# Lab 0

About A.I.R.

# Useful information

Alex Becheru
irlab@becheru.net
irlab.becheru.net

Common Sense

+

!!! attendance is mandatory !!!

at least 8 out of 10 labs

# Grades

Bad <5

Good = 5 / 6

Very Good = 7 / 8

Excellent = 9 / 10

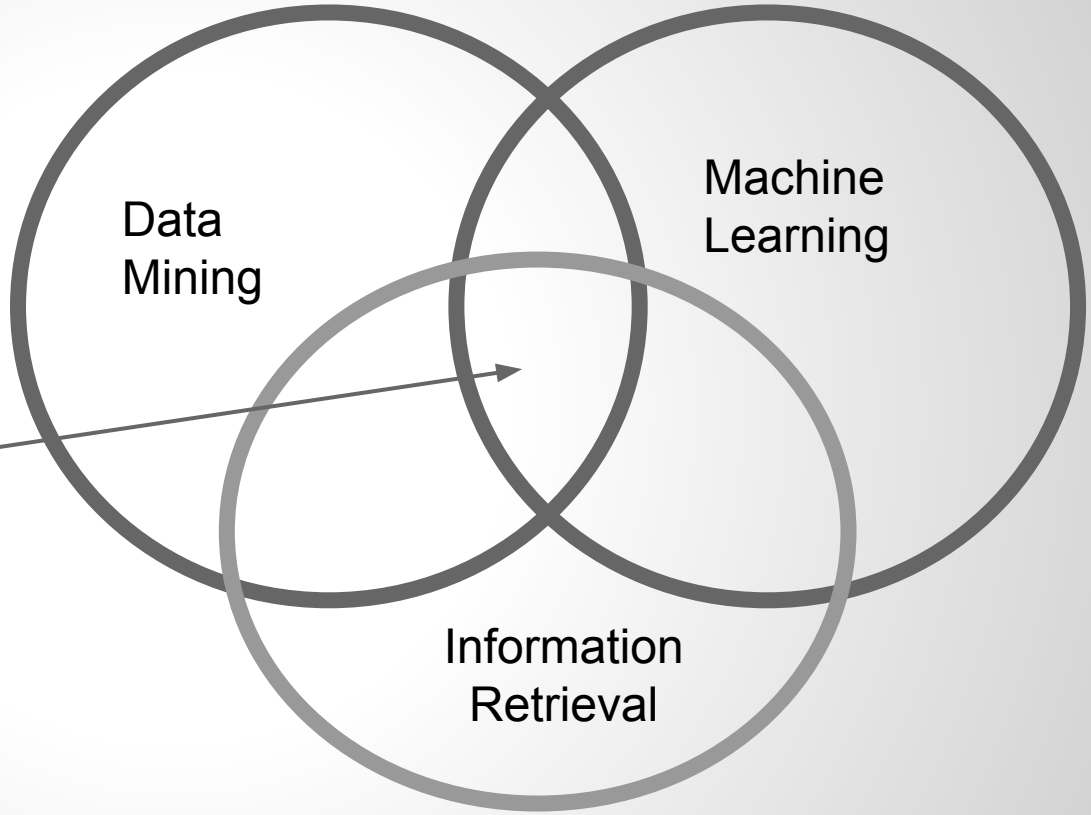
# Projects

mihaescu@software.ucv.ro

!!! Please complete the
I know you form !!!

# What are you going to learn in this course?

Fundamental
algorithms of
computer intelligence



# What is Information Retrieval?

The activity of obtaining information resources relevant to an information need from a collection of information resources

# What is not IR

- searching a particular recipe on Google
- seeking a medical register to analyse the record of a patient
- looking up spa resorts in the Golden Pages

# Human in vitro fertilisation

- take several eggs from the female
- take sperm from the male
- fertilise the eggs with the sperm -> several embryos
- *select the embryos with the best chance of survival*
- put the selected embryos back in the woman's uterus

# Selecting embryos?

- each embryo has 60 recorded features
- for a human embryologist it is hard to asses all 60 features

# Dairy farmers

- you have 1 000 000 cows
- 80 % have to be kept for the milk production
- 20 %  have to be turned into barbecue meat
- which cows should you keep in order to maximise your profit ?
- each cow has about 700 attributes:

- age
- undesirable temperament
- pregnant or not
- etc ...

# Detecting stolen credit cards

Ramon C. Barquin case

- Ramon receives a call from the bank
- the bank asks him if he had lost his cards
- Ramon searches his cards but he can not find them

How did the bank knew that he had lost his cards?

# IR helped the bank !!!

- An automated IR algorithm classified his credit card history:
  - places he has been
  - phone calls
  - other payments
- The algorithm classified them in
  - usual events
  - unusual events
- his latest phone calls (payed with the credit card) were unusual
- the bank personnel was notified by the algorithm

# Supermarkets

How does a supermarket arrange products ?

- why tea is near the cat products?
- why beer is near the car section?

IR helped again !!!

- the supermarket collected data on what was bought by each client
- they classified each shopper into different types
- they created special sections for each type of client

What are the main techniques that IR uses?

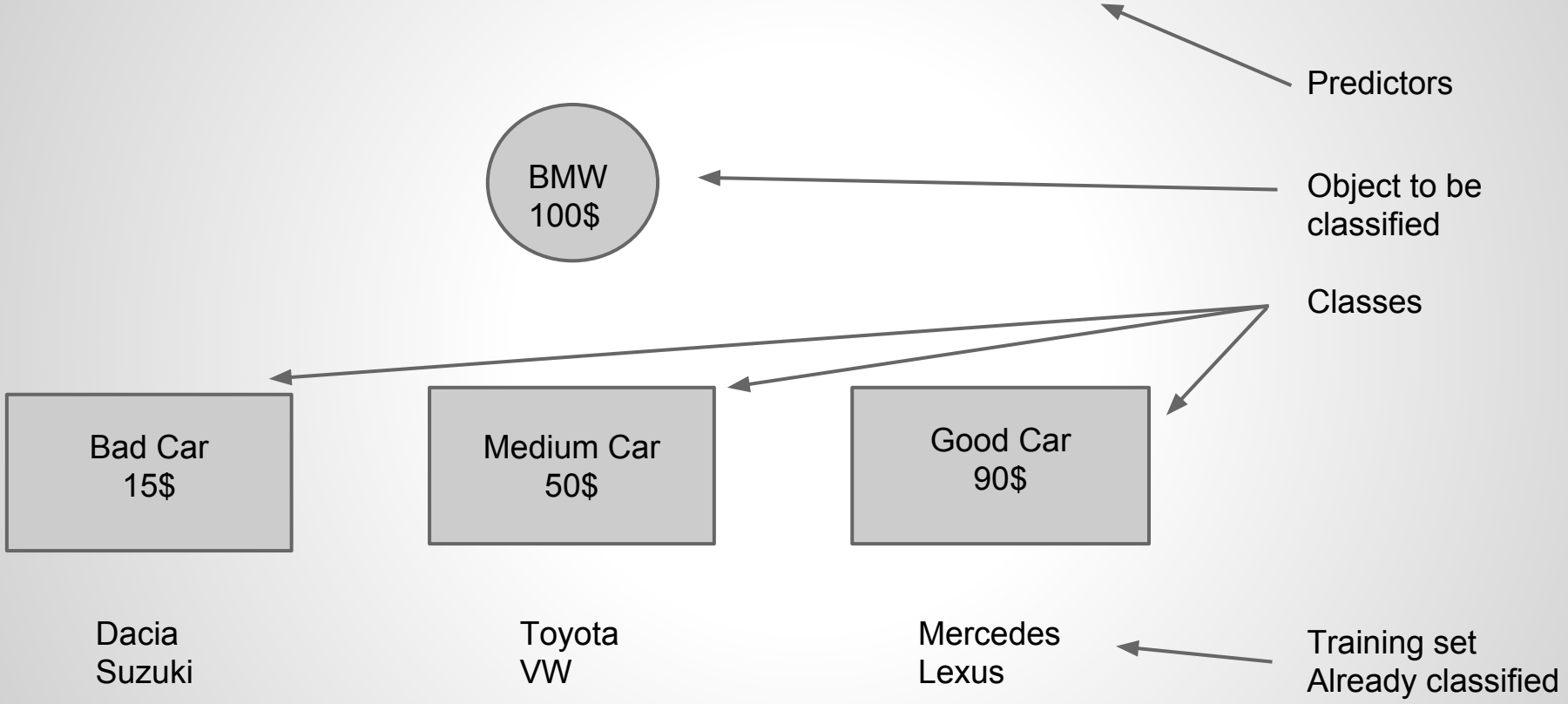
# Classification

The process of placing a specific object/concept in a set of categories

# Fundamental components

- Class - set of  classes
- Predictors - the attributes based on which the classification will be made
- Training set - a object set already classified
- Testing set - used to test the accuracy

# Classify a car maker based on the average price of a car



# Cluster analysis

A method to divide a set of data into several groups (clusters) based on certain predetermined similarities.

Inside a cluster objects are more alike than those that are not in the that cluster

# Fundamental components

- Predictors - the attributes based on which the classification will be made
- Testing set - used to test the accuracy

- The class types are not pre-determined here
- No training set

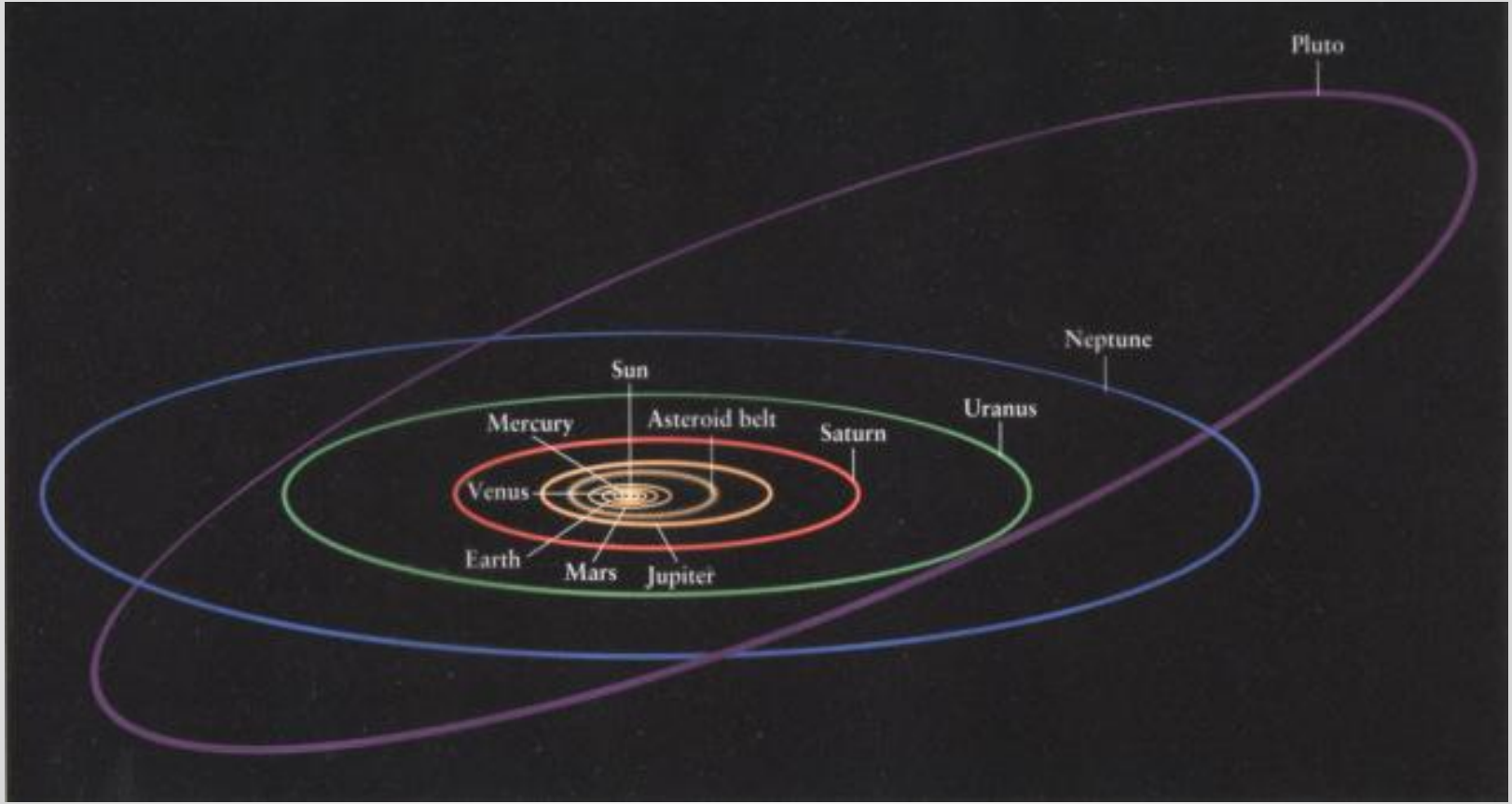
# Clustering applications

How many planets are in our solar system?

# Pluto is not a planet any more

- there are 8 planets, not 9
- Pluto is now a dwarf-planet
- a clustering algorithm did not put Pluto in the same cluster as the rest of the planets

# Pluto does not have the same characteristics as the other planets




Pluto
"Xena"
(2003 UB313)
The Moon
2,275 km
(1,422 miles)
2,384 km
(1,490 miles)
3,476 km
(2,160 miles)
Earth
12,756 km
(7,920 miles)

# Association Rule Discovery

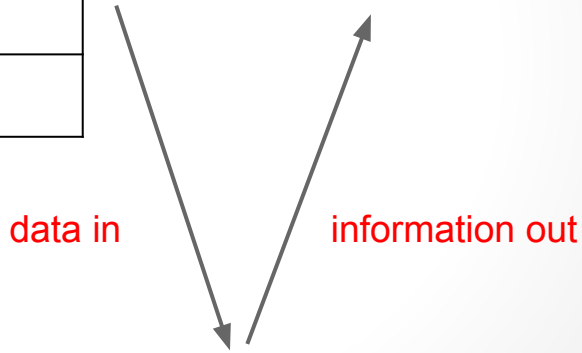
The process of identifying the rules of dependence between different groups of phenomena

What other products besides beer do men buy oftenly ?

# Beer and diaper case

| Items purchased by man with age 25 to 35, during Weekends |
| --- |
| bread, beer, diapers |
| beer, coke, diaper, milk |
| beer, bread, diaper, milk |

The supermarket put the beer shelves near the diaper shelves




If a man bought beer in weekends and he he is of age 25 to 35
-> then there is a high probability we would buy diapers.

# What is this lab about?

A practical approach on how to use Weka for Information Retrieval

# What is Weka

The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools.

# What are the steps of working with Weka

- preprocess the data
- apply IR algorithms
- get the rules / information
- test the rules

# What will you learn from this laboratory?

- load and look at data
- preprocess data
- visualise and explore data
- apply and understand IR algorithms
- interpret output
- be aware of common flaws with IR

# Homework

- Search on the internet for the “beer and diaper case”.
- Find out why men buy diapers.

Questions, Observations, Remarks

???