CENG574
Statistical Data Analysis

Introduction

Department of Computer Engineering
Middle East Technical University

Outline

General Info

- **Instructor:** Hüseyin Aydın (email: huseyin@ceng.metu.edu.tr · Office: B211)
- **Office Hours:** By appointment
- **Main Reference Book:** E. Alpaydın, Introduction to Machine Learning, The MIT Press. [3rd Edition (2014) or 4th edition (2020)]

Course Objectives & Prerequisites

- **Objectives:**
  - The objective of this course is to introduce the concepts and techniques of *clustering* and *multivariate and exploratory data analysis*.
  - This course also offers an opportunity to perform data analysis by using data *visualization*, *projection* and *embedding*.
  - Practice over fields of applications: data streaming etc.
- **Prerequisites:** Knowledge of programming, probability and linear algebra.

Course Outline

Week	Date	Topics	Assignments
1	Feb 20	Introduction	#1
2	Feb 27	Input representation; distance metrics and similarity measures	#2
3	Mar 6	Probability and linear algebra; linear projections of data; PCA	
4	Mar 13	MDS; clustering	#3
5	Mar 20	No Lecture	
6	Mar 27	Hierarchical clustering; k-means clustering	#4
7	Apr 3	Evaluation and validity of clusters	#5
8	Apr 10	Overview and discussion of covered topics	
9	Apr 17	Midterm Exam (in-class)	
10	Apr 24	Other clustering algorithms	
11	May 1	No Lecture	
12	May 8	Non-linear projections	#6
13	May 15	Presentations of dataset analysis prelim. work and discussion	
14	May 22	Data stream analysis	
15	May 29	No Lecture	
16	Jun 5	Presentations of the final analysis of datasets	

Why should you take this course?

- To be able to apply exploratory analysis over data sets
  - To present a proper visualization for the data
  - To interpret the data for further processing
- This preprocessing of data is necessary for construction or selection of Machine Learning models, yet our approach to those models by themselves are superficial.

Why should you not take this course?

- To learn Python and/or R
- To learn statistics and/or ML

Grading

- Attendance and class participation (5 %)
- Assignments (25 %): #1-2: 3 % each · #3-5: 5 % each · #6: 4 %
- Midterm exam (20 %)
- Final dataset analysis (20 %)
- Final exam (30 %)

Other Policies

- Assignments will be done on individual basis.
- Final dataset analysis will be performed in a team setting of 2 persons.
- `Python` is the main programming language for the applied part of the course, yet `R` is also accepted.
- You have a total of 4 days of late submission for the assignments.
- Communication is via ODTUClass (`https://odtuclass.metu.edu.tr`).
- All submissions must reflect your own effort. Using AI agents for code/text generation is prohibited.

Registration Requests & Profiling

- Please fill the Google form whether you've already registered or not:

Wisconsin Breast Cancer Database

Number of Instances: 699

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(benign 2, malignant 4)

Wisconsin Breast Cancer Database

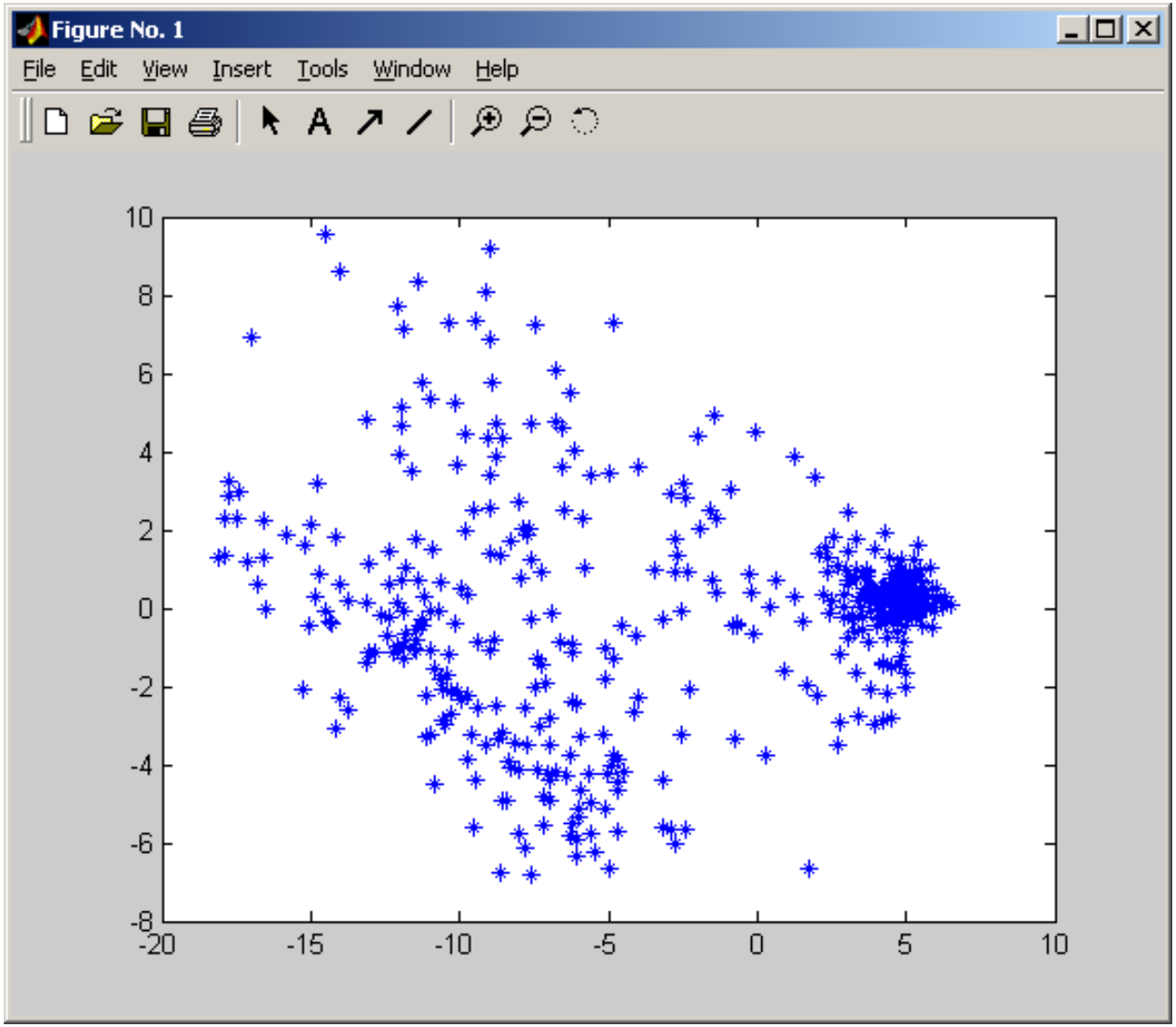
Number of Instances: 699

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(benign 2, malignant 4)

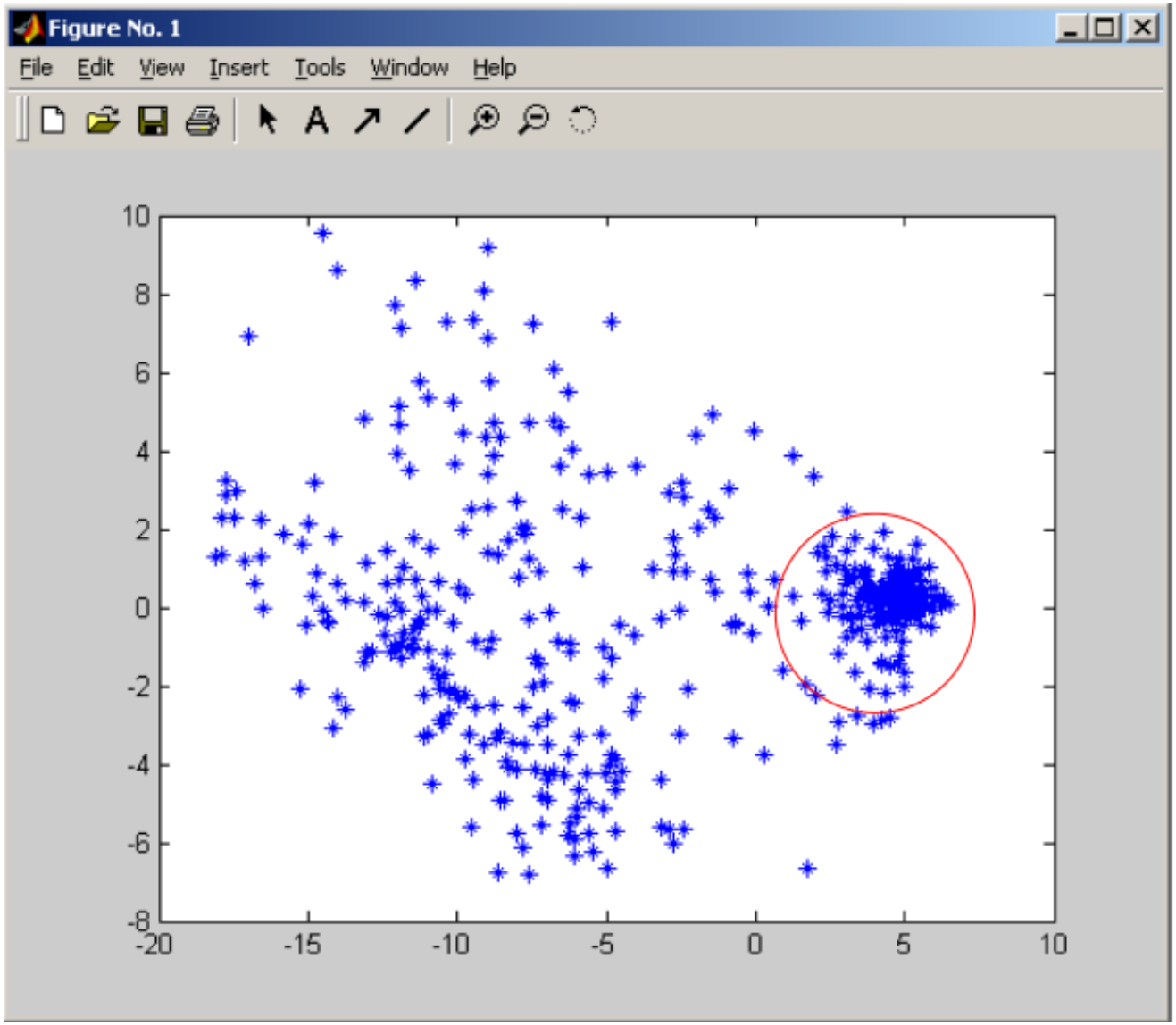
Wisconsin Breast Cancer Database

5,1,1,1,2,1,3,1,1,2
5,4,4,5,7,10,3,2,1,2
3,1,1,1,2,2,3,1,1,2
6,8,8,1,3,4,3,7,1,2
4,1,1,3,2,1,3,1,1,2
8,10,10,8,7,10,9,7,1,4
1,1,1,1,2,10,3,1,1,2
2,1,2,1,2,1,3,1,1,2
2,1,1,1,2,1,1,1,5,2
4,2,1,1,2,1,2,1,1,2
1,1,1,1,1,1,3,1,1,2
2,1,1,1,2,1,2,1,1,2
5,3,3,3,2,3,4,4,1,4
1,1,1,1,2,3,3,1,1,2
8,7,5,10,7,9,5,5,4,4
7,4,6,4,6,1,4,3,1,4
4,1,1,1,2,1,2,1,1,2
4,1,1,1,2,1,3,1,1,2
10,7,7,6,4,10,4,1,2,4
6,1,1,1,2,1,3,1,1,2
7,3,2,10,5,10,5,4,4,4
10,5,5,3,6,7,7,10,1,4
3,1,1,1,2,1,2,1,1,2
1,1,1,1,2,1,3,1,1,2
5,2,3,4,2,7,3,6,1,4

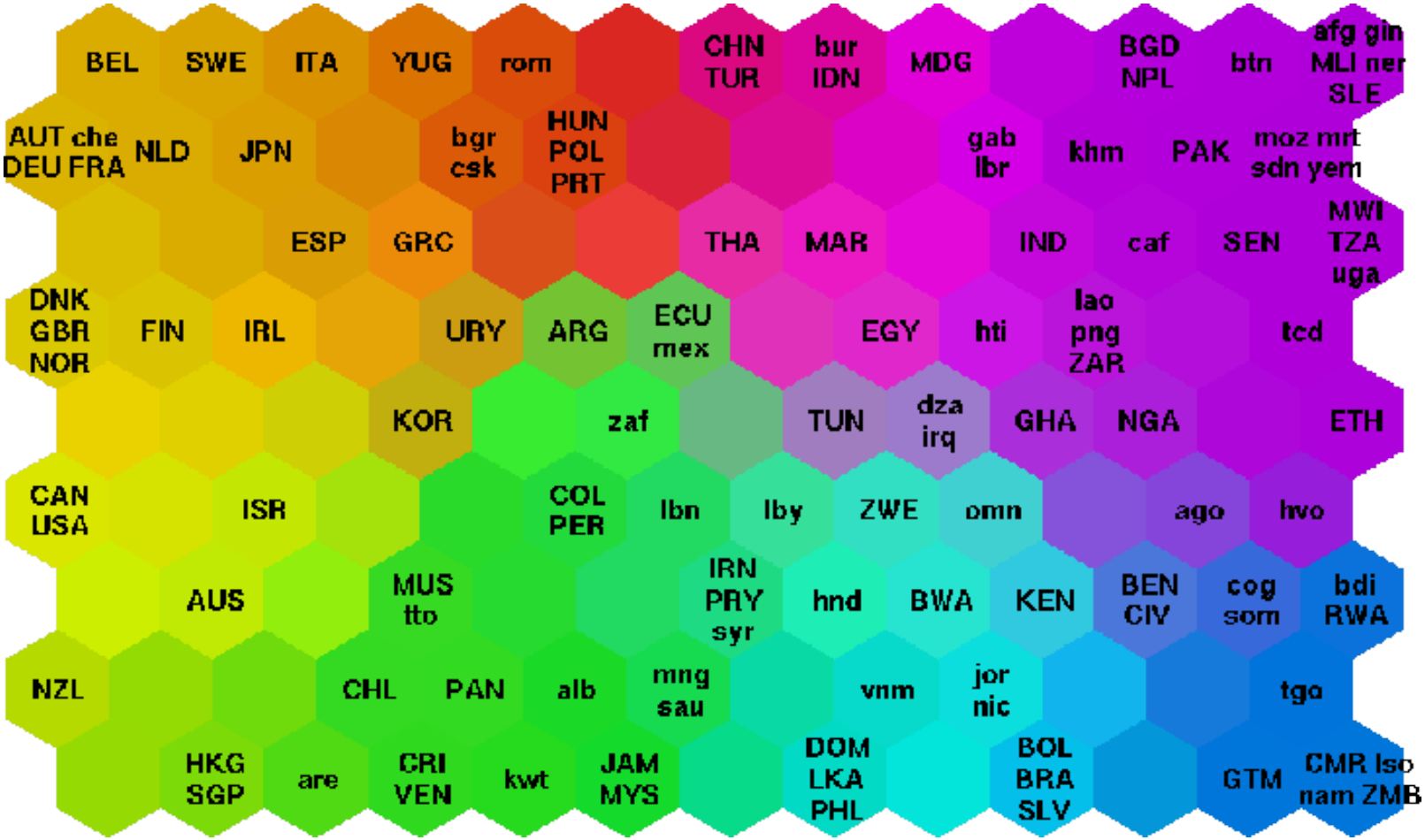
Projection by PCA



Projection by PCA

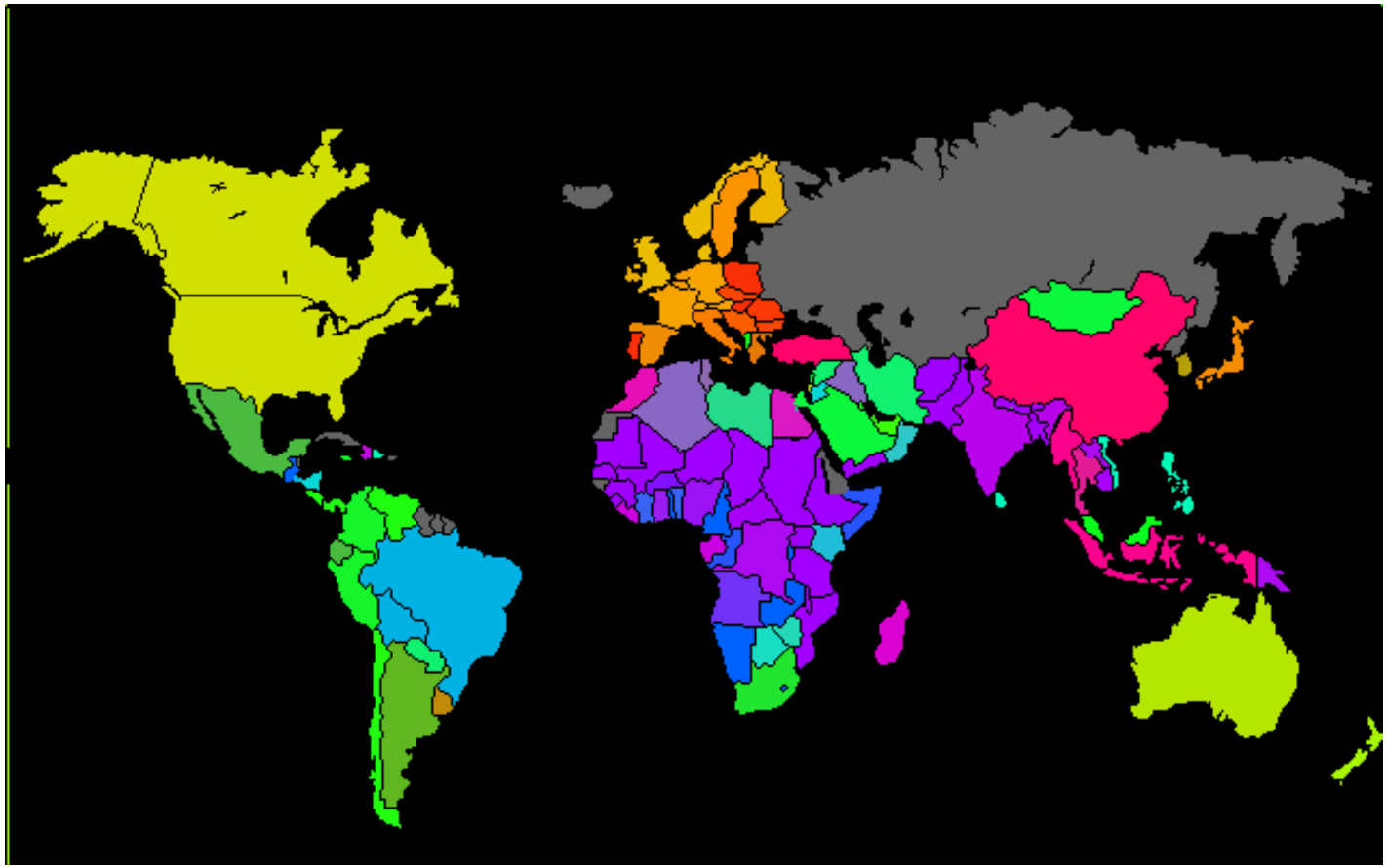


Poverty Map



From: Helsinki University of Technology, Finland

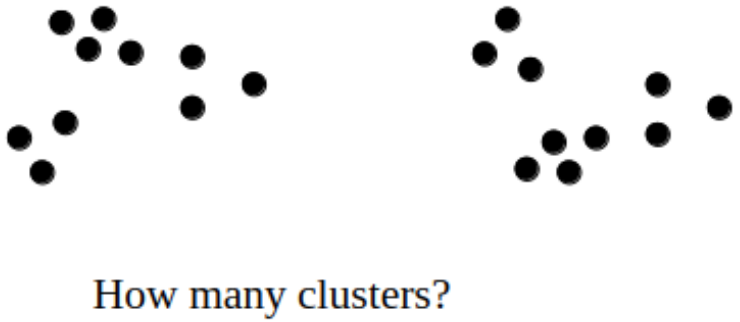
Poverty Map



From: Helsinki University of Technology, Finland

Clustering

Notion of a cluster can be ambiguous:



How many clusters?

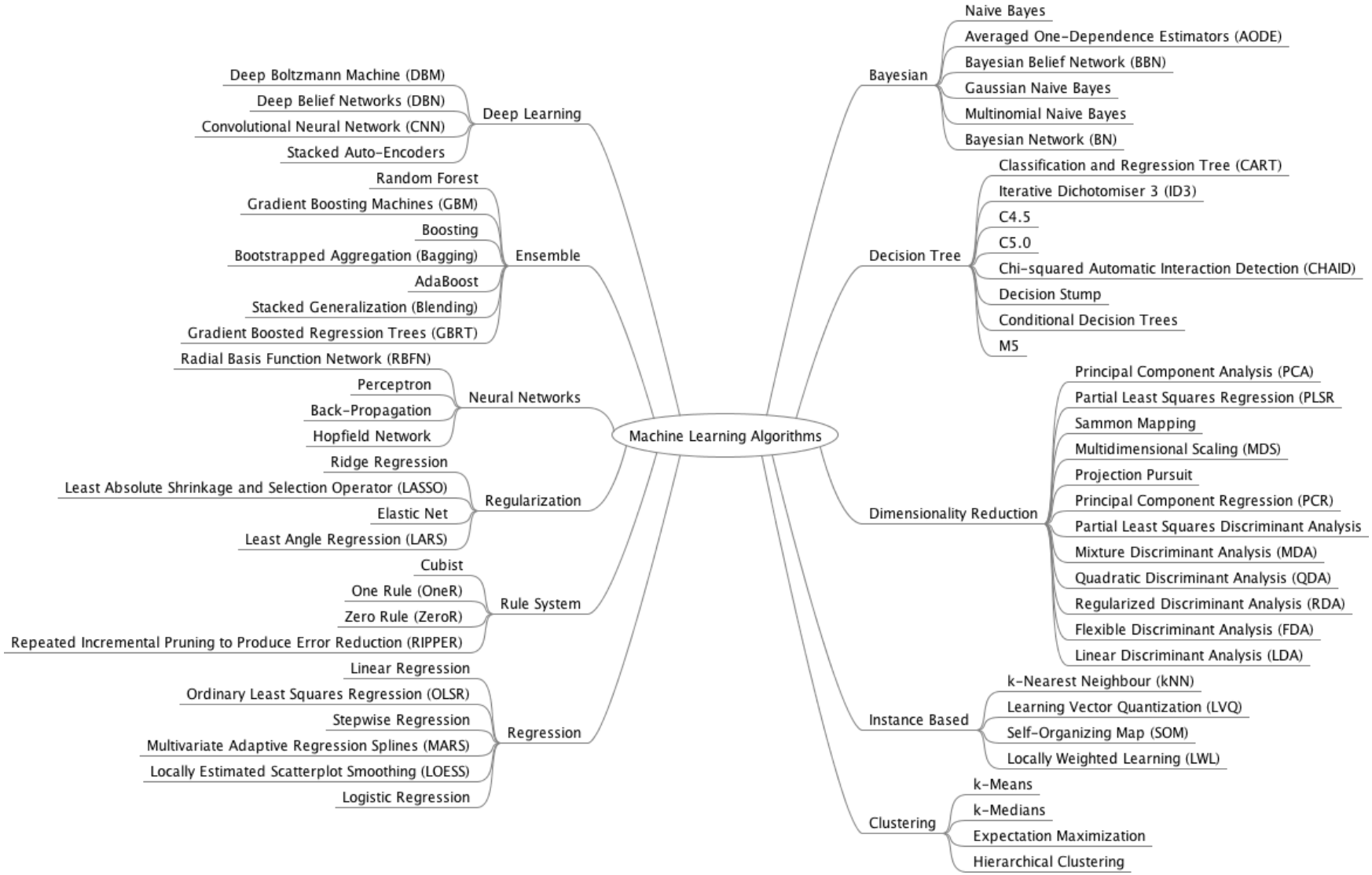
Six Clusters

Two Clusters

Four Clusters

From: Introduction to Data Mining by Tan, Steinbach, Kumar

A mindmap of ML algorithms



Resources: Datasets

- OpenML `https://www.openml.org/search?type=data`
- UCI KDD Archive: `http://kdd.ics.uci.edu/summary.data.application.html`
- Statlib: `http://lib.stat.cmu.edu/`
- Kaggle: `https://www.kaggle.com/`
- Delve: `http://www.cs.utoronto.ca/~delve/`

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Pattern Recognition
- Pattern Recognition Letters
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on Artificial Neural Networks (ICANN)
- International Conference on AI & Statistics (AISTATS)
- International Conference on Pattern Recognition (ICPR)
- ...

Questions

Questions?