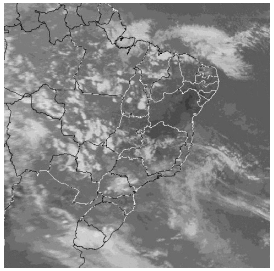
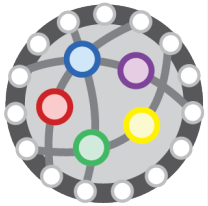


CAP-394
INTRODUCTION TO DATA SCIENCE

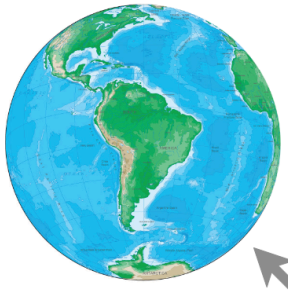
Rafael Santos – rafael.santos@inpe.br
Gilberto Ribeiro – gilberto.queiroz@inpe.br
www.lac.inpe.br/~rafael.santos/cap394.html

Updated in 2019



About this Lecture

Where are we?



Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

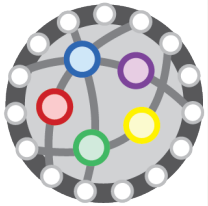
Statistical Models

Create Data Product

Visualize and Report

Make Decisions

Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.



Basic Machine Learning Concepts

Machine Learning Categories

- Supervised:
  - We have data with labels or associated values we want to predict from other data (predictors).
  - We also have data for which we have no labels or outcomes.
  - We need to map predictors -> labels.
  - **Classification, regression.**
- Unsupervised:
  - We have data that we want to segment or cluster together based on similarity or distances in feature space.
  - **Clustering.**
- Reinforcement Learning.

Many algorithms!

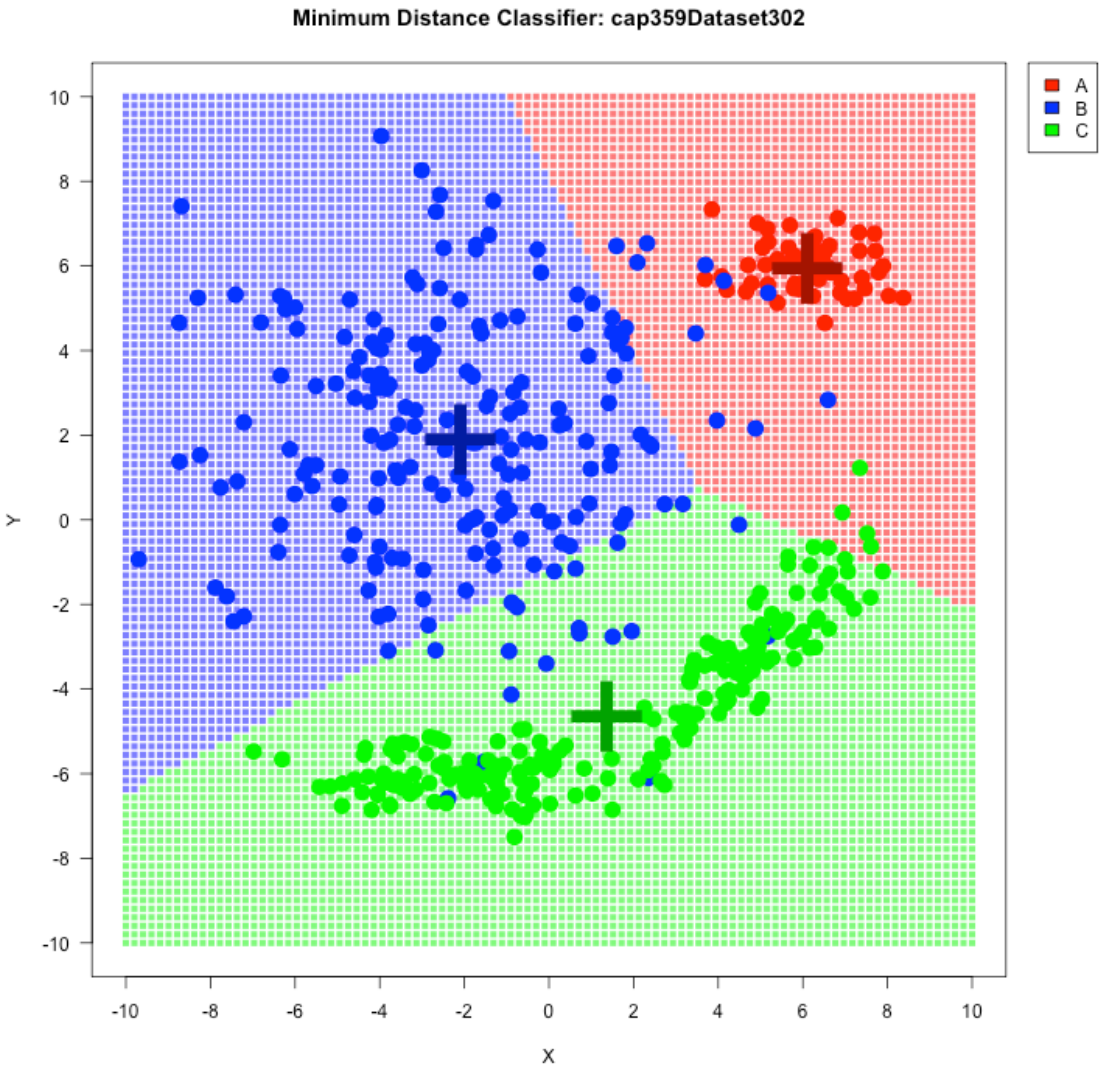
- Linear Regression, Neural Networks (MLPs), Extreme Gradient Boosting (XGBoost), Long Short Term Memory (LSTM), many others.
- Decision Trees, Nearest Neighbors, Minimum Distance, Support Vector Machines, Neural Networks (MLPs), Random Forests, many others.
- Hierarchical Clustering, K-Means, DBScan, Self-Organizing Maps, many others.
- Lots of variants, mixing also allowed!
- We'll play with the basics only.

Classification: Minimum Distance

- Decision Boundary with labeled points and classes' prototypes

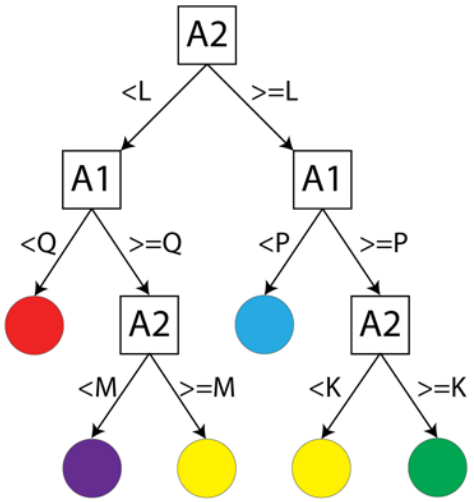
	Classified as		
	A	B	C
A	50	0	0
B	11	170	19
C	4	0	196

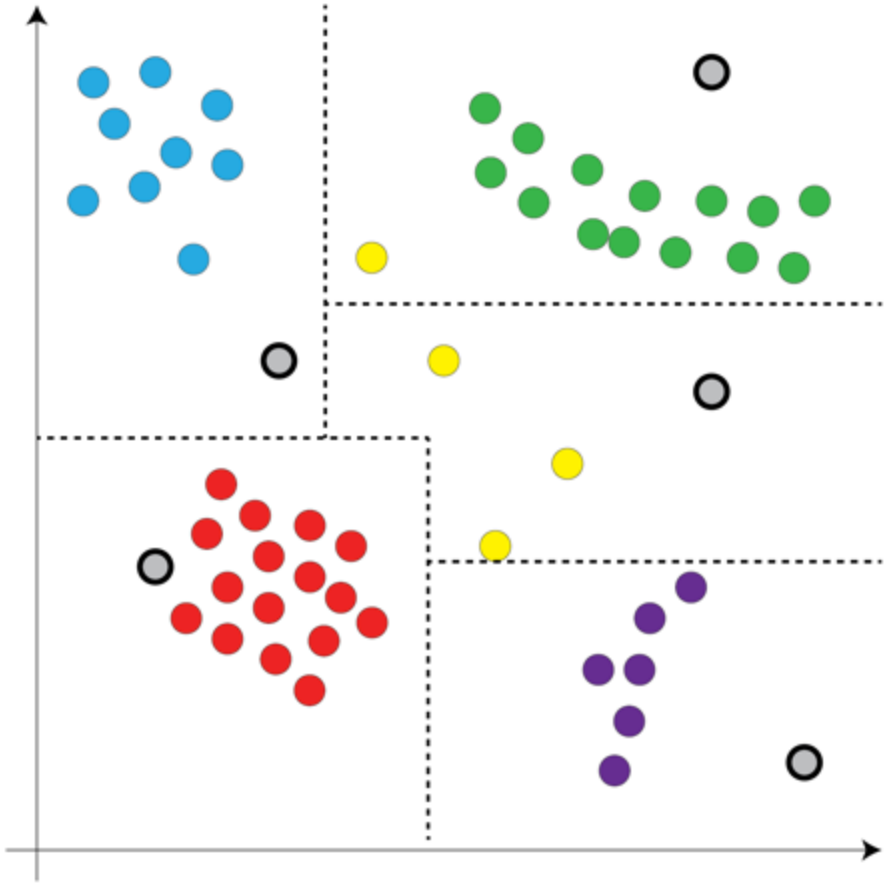
Accuracy: 92.444%



Classification: Decision Tree

- Model is the set of decision rules that best separates the classes.
- Class is determined from evaluation of the rules.

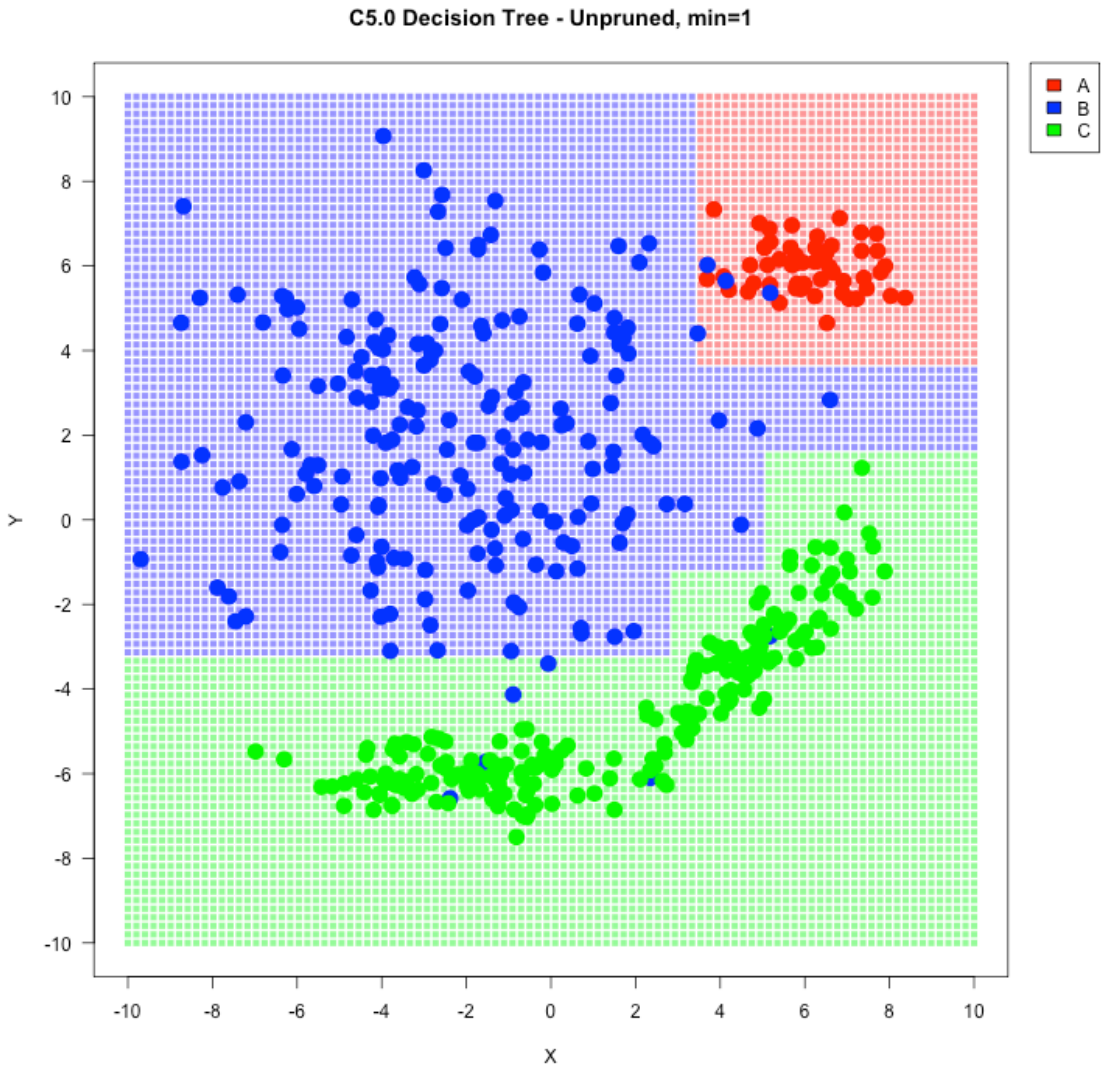




Classification: Decision Tree

	Classified as		
	A	B	C
A	50	0	0
B	3	191	6
C	0	0	200

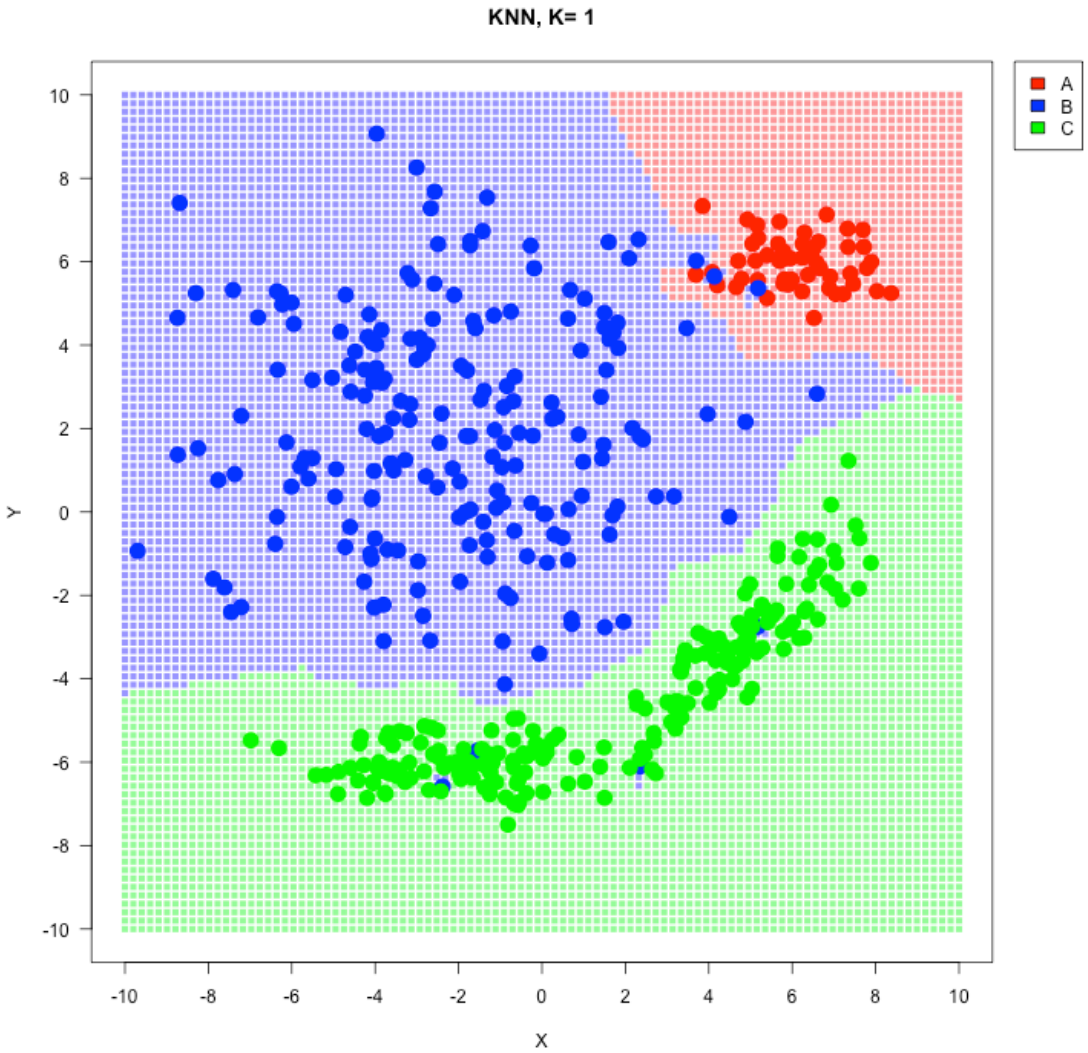
Accuracy: 0.98



Classification: Nearest Neighbors

- K=1
- A perfect classifier? Accuracy = 1

	Classified as		
	A	B	C
A	50	0	0
B	0	200	0
C	0	0	200

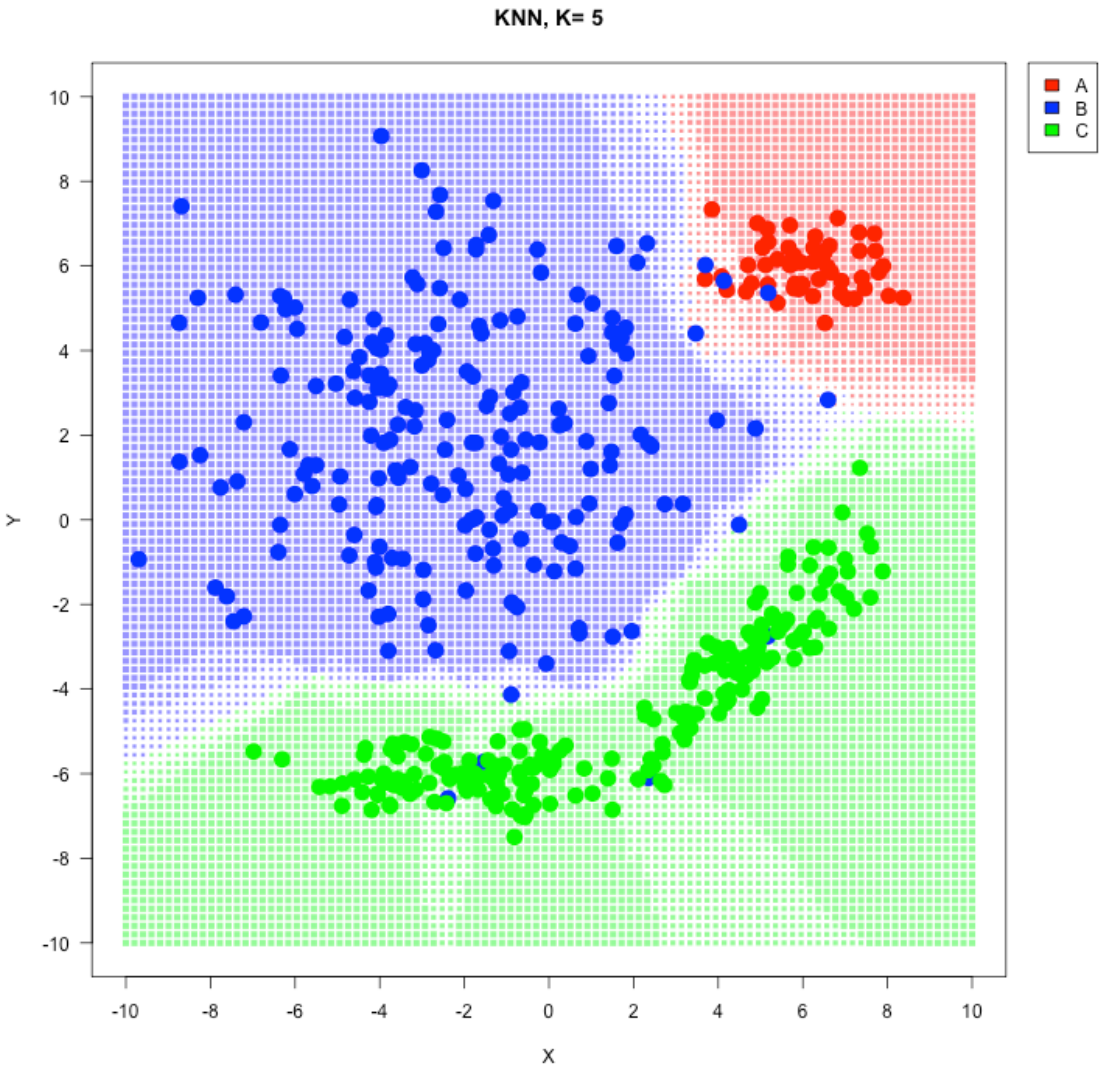


Classification: Nearest Neighbors

- K=5

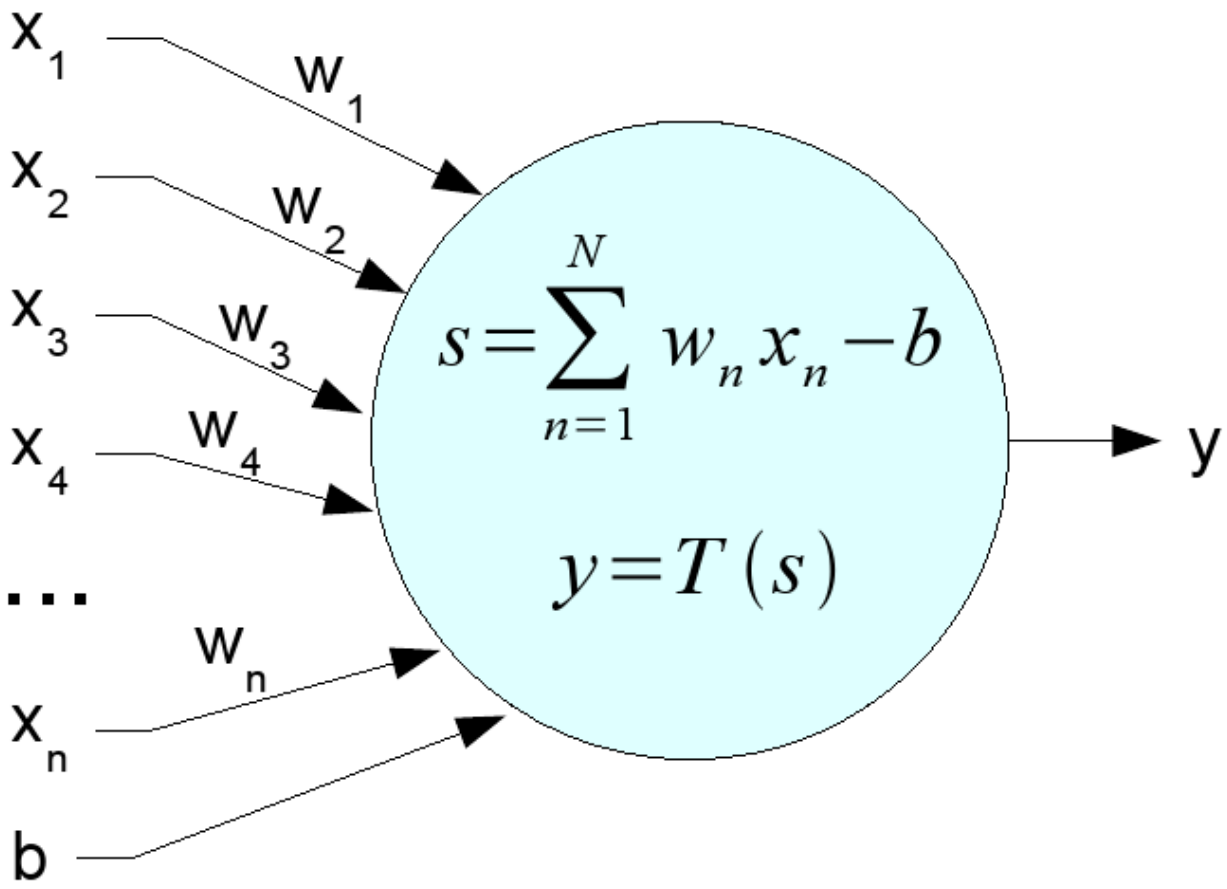
Accuracy: 0.98

	Classified as		
	A	B	C
A	50	0	0
B	4	191	5
C	0	0	200



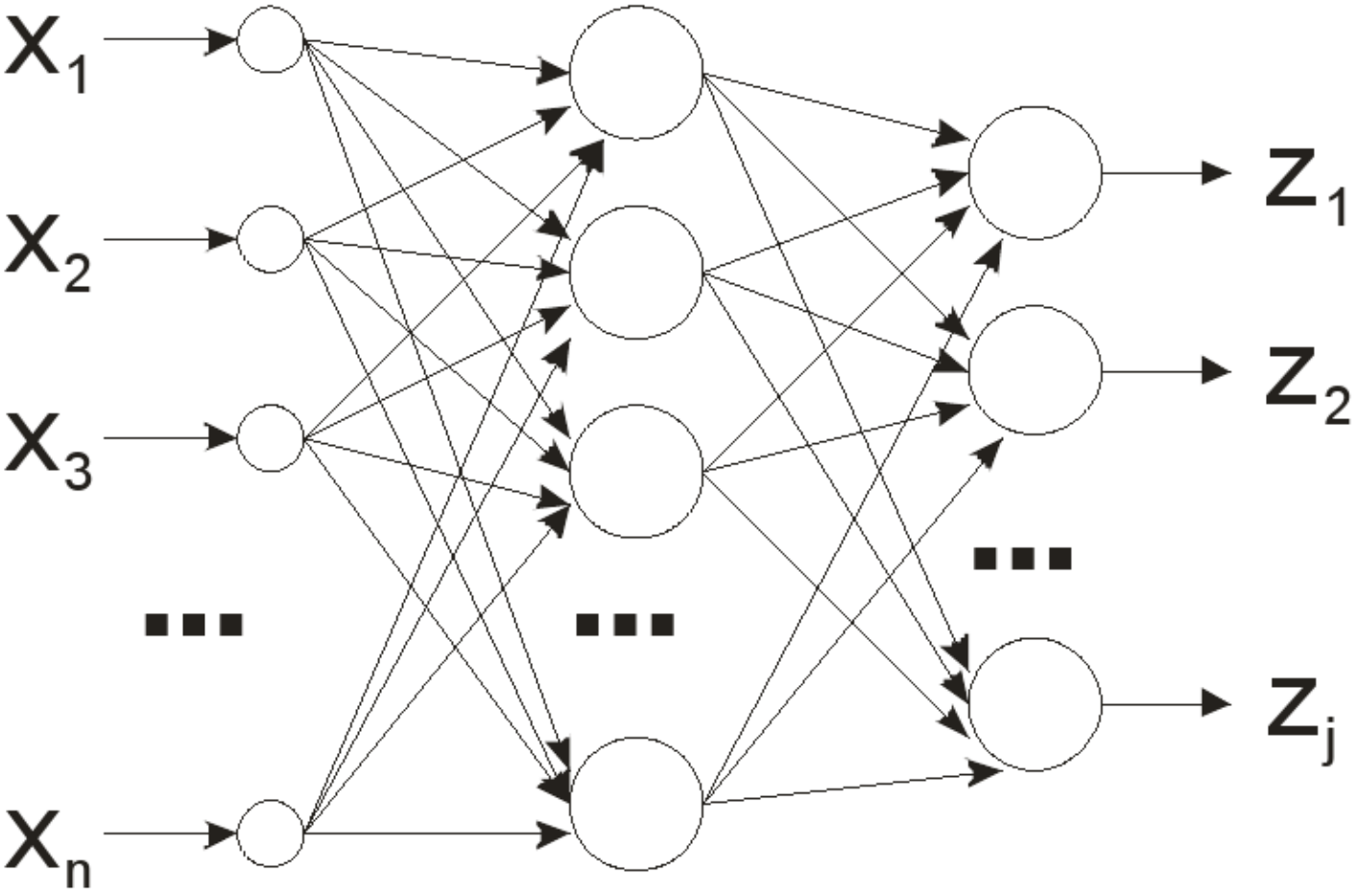
Classification: Neural Networks (MLPs)

- Perceptron

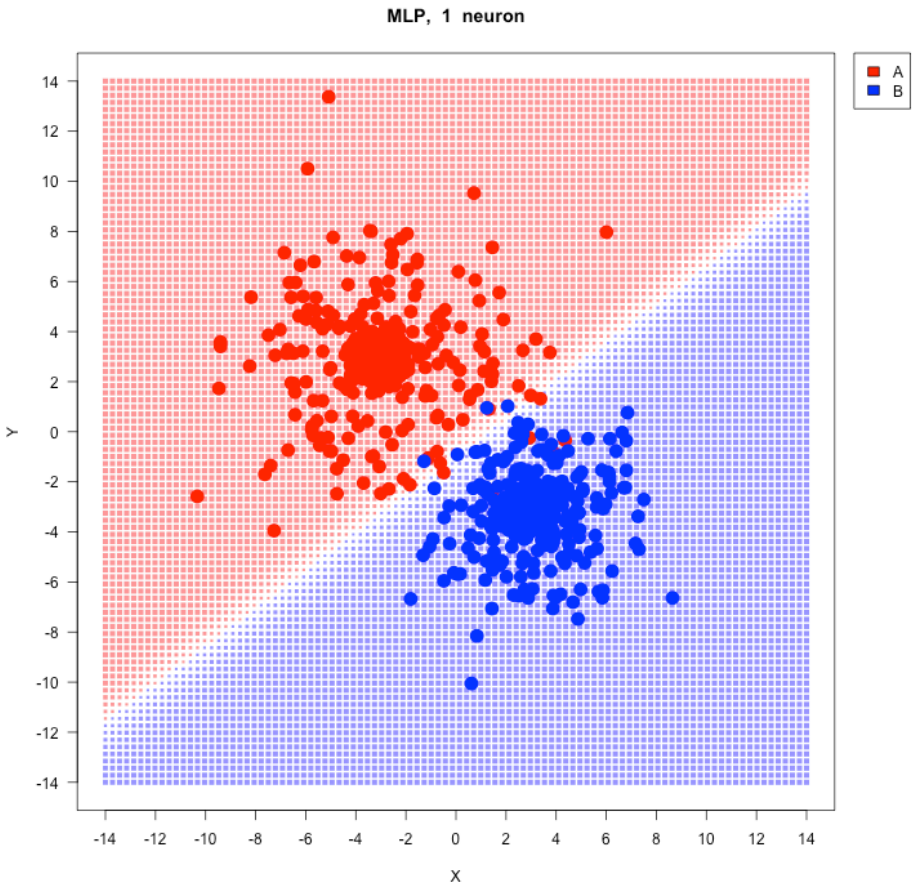


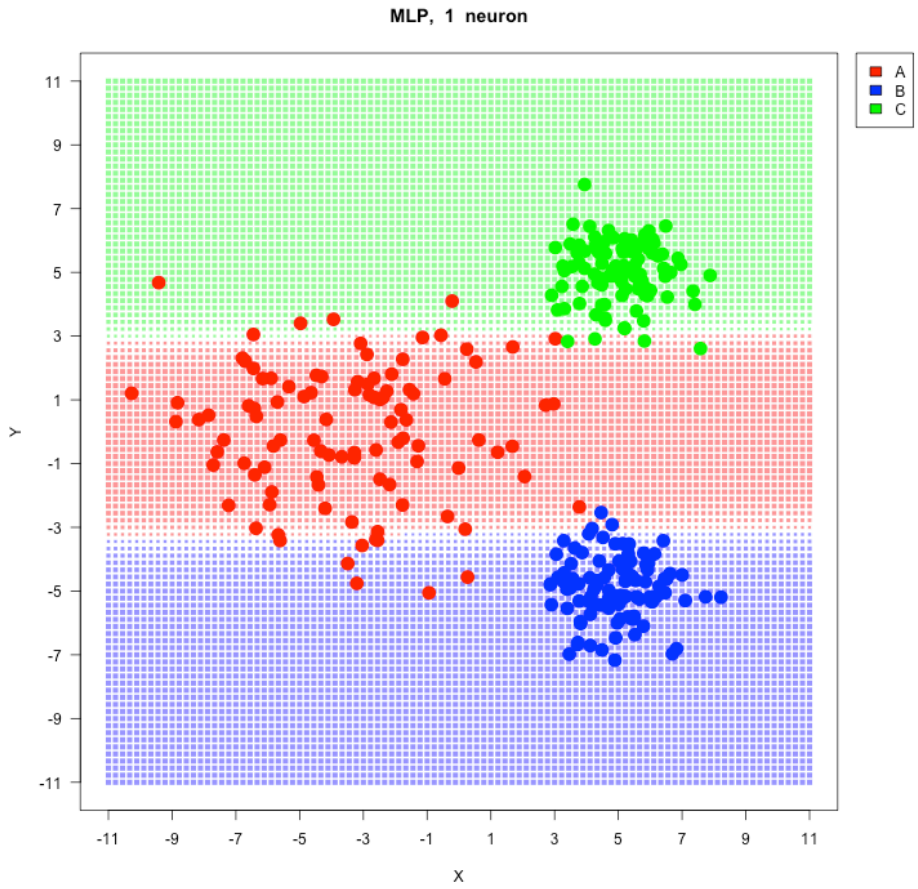
Classification: Neural Networks (MLPs)

- Multilayer Perceptron



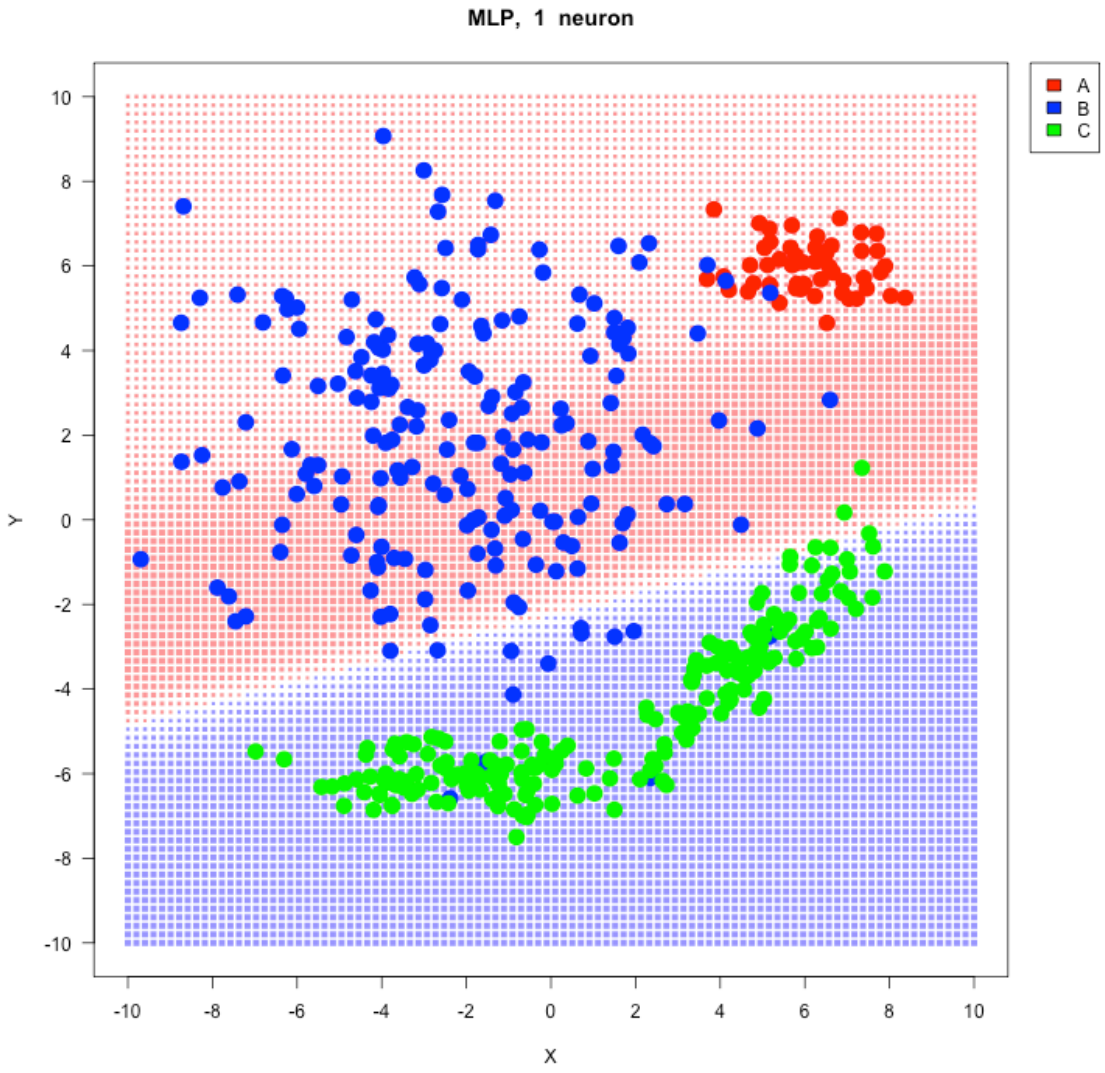
Classification: MLPs with a single neuron/hidden layer





Classification: MLPs with a single neuron/hidden layer

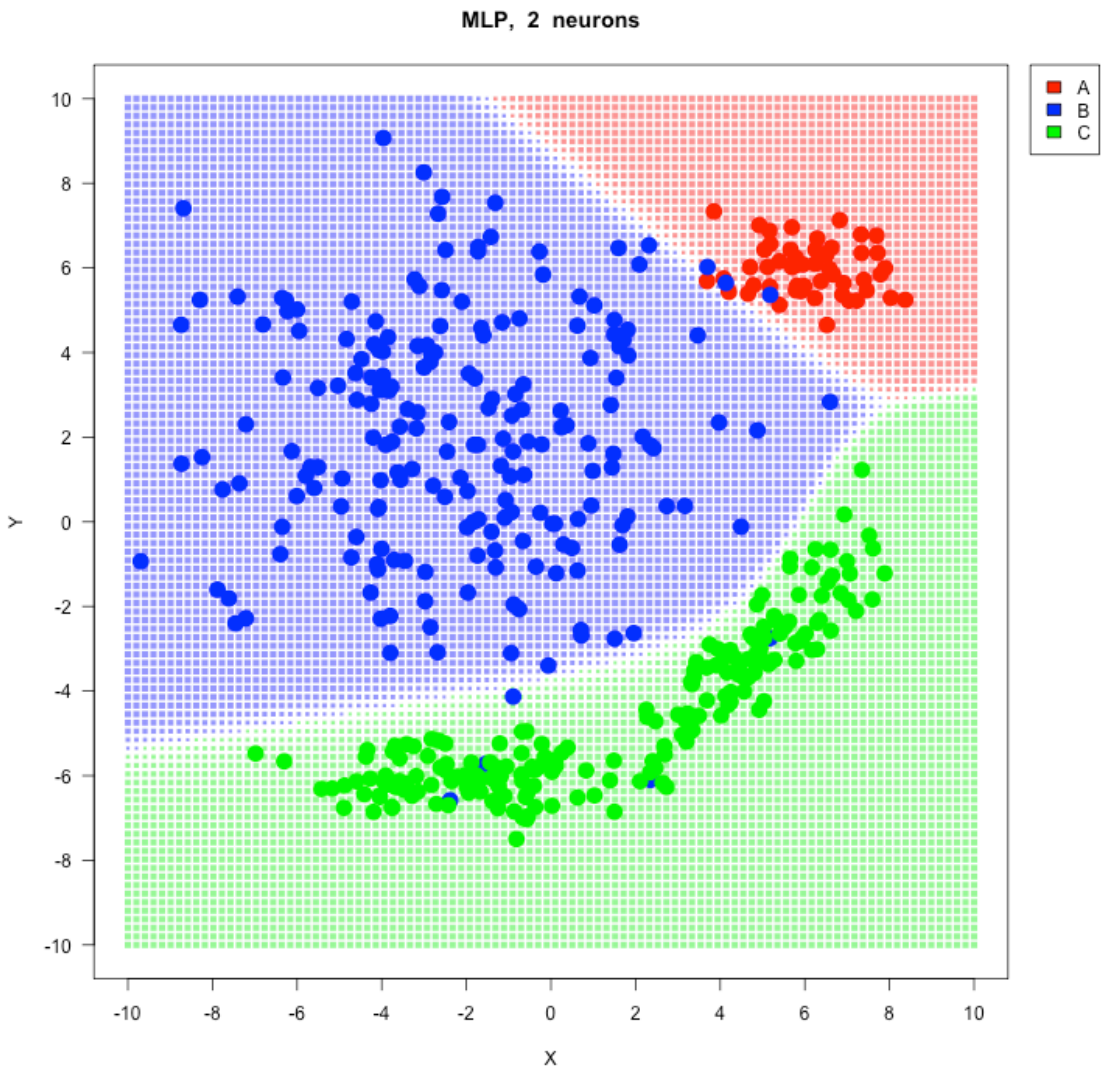
- One neuron cannot separate three classes with non-parallel hyperplanes!



Classification: Neural Networks (MLPs)

- 2 Neurons in the hidden layer
- Accuracy: 0.9777778

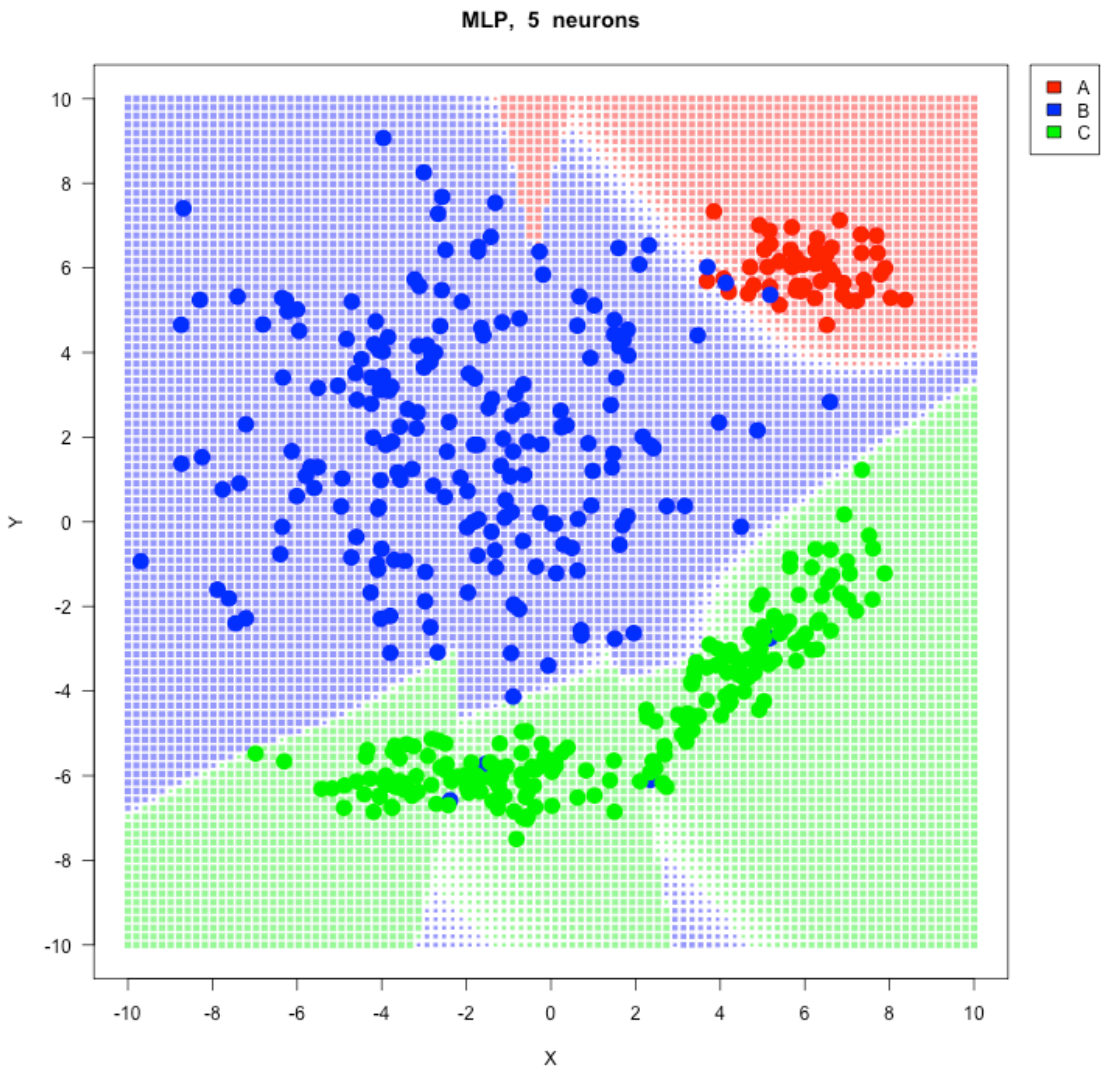
	Classified as		
	A	B	C
A	48	2	0
B	3	192	5
C	0	0	200



Classification: Neural Networks (MLPs)

- 5 Neurons in the hidden layer
- Accuracy: 0.9822222

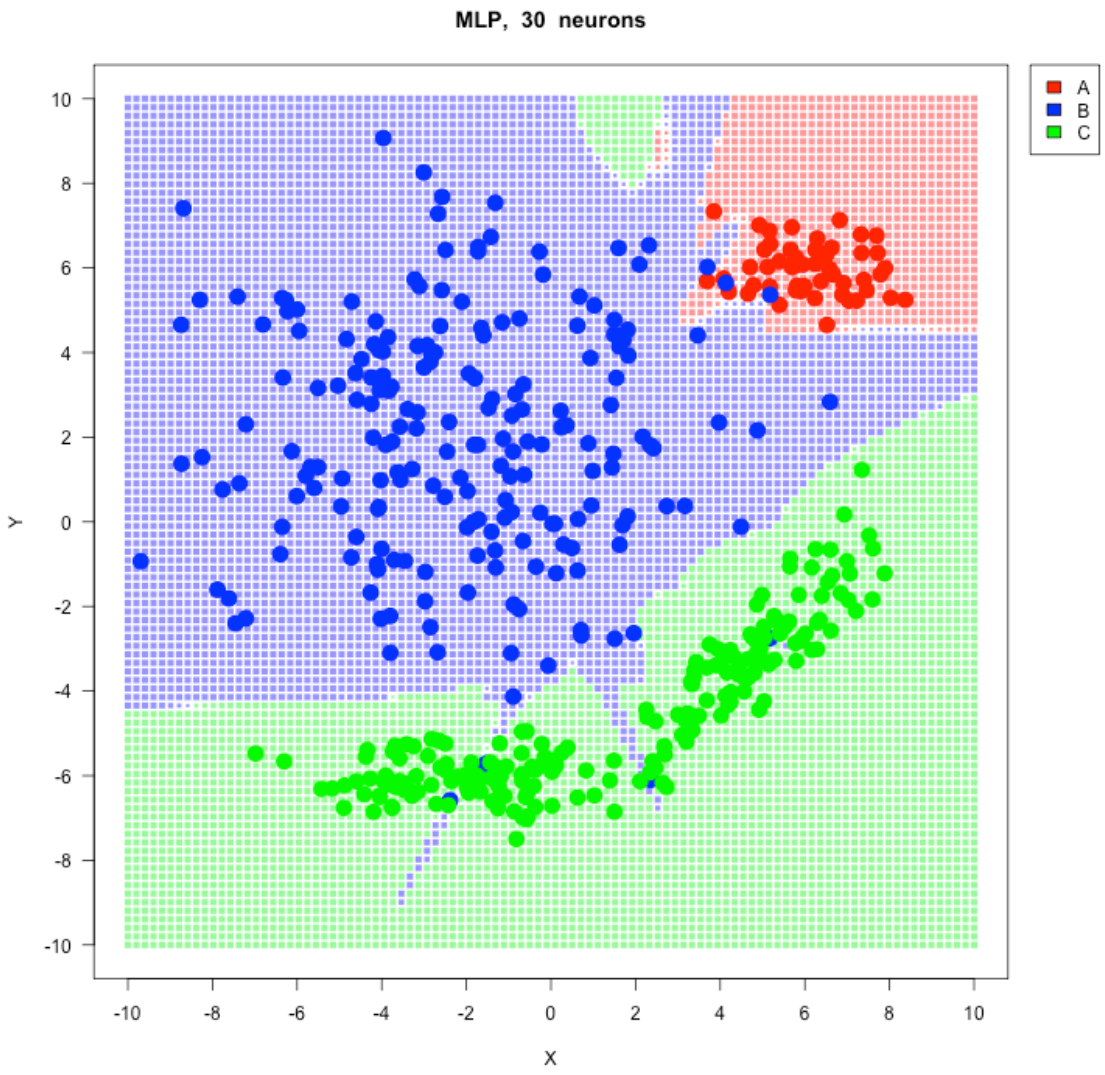
	Classified as		
	A	B	C
A	49	1	0
B	3	193	4
C	0	0	200



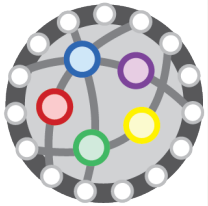
Classification: Neural Networks (MLPs)

- 60 Neurons in the hidden layer
- Accuracy: 0.9933333

	Classified as		
	A	B	C
A	50	0	0
B	1	197	2
C	0	0	200



Introduction to Data Science

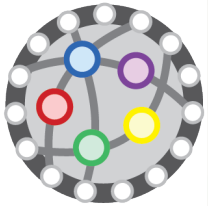


Machine Learning in R

“R Programming”

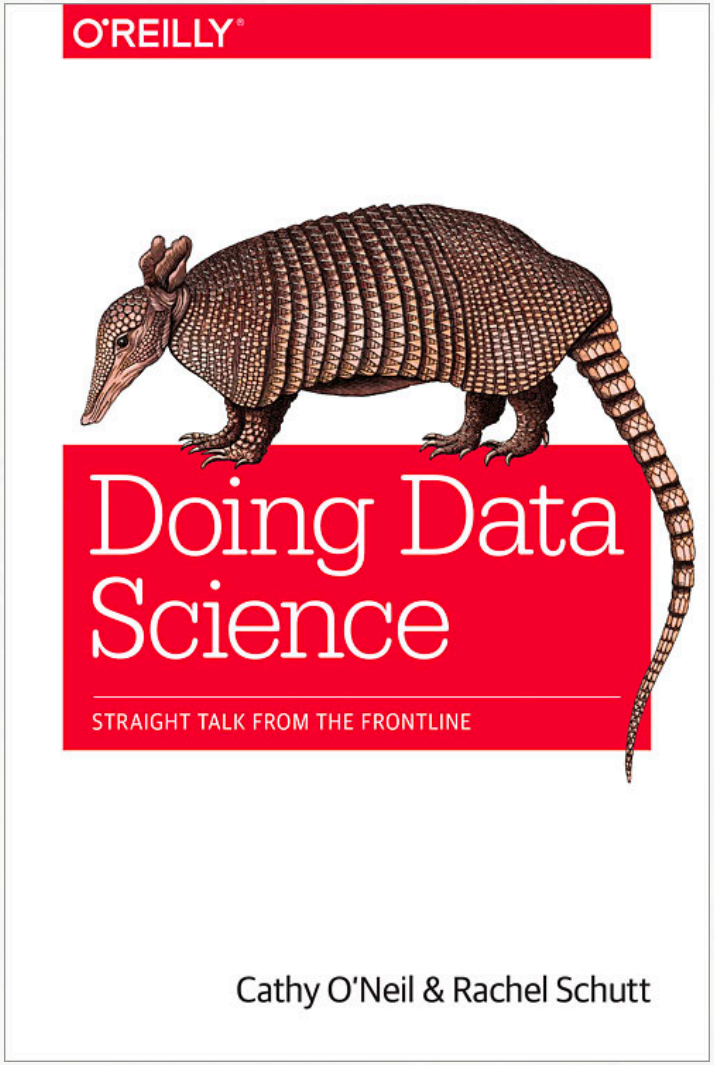
EDA in R

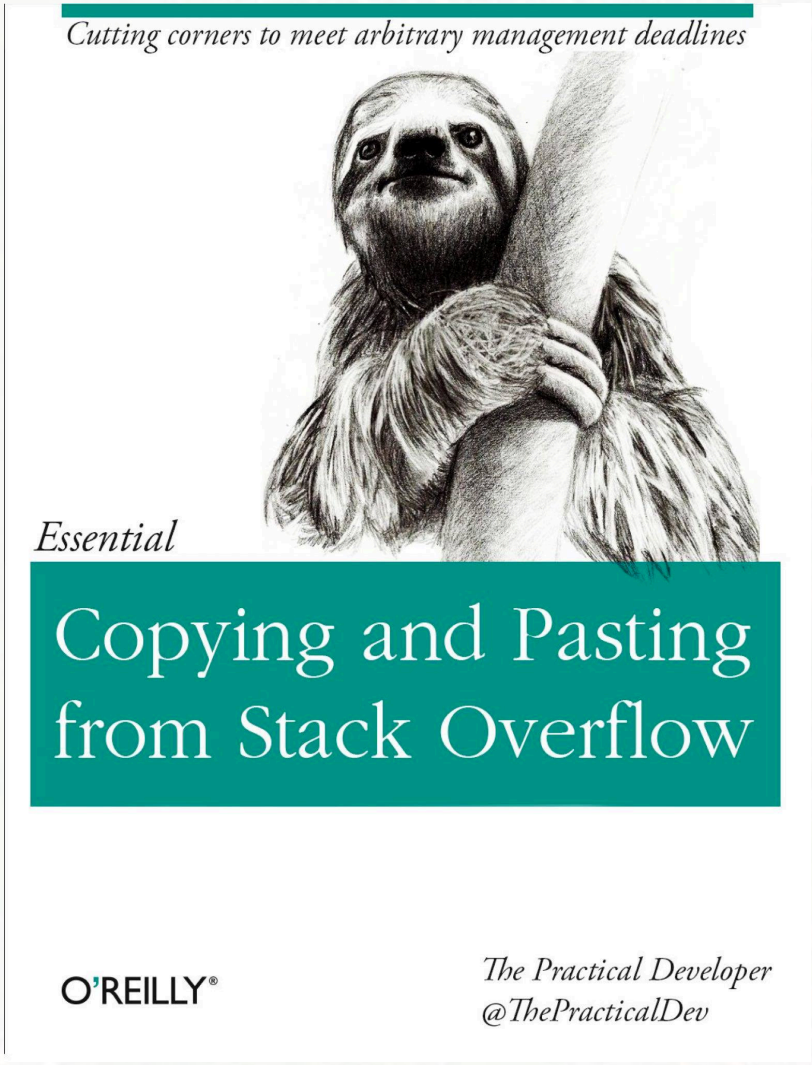
- Let's switch to a browser: http://www.lac.inpe.br/~rafael.santos/r.html



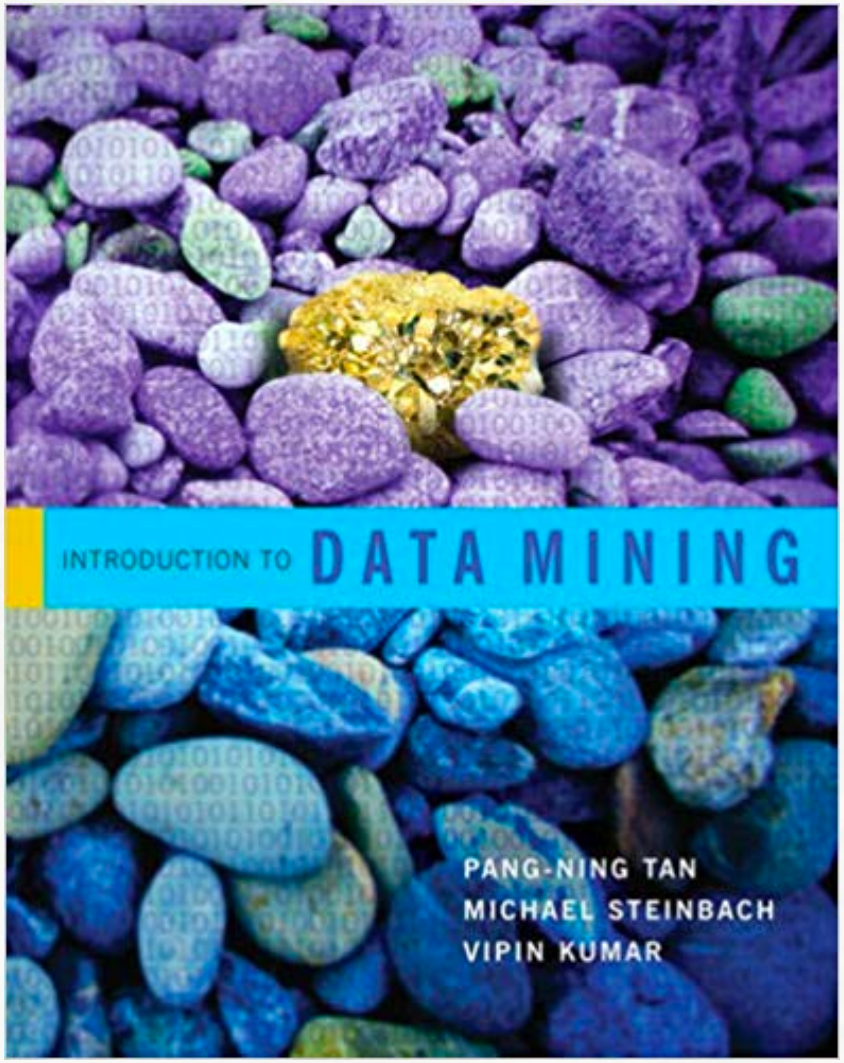
References

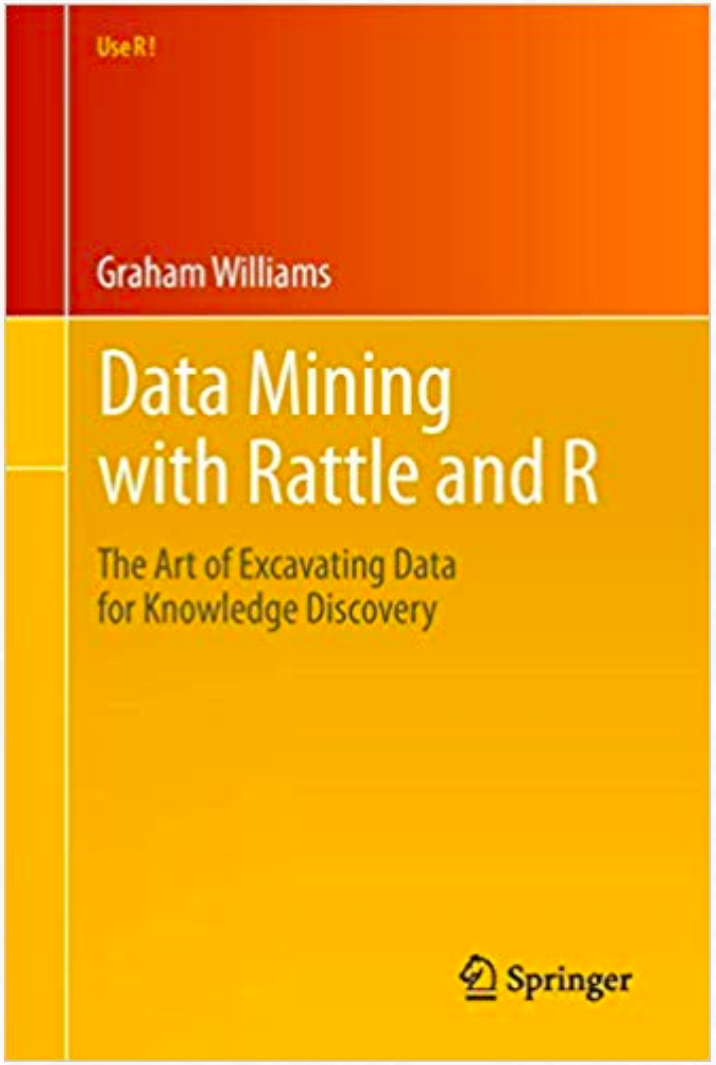
References


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Cathy O'Neil & Rachel Schutt

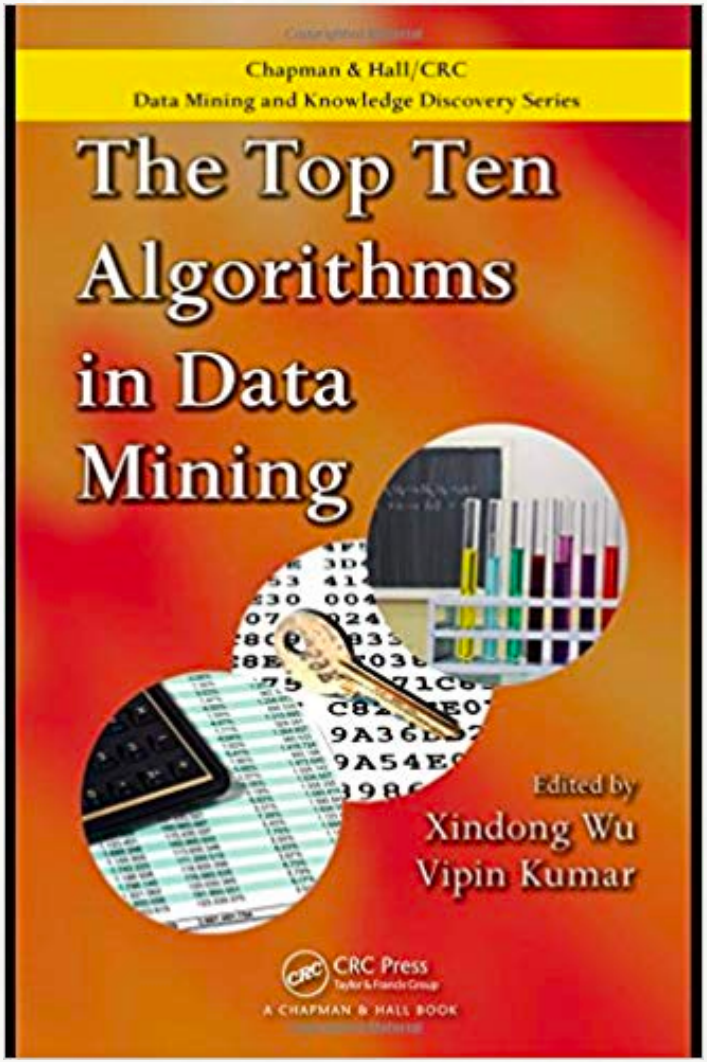

Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev

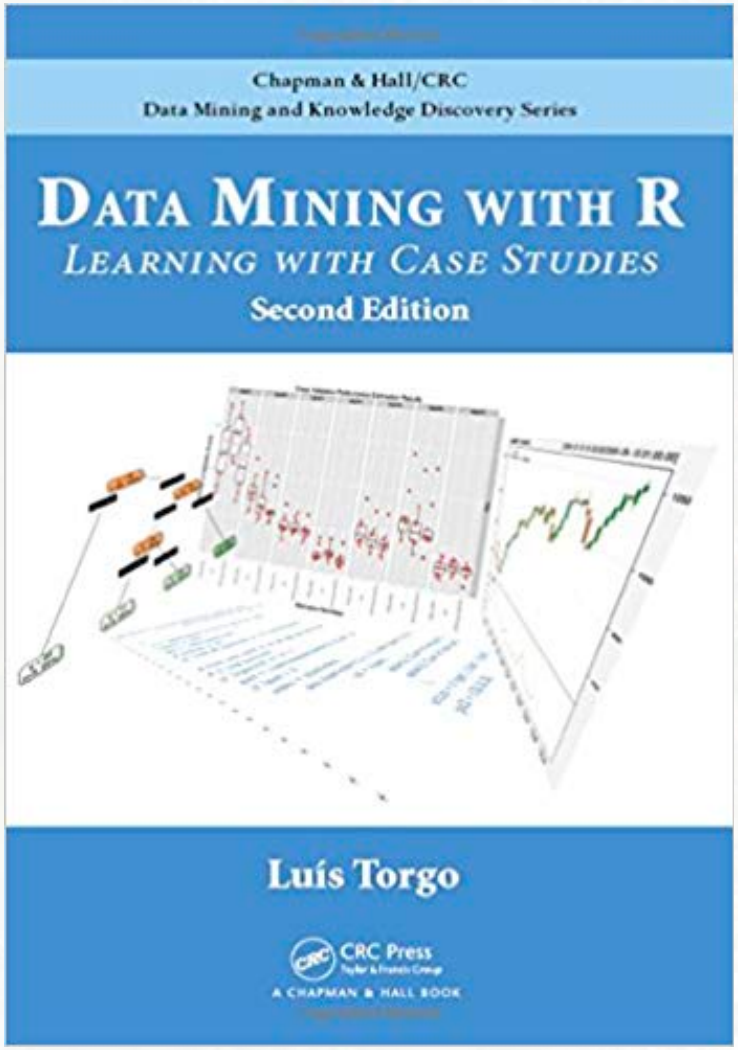
References


INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR


UseR!
Graham Williams
Data Mining with Rattle and R
The Art of Excavating Data for Knowledge Discovery
Springer

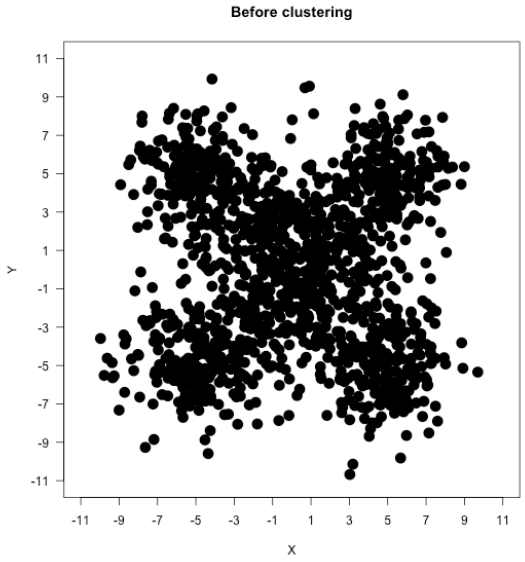
References

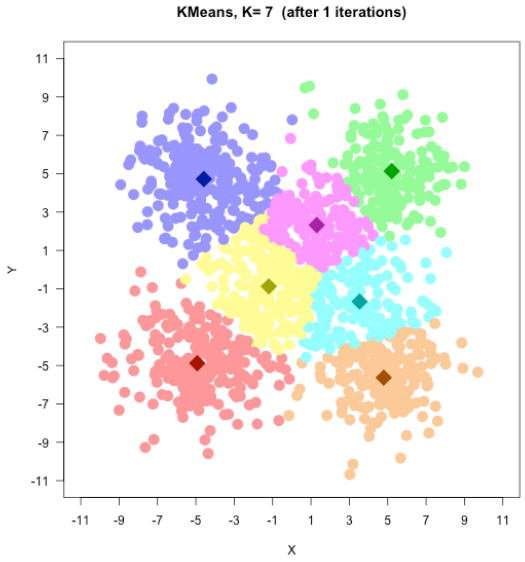

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
The Top Ten Algorithms in Data Mining
Edited by
Xindong Wu
Vipin Kumar
CRC Press
A CHAPMAN & HALL BOOK

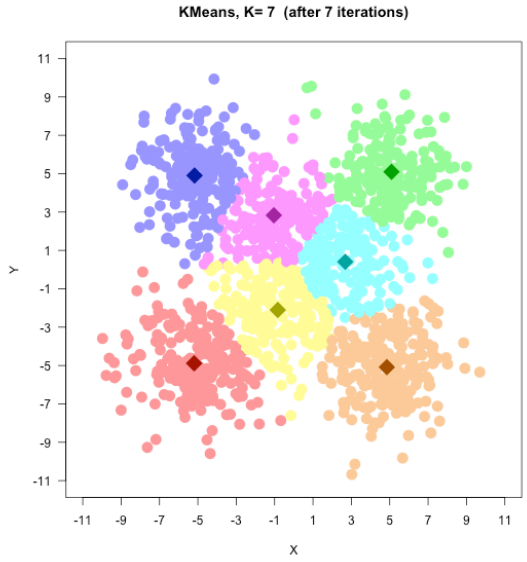

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
DATA MINING WITH R
LEARNING WITH CASE STUDIES
Second Edition
Luís Torgo
CRC Press
A CHAPMAN & HALL BOOK

Clustering: K-Means

- Simulation with artificial dataset and K=7

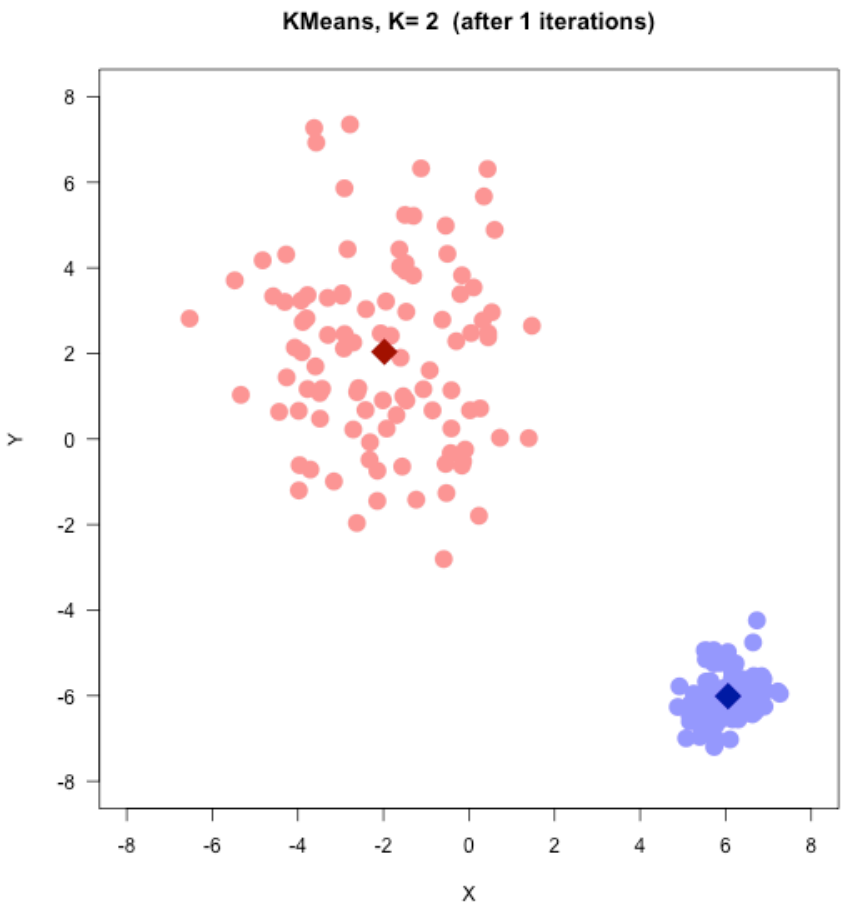


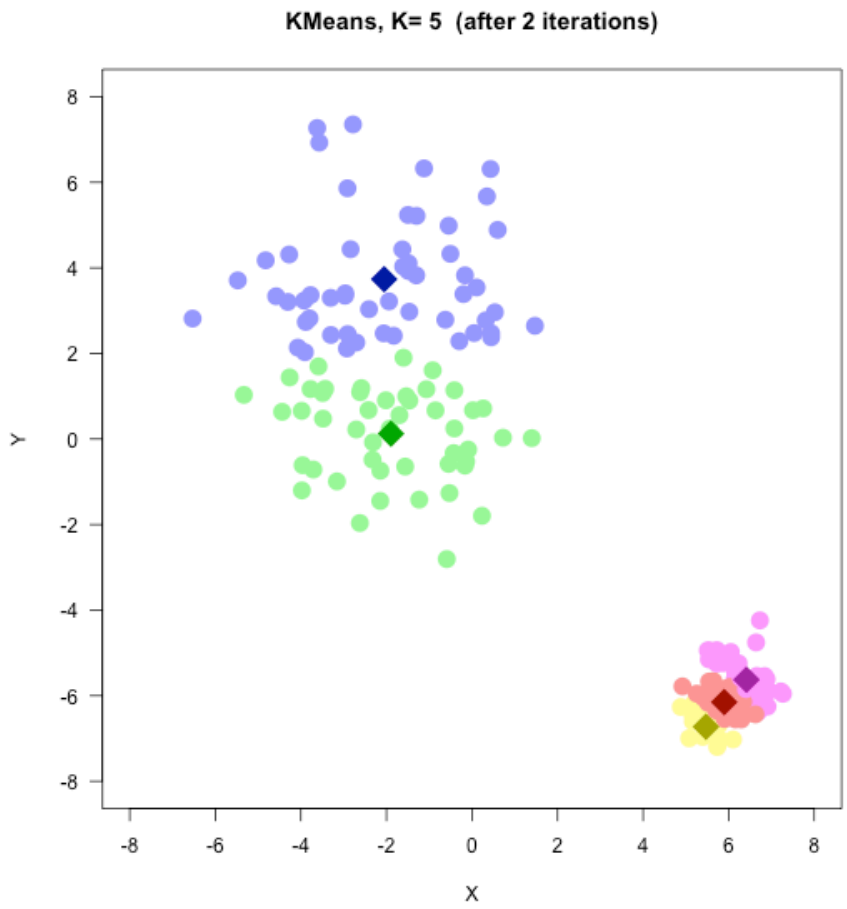




Clustering: K-Means

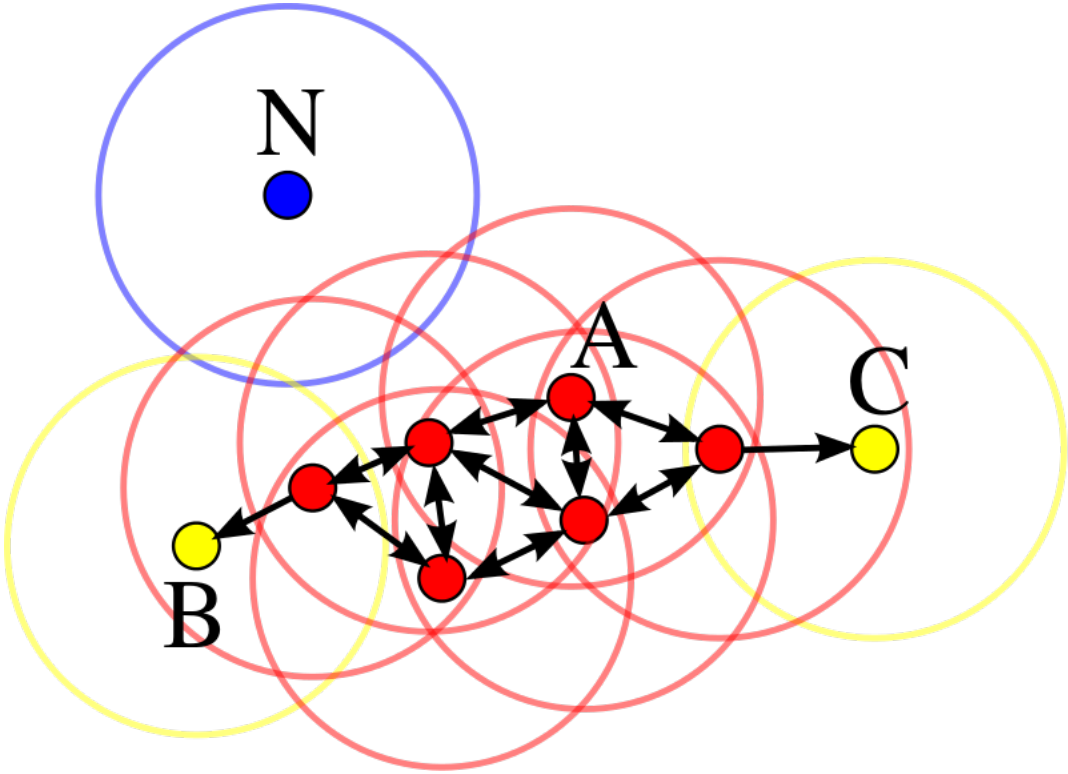
- Implementation in R is very efficient (quick convergence)





Clustering: DBScan

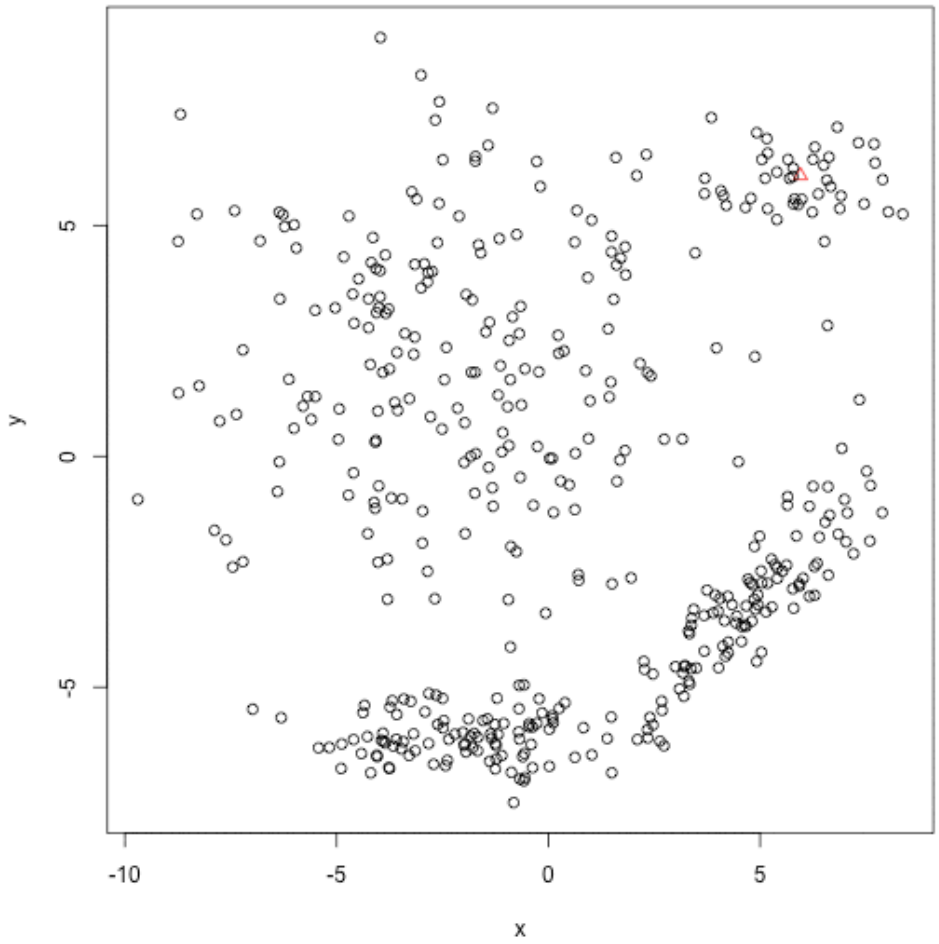
- Identification of density-based clusters:
  - Find core points
  - Test reachability
  - Identify noise



Wikipedia

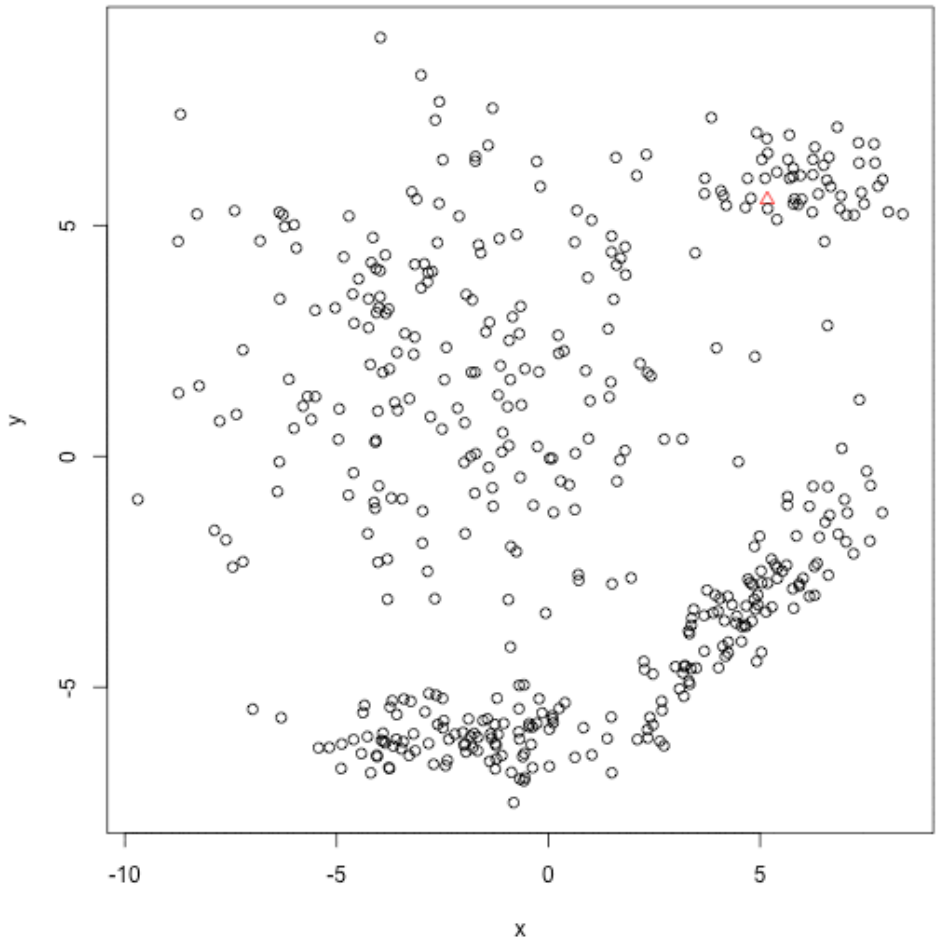
Clustering: DBScan

eps=0.3



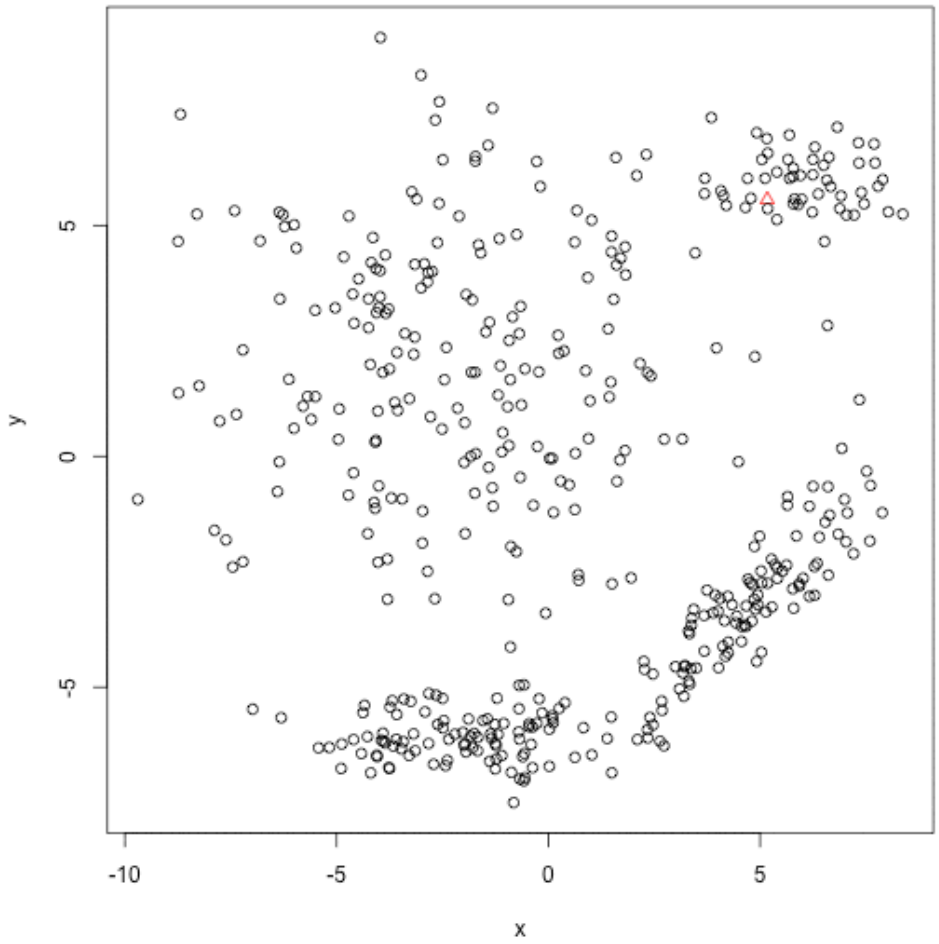
Clustering: DBScan

eps=1.0



Clustering: DBScan

eps=2.0



Hierarchical Clustering

- Methods that create several partitions in the data:
- Top-down: starts with all data in a single cluster, partitions every cluster until each data is in a single cluster.
- Bottom-up: starts with each data in a single cluster, merges data+clusters until all data is in a single cluster.

Hierarchical Clustering

	X	Y
1	0,25	0,27
2	0,32	0,91
3	0,33	0,80
4	0,18	0,33

	1	2	3	4
1	0,000	**0,644**	**0,536**	**0,092**
2	0,644	0,000	**0,110**	**0,597**
3	0,536	0,110	0,000	**0,493**
4	0,092	0,597	0,493	0,000

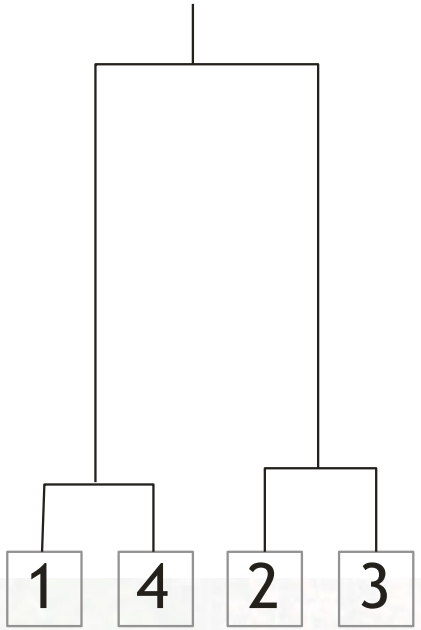
	X	Y
1+4	0,22	0,30
2	0,32	0,91
3	0,33	0,80

	1+4	2	3
1+4	0,000	**0,619**	**0,513**
2	0,619	0,000	**0,110**
3	0,513	0,110	0,000

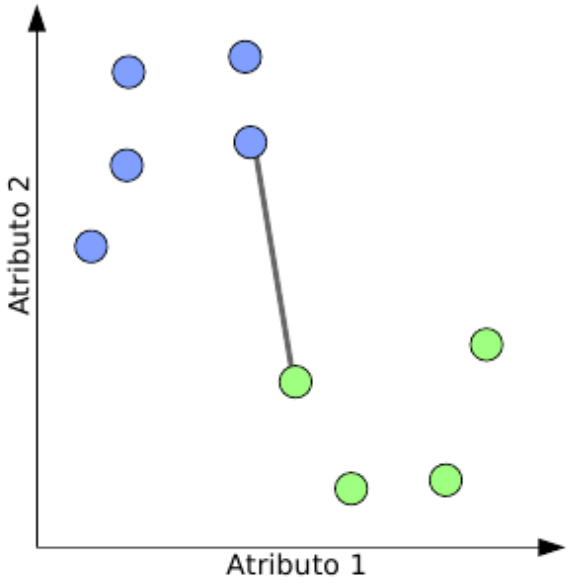
	X	Y
1+4	0,22	0,30
2+3	0,33	0,86

	1+4	2+3
1+4	0,000	**0,566**
2+3	0,566	0,000

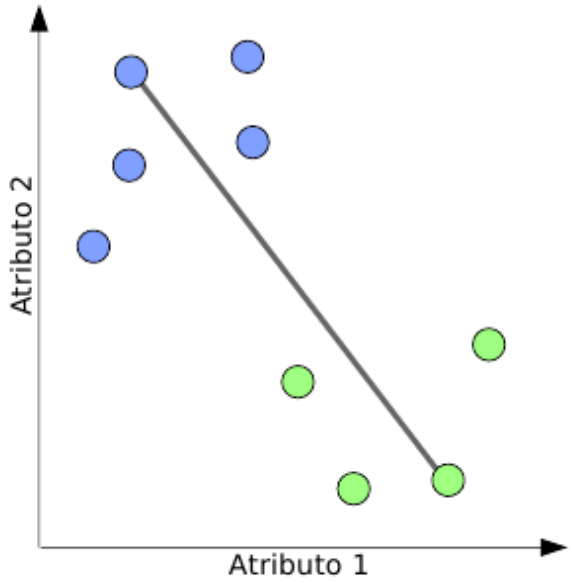
	X	Y
1+2+3+4	0,27	0,58



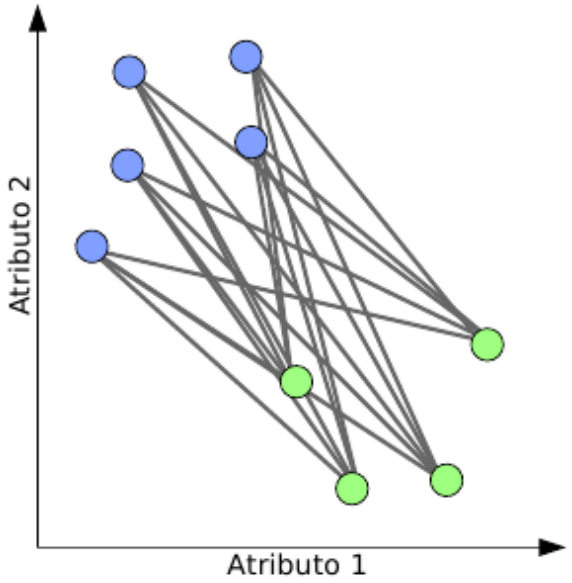
Hierarchical Clustering



Single Linkage

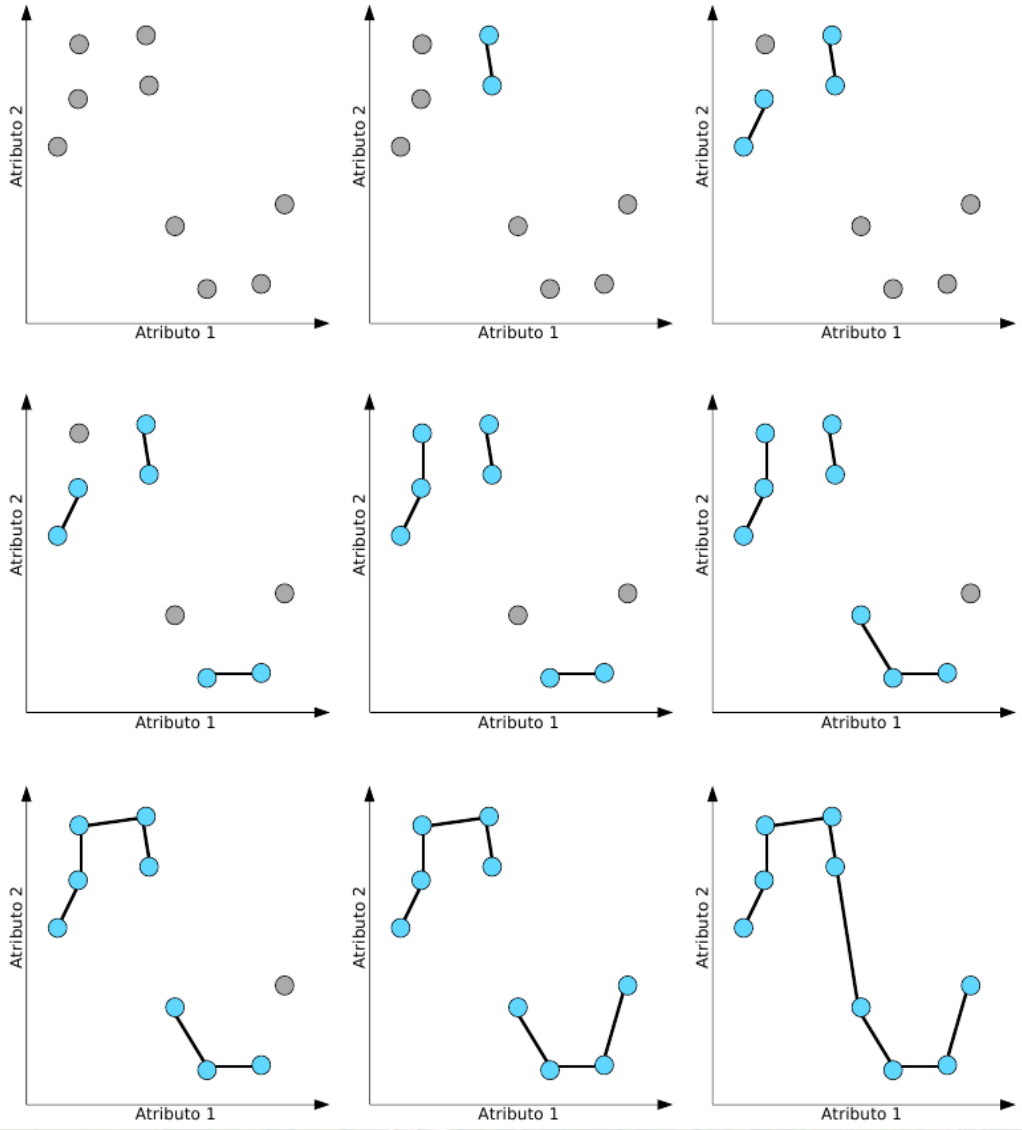


Complete Linkage

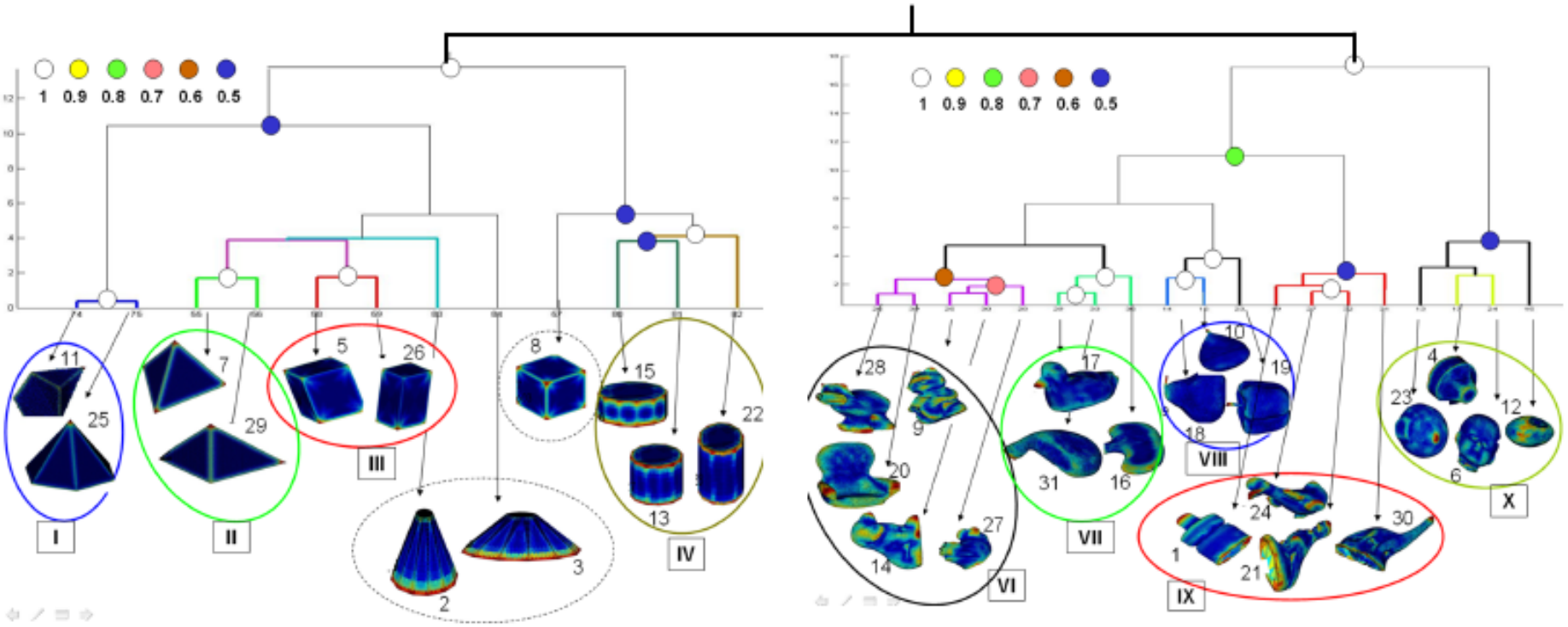


Average Linkage

Hierarchical Clustering

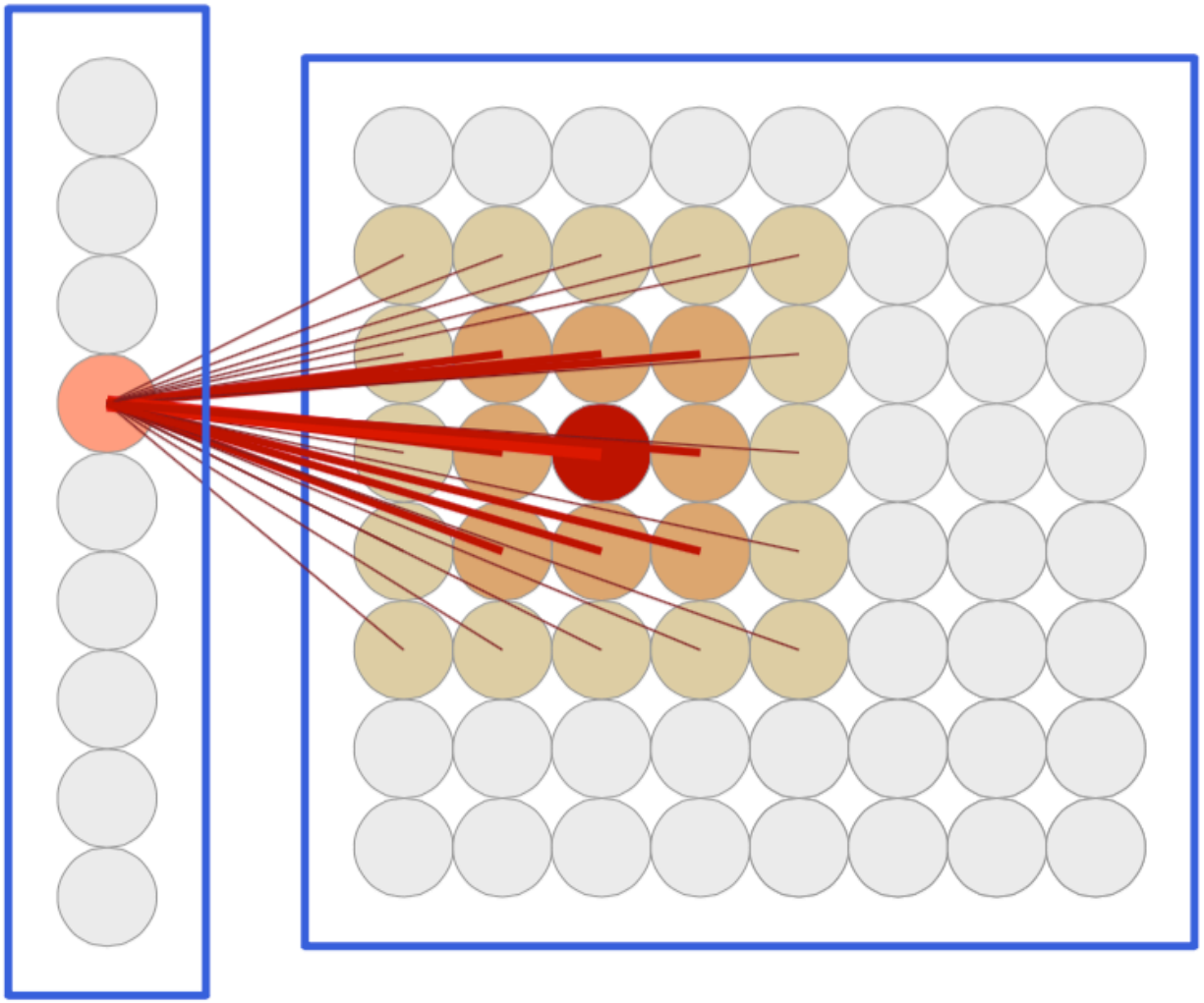


Hierarchical Clustering

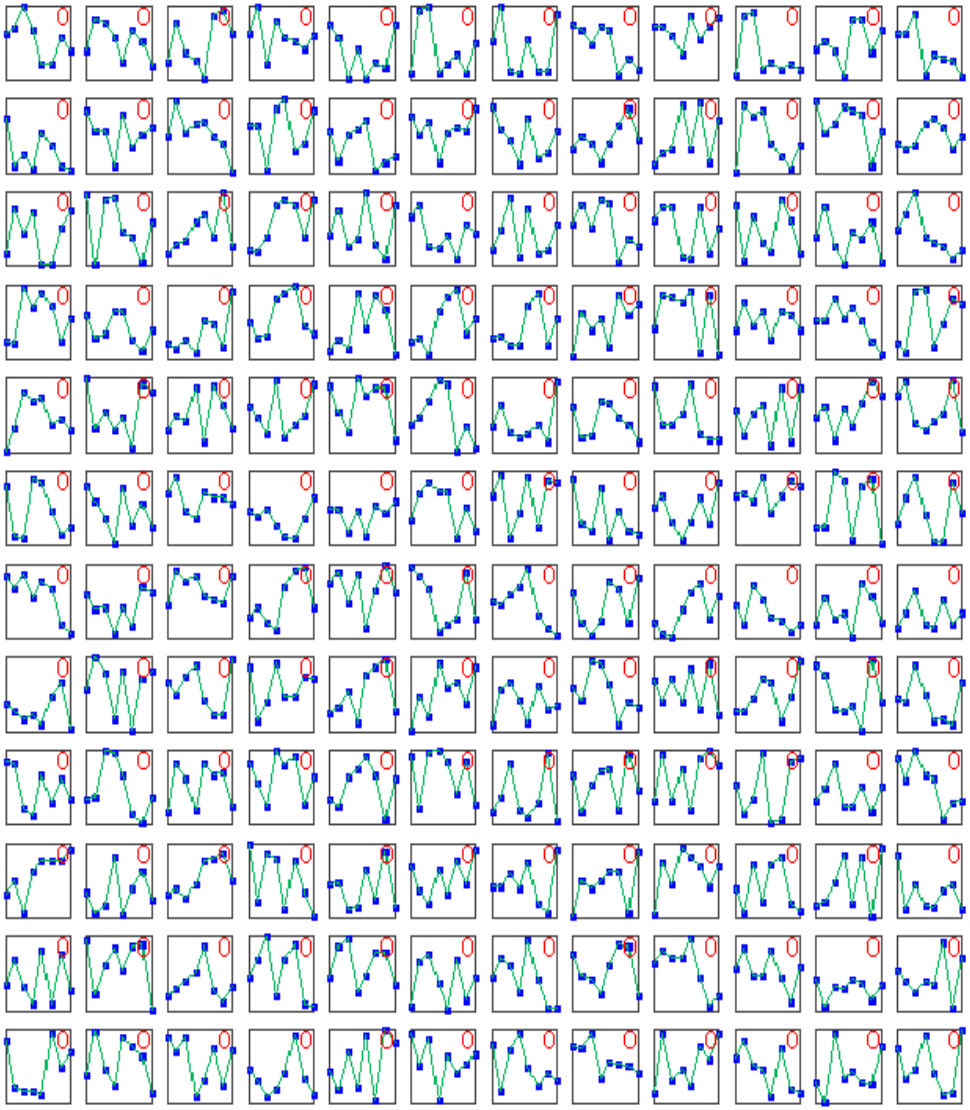


Using non local features for 3D shape grouping. Antonio Adán, Miguel Adán, Santiago Salamanca, and Pilar Merchán. LNCS 5432, 2008.

Self-Organizing Map



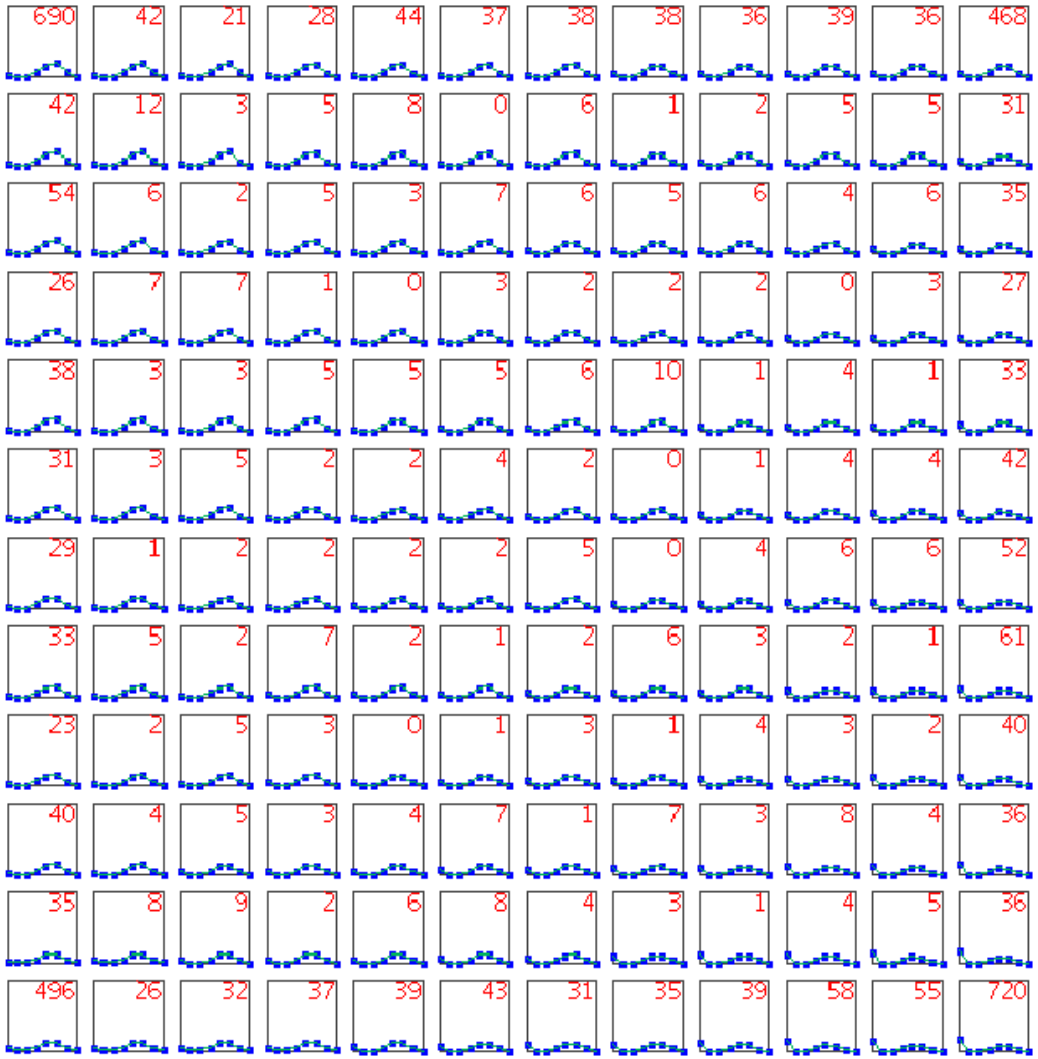
Self-Organizing Map



12x12 SOM,
Dados em 8 dimensões.

T=0
R=25
Lr=0.9

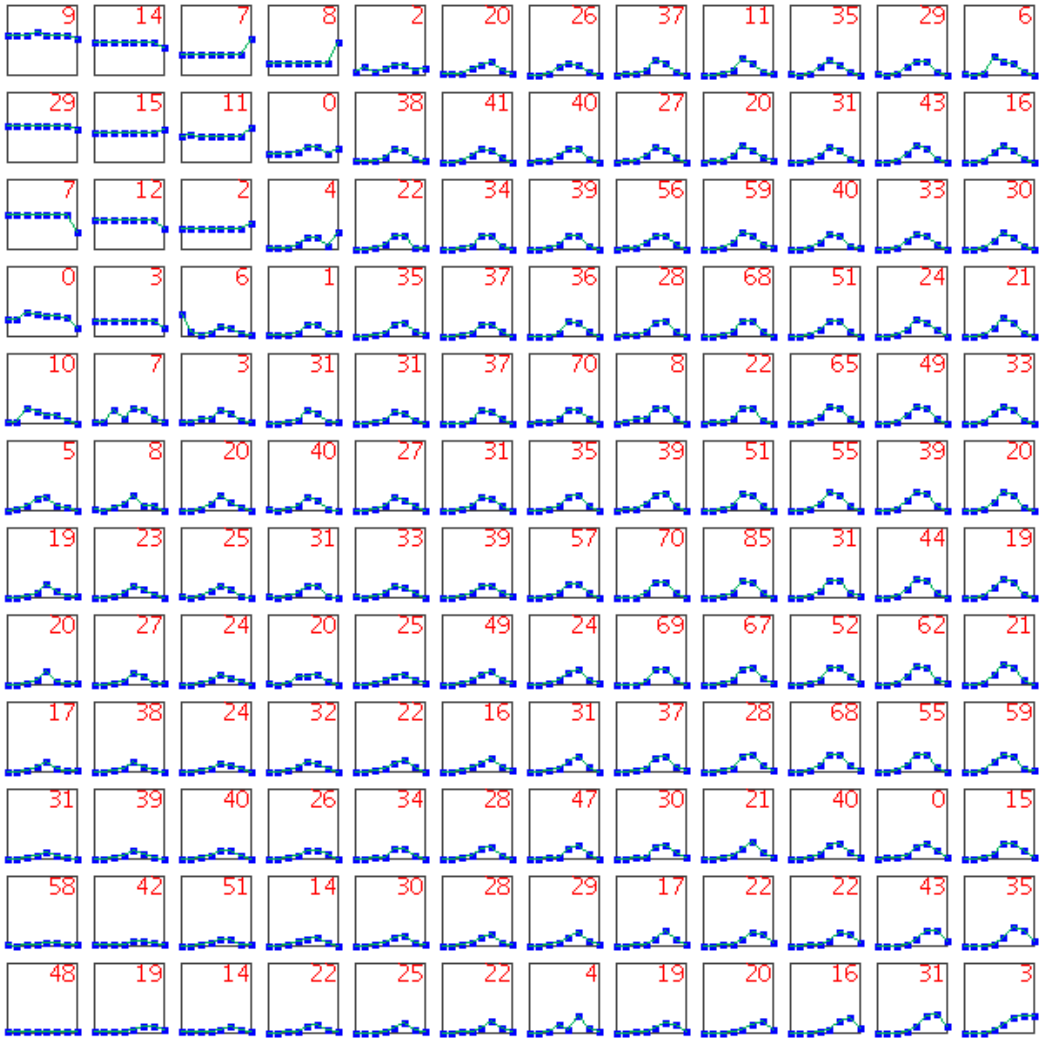
Self-Organizing Map



12x12 SOM,
Dados em 8 dimensões.

T=40
R=16.7
Lr=0.74

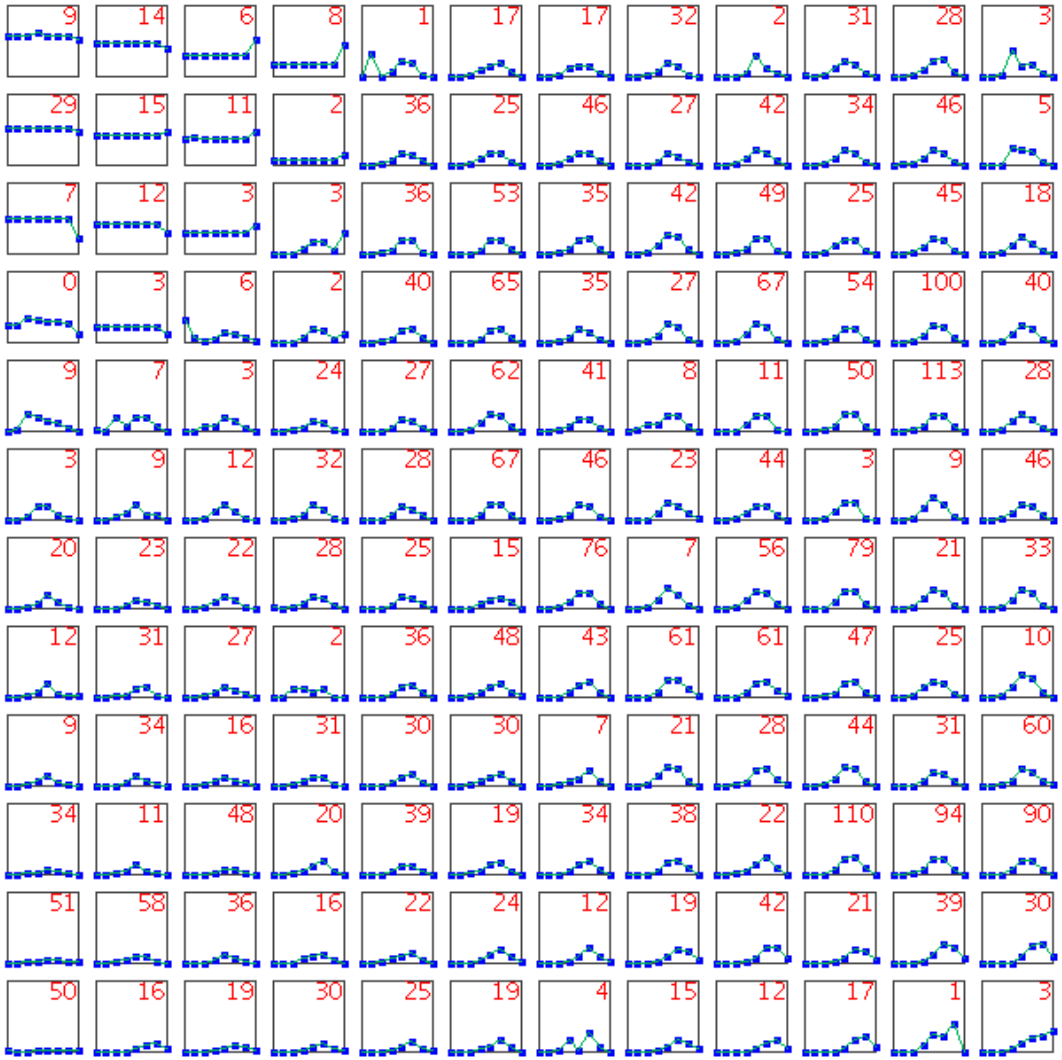
Self-Organizing Map



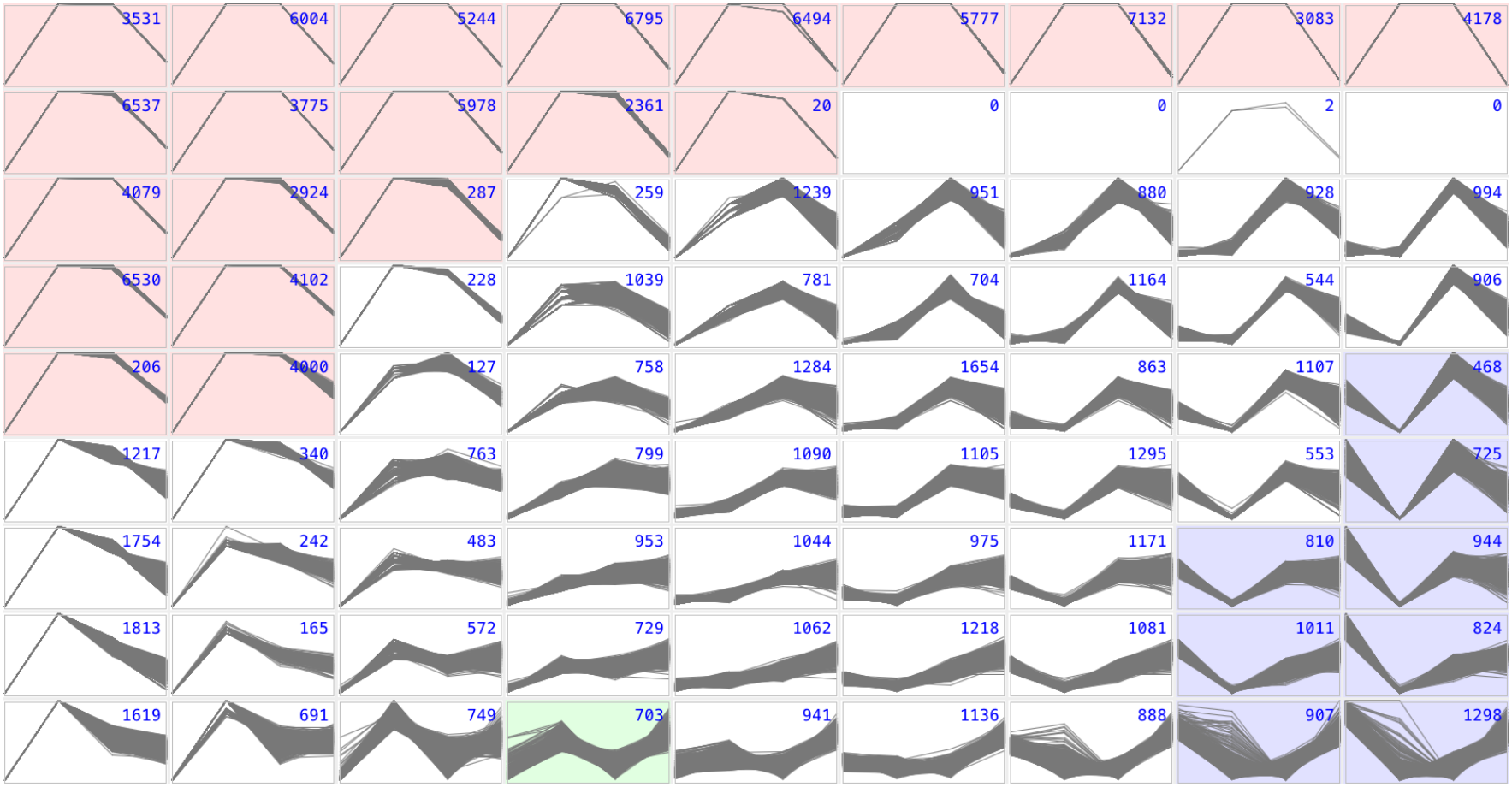
12x12 SOM,
Dados em 8
dimensões.

T=320
R=1
Lr=0.18

Self-Organizing Map



Self-Organizing Map



Visualization of Citizen Science Volunteers' Behaviors with Data from Usage Logs, Alessandra Marli M. Morais, Rafael D.C. Santos, M. Jordan Raddick, Computing in Science and Engineering, July/August 2015