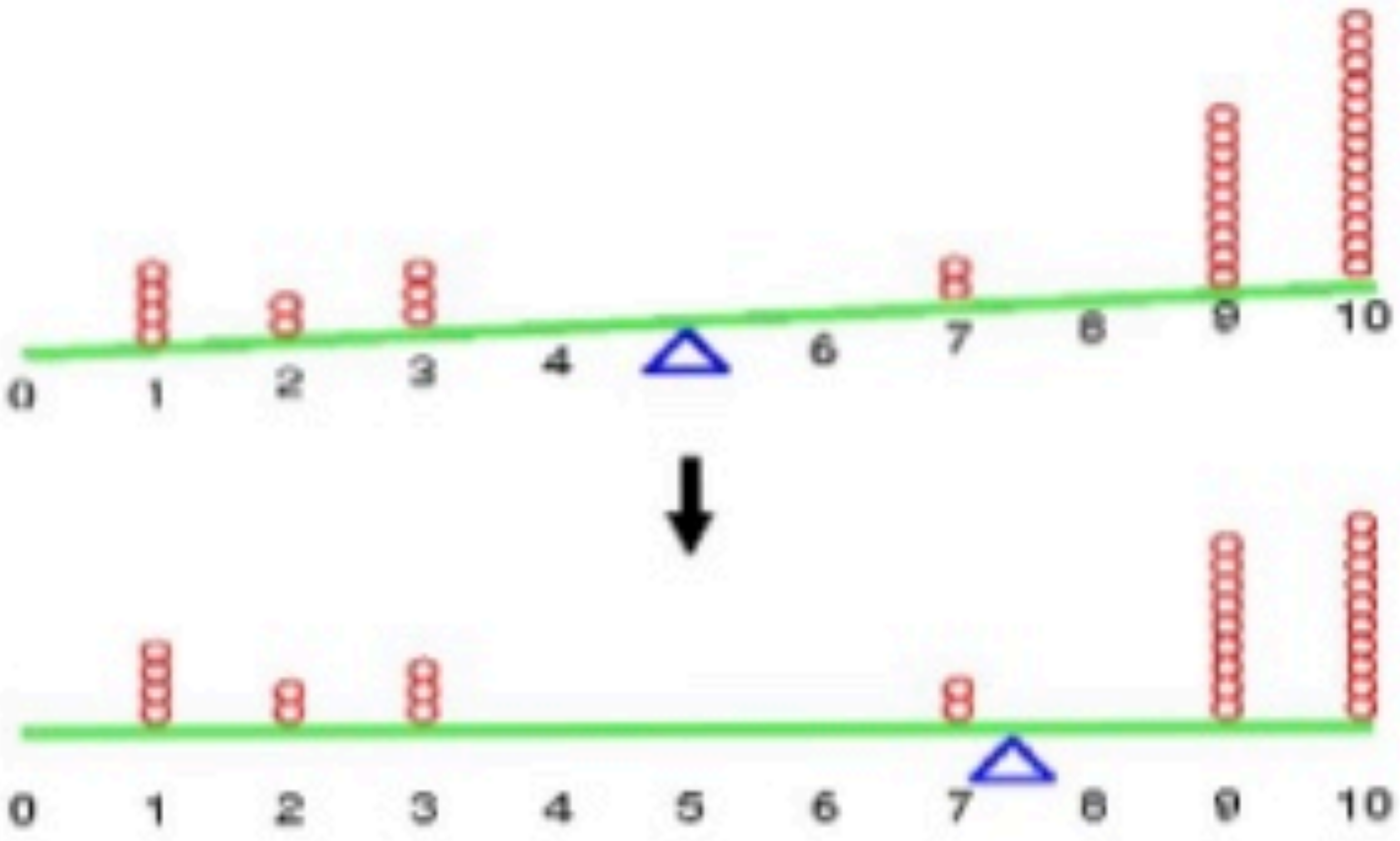
$$\mu_{A \to B}(x) = \mu_A(x) \times \mu_B(x)$$

4. De-fuzzification

When A has some different possibility, we determine most possible value of A by calculating the center of gravity of these membership values.

$$\sum_i \mu_{A_i} \times (x - x_i) = 0$$

E.g.



$$4(x-1)+2(x-2)+3(x-3)+2(x-7)+9(x-9)+13(x-10) = 0$$

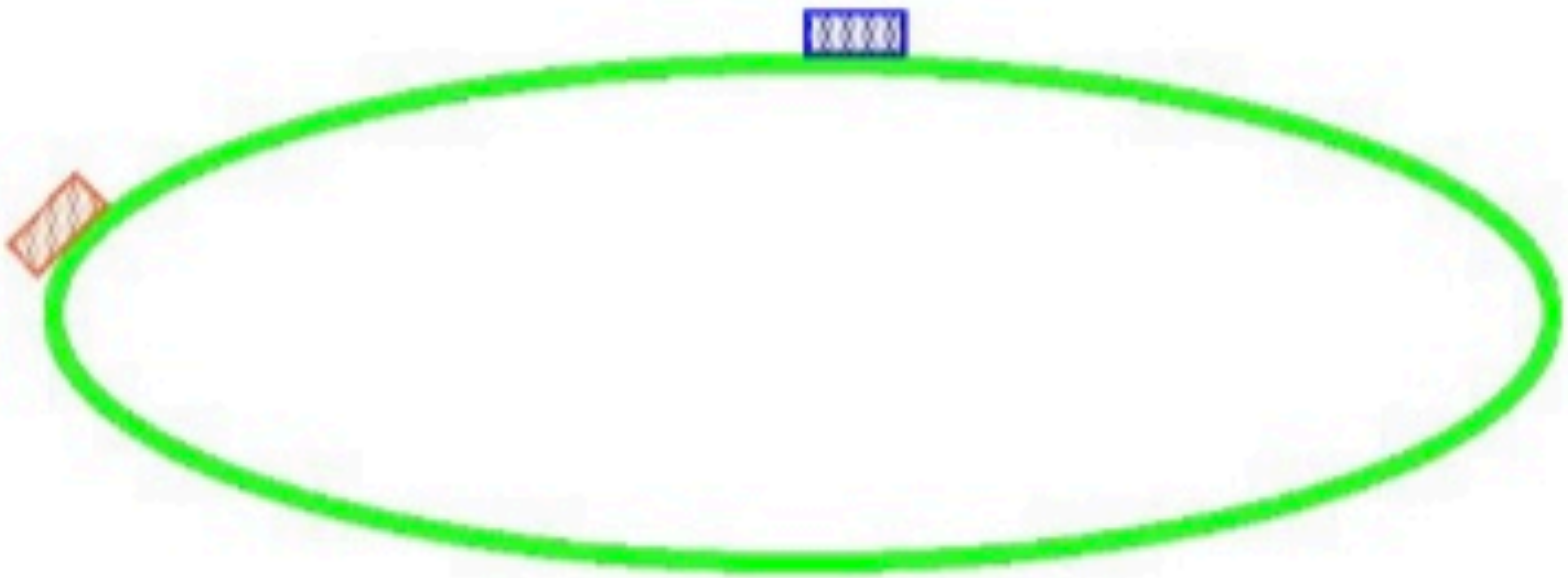
$$33x = 242$$

$$x = 7.33$$

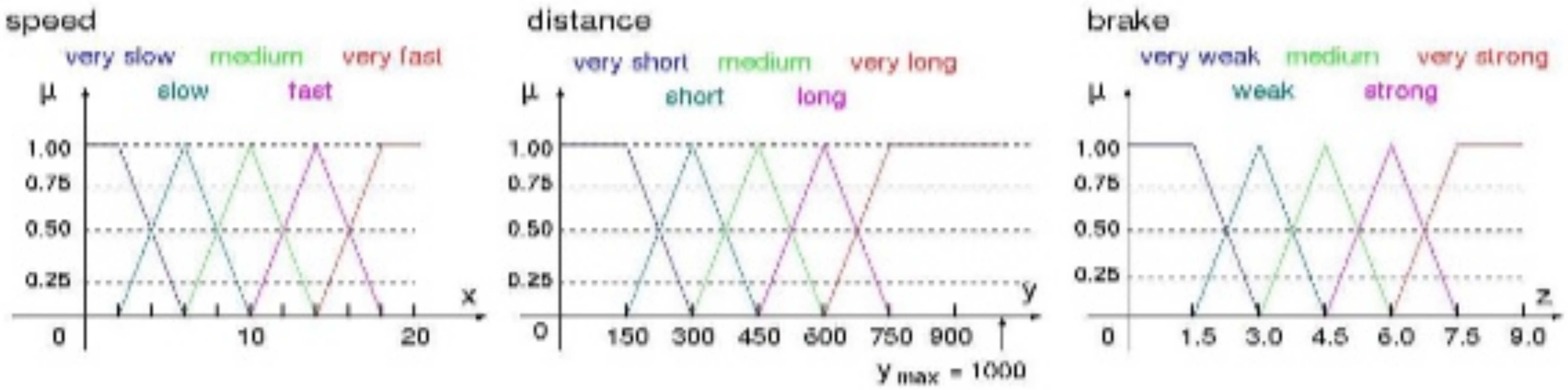
II. Fuzzy Controller

Controll two metro cars

Let's create a virtual metro system with 2 cars on a loop line with 1000 pixels. Each car has a pair of 3 parameters of speed x, distance to the car in front y and strength of brake z.



Membership function of Speed, Distance and Brake assumed here.



An example of Membership value of one rule

Membership value of brake = 0,1,2,3,4,5,6,7,8,9 when speed = 20 and distance = 650 under the rule

IF speed = medium AND distance = long THEN brake = medium.

Rule: If speed is medium, distance is long, then break is

Speed	Mu	Distance	Mu	Break	Mu	TotalM
0	0	0	0	0	0	0
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
40	0	1000	0	0	0	0

From the work by Korol Andrey (2015 Fall)

An example of Membership value of one rule

Membership value of brake = 0,1,2,3,4,5,6,7,8,9 when speed = 20 and distance = 650 under the rule

IF speed = medium AND distance = long THEN brake = medium.

Rule: If speed is medium, distance is long, then break is

Speed	Mu	Distance	Mu	Break	Mu	TotalM
0	0	0	0	0	0	0
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
[illegible]	[illegible]	[illegible]	[illegible]	0	0	0
[illegible]	[illegible]	[illegible]	[illegible]	1	0	0
[illegible]	[illegible]	[illegible]	[illegible]	2	0	0
[illegible]	[illegible]	[illegible]	[illegible]	3	0	0
[illegible]	[illegible]	[illegible]	[illegible]	4	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	6	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	7	0	0
[illegible]	[illegible]	[illegible]	[illegible]	8	0	0
[illegible]	[illegible]	[illegible]	[illegible]	9	0	0
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
40	0	1000	0	[illegible]	0	0

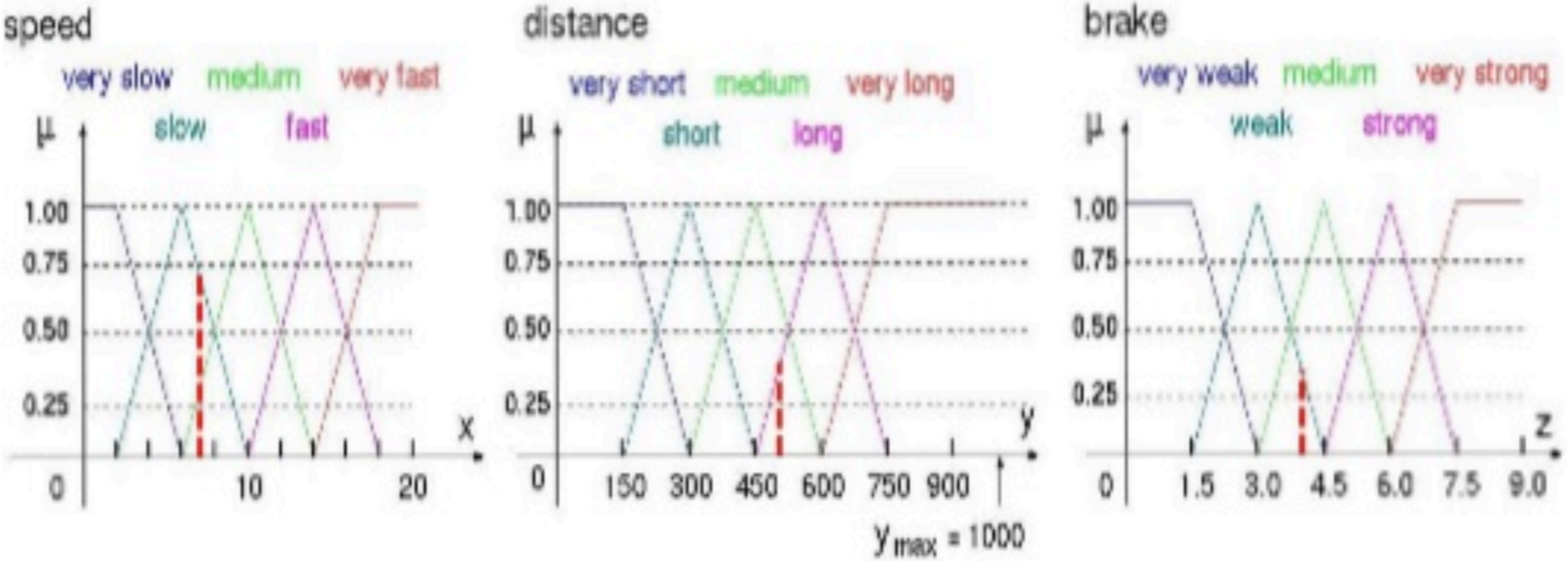
From the work by Korol Andrey (2015 Fall)

Membership value of two rules

Error included below. Later will be corrected.

IF x = slow AND y = long THEN z = weak

Assume now x = 7, y = 500, z = 4



Then the membership value of this rule is → (0.72 + 0.35) x 0.31 = 0.3317

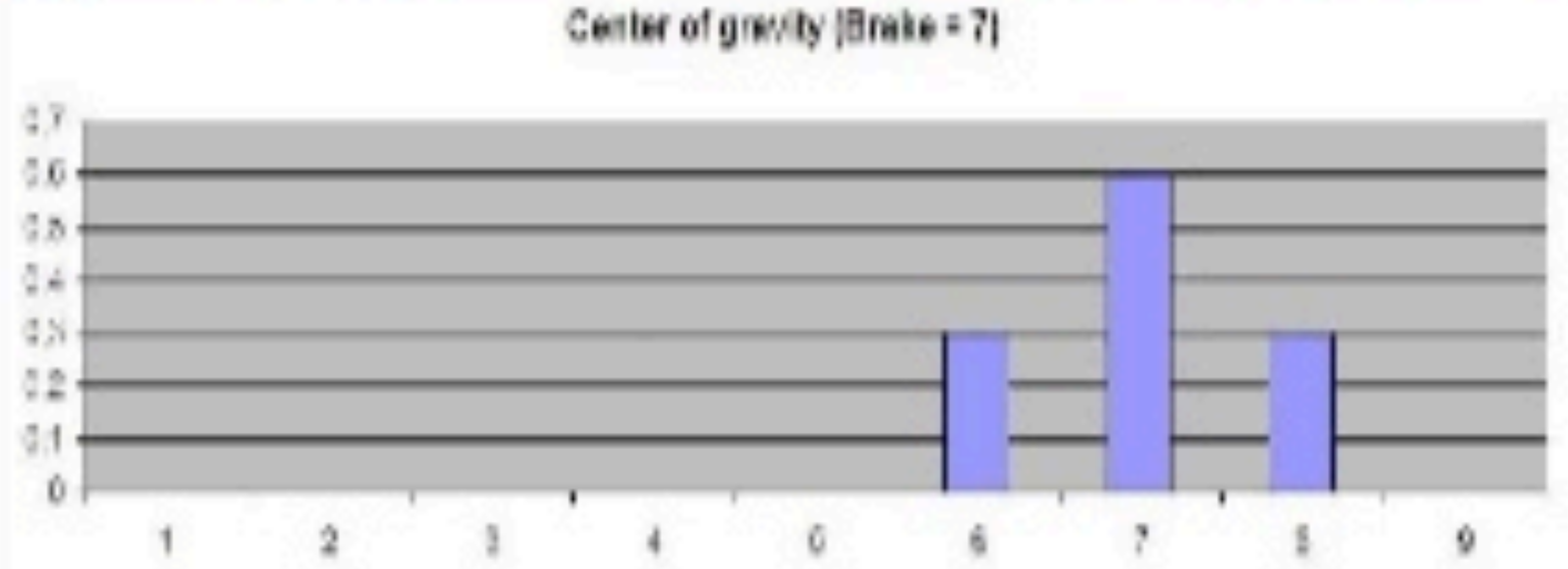
Membership value of 3 rules for a pair of speed & distance

Speed	Distance	Brake	Rule 1: IF x=medium AND y=small THEN z=strong				Rule 2: IF x=medium AND y=medium THEN z=medium				Rule 3: IF x=medium AND y=large THEN z=week				Max of rules
			mSp1	mDs1	mBr1	min(mSp,mDs)*mBr	mSp2	mDs2	mBr2	min(mSp,mDs)*mBr	mSp3	mDs3	mBr3	min(mSp,mDs)*mBr	
		0	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	0	0	0
		1	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	0	0	0
		2	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	0,25	0.1875	0,1875
		3	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	1	0,75	0.75
		4	0,75	0	0	0	0.75	0.25	0,75	0,1875	0,75	0,75	0,25	0,1875	0,1875
11,00	550,00	5	0,75	0	0,3	0	0.75	0.25	0,75	0,1875	0,75	0,75	0	0	0,1875
		6	0,75	0	1	0	0.75	0.25	0	0	0,75	0,75	0	0	0
		7	0,75	0	0,3	0	0.75	0.25	0	0	0,75	0,75	0	0	0
		8	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	0	0	0
		9	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	0	0	0
		10	0,75	0	0	0	0.75	0.25	0	0	0,75	0,75	0	0	0

From the work by Yulia Bogutskaya (2016 Fall)

Defuzzified value of break for a pair of a speed and a distance

			Speed is very slow AND Distance is very short THEN Brake is strong			Speed is very slow AND Distance is short THEN Brake is strong			Speed is medium AND Distance is short THEN Brake is very strong			
Speed	Distance	Brake	µ1 Speed	µ1 Distance	µ1 Brak	µ2 Speed	µ2 Distance	µ3 Brake	µ3 Speed	µ3 Distance	µ3 Brake	Result
0	150	0	1	0.4	0	1	0.6	0	0	0.6	0	0
0	150	1	1	0.4	0	1	0.6	0	0	0.6	0	0
0	150	2	1	0.4	0	1	0.6	0	0	0.6	0	0
0	150	3	1	0.4	0	1	0.6	0	0	0.6	0	0
0	150	4	1	0.4	0	1	0.6	0	0	0.6	0	0
0	150	5	1	0.4	0	1	0.6	0	0	0.6	0	0
0	150	6	1	0.4	0.5	1	0.6	0.5	0	0.6	0	0.3
0	150	7	1	0.4	1	1	0.6	1	0	0.6	0	0.6
0	150	8	1	0.4	0.5	1	0.6	0.5	0	0.6	0.5	0.3
0	150	9	1	0.4	0	1	0.6	0	0	0.6	1	0



From the work by Kuchur Alexander (2015 Fall)

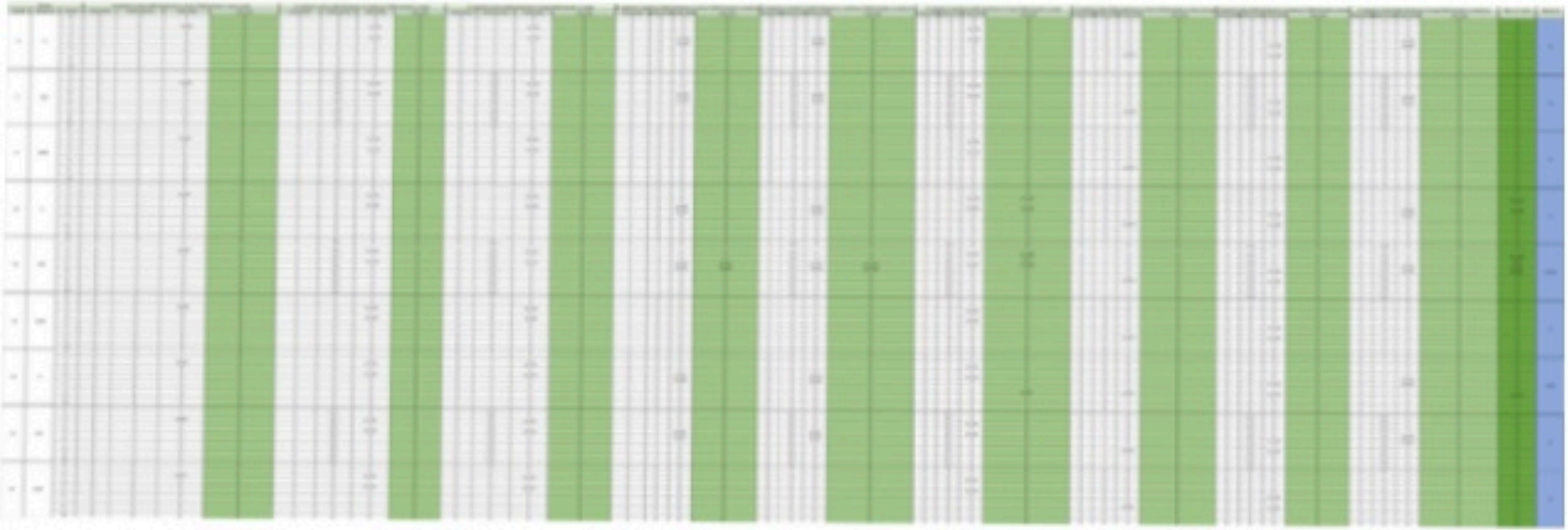
Membership value of 3 rules for 3 pairs of speed & distance

Speed	Distance	Brake	Rule 1: IF x=medium AND y=small THEN z=strong				Rule 2: IF x=medium AND y=medium THEN z=medium				Rule 3: IF x=medium AND y=large THEN z=weak				Max of rules	Balance
			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		
		0	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		1	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		2	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	0.25	0.125	0.125	
		3	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	1	0.5	0.5	
		4	0.75	0	0	0	0.75	0.5	0.75	0.375	0.75	0.5	0.25	0.125	0.375	
11.00	500.00	5	0.75	0	0.5	0	0.75	0.5	0.75	0.375	0.75	0.5	0	0	0.375	3.727273
		6	0.75	0	1	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		7	0.75	0	0.5	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		8	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		9	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		10	0.75	0	0	0	0.75	0.5	0	0	0.75	0.5	0	0	0	
		0	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		1	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		2	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	0.25	0.1875	0.1875	
		3	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	1	0.75	0.75	
		4	0.75	0	0	0	0.75	0.25	0.75	0.1875	0.75	0.75	0.25	0.1875	0.1875	
11.00	550.00	5	0.75	0	0.5	0	0.75	0.25	0.75	0.1875	0.75	0.75	0	0	0.1875	2.285714
		6	0.75	0	1	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		7	0.75	0	0.5	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		8	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		9	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		10	0.75	0	0	0	0.75	0.25	0	0	0.75	0.75	0	0	0	
		0	0.75	0	0	0	0.75	0	0	0	0.75	1	0	0	0	
		1	0.75	0	0	0	0.75	0	0	0	0.75	1	0	0	0	
		2	0.75	0	0	0	0.75	0	0	0	0.75	1	0.25	0.1875	0.1875	
		3	0.75	0	0	0	0.75	0	0	0	0.75	1	1	0.75	0.75	
		4	0.75	0	0	0	0.75	0	0.75	0	0.75	1	0.25	0.1875	0.1875	
11.00	600.00	5	0.75	0	0.5	0	0.75	0	0.75	0	0.75	1	0	0	0	[illegible]
		6	0.75	0	1	0	0.75	0	0	0	0.75	1	0	0	0	
		7	0.75	0	0.5	0	0.75	0	0	0	0.75	1	0	0	0	
		8	0.75	0	0	0	0.75	0	0	0	0.75	1	0	0	0	
		9	0.75	0	0	0	0.75	0	0	0	0.75	1	0	0	0	
		10	0.75	0	0	0	0.75	0	0	0	0.75	1	0	0	0	

From the work by Yulia Bogutskaya (2016 Fall)

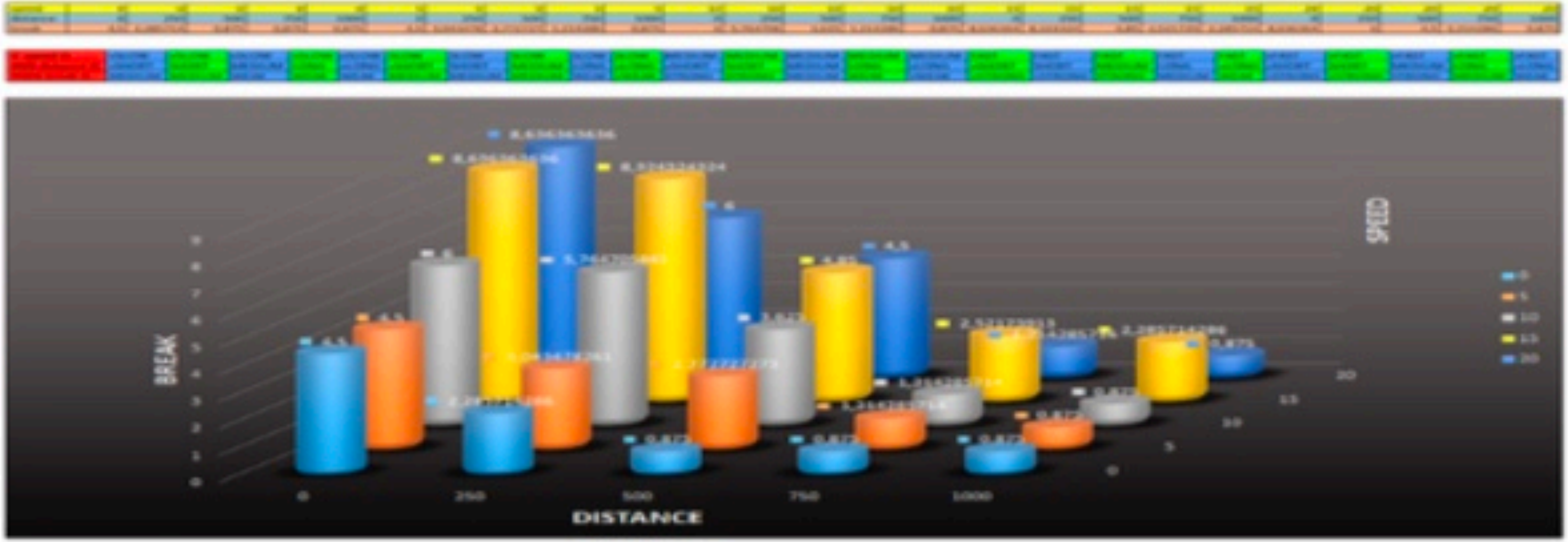
Membership function of 25 rules

Too small to be visible but all combination of speed, distance and brake.



From the work by Lishko Aleksandr (2016 Fall)

6. 3-D bar-graph of speed-distance-brake with 25 rules

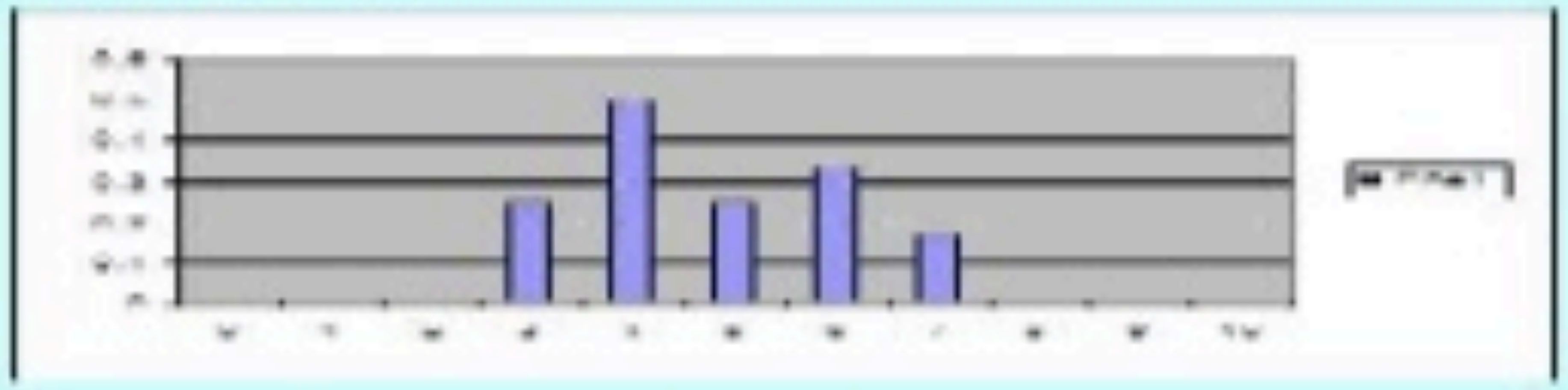


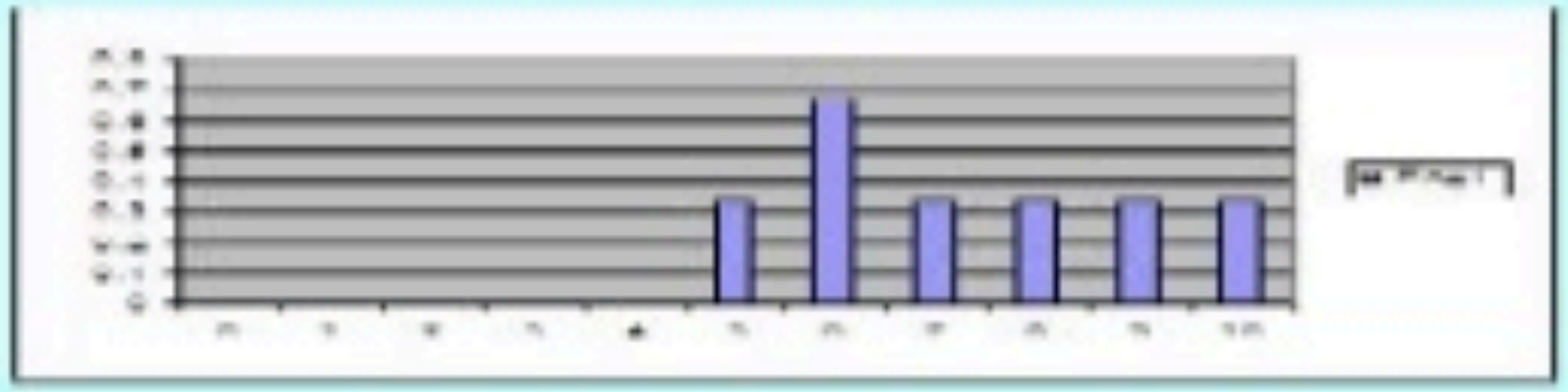
From the work by Bokhanov Evgenii (2015 Fall)

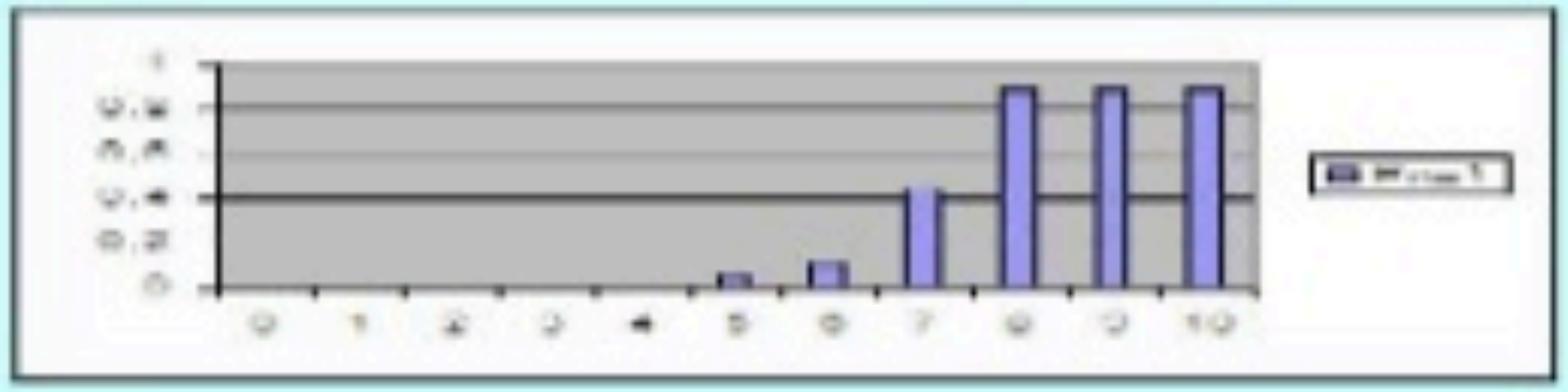
3-D surface of speed-distance-brake with limited domain

An example of how to draw for a fixed speed and three diferent value of distances

Distance	Speed	Brake	Max	Center of gravity	Final brake
400	12	0	0		4,8
400	12	1	0		4,8
400	12	2	0		4,8
400	12	3	0,25		4,8
400	12	4	0,5		4,8
400	12	5	0,25		4,8
400	12	6	0,333333		4,8
400	12	7	0,166667		4,8
400	12	8	0		4,8
400	12	9	0		4,8
400	12	10	0		4,8
400	24	0	0		7,3
400	24	1	0		7,3
400	24	2	0		7,3
400	24	3	0		7,3
400	24	4	0		7,3
400	24	5	0,333333		7,3
400	24	6	0,666667		7,3
400	24	7	0,333333		7,3
400	24	8	0,333333		7,3
400	24	9	0,333333		7,3
400	24	10	0,333333		7,3
400	36	0	0		8,5
400	36	1	0		8,5
400	36	2	0		8,5
400	36	3	0		8,5
400	36	4	0		8,5
400	36	5	0,0555556		8,5
400	36	6	0,111111		8,5
400	36	7	0,444444		8,5
400	36	8	0,888889		8,5
400	36	9	0,888889		8,5
400	36	10	0,888889		8,5



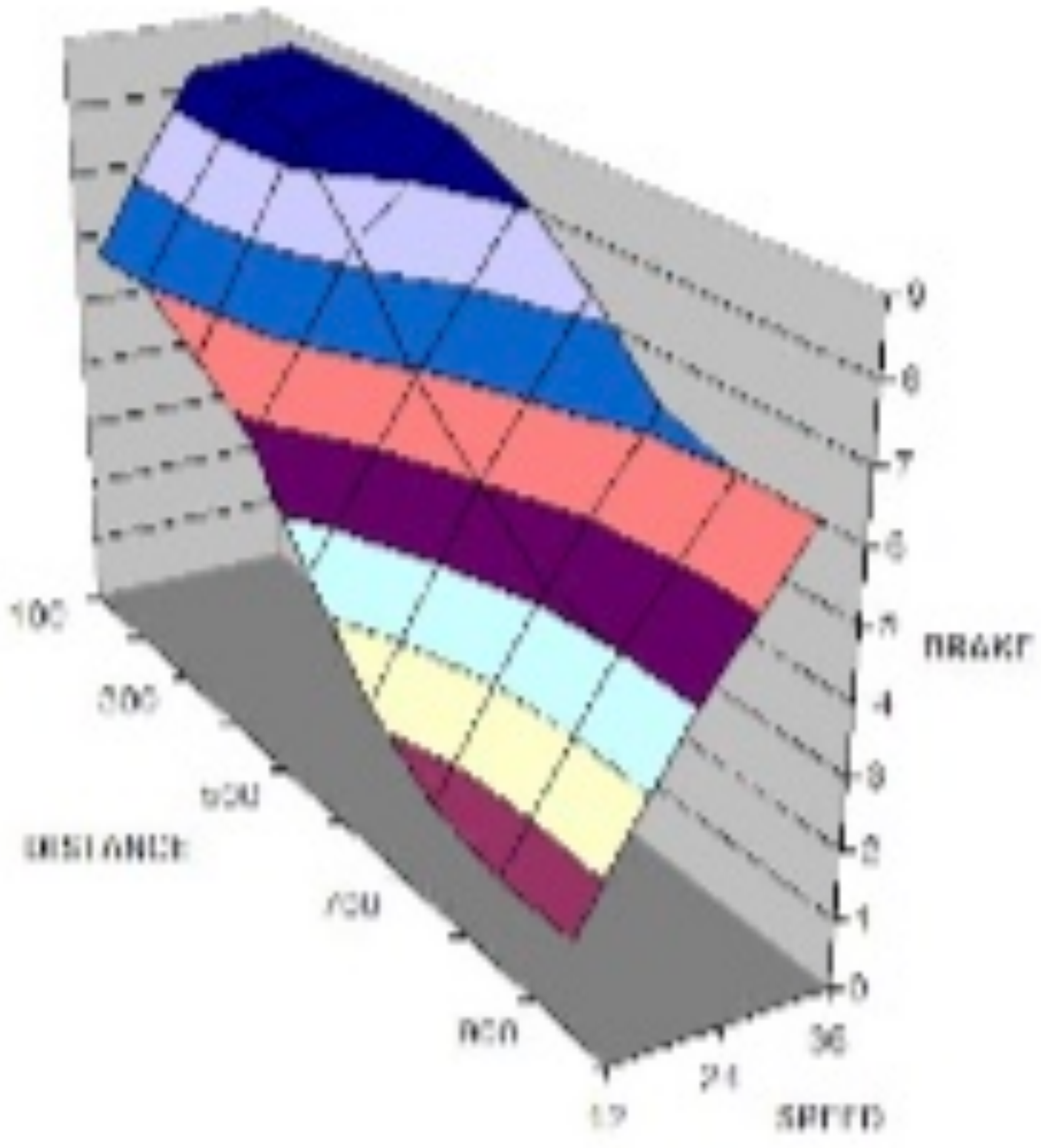




From the work by Bokhanov Evgenii (2015 Fall)

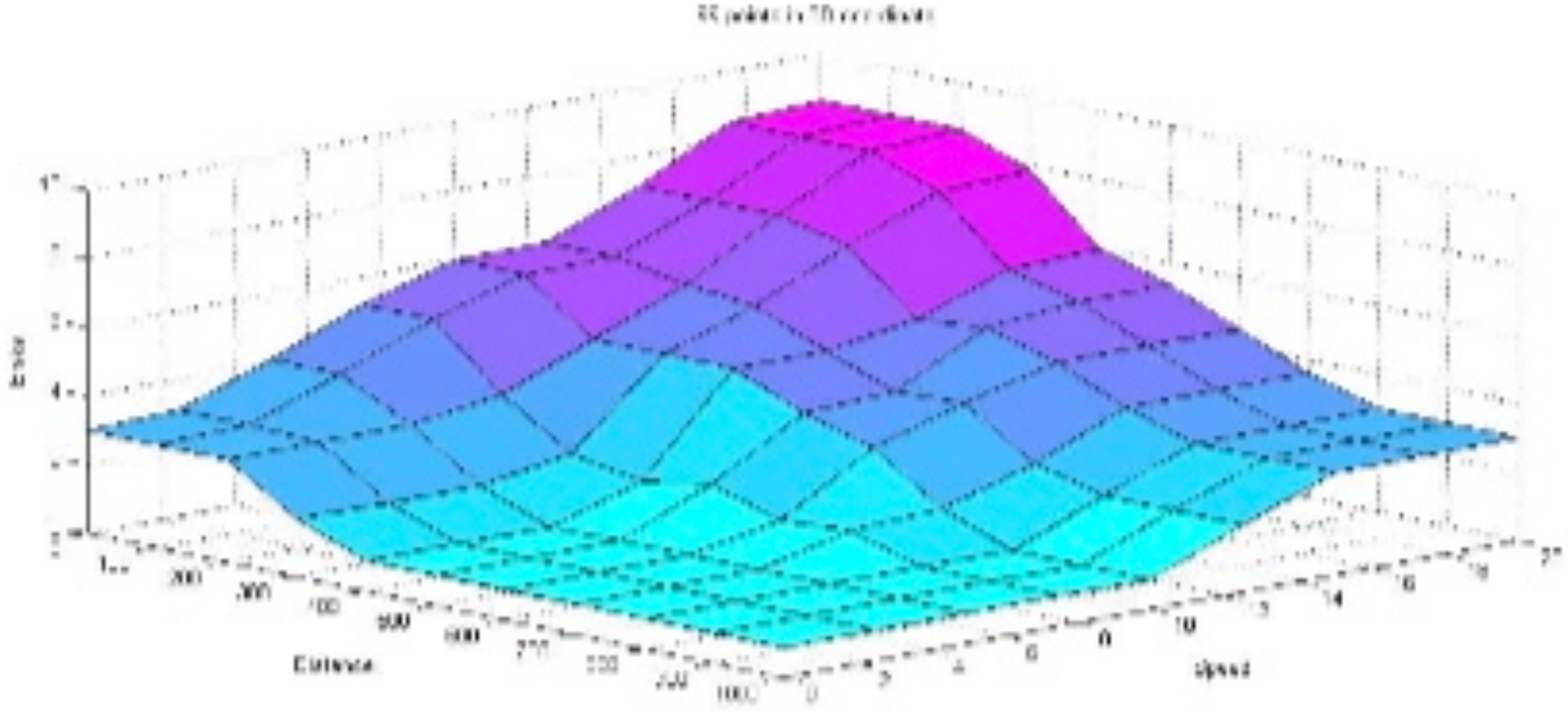
3-D surface of speed-distance-brake with limited domain (continued)

Distance	Speed	Final Brake
100	12	6
100	24	8.5
100	36	8.7
200	12	6
200	24	8.5
200	36	8.7
300	12	5.1
300	24	6.4
300	36	6.7
400	12	4.8
400	24	7.0
400	36	8.8
500	12	3.4
500	24	5.8
600	36	8.1
600	12	2.2
600	24	5
600	36	7.5
700	12	1.5
700	24	4.2
700	36	5.5
800	12	1.9
800	24	4
800	36	8
900	12	1.0
900	24	1
900	36	6



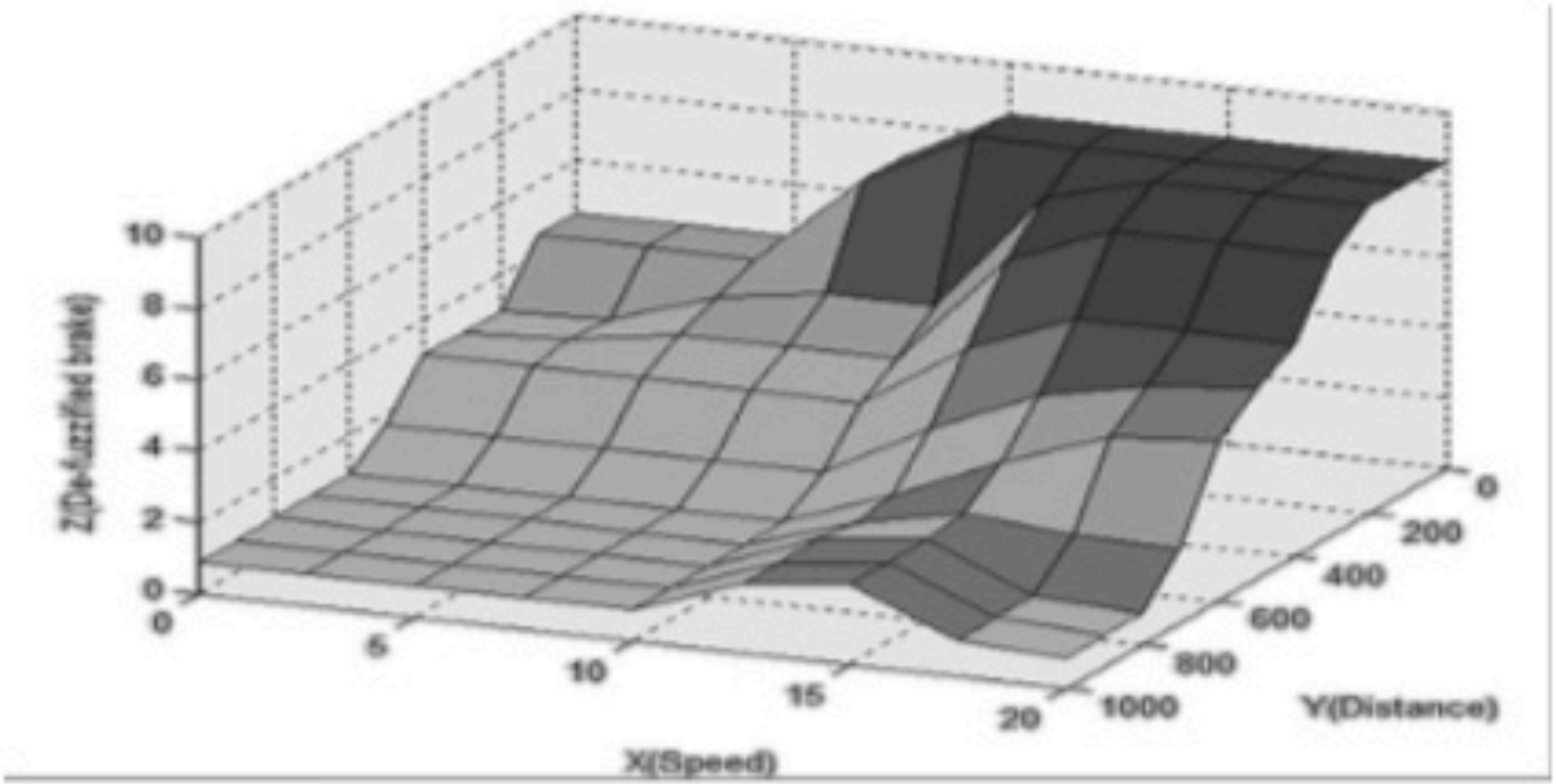
From the work by Bokhanov Evgenii (2015 Fall)

A 3-D surface of speed-distance-brake over whole domain



From the work by Yulia Bogutskaya (2016 Fall)

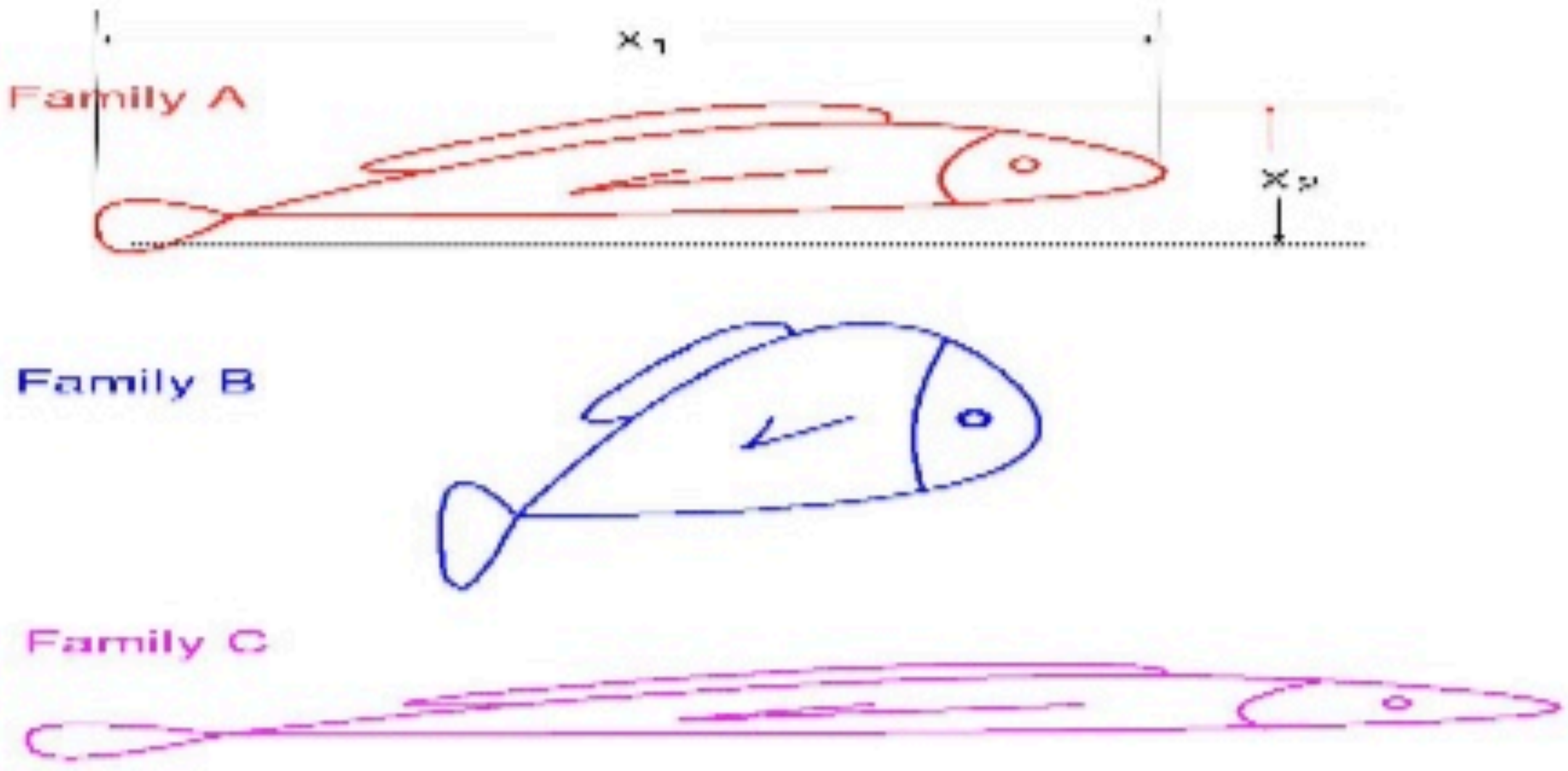
Another 3-D surface of speed-distance-brake over whole domain



From the work by Kolesnikov Dmitry (2016 Fall)

III. Fuzzy Classification

An example of classification - 3 families of fish



Rules to classify as an example

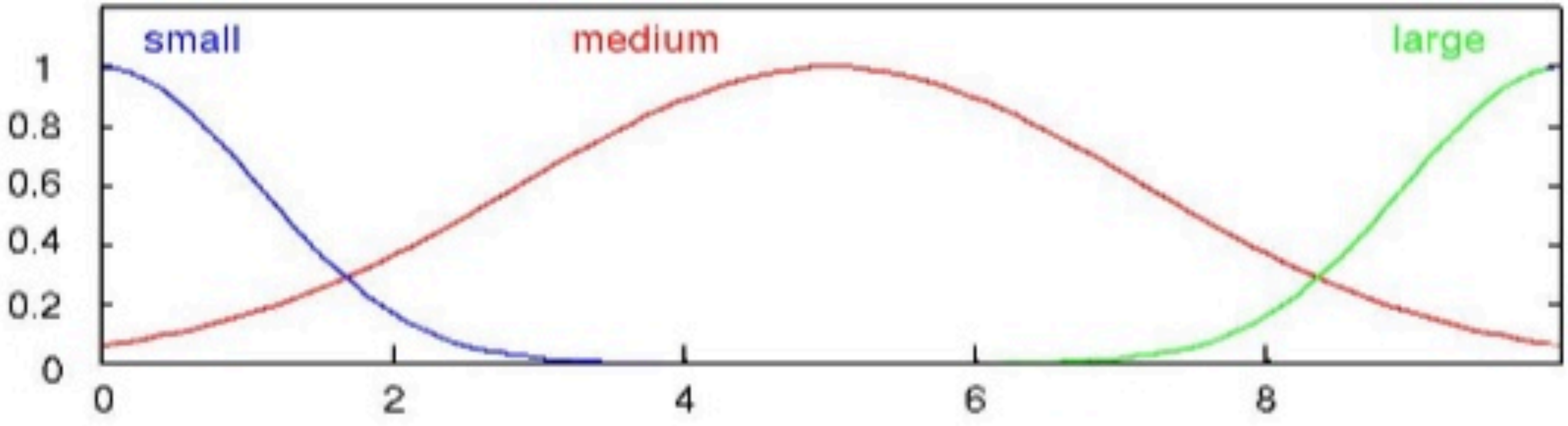
R_1: IF X_1 = medium AND X_2 = small THEN A

R_2: IF X_1 = small AND X_2 = medium THEN B

R_3: IF X_1 = large AND X_2 = small THEN C

Memership function for the size of two parts

$$\mu(x) = \exp\{-\frac{(x-(avg)^2)}{(std)^2}\}$$



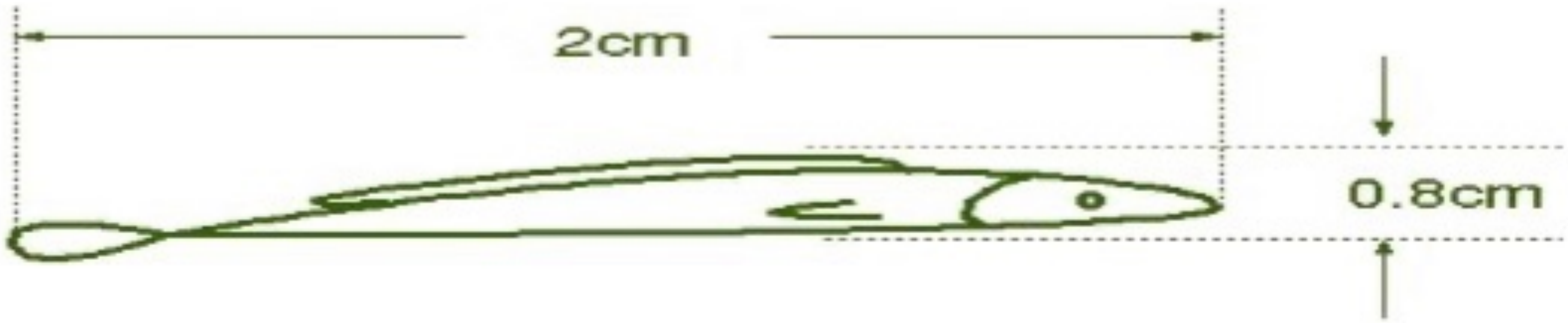
How to estimate avg and std from dataset

How we specify avg and std for each of membership function from dataset given?

Algorithm 1

1. *Select maximum data + minimum data + other randomly chose $5N - 2$ data.*
2. *Sort these $5N$ data from small to large in each attribute.*
3. *Devide the data in each attribute into 5 groups, that is, very small, small, medium, large, and very large.*
4. *Calculate average and stndard deviation in eact devision.*

Question: Which family is this new fish?



Takagi Sugeno Formula

R_k: If x_1 is A_1^k, and x_2 is A_2^k and $\cdots$ and x_N is A_N^k then y is g^k.

Takagi-Sugeno rules: Estimation of a single input

Estimation of y for an input $\mathbf{x} = (x_1, x_2, \cdots, x_N)$

$$y_j = \frac{\sum_{k=1}^{H}(M_k(\mathbf{x}) \cdot g_k)}{\sum_{k=1}^{H} M_k(\mathbf{x})}$$

where

$$M_k(\mathbf{x}) = \prod_{i=1}^{N} \mu_{ik}(x_i)$$

where μ_{ik} is i-th attribute of k-th rule

A rule set to classify

R_k: IF $x_1 = A_1^k$ AND $x_2 = A_2^k$ AND $\cdots$ $x_N = A_N^k$ THEN $y = g^k$

E.g.

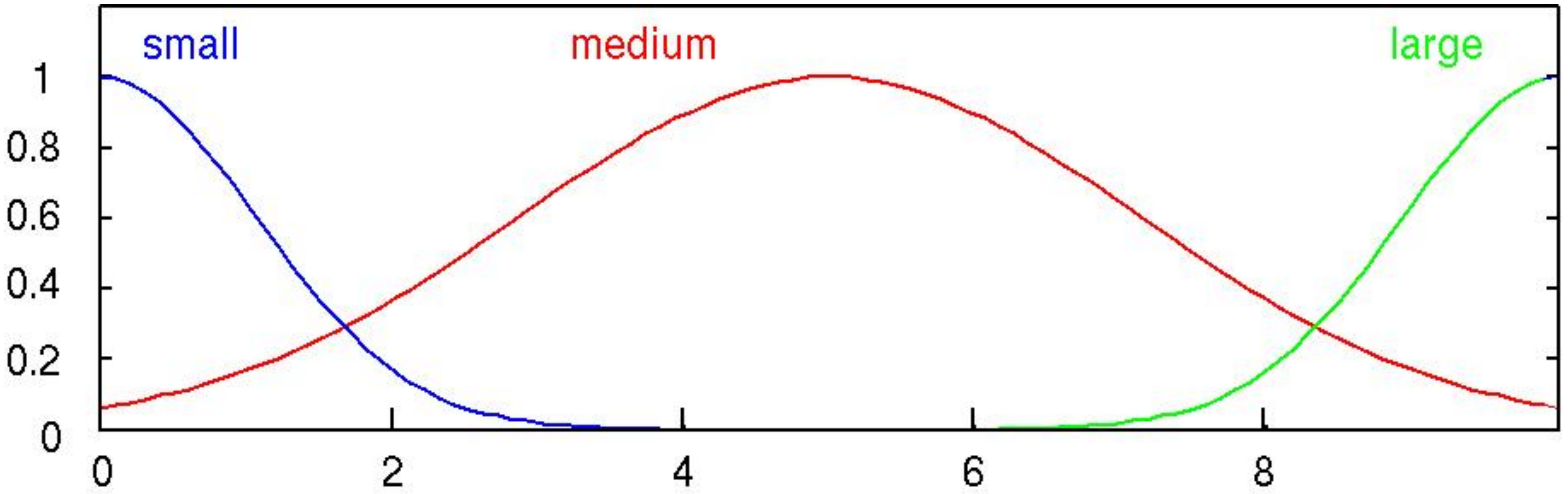
R_1: IF x_1 = medium AND x_2 = small THEN $y = 1$

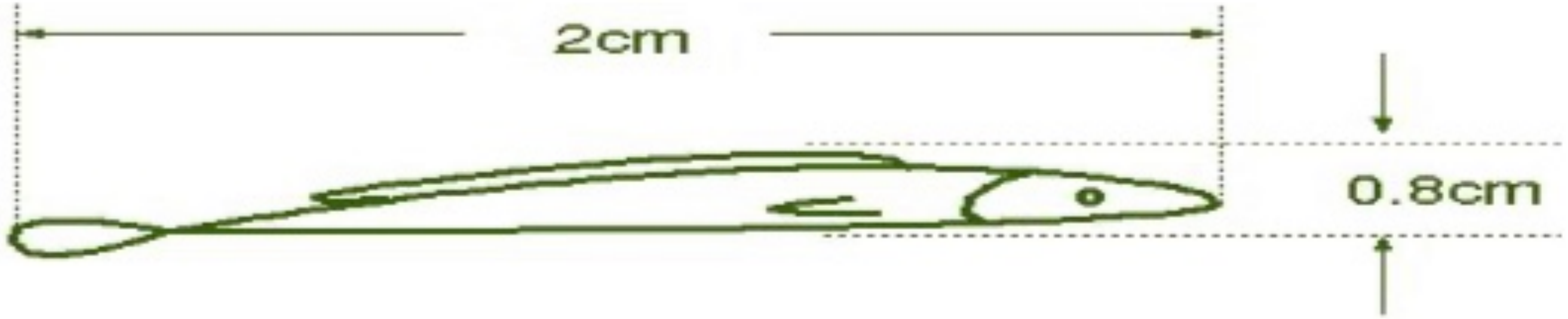
R_2: IF x_1 = small AND x_2 = medium THEN $y = 2$

R_3: IF x_1 = large AND x_2 = small THEN $y = 3$

Membership functions for x_1 and x_2

$$\mu(x) = \exp\{-(x-(avg)^2)/(std)^2\}$$





Takagi-Sugeno rules

Estimation of y for $x = (x_1, x_2, \cdots x_N)$

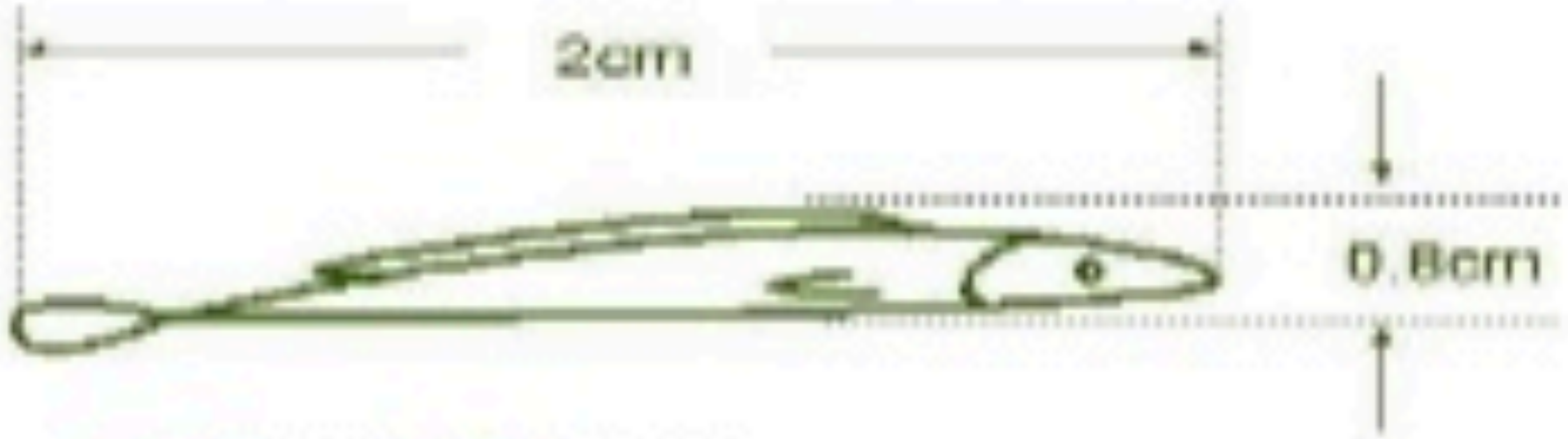
$$y_j = \{\Sigma_{k=1}^{H} (M_k(x) \cdot g_k\} / \{\Sigma_{k=1}^{H} (M_k(x)\}$$

where

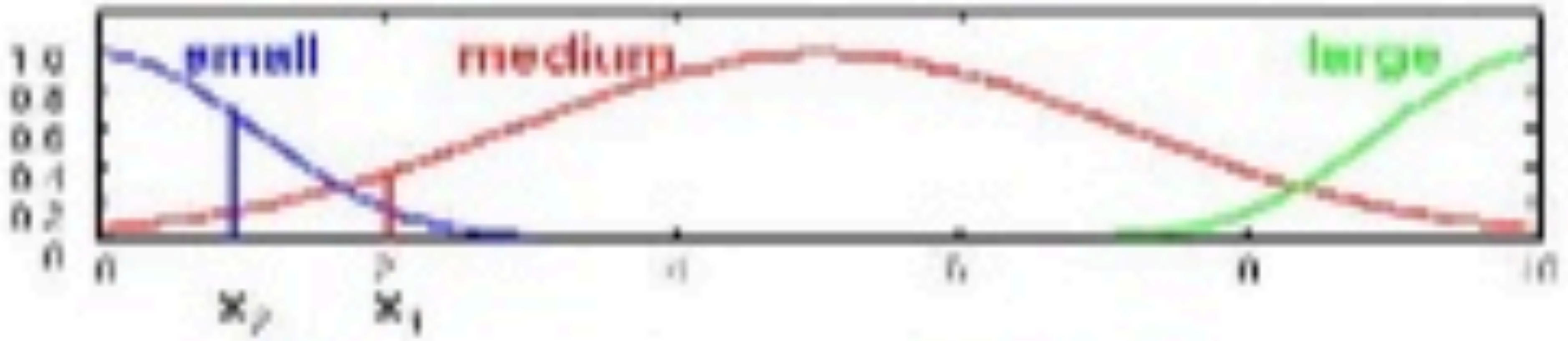
$$M_k(x) = \prod_{j=1}^{M} \mu_{ik}(x_i)$$

where μ_{ik} is membership of i-th attribute A_i^k of the k-th rule

Three rules to classify

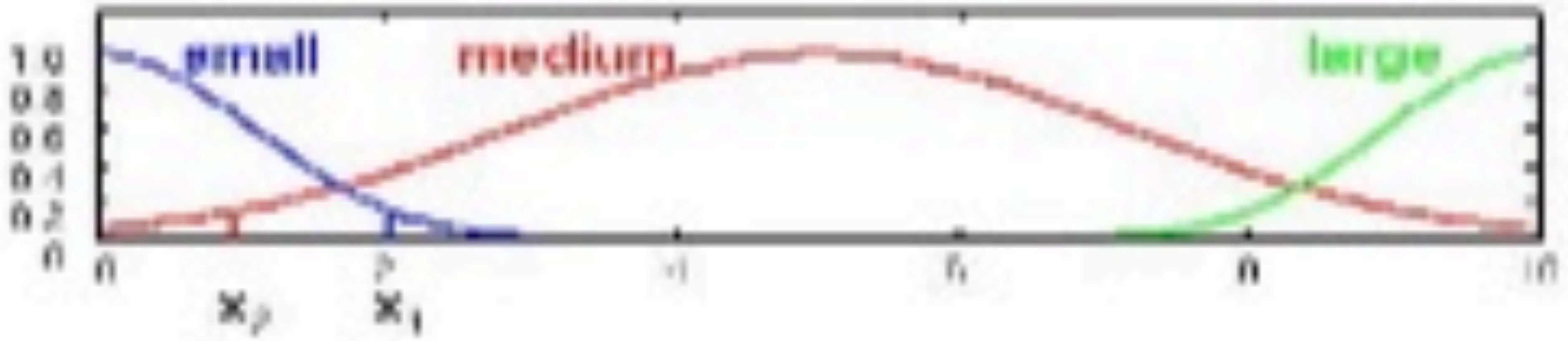


R_1: IF X1 = medium AND X2 = small THEN y = 1



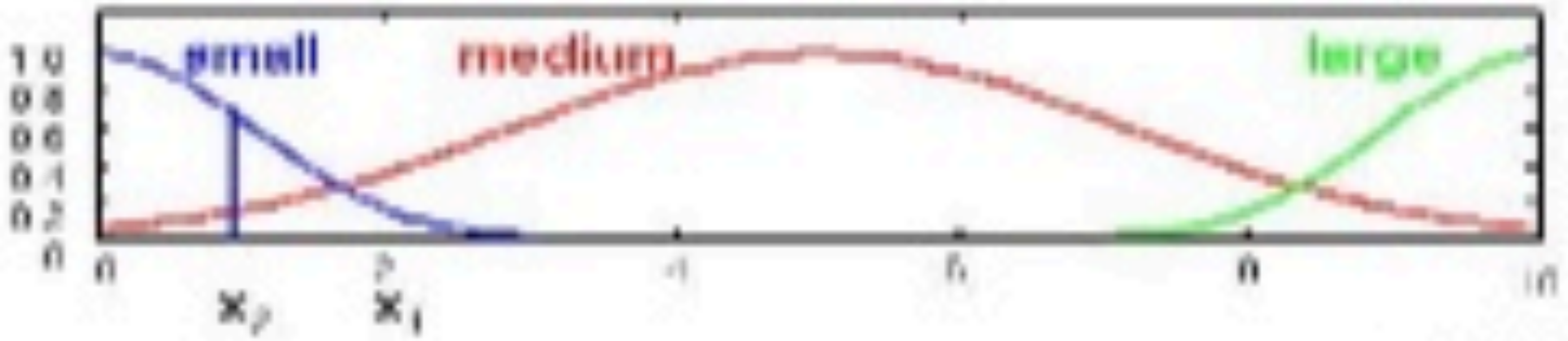
M_1 = 0.39 x 0.41 = 0.16

R_2: IF X1 = small AND X2 = medium THEN y = 2



M_2 = 0.19 x 0.19 = 0.03

R_3: IF X1 = large AND X2 = small THEN y = 3



M_3 = 0.01 x 0.71 = 0.01

$$y = \frac{0.16 \times 1 + 0.03 \times 2 + 0.01 \times 3}{0.16 + 0.03 + 0.01} = \frac{0.25}{0.2} = 1.25$$

A benchmark – Iris database

Iris flower dataset (taken from University of California Urvine Machine Learning Repository) consists of three species of iris flower

setosa, versicolor and *virginica.*

Each sample represents four attributes of the iris flower

sepal-length, sepal-width, petal-length, and *petal-width.*

Iris Flower Database to design

Setosa				Versicolor				Virginica			
x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0.56	0.66	0.20	0.08	0.84	0.66	0.67	0.52	0.85	0.57	0.84	0.72
0.62	0.70	0.22	0.04	0.66	0.61	0.57	0.56	0.91	0.82	0.88	1.00
0.68	0.84	0.22	0.08	0.63	0.45	0.51	0.40	0.82	0.73	0.74	0.80
0.61	0.77	0.23	0.08	0.75	0.68	0.61	0.60	0.81	0.61	0.77	0.76
0.61	0.68	0.20	0.04	0.76	0.50	0.58	0.40	0.86	0.68	0.80	0.84
0.54	0.68	0.16	0.04	0.77	0.66	0.68	0.56	0.72	0.57	0.72	0.80
0.73	0.91	0.17	0.08	0.71	0.66	0.52	0.52	0.73	0.64	0.74	0.96
0.72	1.00	0.22	0.16	0.85	0.70	0.64	0.56	0.81	0.73	0.77	0.92
0.68	0.89	0.19	0.16	0.71	0.68	0.65	0.60	0.82	0.68	0.80	0.72
0.65	0.80	0.20	0.12	0.73	0.61	0.59	0.40	0.97	0.86	0.97	0.88
0.72	0.86	0.25	0.12	0.78	0.50	0.65	0.60	0.97	0.59	1.00	0.92
0.65	0.86	0.22	0.12	0.71	0.57	0.57	0.44	0.76	0.50	0.72	0.60
0.68	0.77	0.25	0.08	0.75	0.73	0.70	0.72	0.87	0.73	0.83	0.92
0.65	0.84	0.22	0.16	0.77	0.64	0.58	0.52	0.71	0.64	0.71	0.80
0.58	0.82	0.14	0.08	0.80	0.57	0.71	0.60	0.97	0.64	0.97	0.80
0.65	0.75	0.25	0.20	0.77	0.64	0.68	0.48	0.80	0.61	0.71	0.72
0.61	0.77	0.28	0.08	0.81	0.66	0.62	0.52	0.85	0.75	0.83	0.84

Avg and std of each column

	Setosa				Versicolor				Virginica			
	x1	x2	x3	x4	x1	x2	x3	x4	x1	x2	x3	x4
	0.56	0.66	0.2	0.08	0.84	0.66	0.67	0.52	0.85	0.57	0.84	0.72
	0.62	0.7	0.22	0.04	0.66	0.61	0.57	0.56	0.91	0.82	0.88	1
	0.68	0.84	0.22	0.08	0.63	0.45	0.51	0.4	0.82	0.73	0.74	0.8
	0.61	0.77	0.23	0.08	0.75	0.68	0.61	0.6	0.81	0.61	0.77	0.76
	0.61	0.68	0.2	0.04	0.76	0.5	0.58	0.4	0.86	0.68	0.8	0.84
	0.54	0.68	0.16	0.04	0.77	0.66	0.68	0.56	0.72	0.57	0.72	0.8
	0.73	0.91	0.17	0.08	0.71	0.66	0.52	0.52	0.73	0.64	0.74	0.96
	0.72	1	0.22	0.16	0.85	0.7	0.64	0.56	0.81	0.73	0.77	0.92
	0.68	0.89	0.19	0.16	0.71	0.68	0.65	0.6	0.82	0.68	0.8	0.72
	0.65	0.8	0.2	0.12	0.73	0.61	0.59	0.4	0.97	0.86	0.97	0.88
	0.72	0.86	0.25	0.12	0.78	0.5	0.65	0.6	0.97	0.59	1	0.92
	0.65	0.86	0.22	0.12	0.71	0.57	0.57	0.44	0.76	0.5	0.72	0.6
	0.68	0.77	0.25	0.08	0.75	0.73	0.7	0.72	0.87	0.73	0.83	0.92
	0.65	0.84	0.22	0.16	0.77	0.64	0.58	0.52	0.71	0.64	0.71	0.8
	0.58	0.82	0.14	0.08	0.8	0.57	0.71	0.6	0.97	0.64	0.97	0.8
	0.65	0.75	0.25	0.2	0.77	0.64	0.68	0.48	0.8	0.61	0.71	0.72
	0.61	0.77	0.28	0.08	0.81	0.66	0.62	0.52	0.85	0.75	0.83	0.84
Avg:	0.64	0.80	0.21	0.10	0.75	0.62	0.62	0.53	0.84	0.67	0.81	0.82
Deviation:	0.04	0.07	0.03	0.04	0.04	0.06	0.05	0.07	0.07	0.08	0.08	0.08

added by Evgene Borisiuk (on 05 February 2019)

Iris Flower Database to validate

Setosa				Versicolor				Virginica			
x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
0.65	0.80	0.20	0.08	0.89	0.73	0.68	0.56	0.80	0.75	0.87	1.00
0.62	0.68	0.20	0.08	0.81	0.73	0.65	0.60	0.73	0.61	0.74	0.76
0.59	0.73	0.19	0.08	0.87	0.70	0.71	0.60	0.90	0.68	0.86	0.84
0.58	0.70	0.22	0.08	0.70	0.52	0.58	0.52	0.80	0.66	0.81	0.72
0.63	0.82	0.20	0.08	0.82	0.64	0.67	0.60	0.82	0.68	0.84	0.88
0.68	0.89	0.25	0.16	0.72	0.64	0.65	0.52	0.96	0.68	0.96	0.84
0.58	0.77	0.20	0.12	0.80	0.75	0.68	0.64	0.62	0.57	0.65	0.68
0.63	0.77	0.22	0.08	0.62	0.55	0.48	0.40	0.92	0.66	0.91	0.72

Wine dataset to design rules

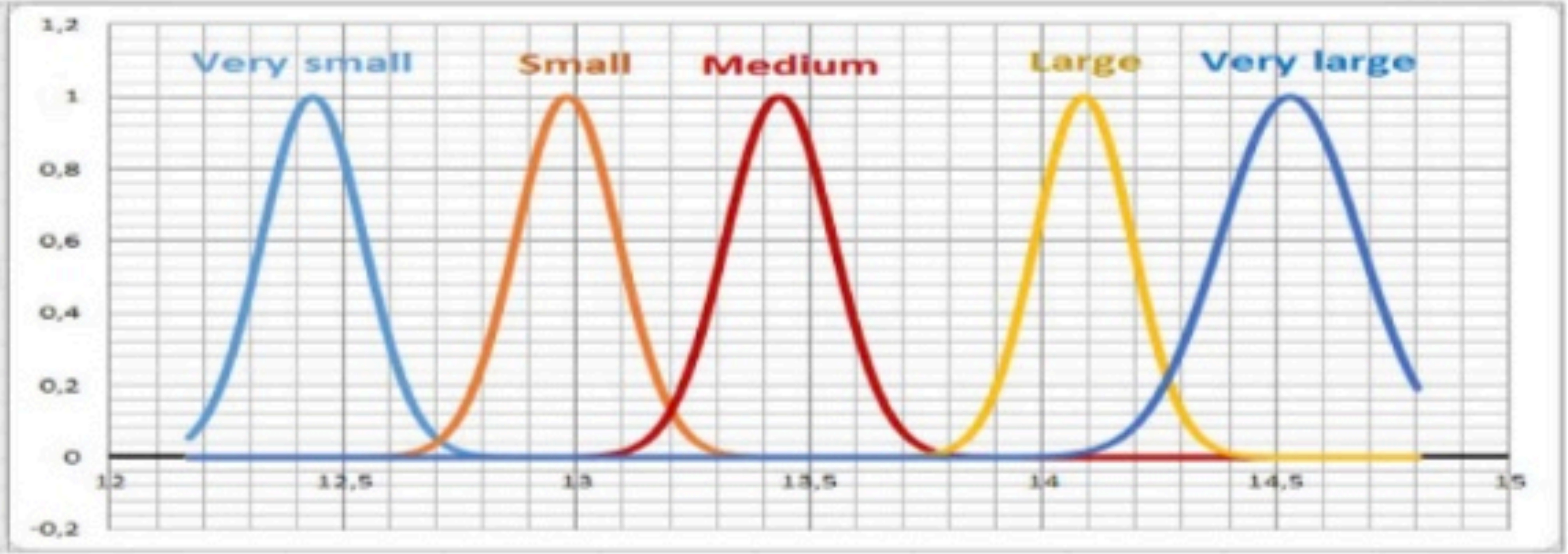
class	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	14,23	1,71	2,43	15,6	127	2,8	3,06	0,28	2,29	5,64	1,04	3,92	1065
	13,2	1,78	2,14	11,2	100	2,65	2,76	0,26	1,28	4,38	1,05	3,4	1050
	13,16	2,36	2,67	18,6	101	2,8	3,24	0,3	2,81	5,68	1,03	3,17	1185
	14,37	1,95	2,5	16,8	113	3,85	3,49	0,24	2,18	7,8	0,86	3,45	1480
	13,24	2,59	2,87	21	118	2,8	2,69	0,39	1,82	4,32	1,04	2,93	735
	14,2	1,76	2,45	15,2	112	3,27	3,39	0,34	1,97	6,75	1,05	2,85	1450
	14,39	1,87	2,45	14,6	96	2,5	2,52	0,3	1,98	5,25	1,02	3,58	1290
	14,06	2,15	2,61	17,6	121	2,6	2,51	0,31	1,25	5,05	1,06	3,58	1295
	14,83	1,64	2,17	14	97	2,8	2,98	0,29	1,98	5,2	1,08	2,85	1045
	13,86	1,35	2,27	16	98	2,98	3,15	0,22	1,85	7,22	1,01	3,55	1045
2	12,37	0,94	1,36	10,6	88	1,98	0,57	0,28	0,42	1,95	1,05	1,82	520
	12,33	1,1	2,28	16	101	2,05	1,09	0,63	0,41	3,27	1,25	1,67	680
	12,64	1,36	2,02	16,8	100	2,02	1,41	0,53	0,62	5,75	0,98	1,59	450
	13,67	1,25	1,92	18	94	2,1	1,79	0,32	0,73	3,8	1,23	2,46	630
	12,37	1,13	2,16	19	87	3,5	3,1	0,19	1,87	4,45	1,22	2,87	420
	12,17	1,45	2,53	19	104	1,89	1,75	0,45	1,03	2,95	1,45	2,23	355
	12,37	1,21	2,56	18,1	98	2,42	2,65	0,37	2,08	4,6	1,19	2,3	678
	13,11	1,01	1,7	15	78	2,98	3,18	0,26	2,28	5,3	1,12	3,18	502
	12,37	1,17	1,92	19,6	78	2,11	2	0,27	1,04	4,68	1,12	3,48	510
	13,34	0,94	2,36	17	110	2,53	1,3	0,55	0,42	3,17	1,02	1,93	750
3	12,86	1,35	2,32	18	122	1,51	1,25	0,21	0,94	4,1	0,76	1,29	630
	12,88	2,99	2,4	20	104	1,3	1,22	0,24	0,83	5,4	0,74	1,42	530
	12,81	2,31	2,4	24	98	1,15	1,09	0,27	0,83	5,7	0,66	1,36	560
	12,7	3,55	2,36	21,5	106	1,7	1,2	0,17	0,84	5	0,78	1,29	600
	12,51	1,24	2,25	17,5	85	2	0,58	0,6	1,25	5,45	0,75	1,51	650
	12,6	2,46	2,2	18,5	94	1,62	0,66	0,63	0,94	7,1	0,73	1,58	695
	12,25	4,72	2,54	21	89	1,38	0,47	0,53	0,8	3,85	0,75	1,27	720
	12,53	5,51	2,64	25	96	1,79	0,6	0,63	1,1	5	0,82	1,69	515
	13,49	3,59	2,19	19,5	88	1,62	0,48	0,58	0,88	5,7	0,81	1,82	580
	12,84	2,96	2,61	24	101	2,32	0,6	0,53	0,81	4,92	0,89	2,15	590

From the work by Savchuk Artem (2016 Fall)

Two sets of membership function from 13 attributes (1)

Membership functions for attribute x1(Alcohol):

	very small	small	medium	large	very large
average	12,43	12,98	13,435	14,0875	14,53
std	0,024	0,024	0,0263	0,0213	0,045

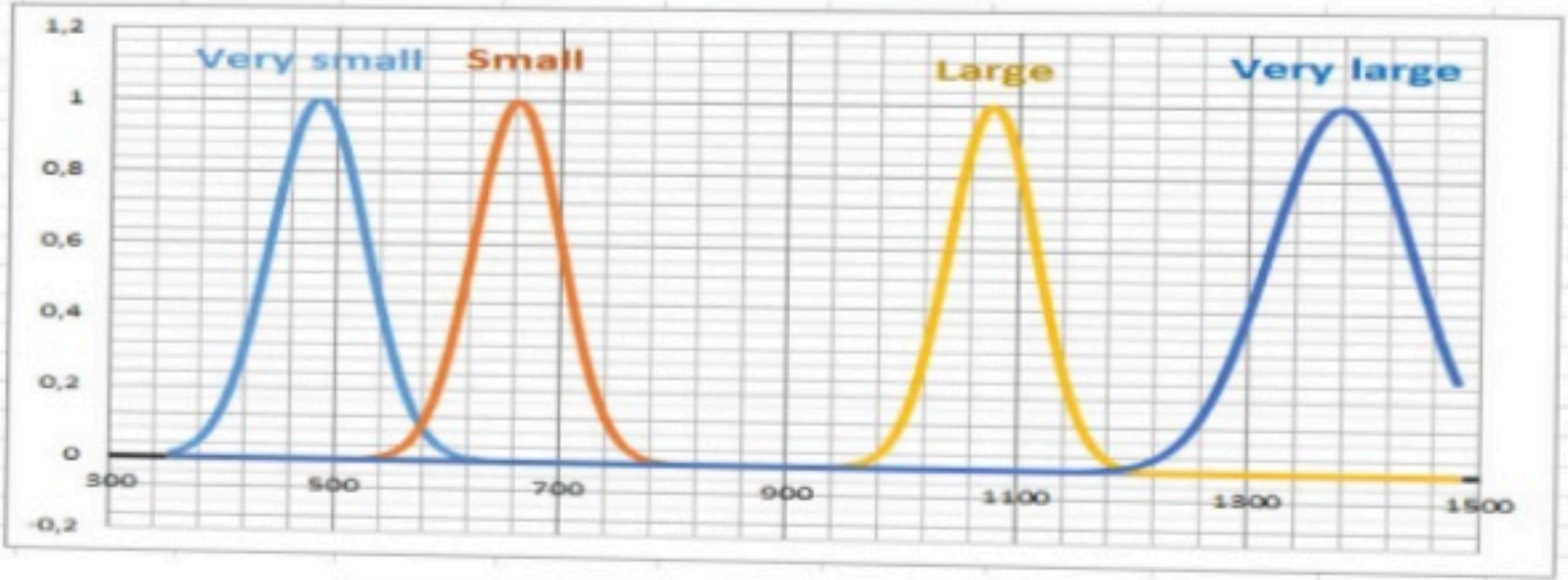


From the work by Savchuk Artem (2016 Fall)

Two sets of membership function from 13 attributes (2)

Membership functions for attribute x13 (Proline):

	very small	small	medium	large	very large
average	468,67	661,5	0	1078	1378,75
std	3670,89	3003,91	0	2916	7554,69



From the work by Savchuk Artem (2016 Fall)

Rules to classify a wine dataset

#	If X1	AND X2	AND X3	AND X4	AND X5	AND X6	AND X7	AND X8	AND X9	AND X10	AND X11	AND X12	AND X13	Then
1	large	small	large	very small	small	large	very large	small	large	very large	large	medium	very large	A
2	very large	small	large	medium	medium	medium	large	very small	large	medium	large	large	large	A
3	very small	very small	medium	very large	very small	small	medium	small	large	medium	large	very small	small	B
4	medium	very small	small	medium	medium	medium	small	large	very small	small	large	small	small	B
5	small	medium	large	large	small	small	very small	medium	small	very large	small	medium	very small	C
6	very small	small	very large	very large	large	very small	very small	large	very small	medium	small	small	small	C
7	very large	large	small	small	very large	very large	small	large	very large	large	medium	very large	medium	Other

From the work by Savchuk Artem (2016 Fall)

Wine data for validation

class	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	14,1	2,16	2,3	18	105	2,95	3,32	0,22	2,38	5,75	1,25	3,17	1510
	14,12	1,48	2,32	16,8	95	2,2	2,43	0,26	1,57	5	1,17	2,82	1280
	13,75	1,73	2,41	16	89	2,6	2,76	0,29	1,81	5,6	1,15	2,9	1320
	14,75	1,73	2,39	11,4	91	3,1	3,69	0,43	2,81	5,4	1,25	2,73	1150
	14,38	1,87	2,38	12	102	3,3	3,64	0,29	2,96	7,5	1,2	3	1547
2	12,21	1,19	1,75	16,8	151	1,85	1,28	0,14	2,5	2,85	1,28	3,07	718
	12,29	1,61	2,21	20,4	103	1,1	1,02	0,37	1,46	3,05	0,906	1,82	870
	13,86	1,51	2,67	25	86	2,95	2,86	0,21	1,87	3,38	1,36	3,16	410
	13,49	1,66	2,24	24	87	1,88	1,84	0,27	1,03	3,74	0,98	2,78	472
	12,99	1,67	2,6	30	139	3,3	2,89	0,21	1,96	3,35	1,31	3,5	985
3	12,93	2,81	2,7	21	96	1,54	0,5	0,53	0,75	4,6	0,77	2,31	600
	13,36	2,56	2,35	20	89	1,4	0,5	0,37	0,64	5,6	0,7	2,47	780
	13,52	3,17	2,72	23,5	97	1,55	0,52	0,5	0,55	4,35	0,89	2,06	520
	13,62	4,95	2,35	20	92	2	0,8	0,47	1,02	4,4	0,91	2,05	550
	12,25	3,88	2,2	18,5	112	1,38	0,78	0,29	1,14	8,21	0,65	2	855

From the work by Savchuk Artem (2016 Fall)

Result of validate rules

No.	Family A	Family B	Family C	Evaluation
#1	A	B	C	Good
#2	A	C	C	Not Good
#3	A	A	C	Not Good
#4	A	B	C	Good
#5	A	Other	C	Not Good
Success rate	100%	40,00%	100%	40%

From the work by Savchuk Artem (2016 Fall)

Practice 3

To design classification rules

1. Select one dataset from those given
2. Create table with raw data from upper half of the dataset
   by sorting data in each column from the smallest to the largest
3. Each colum is devided into 5 categories: VS, S, M, L and VL by (max-min)/5
4. Define Gaussian membership functions μ_{ij} for $i = 1, 2, \ldots, N$ and $j = 1, 2, 3, 4, 5$
   where N is the number of attributes
   and
   $j = 1, 2, 3, 4, 5$ means VS, S, M, L, VL, respectively
5. Translate the table in 2. into fuzzy variables VS, S, M, L, VL
6. Create P rules from upper half of the dataset

To validate the rules

7. Apply Takagi-Sugeno formula to lower half of the dataset
8. Create table with columns being $x_1, x_2, \cdots x_N$ real class, calculated class
9. Calculate overall success rate

From the work by Savchuk Artem (2016 Fall)

III. Time Series Prediction

Forecasting values from their history

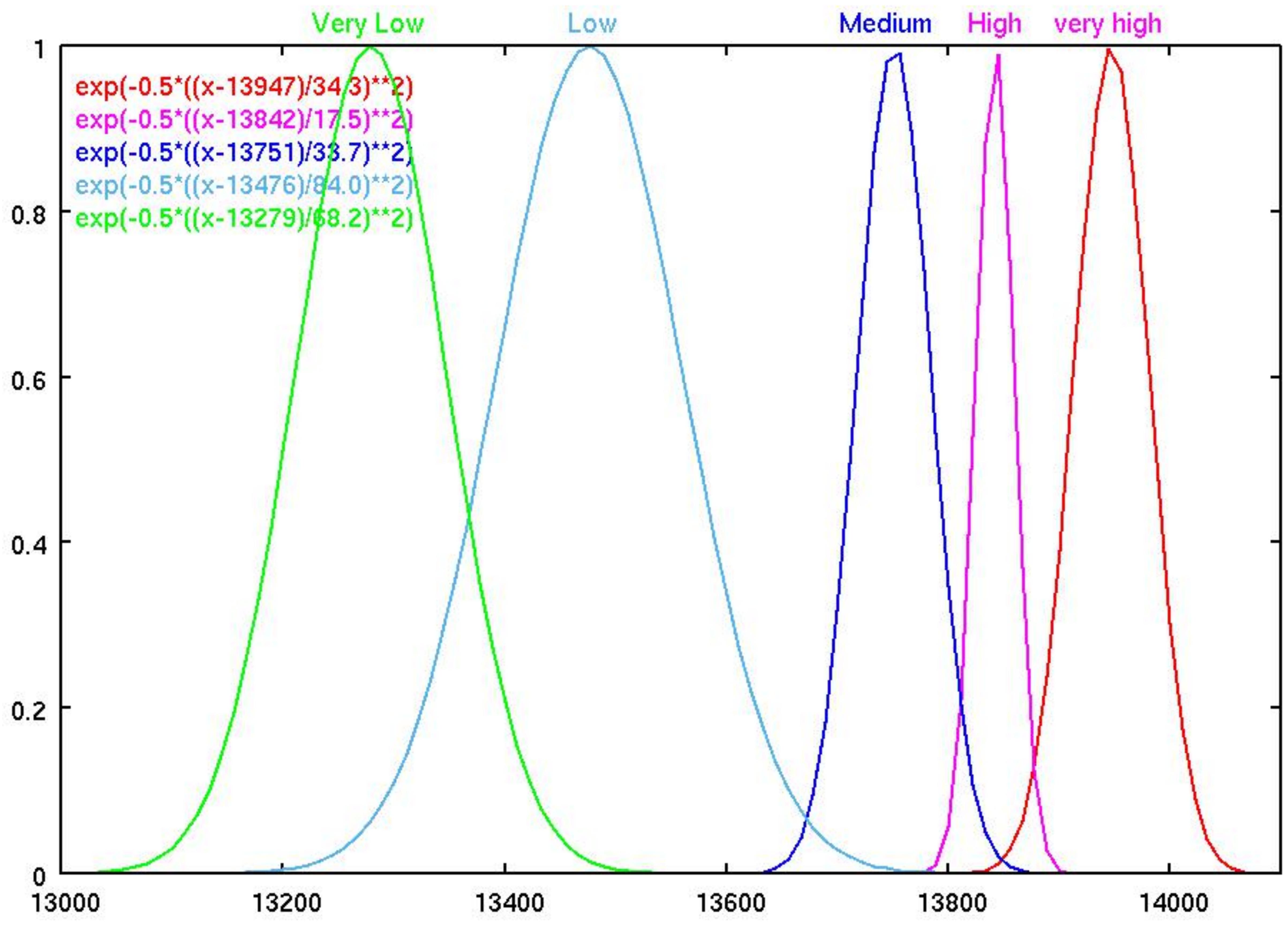
Assume y(t) is a valuable y at time t
such as price of a stock during a day

R_i: If $y(t-1)$ is A_1^i and $y(t-2)$ is A_2^i and $\cdots$ and $y(t-n+1)$ is A_n^i then $y(t)$ is g^i.

An example of time-series forecasting

Create your set own set of rules such as
If y(t-1) = medium AND y(t-2) = law AND y(t-3) = medium they y(t) =2

Original time series	Sorted original and categorize		Human language of orginal
13725	14000	Very High avg = 13947 std - 34.3	medium
13816	13943		high
13739	13943		medium
13403	13943		law
13443	13904		law
13425	13851	High avg = 13842 std = 17.5	law
13305	13851		very law
13271	13851		very law
13658	13816		law
13904	13786	Medium avg = 13750 std = 33.7	very high
13469	13786		law
13182	13786		very law
13463	13739		law
13786	13725		medium
13943	13717		very high
13851	13717		high
13786	13658	Law avg = 13476 std = 84.0	medium
13717	13474		medium
13943	13469		very high
13851	13463		high
13362	13443		very law
13212	13425		very law
13358	13403		very law
13265	13362	Very Law avg = 13279 std = 68	very law
13474	13358		law
13786	13305		medium
13717	13271		medium
13943	13265		very high
13851	13212		high
14000	13182		very high



Then estimate y(t+1), y(t+2), …

By using Takagi-Sugeno Formula

$$y_j = \frac{\sum_{k=1}^{H}(M_k(\mathbf{x}) \cdot g_k)}{\sum_{k=1}^{H} M_k(\mathbf{x})}$$

where

$$M_k(\mathbf{x}) = \prod_{i=1}^{N} \mu_{ik}(x_i)$$

where μ_{ik} is i-th attribute of k-th rule

Create your own set of rules!

such as

IF y(t-1) = low AND y(t-2) = low AND y(t-3) = low THEN y(t) =2 (low)
OR
IF y(t-1) = low AND y(t-2) = medium AND y(t-3) = high THEN y(t) =5 (very high)
OR
……

y(t-1) = 13449, y(t-2) = 13352, y(t-3) = 13029, y(t-4) = 13227, y(t-5) = 13918

Try to predict y(t-2), y(t-1) and compare real vlaue
Also predict y(t)

Forecasting a value at time t from some related items at time t

E.G.

Forecasting BRY today from EUR, USD, RUB and JPY today

$$R_i\text{: If } x_1(t) \text{ is } A_1^i \text{ and } x_2(t) \text{ is } A_2^i \text{ and } \cdots \text{ and } x_n(t) \text{ is } A_n^i \text{ then } y(t) \text{ is } g^i.$$

Time	EUR	USD	RUB	JPY	BRY
t					
t-1					
t-2					
t-3					
t-4					
t-5					
...					

Yet another forecasting

year	enrolled	actual change	F(t) (degree of membership by vector representation) big decrease	decrease	no change	increase	big increase	too big increase	predicted change
1971	13055								
1972	13563	+508	0	0	0	0.5	1	0.5	
1973	13867	+304	0	0	0	1	0.5	0	
1974	14696	+829	0	0	0	0	0.5	1	
1975	15460	+764							
1976	15311	-149			....				
1977	15603	+292							⋮
1978	15861	+258							
1979	16807	+946							
1980	16919	+112							
1981	16388	-531							?
1982	15433	-955							
1983	15497	+64							
1984	15145	-352							
1985	15163	+18							⋮
1986	15984	+821							
1987	16859	+875							
1988	18150	+1291							
1989	18970	+820							
1990	19328	+358							
1991	19337	+9							
1992	18876	-461							

very very small very small small ...... big very big too big

$$F(t) = [\ 1 \quad 0.5 \quad 0 \quad 0 \quad \cdots\cdots \quad 0 \quad 0 \quad 0\ ]$$

$$C(t) = F(t-1) = [\ c_1 \quad \cdots\cdots \quad c_m\ ]$$

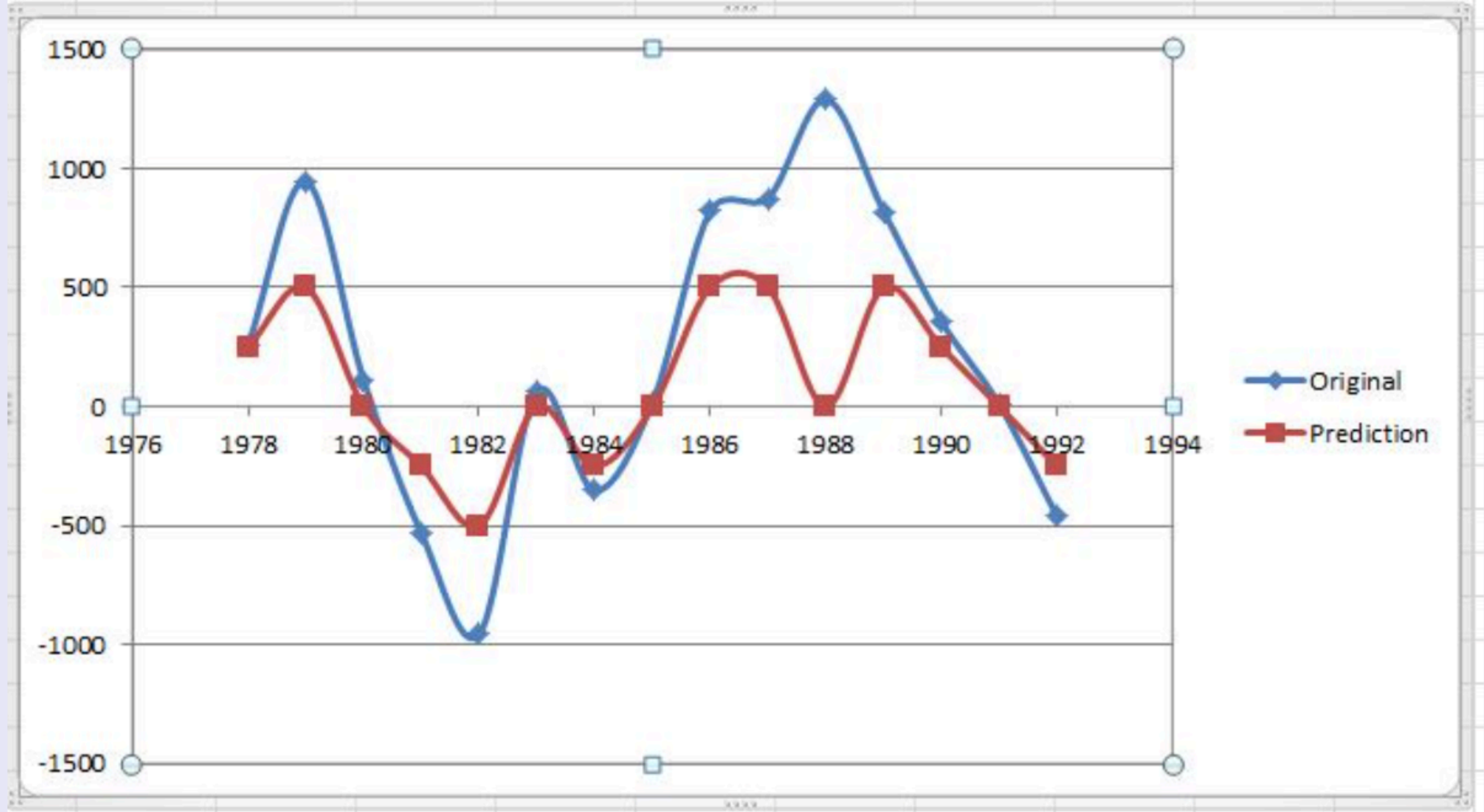
$$O(t) = \begin{bmatrix} F(t-2) \\ \vdots \\ F(t-w-1) \end{bmatrix} => \begin{bmatrix} o_{11} & \cdots\cdots & o_{1m} \\ & \vdots & \\ o_{w1} & \cdots\cdots & o_{wm} \end{bmatrix}$$

$$R(t) = \begin{bmatrix} c_1 \cdot o_{11} & \cdots\cdots & c_m \cdot o_{1m} \\ & \vdots & \\ c_w \cdot o_{w1} & \cdots\cdots & c_w \cdot o_{wm} \end{bmatrix} => \begin{bmatrix} R_{11} & & R_{1m} \\ & \vdots & \\ R_{w1} & \cdots\cdots & R_{wm} \end{bmatrix}$$

$$F(t) = [\ \max(R_{11}, R_{21}, \cdots, R_{w1}) \quad \cdots\cdots \quad \max(R_{1w}, R_{22}, \cdots, R_{ww})\ ]$$

An examples of prediction ... By Soloduha Pavel

	big decrease	low decrease	not change	low increase	medium increase	big increase	Prediction	Original
1971								
1972								508
1973								304
1974								829
1975								764
1976								-149
1977								292
1978	0	0	0,5	0,5	0	0	250	258
1979	0	0	0	0,25	0	0	500	946
1980	0	0,25	1	0,25	0	0	0	112
1981	0	0,5	0,5	0	0	0	-250	-531
1982	0	0,25	0	0	0	0	-500	-955
1983	0	0,25	1	0,25	0	0	0	64
1984	0	0,5	0,5	0	0	0	-250	-352
1985	0	0,25	1	0,25	0	0	0	18
1986	0	0	0	0,25	0	0	500	821
1987	0	0	0	0,25	0	0	500	875
1988	0	0	0	0	0	0	0	1291
1989	0	0	0	0,25	0	0	500	820
1990	0	0	0,5	0,5	0	0	250	358
1991	0	0,25	1	0,25	0	0	0	9
1992	0	0,5	0,5	0	0	0	-250	-461



IV. Fuzzy Clustering

Another fuzzy arithmetic

Fuzzy relation

$\star$ Example 1 ... $X = \{1, 2\}$, $Y = \{2, 3, 4\}$, $R : X < Y$

Let's think of it as a *crisp* logic, that is, the value is 1 (yes) or 0 (no). Then membership function of this relation will be:

$X \setminus Y$	2	3	4
1	1	0	0
2	1	1	0
3	0	0	1

Then what about the relation $R : x \approx y$. Let's think of this example with *fuzzy* logic.

$\star$ Example 2 ... $X = \{1, 2\}$, $Y = \{2, 3, 4\}$, $R : X \approx Y$

$X \setminus Y$	2	3	4
1	2/3	1/3	0
2	1	2/3	1/3
3	2/3	1	2/3

$\star$ Example 3 ... X = {Brest, London, BuenosAires} Y=Tokyo, NewYork, Minsk, Johanesburg R: very far.

$X \setminus Y$	Tokio	New York	Minsk	Johanesburg
Brest				
London				
Buenos Aires				

Or

$X = \{\text{green, yellow, red}\}$, $Y = \{\text{unripe, semiripe, ripe}\}$.

A red apple may be ripe, a yellow apple is probably seme-ripe, and a green apple is most likely unripe.

$R_1 =$

X \ Y	unripe	semi-ripe	ripe
green	1	0.5	0
yellow	0.3	1	0.4
red	0	0.2	1

Yet another similar fuzzy relations

Y = {unripe, semi-ripe, ripe}, Z = {sour, sour-sweet, sweet}

Let's call this relation R_2

R_2 =

Y \ Z	sour	sour-sweet	sweet
unripe	0.8	0.5	0.1
semi-ripe	0.1	0.7	0.5
ripe	0.2	0.3	9

We combine these two relations R1 and R2 by the formula

$$\mu_R(x,z) \geq \max_{y\in X}\{\min\{\mu_R(x,y),\mu_R(y,z)\}\},$$

$R_1 =$

X \ Y	unripe	semi-ripe	ripe
green	1	0.5	0
yellow	0.3	1	0.4
red	0	0.2	1

$R_2 =$

Y \ Z	sour	sour-sweet	sweet
unripe	0.8	0.5	0.1
semi-ripe	0.1	0.7	0.5
ripe	0.2	0.3	9

$R_3 =$

X \ Z	sour	sour-sweet	sweet
green	0.8	0.5	0.5
yellow	0.3	0.7	0.5
red	0.2	0.3	0.9

Expression by our daily language

This relation could be expressed by our daily language like

"If tomato is red then it's most likely sweet , possibly sour-sweet, and unlikely sour."

"If tomato is yellow then probably it's sour-sweet , possibly sour, maybe sweet."

"If tomato is green then almost always sour, less likely sour-sweet, unlikely sweet."

Or, we could say:

"Now tomato is more or less red, then what is taste like?"

An exercise

A \ B	d	e	f
a	1	0.5	0
b	0.3	1	0.4
c	0	0.2	1

o

B \ C	g	h	i	j
d	0.8	0.2	1	0.9
e	0.3	1	0.6	0.4
f	1	0.2	0	0.4

= ?

Clustering by similarity

An example of a forign alphabet we usually don't know ... By Ilya Marchanka
Original matrix R^(0) of how resemble with each other

Source matrix :

Arabic	0	1	2	3	4	5	6	7	8	9
Bengali	০	১	২	৩	৪	৫	৬	৭	৮	৯
০	1	0,3	0,1	0,7	0,6	0,5	0,4	0,4	0,5	0,4
১	0,3	1	0,8	0,4	0,4	0,2	0,6	0,1	0,3	0,5
২	0,1	0,8	1	0,4	0,3	0,1	0,2	0,1	0,1	0,3
৩	0,7	0,4	0,4	1	0,5	0,4	0,8	0,1	0,2	0,6
৪	0,6	0,4	0,3	0,5	1	0,6	0,3	0,8	0,5	0,3
৫	0,5	0,2	0,1	0,4	0,6	1	0,3	0,1	0,2	0,6
৬	0,4	0,6	0,2	0,8	0,3	0,3	1	0,5	0,4	0,7
৭	0,4	0,1	0,1	0,1	0,8	0,1	0,5	1	0,5	0,1
৮	0,5	0,3	0,1	0,2	0,5	0,2	0,4	0,5	1	0,4
৯	0,4	0,5	0,3	0,6	0,3	0,6	0,7	0,1	0,4	1

Clustering by Fuzzy Relation of Proximity

Algorithm 2 *0. Initialize I and C*

1. *Calculate a max-min similarity-relation* $R^{(0)} = [a_{ij}]$
2. *Set* $a_{ij} = 0$ *for all* $a_{ij} < \alpha$ *and* $i = j$
3. *Select* s *and* t *such that* $a_{st} = \max\{a_{ij} | i < j AND i, j \in I\}$. *When tie, select one of these pairs at random*

   WHILE $a_{st} \neq 0$ *DO put* s *and* t *into the same cluster* $C = \{s, t\}$ *ELSE 4.*
   ELSE all indices $\in I$ *into separated clusters and STOP*

4. *Choose* $u \in I - C$ *such that*

$$\sum_{i \in C} a_{iu} = \max_{j \in I-C} \{\sum_{i \in C} a_{ij} | a_{ij} \neq 0\}$$

   When a tie, select one such u at random.

   WHILE such a u exists, put u into $C = \{s, t, u\}$ *and REPEAT 4.*

5. *Let* $I = I - C$ *and GOTO 3.*

0. Initialize I = {1,2,3,…,N} and C = { }

Find maximum a_{ij} (if some are equal, select one at random)

$$\begin{matrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} & a_{19} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} & a_{28} & a_{29} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} & a_{38} & a_{39} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} & a_{48} & a_{49} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} & a_{59} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} & a_{68} & a_{69} \\ a_{71} & a_{72} & a_{73} & a_{74} & a_{75} & a_{76} & a_{77} & a_{78} & a_{79} \\ a_{81} & a_{82} & a_{83} & a_{84} & a_{85} & a_{86} & a_{87} & a_{88} & a_{89} \\ a_{91} & a_{92} & a_{93} & a_{94} & a_{95} & a_{96} & a_{97} & a_{98} & a_{99} \end{matrix} \rightarrow$$

Look for

Maximum a_{ij} was a_{st}

Put s and t to C{ }

Assume now e.g. a_{47}

So

$$C = \{4, 7\}$$

$$I = \{1, 2, 3, 5, 6, 8, 9\}$$

Then calculate $\max \{ \sum_{i=1}^{2} (a_{is} + a_{it}) \}$

I.e. $\max \{ (a_{sj} + a_{tj}),\ (a_{sj} + a_{tj}),\ \cdots,\ (a_{sj} + a_{tj}) \}$

If multiple such j then select one at random

$$
\begin{matrix}
a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} & a_{19} \\
a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} & a_{28} & a_{29} \\
a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} & a_{38} & a_{39} \\
a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} & a_{48} & a_{49} \\
a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} & a_{59} \\
a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} & a_{68} & a_{69} \\
a_{71} & a_{72} & a_{73} & a_{74} & a_{75} & a_{76} & a_{77} & a_{78} & a_{79} \\
a_{81} & a_{82} & a_{83} & a_{84} & a_{85} & a_{86} & a_{87} & a_{88} & a_{89} \\
a_{91} & a_{92} & a_{93} & a_{94} & a_{95} & a_{96} & a_{97} & a_{98} & a_{99}
\end{matrix}
$$

Assume now e.g. $a_{94} + a_{97}$ is such maximum

Then put 9 into C

$C = \{4, 7, 9\}$

$I = \{1, 2, 3, 5, 6, 8\}$

Choose u from

$$\text{such that } \sum_{i \in C} a_{iu} \max_{j \in \{I-C\}} \{ \sum_{i \in C} a_{ij} \mid a_{ij} \neq 0 \}$$

If multiple such u then select one at random

$$\begin{matrix}
a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} & a_{19} \\
a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} & a_{28} & a_{29} \\
a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} & a_{38} & a_{39} \\
a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} & a_{48} & a_{49} \\
a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} & a_{59} \\
a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} & a_{68} & a_{69} \\
a_{71} & a_{72} & a_{73} & a_{74} & a_{75} & a_{76} & a_{77} & a_{78} & a_{79} \\
a_{81} & a_{82} & a_{83} & a_{84} & a_{85} & a_{86} & a_{87} & a_{88} & a_{89} \\
a_{91} & a_{92} & a_{93} & a_{94} & a_{95} & a_{96} & a_{97} & a_{98} & a_{99}
\end{matrix}$$

Assume now $a_{24} + a_{27} + a_{29}$ is such maximum

Put 2 into C

$C = \{2, 4, 7, 9\}$

$I = \{1, 3, 5, 6, 8\}$

Repeat this prosedure

$$\begin{matrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} & a_{19} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} & a_{28} & a_{29} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} & a_{38} & a_{39} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} & a_{48} & a_{49} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} & a_{59} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} & a_{68} & a_{69} \\ a_{71} & a_{72} & a_{73} & a_{74} & a_{75} & a_{76} & a_{77} & a_{78} & a_{79} \\ a_{81} & a_{82} & a_{83} & a_{84} & a_{85} & a_{86} & a_{87} & a_{88} & a_{89} \\ a_{91} & a_{92} & a_{93} & a_{94} & a_{95} & a_{96} & a_{97} & a_{98} & a_{99} \end{matrix}$$

Assume now

$$a_{11} + a_{12} + a_{14} + a_{17} + a_{19}$$

is such maximum

Put 2 into C

$$C = \{1, 2, 4, 7, 9\}$$

$$I = \{3, 5, 6, 8\}$$

Till a_{ij} included is/are 0 Or no such maximum

a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	a_{37}	a_{38}	a_{39}
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	a_{47}	a_{48}	a_{49}
a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	a_{57}	a_{58}	a_{59}
a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	a_{66}	a_{67}	a_{68}	a_{69}
a_{71}	a_{72}	a_{73}	a_{74}	a_{75}	a_{76}	a_{77}	a_{78}	a_{79}
a_{81}	a_{82}	a_{83}	a_{84}	a_{85}	a_{86}	a_{87}	a_{88}	a_{89}
a_{91}	a_{92}	a_{93}	a_{94}	a_{95}	a_{96}	a_{97}	a_{98}	a_{99}

If

$$a_{11} + a_{12} + a_{14} + a_{17} + a_{19}$$

is such maximum like previous slide

but incase $a_{11} = 0$ for example

Stop

And start again from the beginning

searching for another cluster

with

$$I = \{2, 4. 7, 9\}$$

$$C = \{ \}$$

Example: Let's Start with the following $R^{(0)}$,

$$R^{(0)} = \begin{bmatrix} 1 & .7 & .5 & .8 & .6 & .6 & .5 & .9 & .4 & .5 \\ .7 & 1 & .3 & .6 & .7 & .9 & .4 & .8 & .6 & .6 \\ .5 & .3 & 1 & .5 & .5 & .4 & .1 & .4 & .7 & .6 \\ .8 & .6 & .5 & 1 & .7 & .5 & .5 & .7 & .5 & .6 \\ .6 & .7 & .5 & .7 & 1 & .6 & .4 & .5 & .8 & .9 \\ .6 & .9 & .4 & .5 & .6 & 1 & .3 & .7 & .7 & .5 \\ .5 & .4 & .1 & .5 & .4 & .3 & 1 & .6 & .2 & .3 \\ .9 & .8 & .4 & .7 & .5 & .7 & .6 & 1 & .4 & .4 \\ .4 & .6 & .7 & .5 & .8 & .7 & .2 & .4 & 1 & .7 \\ .5 & .6 & .6 & .6 & .9 & .5 & .3 & .4 & .7 & 1 \end{bmatrix}$$

Now assumming $\alpha = 0.55$ apply [1.] and [2.]

$$\begin{bmatrix} 0 & .7 & 0 & .8 & .6 & .6 & 0 & .9 & 0 & 0 \\ .7 & 0 & 0 & .6 & .7 & .9 & 0 & .8 & .6 & .6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7 & .6 \\ .8 & .6 & 0 & 0 & .7 & 0 & 0 & .7 & 0 & .6 \\ .6 & .7 & 0 & .7 & 0 & .6 & 0 & 0 & .8 & .9 \\ .6 & .9 & 0 & 0 & .6 & 0 & 0 & .7 & .7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .6 & 0 & 0 \\ .9 & .8 & 0 & .7 & 0 & .7 & .6 & 0 & 0 & 0 \\ 0 & .6 & .7 & 0 & .8 & .7 & 0 & 0 & 0 & .7 \\ 0 & .6 & .6 & .6 & .9 & 0 & 0 & 0 & .7 & 0 \end{bmatrix}$$

Then repeat $R^{(n+1)} = R^{(n)} \circ R^{(n)}$ till $R^{(n)} = R^{(n+1)}$.

$$R^{(n)} = \begin{bmatrix} 1 & .2 & .5 & .8 & .6 & .2 & .3 & .9 & .4 & .3 \\ .2 & 1 & .3 & .6 & .7 & .9 & .2 & .8 & .3 & .2 \\ .5 & .3 & 1 & .5 & .3 & .4 & .1 & .3 & .7 & .6 \\ .8 & .6 & .5 & 1 & .7 & .3 & .5 & .4 & .1 & .3 \\ .6 & .7 & .3 & .7 & 1 & .2 & .4 & .5 & .8 & .9 \\ .2 & .9 & .4 & .3 & .2 & .4 & .1 & .3 & .7 & .2 \\ .3 & .2 & .1 & .5 & .4 & .1 & 1 & .6 & .1 & .3 \\ .9 & .8 & .3 & .4 & .5 & .3 & .6 & 1 & 0 & .2 \\ .4 & .3 & .7 & .1 & .8 & .7 & .1 & 0 & 1 & .1 \\ .3 & .2 & .6 & .3 & .9 & .2 & .3 & .2 & .1 & 1 \end{bmatrix}$$

First, set $I = \{1,2,3,4,5,6,7,8,9,10\}$ and $C = \{\ \}$. Then

Step 3. Now $a_{18} = a_{26} = a_{5\ 10} = 0.9$ are maximum and a_{18} is randomly selected. Then $C = \{1,8\}$.

Step 4. $a_{12} + a_{82} = a_{14} + a_{84} = 1.5$ are maximum and $j = 4$ is randomly selected. Then $C = \{1,8,4\}$.

Step 4. $a_{12} + a_{42} + a_{82} = 2.1$ is maximum, then $C = \{1,8,4,2\}$.

Step 4. There are no j such that $a_{1j} + a_{2j} + a_{4j} + a_{8j}$ is maximum. Then final $C = \{1,8,4,2\}$.

- ⋆ $a_{16} + a_{26} + a_{46} + a_{86} = 0.6 + 0.9 + 0 + 0.7 = 2.2$ seems maximum but actually not because $a_{46} = 0$

Note that $\sum_{i \in C} a_{iu} = \max_{j \in I \setminus C} \{ \sum_{i \in C} a_{ij} | a_{ij} \neq 0 \}$

Step 5. Let $I = \{3,5,6,7,9,10\}$

Step 3. $a_{5\ 10} = 0.9$ is maximum. Then renew C as $\{5,10\}$.

Step 4. $a_{59} + a_{10\ 9} = 1.5$ is maximum. Then $C = \{5,10,9\}$.

Step 4. There are no j in $\{3,6,9\}$ such that $a_{5j} + a_{9j} + a_{10j}$ is maximum. Then final $C = \{5,10,9\}$.

Step 5. Let $I = \{3,6,7\}$.

Step 3. Now $a_{36} = a_{37} = a_{67} = 0$. Then $\{3\}$, $\{6\}$and $\{7\}$are three separated clusters. In fact,

$$\begin{bmatrix} a_{33} & a_{36} & a_{37} \\ a_{63} & a_{66} & a_{67} \\ a_{73} & a_{76} & a_{77} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

So $\sum_{i\in\{3,6,7\}} a_{iu} = \max_{j\in\{3,6,7\}}\{\sum_{i\in C} a_{ij} | a_{ij} \neq 0\}$ does not exit any more.

In this way, when $\alpha = 0.55$, we have 5 clasters $\{1,8,4,2\}$,$\{5,10,9\}$, $\{3\}$, $\{6\}$ and $\{7\}$ are obtained.

Summary-1

By repeating $R^{(n+1)} = R^{(n)} \circ R^{(n)}$ till $R^{(n)} = R^{(n+1)}$. In this way, similarity-relation $R^{(n)}$ will be calculated as:

$$R^{(n)} = \begin{bmatrix} 1 & .2 & .5 & .8 & .6 & .2 & .3 & .9 & .4 & .3 \\ .2 & 1 & .3 & .6 & .7 & .9 & .2 & .8 & .3 & .2 \\ .5 & .3 & 1 & .5 & .3 & .4 & .1 & .3 & .7 & .6 \\ .8 & .6 & .5 & 1 & .7 & .3 & .5 & .4 & .1 & .3 \\ .6 & .7 & .3 & .7 & 1 & .2 & .4 & .5 & .8 & .9 \\ .2 & .9 & .4 & .3 & .2 & .4 & .1 & .3 & .7 & .2 \\ .3 & .2 & .1 & .5 & .4 & .1 & 1 & .6 & .1 & .3 \\ .9 & .8 & .3 & .4 & .5 & .3 & .6 & 1 & 0 & .2 \\ .4 & .3 & .7 & .1 & .8 & .7 & .1 & 0 & 1 & .1 \\ .3 & .2 & .6 & .3 & .9 & .2 & .3 & .2 & .1 & 1 \end{bmatrix}$$

Now apply [1.] and [2.]

$$\begin{bmatrix} 0 & .7 & 0 & .8 & .6 & .6 & 0 & .9 & 0 & 0 \\ .7 & 0 & 0 & .6 & .7 & .9 & 0 & .8 & .6 & .6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7 & .6 \\ .8 & .6 & 0 & 0 & .7 & 0 & 0 & .7 & 0 & .6 \\ .6 & .7 & 0 & .7 & 0 & .6 & 0 & 0 & .8 & .9 \\ .6 & .9 & 0 & 0 & .6 & 0 & 0 & .7 & .7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .6 & 0 & 0 \\ .9 & .8 & 0 & .7 & 0 & .7 & .6 & 0 & 0 & 0 \\ 0 & .6 & .7 & 0 & .8 & .7 & 0 & 0 & 0 & .7 \\ 0 & .6 & .6 & .6 & .9 & 0 & 0 & 0 & .7 & 0 \end{bmatrix}$$

$$R^{(0)} = \begin{bmatrix} 1 & .7 & .5 & .8 & .6 & .6 & .5 & .9 & .4 & .5 \\ .7 & 1 & .3 & .6 & .7 & .9 & .4 & .8 & .6 & .6 \\ .5 & .3 & 1 & .5 & .5 & .4 & .1 & .4 & .7 & .6 \\ .8 & .6 & .5 & 1 & .7 & .5 & .5 & .7 & .5 & .6 \\ .6 & .7 & .5 & .7 & 1 & .6 & .4 & .5 & .8 & .9 \\ .6 & .9 & .4 & .5 & .6 & 1 & .3 & .7 & .7 & .5 \\ .5 & .4 & .1 & .5 & .4 & .3 & 1 & .6 & .2 & .3 \\ .9 & .8 & .4 & .7 & .5 & .7 & .6 & 1 & .4 & .4 \\ .4 & .6 & .7 & .5 & .8 & .7 & .2 & .4 & 1 & .7 \\ .5 & .6 & .6 & .6 & .9 & .5 & .3 & .4 & .7 & 1 \end{bmatrix}$$

$$\mu_R(x,z) \geq \max_{y \in X}\{\min\{\mu_R(x,y), \mu_R(y,z)\}\}$$

Summary-2

$$\begin{bmatrix} 0 & .7 & 0 & .8 & .6 & .6 & 0 & .9 & 0 & 0 \\ .7 & 0 & 0 & .6 & .7 & .9 & 0 & .8 & .6 & .6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7 & .6 \\ .8 & .6 & 0 & 0 & .7 & 0 & 0 & .7 & 0 & .6 \\ .6 & .7 & 0 & .7 & 0 & .6 & 0 & 0 & .8 & .9 \\ .6 & .9 & 0 & 0 & .6 & 0 & 0 & .7 & .7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .6 & 0 & 0 \\ .9 & .8 & 0 & .7 & 0 & .7 & .6 & 0 & 0 & 0 \\ 0 & .6 & .7 & 0 & .8 & .7 & 0 & 0 & 0 & .7 \\ 0 & .6 & .6 & .6 & .9 & 0 & 0 & 0 & .7 & 0 \end{bmatrix}$$

First, set $I = \{1,2,3,4,5,6,7,8,9,10\}$ and $C = \{\ \}$.

Then

3. Now $a_{18} = a_{26} = a_{5\ 10} = 0.9$ are maximum and a_{18} is randomly selected. Then $C = \{1,8\}$.

4. $a_{12} + a_{82} = a_{14} + a_{84} = 1.5$ are maximum and $j = 4$ is randomly selected. Then $C = \{1,8,4\}$.

4. $a_{12} + a_{42} + a_{82} = 2.1$ is maximum, then $C = \{1,8,4,2\}$.

4. There are no j such that $a_{1j} + a_{2j} + a_{4j} + a_{8j}$ is maximum. Then final $C = \{1,8,4,2\}$.

   ⋆ $a_{16} + a_{26} + a_{46} + a_{86} = 0.6 + 0.9 + 0 + 0.7 = 2.2$ seems maximum but actually not because $a_{46} = 0$

   Note that $\sum_{i \in C} a_{iu} = \max_{j \in I \setminus C}\{\sum_{i \in C} a_{ij} | a_{ij} \neq 0\}$

5. Let $I = \{3,5,6,7,9,10\}$

3. $a_{5\ 10} = 0.9$ is maximum. Then renew C as $\{5,10\}$.

4. $a_{59} + a_{10\ 9} = 1.5$ is maximum. Then $C = \{5,10,9\}$.

4. There are no j in $\{3,6,9\}$ such that $a_{5j} + a_{9j} + a_{10j}$ is maximum. Then final $C = \{5,10,9\}$.

5. Let $I = \{3,6,7\}$.

3. Now $a_{36} = a_{37} = a_{67} = 0$. Then $\{3\}$, $\{6\}$and $\{7\}$are three separated clusters. In fact,

$$\begin{bmatrix} a_{33} & a_{36} & a_{37} \\ a_{63} & a_{66} & a_{67} \\ a_{73} & a_{76} & a_{77} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

So $\sum_{i \in \{3,6,7\}} a_{iu} = \max_{j \in \{3,6,7\}}\{\sum_{i \in C} a_{ij} | a_{ij} \neq 0\}$ does not exit any more.

In this way, when $\alpha = 0.55$, we have 5 clasters $\{1,8,4,2\}$,$\{5,10,9\}$, $\{3\}$, $\{6\}$ and $\{7\}$ are obtained.

Let's take 6 KANJ as an example

$R^{\text{similarity}} =$

	由	目	田	旦	日	中
由	1	0.4	0.9	0.3	0.7	0.6
目	0.4	1	0.6	0.7	0.8	0.2
田	0.9	0.6	1	0.4	0.7	0.6
旦	0.3	0.7	0.5	1	0.9	0.1
日	0.7	0.8	0.7	0.9	1	0.1
中	0.6	0.2	0.6	0.1	0.1	1

Calculate R(m+1)=R(m)∘R(m) till R(m+1)=R(m)

$$
\overset{R^{(0)}}{\begin{matrix}
1 & 0.4 & 0.9 & 0.3 & 0.7 & 0.6 \\
0.4 & 1 & 0.6 & 0.7 & 0.8 & 0.2 \\
0.9 & 0.6 & 1 & 0.4 & 0.7 & 0.6 \\
0.3 & 0.7 & 0.5 & 1 & 0.9 & 0.1 \\
0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.1 \\
0.6 & 0.2 & 0.6 & 0.1 & 0.1 & 1
\end{matrix}}
\circ
\overset{R^{(0)}}{\begin{matrix}
1 & 0.4 & 0.9 & 0.3 & 0.7 & 0.6 \\
0.4 & 1 & 0.6 & 0.7 & 0.8 & 0.2 \\
0.9 & 0.6 & 1 & 0.4 & 0.7 & 0.6 \\
0.3 & 0.7 & 0.5 & 1 & 0.9 & 0.1 \\
0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.1 \\
0.6 & 0.2 & 0.6 & 0.1 & 0.1 & 1
\end{matrix}}
=
\overset{R^{(1)}}{\begin{matrix}
1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\
0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\
0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\
0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.5 \\
0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\
0.6 & 0.6 & 0.6 & 0.5 & 0.6 & 1
\end{matrix}}
$$

to continue

$$R^{(1)} \circ R^{(1)} = R^{(2)}$$

$$\begin{matrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.5 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.5 & 0.6 & 1 \end{matrix} \circ \begin{matrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.5 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.5 & 0.6 & 1 \end{matrix} = \begin{matrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.6 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 1 \end{matrix}$$

to continue

$$R^{(2)} \circ R^{(2)} = R^{(3)}$$

$$\begin{bmatrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.6 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.6 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.6 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 1 \end{bmatrix}$$

Now R(2) = R(3)

Now apply
step 1 and step 2

$$R^{(2)} = \begin{matrix} 1 & 0.7 & 0.9 & 0.7 & 0.7 & 0.6 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.6 \\ 0.9 & 0.7 & 1 & 0.7 & 0.7 & 0.6 \\ 0.7 & 0.8 & 0.7 & 1 & 0.9 & 0.6 \\ 0.7 & 0.8 & 0.7 & 0.9 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 1 \end{matrix}$$

Replace all Aii with 0

Replace all Aij < α with 0

$\Downarrow$ Now assume $\alpha = 0.75$

$$\begin{matrix} 0 & 0 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0.8 & 0 \\ 0.9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0.9 & 0 \\ 0 & 0.8 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

0	0	0.9	0	0	0
0	0	0	0.8	0.8	0
0.9	0	0	0	0	0
0	0.8	0	0	0.9	0
0	0.8	0	0.9	0	0
0	0	0	0	0	0

I = {1,2,3,4,5,6}

C = {}

A13=A45=0.9 is maximum

(Don't forget s must be smaller than t)

A45 is randomely selected

Then C={4,5}

Now A42+A52=1.6 is maximum

Then C={2,4,5}

no s such that A2s+A4s+A5s is maximum with $s \in I - C = \{1, 3, 6\}$ and $As \neq 0$

to continue

$$\begin{array}{c|ccc} & (1) & (3) & (6) \\ \hline (1) & 0 & 0.9 & 0 \\ (3) & 0.9 & 0 & 0 \\ (6) & 0 & 0 & 0 \end{array}$$

I = {1,3,6}

C = {}

A13=0.9 is maximum

Then C ={1,3}

no s such that A1s+A3s is maximum

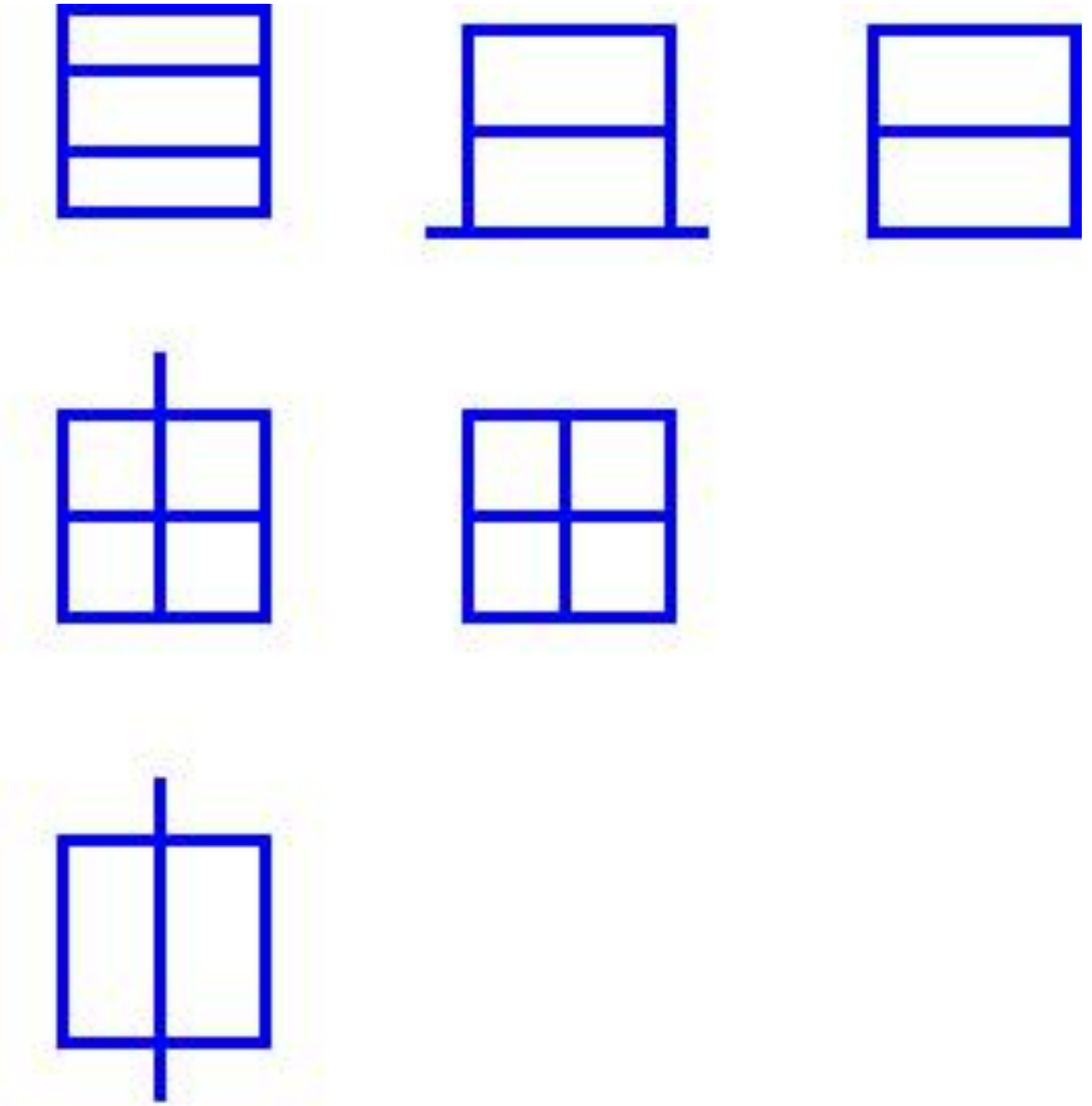
with $s \in I - C = \{6\}$ and $As \neq 0$

Result of clustering

C1 = {2,4,5}

C2 = {1,3}

C3 = {6}



IV. Fuzzy data mining

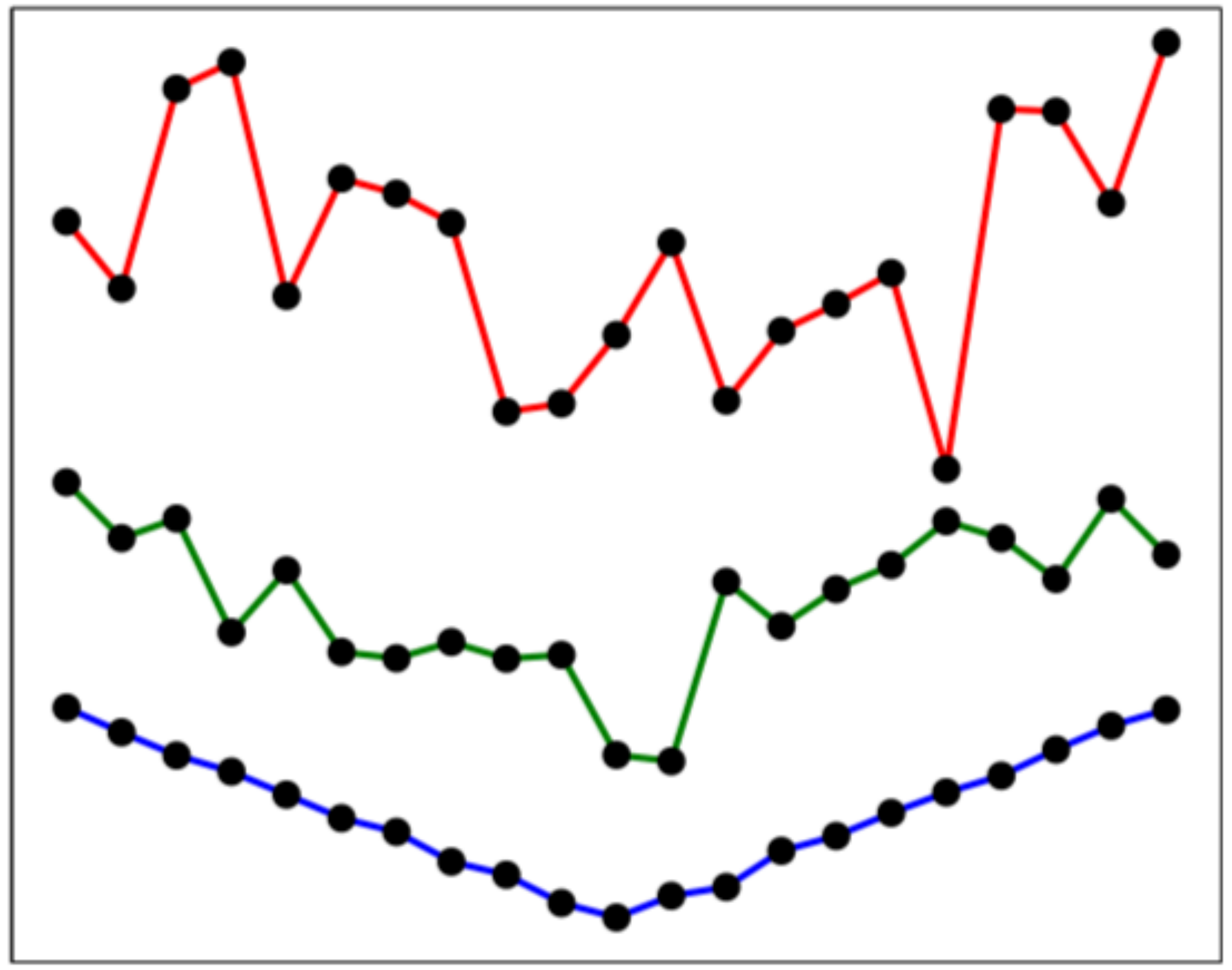
What is Data Mining?

Data mining is an automatic or semi-automatic process
that analyses large amounts of scattered information to make sense of it
and
turn it into knowledge

Data mining

- The process starts with giving a certain input of data to the data mining tools
- It is widely used by organizations in building a marketing strategy, by hospitals for diagnostic tools, etc.
- Today most organizations use data mining for analysis of Big Data
- Also known as “Knowledge Discovery in Databases (KDD)”

Toy Example:
Mining a rule in natural language from data



A profile which is decreasing at the beginning is typically increasing at the end

Application

- Medical Data Analysis
- Phishing Detection
- Invasion Detection
- Commercial Data Analysis

Popular Technique for Data Mining

- Regression
- Association
- Clustering
- Classification
- Outlier analysis
- Decision tree
- Bayes theory
- etc.

Example (1)
Mobile service provider

- From a large amount of data such as billing information, email, text messages, web data transmission, customer service used, the datmining tools can predict predict customers who are looking to change the venders

- With these results a provability score is given. The mobile service providers can provide incentives, offers to customers who are going to change the venders

Example (2)
Retail

- Looking at the purchase history reveals the buying preferences of the customers

- The results help the supermarkets design the placements of products on shelves and offers on items such as coupons on matching products and special discounts on some products.

Example (3)
Market Basket Analysis

- This find the groups of items that are bought together in stores. Analysis of the transactions show the patterns such as which things are bought together often like bread and butter, or which items have higher sales volume on certain days such as beer on Fridays

- This information helps in planning the store layouts, offering a special discount to the items that are less in demand, creating offers such as “Buy 2 get 1 free” or “Get 50% on second purchase” etc

Example (4)
e-Commerce

- Many e commerce sites use data mining over the purchasing history of the customers of the website

- The Amazon etc. show "People also viewed", "Frequently bought together" to the customers who are interacting with the site

Example (5)
Crime prevention

- Data mining detects outliers across a vast amount of data, the criminal data including all details of the crime that has happened. Data mining will study the criminal patterns and criminal trends and predict future events with better accuracy

- The agencies can find out which area is more prone to crime, how much police pesonnel should be deployed, which age group should be targeted, vehicle numbers to be scrutinized, etc.

Example (6)
Research

- Researchers use Data Mining tools to explore the associations between the parameters under research such as environmental conditions like air pollution and the spread of diseases like asthma among people in targeted regions

Commercial and open-source software

- Weka
- Rapid miner
- Orange data miningtools

E.g.
Fuzzy Data Mining
for
Anomaly Detection

Data are applied with fuzzy logic rules
to
classify them
as
normal or malicious

for the purpose
feature sets should be extracted from the row data

Both network traffic and system audit data are used as inputs

Popular two data mining methods

- Association rules and frequency episodes have been used to mine audit data to find normal patterns for anomaly intrusion detection (Lee, Stolfo and Mok (1998) “Mining audit danta tobuild intrusion detection models.” in *Proceedings of the 4th international conference on knowledge discovery and data mining*)

Association rule

- When presented with a set of audit data, the system will mine a set of fuzzy association rules from the data.

- These rules will be considered a high level description of patterns of behavior found in the data.

- For anomaly detection, we mine a set of rules from a data set with no intrusions (termed a reference data set) and use this as a description of normal behavior.

- When considering a new set of audit data, a set of association rules is mined from the new data and the similarity of this new rule set and the reference set is computed. If the similarity is low, then the new data will cause an alarm.

Notice that

- an intrusion that deviates only slightly from a pattern derived from the audit data may not be detected of a small change in normal behavior may cause a false alarm.

False negatives & False positive

- If the system warn when access is normal, it is called Falese negative
- If the system does not warn when access is abnormal, it is called False positive

Examples of features

- The CPU usage time and the connection duration
- The numbere of different TCP/UDP services initiated by the same source host

Examples of output

- **IF** the number different destination addresses during the last 2 seconds was high **THEN** an unusual situation exists

- **IF** the DP is high **THEN** and unusual situation exists

- **IF** the number of SYN flags is low **AND** the number of FIN flags is low **THEN** the number of RST flags is low in a 2 second period