Introduction to Data Mining and Machine Learning Techniques

Lecture 1

Dr. Vassilis S. Kodogiannis
Reader in Computational Intelligence
V.Kodogiannis@westminster.ac.uk

7BUIS008W Data Mining and Machine Learning

Week No	Week	Lecture	Module-Staff	Tutorials
1	22/09/20	Introduction	VK	No tutorials during week-1
2	29/09/20	K-means clustering	VK	Familiarization with R,
3	06/10/20	Hierarchical clustering	VK	Practical/Lab exercises on clustering: (partition algorithms) **06/10/20: CWK1 to be issued**
4	13/10/20	Neural - MLP	VK	Practical/Lab exercises on clustering: (hierarchical algorithms)
5	20/10/20	Neural (unsupervised) + NF	VK	Practical/Lab exercises on Neural Networks
6	27/10/20	No lecture / tutorial – Engagement Week		
7	03/11/20	Support Vector Machine (SVM)	VK	Practical/Lab exercises on Neural Networks & SVM
8	10/11/20	Pattern Mining	PC	Practical / Lab exercises on association analysis
9	17/11/20	Pattern Mining	PC	
10	24/11/20	Predictive Modelling	PC	Practical / Lab exercises on decision trees/ Naïve Bayes **24/11/20: CWK1 to be submitted via BB**
11	01/12/20	Predictive Modelling	PC	**24/11/120: CWK2 to be issued**
12	08/12/20	Ensemble Learning, Review	PC	Practical / Lab exercises Ensemble Learning **07/01/21: CWK2 to be submitted via BB**

Module Staff Dr Panagiotis Chountas (PC), chountp@westminster.ac.uk, Dr Vassilis Kodogiannis (VK) Kodogiv@westminster.ac.uk

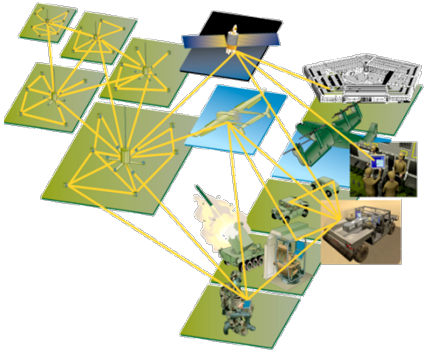
Why Mine Data? Commercial Viewpoint

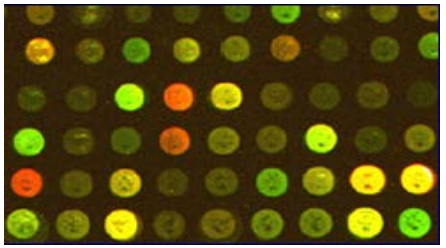
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

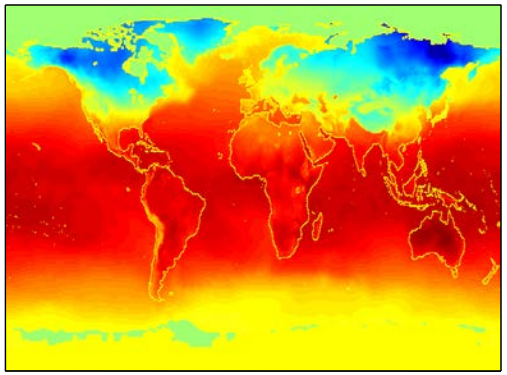
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data





- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

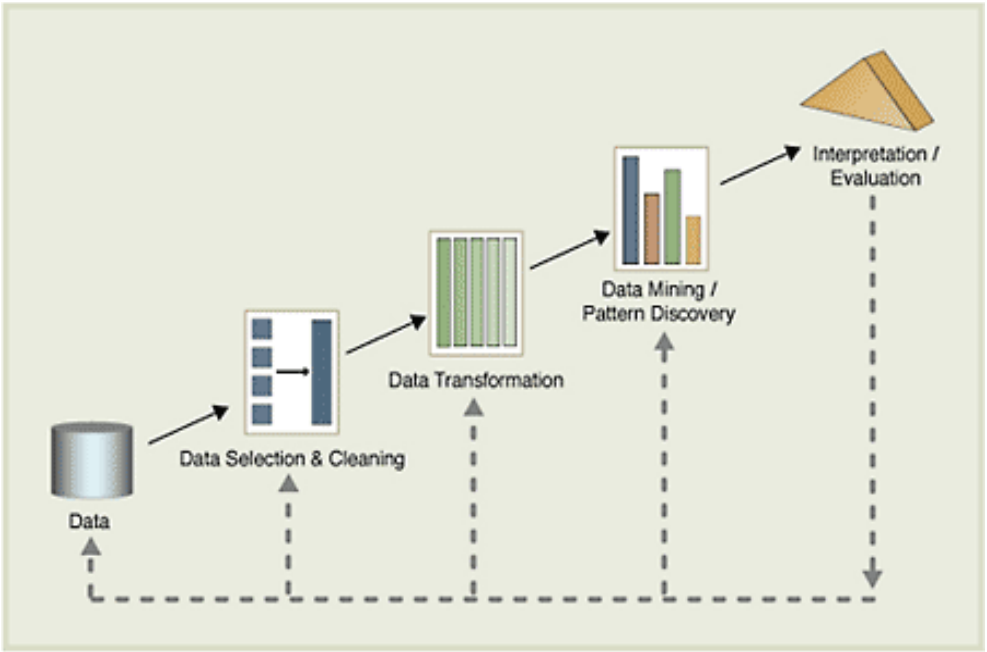


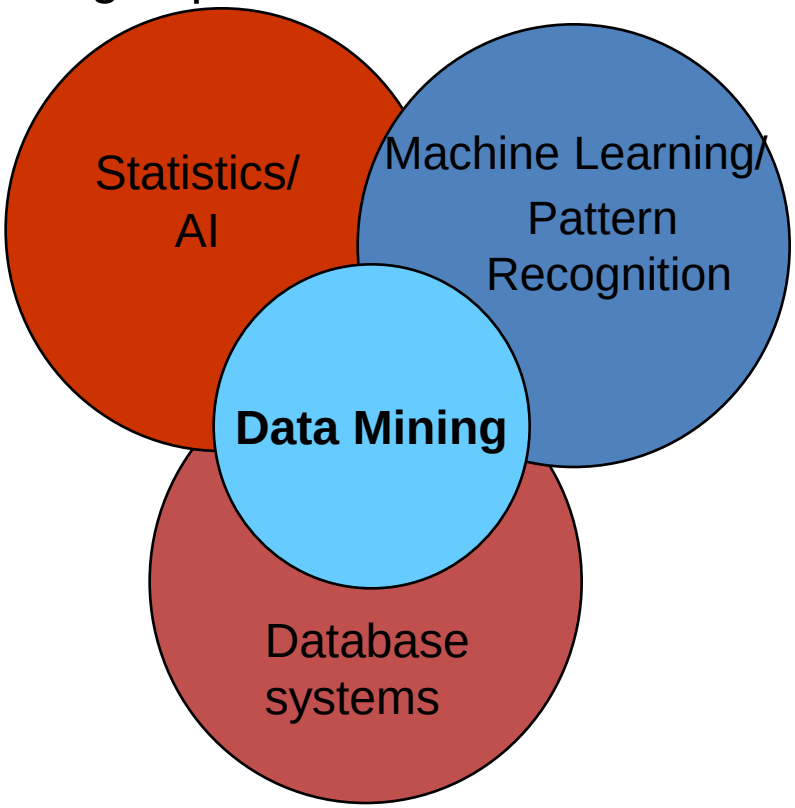
What is Data Mining?

Many Definitions

Non-trivial extraction of implicit, previously unknown and potentially useful information from data

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

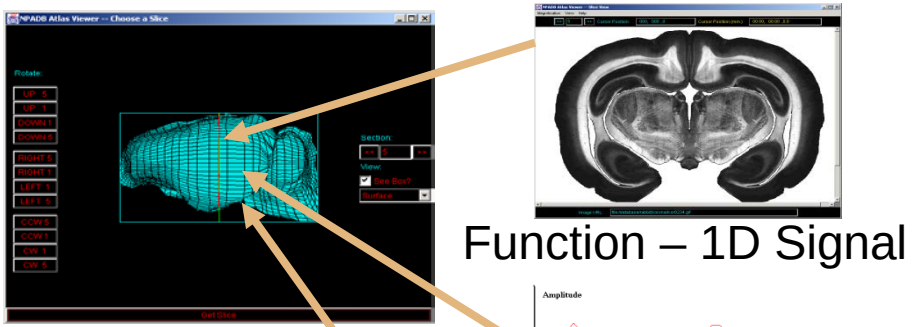




Data Mining: On What Kind of Data?

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced Database Systems
  - Object-Relational
  - Spatial and Temporal
  - Time-Series
  - Multimedia
  - Text
  - Heterogeneous, Legacy, and Distributed
  - WWW

Structure - 3D Anatomy



Function – 1D Signal

Metadata – Annotation

GeneFilter Comparison Report

GeneFilter 1 Name: O2#1 8-20-99adjfinal							GeneFilter 1 Nam N2#1finaladj	
		INTENSITIES						
		RAW	NORMALIZED					
ORF NAME	GENE NAME		CHRM		F	G	R	
YAL001C	TFC3	1	1	A	1	2	12.03	7.38
YBL088C	PET112		2		1	A 1 3		53.21
YBR154C	RPB5	2	1	A	1	4	79.26	78.51
YCL044C		3	1	A	1	5	53.22	44.66

Challenges of Data Mining

- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Challenges with Machine Learning

The field of Machine Learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Machine Learning

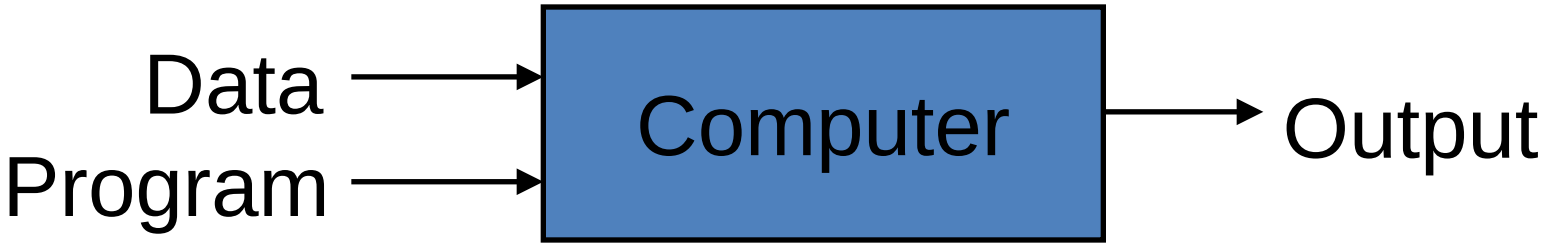
- Definition
  - Field of study that gives computers the ability to learn without being explicitly programmed.
    -- Arthur Samuel (1959).
- Examples:
  - Database mining
    - Large datasets from growth of automation/web.
    - E.g., Web click data, medical records, biology, engineering
  - Applications can't be programed by hand.
    - E.g., Autonomous helicopter, handwriting recognition, most of Natural Language Processing (NLP), Computer Vision.
  - Self-customizing programs
    - E.g., Amazon, Netflix product recommendations
  - Understanding human learning (brain, real AI).

"Machine Learning," Andrew Ng, accessed January 20, 2016, https://www.coursera.org/learn/machine-learning

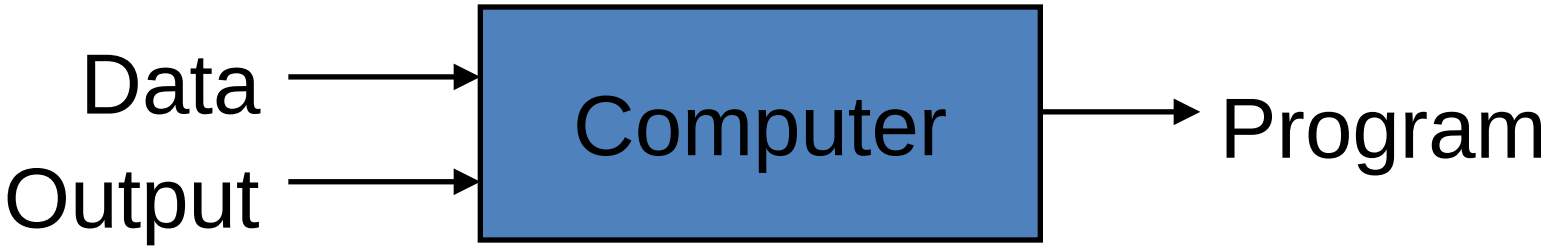
So What Is Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming



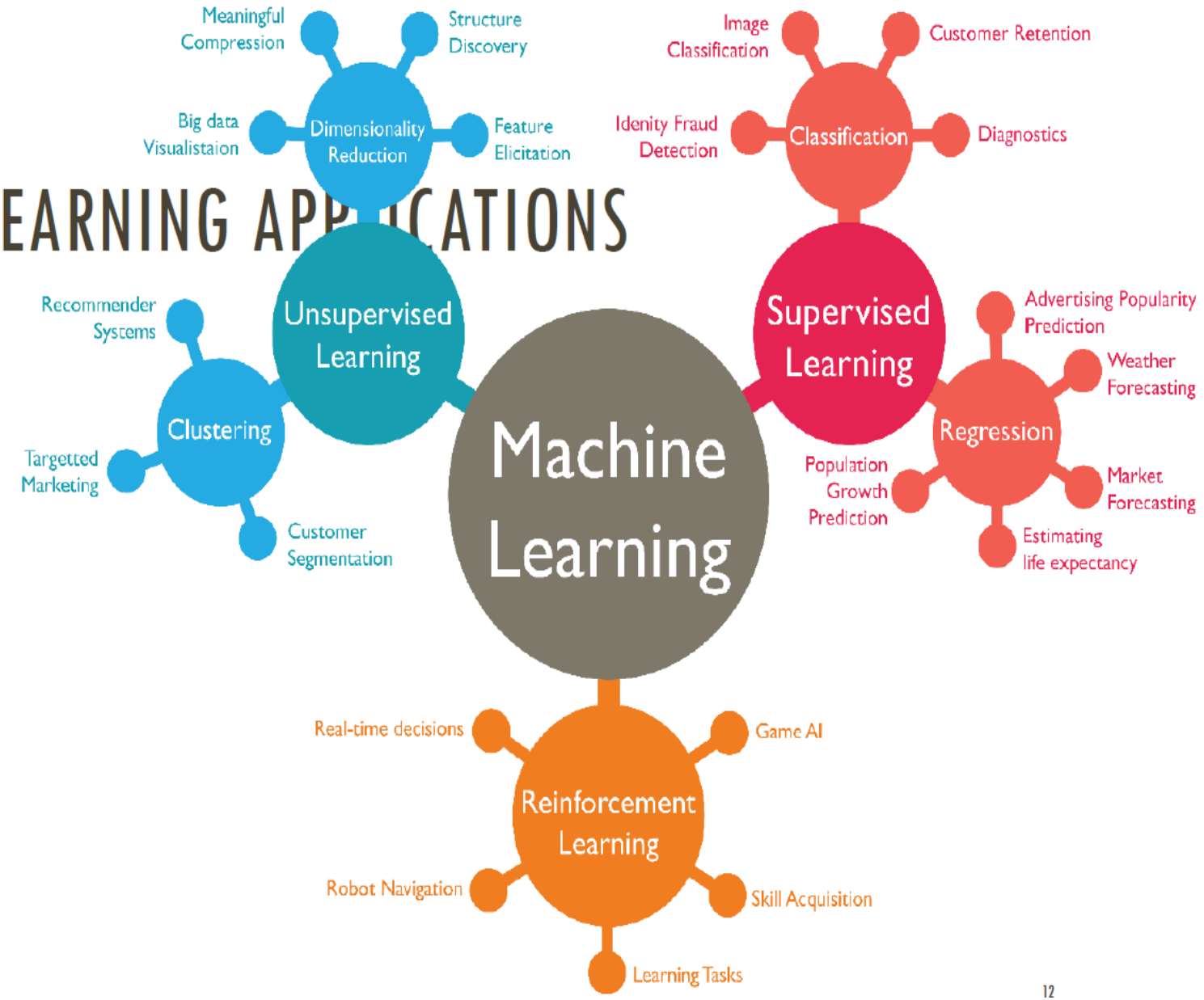
Machine Learning



Types of Learning

- **Supervised learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Semi-supervised learning**
  - Training data includes a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

MACHINE LEARNING APPLICATIONS

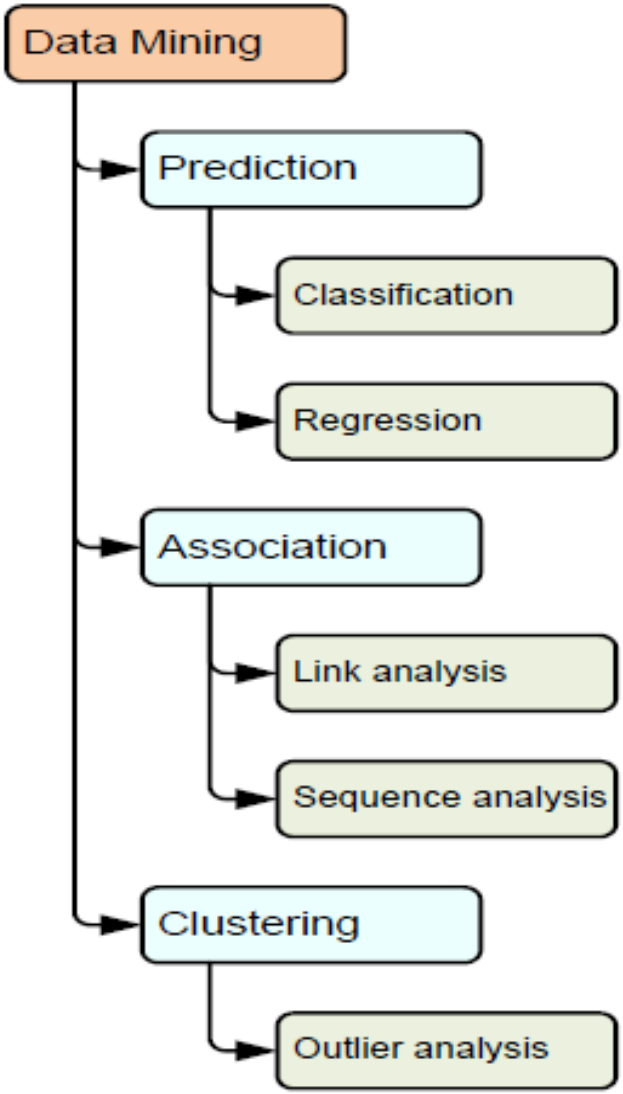


12

Topics to be discussed

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]

A Taxonomy for Data Mining Tasks



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Classification: Definition

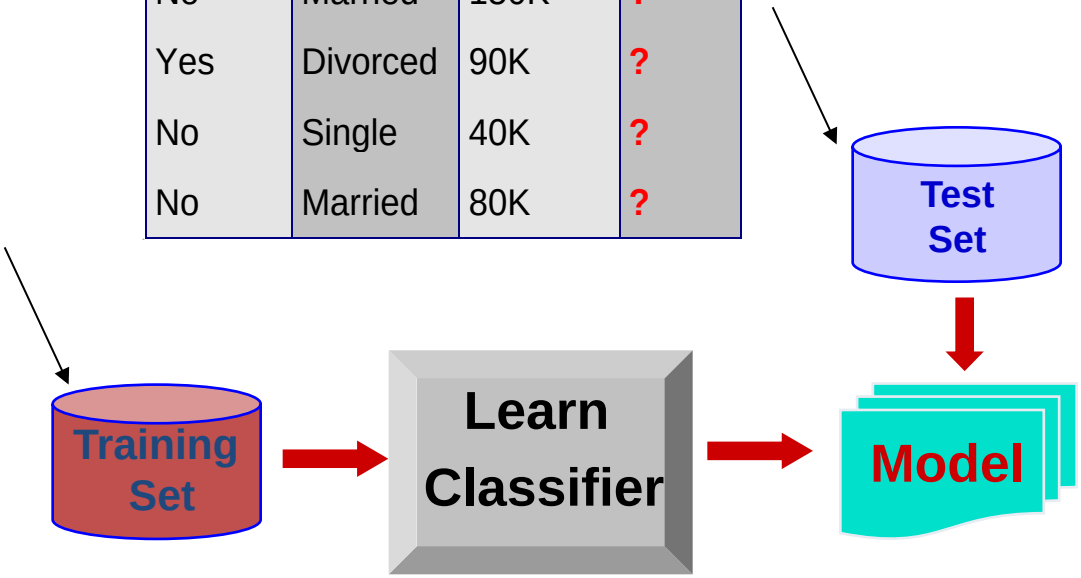
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

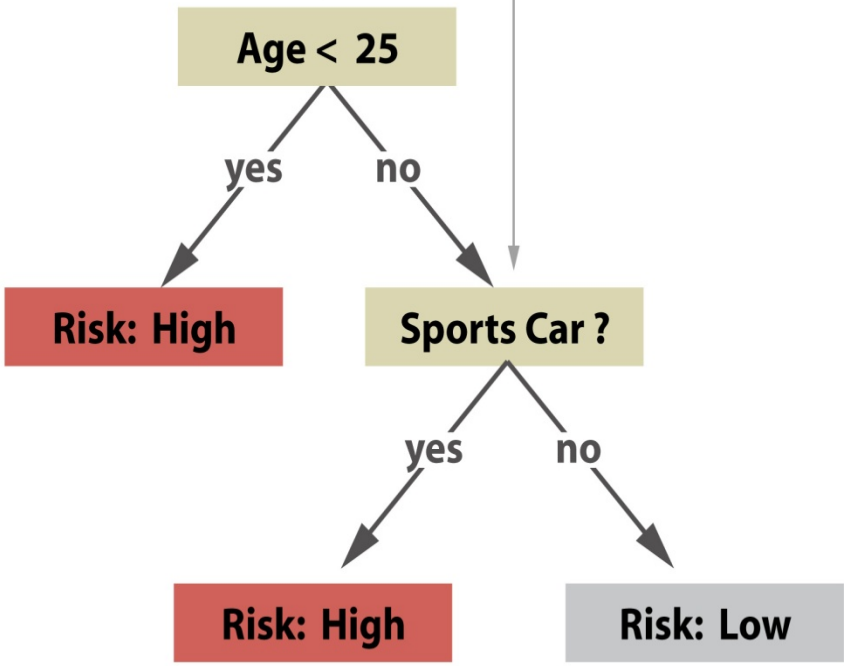
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification via Decision Trees

Insurance Risk Assessment

Age	Car Type	Risk
23	family	High
17	sports	High
43	sports	High
68	family	Low
32	truck	Low
20	family	High



Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

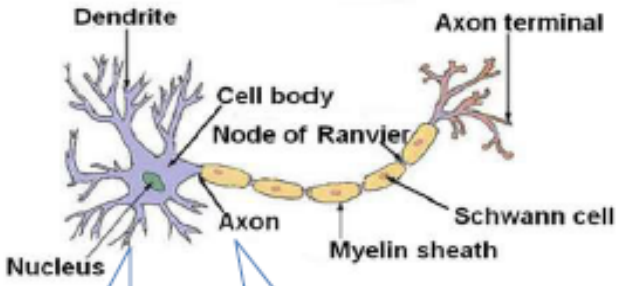
Neuron & Neural Networks

Pedestrian

Car

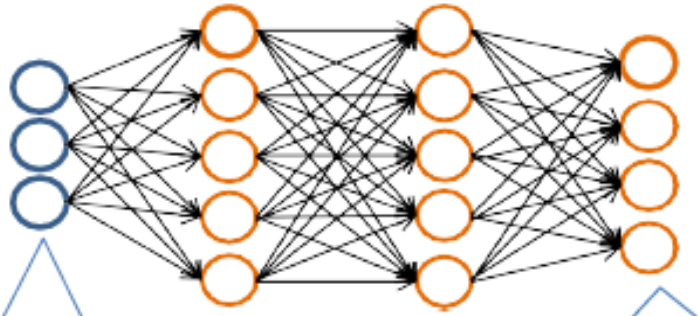
Motorcycle

Truck



Input

Output



Input (Image Pixel)

Output (Judgement)

$h_\Theta(x) \in \mathbb{R}^4$

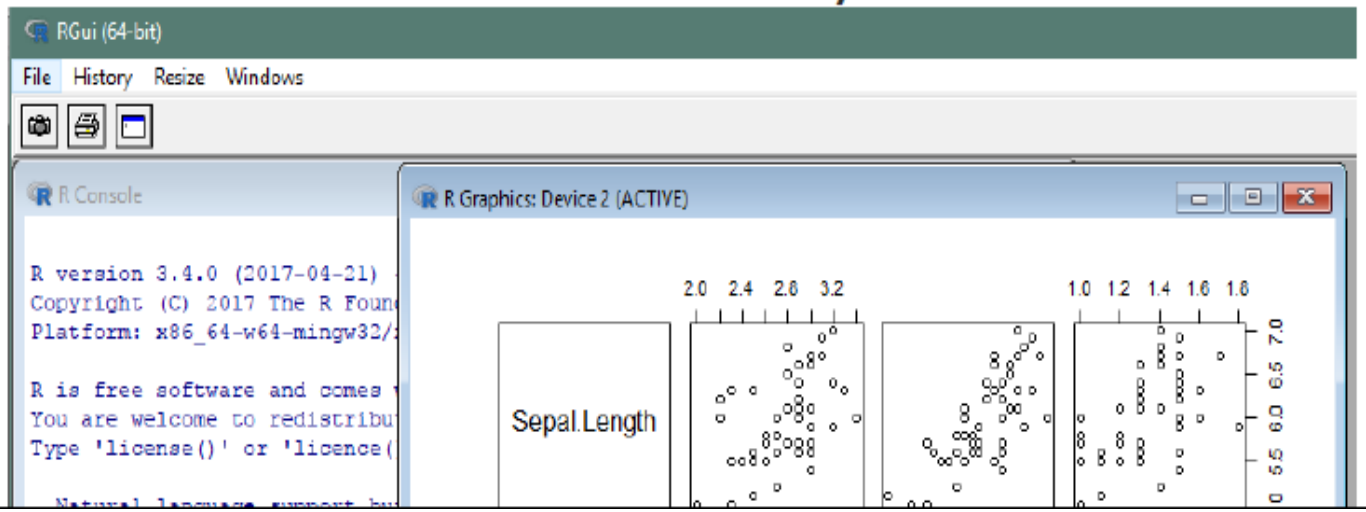
Want $h_\Theta(x) \approx \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$, $h_\Theta(x) \approx \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$, $h_\Theta(x) \approx \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$, etc.

when pedestrian when car when motorcycle

"Machine Learning," Andrew Ng, accessed January 20, 2016, https://www.coursera.org/learn/machine-learning

INTRODUCTION TO R

- Open source programming language and software environment for statistical computing.
  - Used by statisticians and data miners for developing statistical software and data analysis.



ikipedia

R is a free-distributed software and can be downloaded from: *https://cran.r-project.org/*. Versions for Windows, Mac and Linux are available. Make sure you download the latest version (R.4.02).

However, the installation of R language does not include the existence of a suitable interface, from where you are going to write and execute your codes. Therefore you need to download a suitable interface tool and this is the RStudio.

RStudio is an integrated development environment for R with a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging and workspace management. The free version can be downloaded from:

https://rstudio.com/products/rstudio/download/
https://rstudio.com/products/rstudio/download/#download

In order to complete the process, the installation needs to be performed through these two steps (in this specific order):

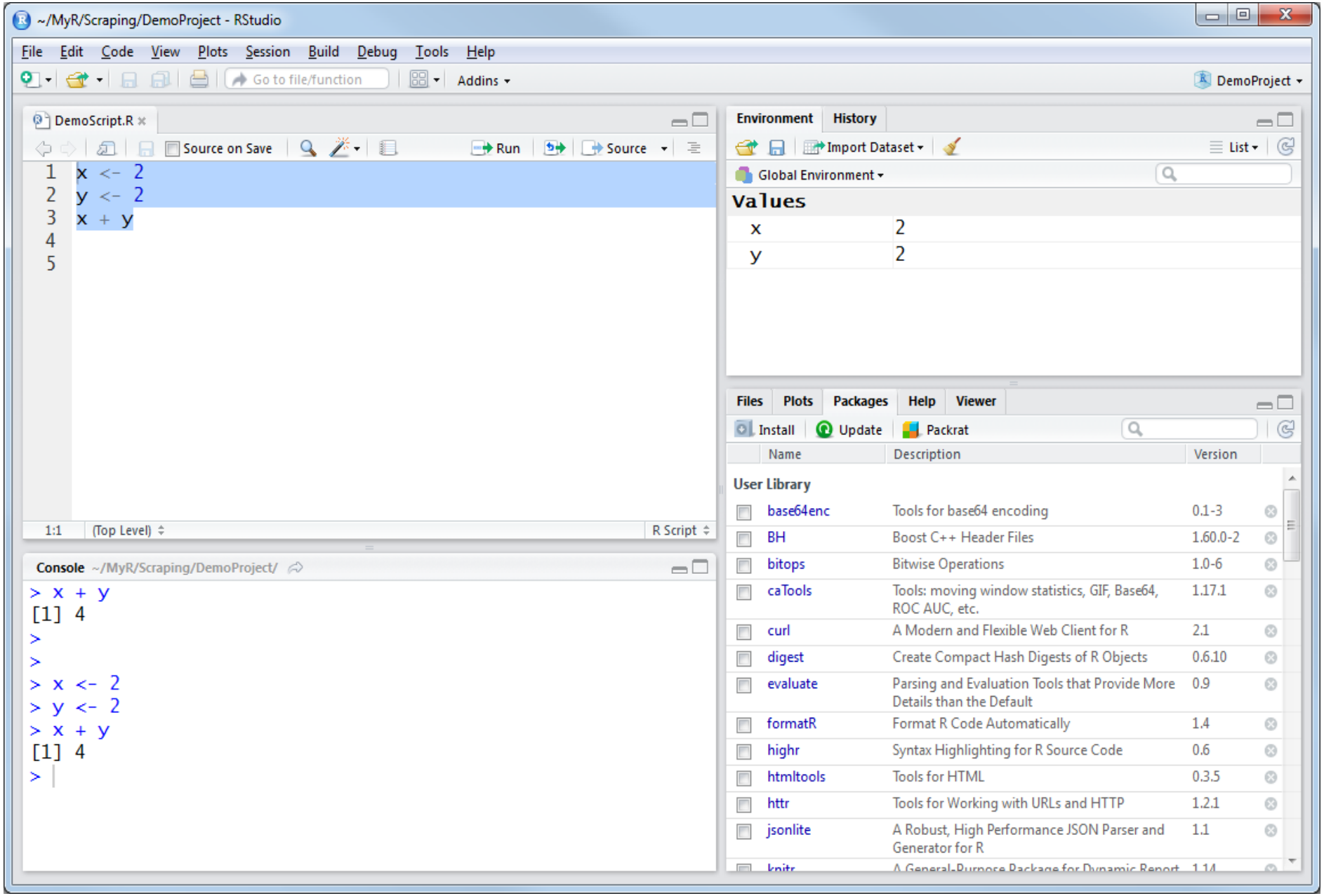
1. Install the R language
2. Install the RStudio (the RStudio will "see" the already installed R language).

Then by pressing the RStudio icon you can start working in R.

Check also the information from university website:

https://support.ecs.westminster.ac.uk/w/index.php/Pub:_RStudio

https://support.ecs.westminster.ac.uk/w/index.php/R_Example_Programs


~/MyR/Scraping/DemoProject - RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
DemoProject
DemoScript.R
Source on Save
Run
Source
1 x <- 2
2 y <- 2
3 x + y
4
5
1:1
(Top Level)
R Script
Console ~/MyR/Scraping/DemoProject/
> x + y
[1] 4
>
>
> x <- 2
> y <- 2
> x + y
[1] 4
>
Environment
History
Import Dataset
List
Global Environment
Values
x 2
y 2
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.60.0-2
bitops Bitwise Operations 1.0-6
caTools Tools: moving window statistics, GIF, Base64, ROC AUC, etc. 1.17.1
curl A Modern and Flexible Web Client for R 2.1
digest Create Compact Hash Digests of R Objects 0.6.10
evaluate Parsing and Evaluation Tools that Provide More Details than the Default 0.9
formatR Format R Code Automatically 1.4
highr Syntax Highlighting for R Source Code 0.6
htmltools Tools for HTML 0.3.5
httr Tools for Working with URLs and HTTP 1.2.1
jsonlite A Robust, High Performance JSON Parser and Generator for R 1.1