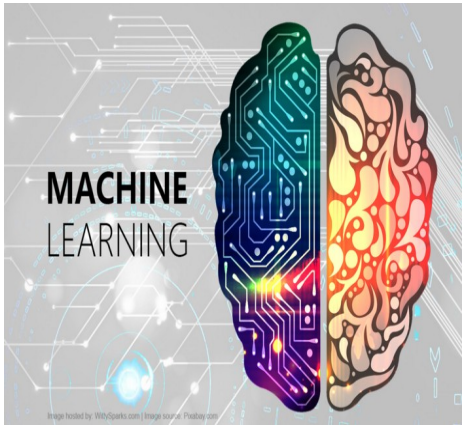
Introduction to Data Mining and Machine Learning Techniques

Lecture 1

Dr. Vassilis S. Kontogiannis
Reader in Computational Intelligence
Email: V.Kontogiannis@westminster.ac.uk
https://scholar.google.co.uk/citations?user=meTTcLAAAAAJ&hl=en&oi=ao



Week No:	Week Starting	Lecture (VK)	Tutorials
1	22/01/2024	Machine Learning & Data Mining overview (Introduction)	Provide information on the software needed for this module. Introduction to R programming with simple exercises using R
2	29/01/2024	Ethics and Bias in Machine Learning. Data Pre-processing, Dimensionality Reduction Techniques	Introduction to R programming with simple exercises using R
3	05/02/2024	Clustering Methods - (Partitioning)	Practical exercises on data pre-processing and PCA (manually and via R)
4	12/02/2024	Clustering Methods - (Hierarchical)	Practical exercises on clustering (manually and via R). Partitioning clustering (kmeans). 15/02/2024: CWK to be issued
5	19/02/2024	Decision Trees	Practical exercises on clustering (manually and via R). Hierarchical clustering.
6	26/02/2024	Engagement Week – No Lecture & Tutorials	
7	04/03/2024	Classification Methods: K-Nearest Neighbour & Naïve Bayes algorithms	Practical exercises on Decision Trees (manually and via R)
8	11/03/2024	Neural Networks & Applications	Practical exercises on NB & K-NN classifiers (manually)
9	18/03/2024	Association Rules	Practical exercises on Neural Networks (MLP) (manually and via R)
10	25/03/2024	Review on the topics covered in this module	Practical exercises on Association Rules (manually and via R)
11	01/04/2024	No Lecture - Easter	Practical exercises on NB & K-NN classifiers (via R)
12	8/04/2024	In-Class Test (8/4/2024) - 15:00-17:00 *(40% of the module total mark)*	Final Clarifications for CW assessment (on-line) 30/04/2024: CWK to be submitted via BB

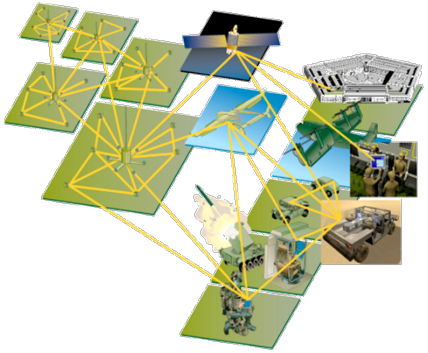
Why Mine Data? Commercial Viewpoint

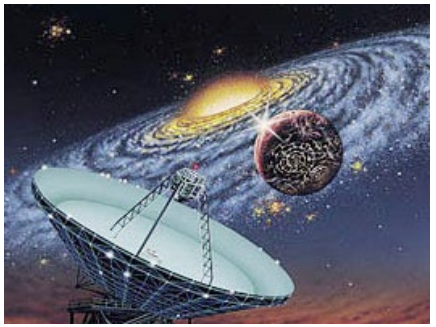
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

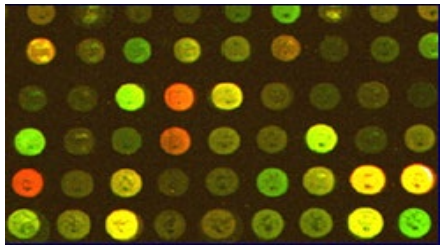
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

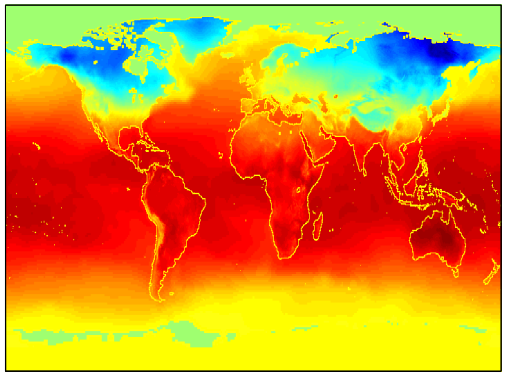
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data







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

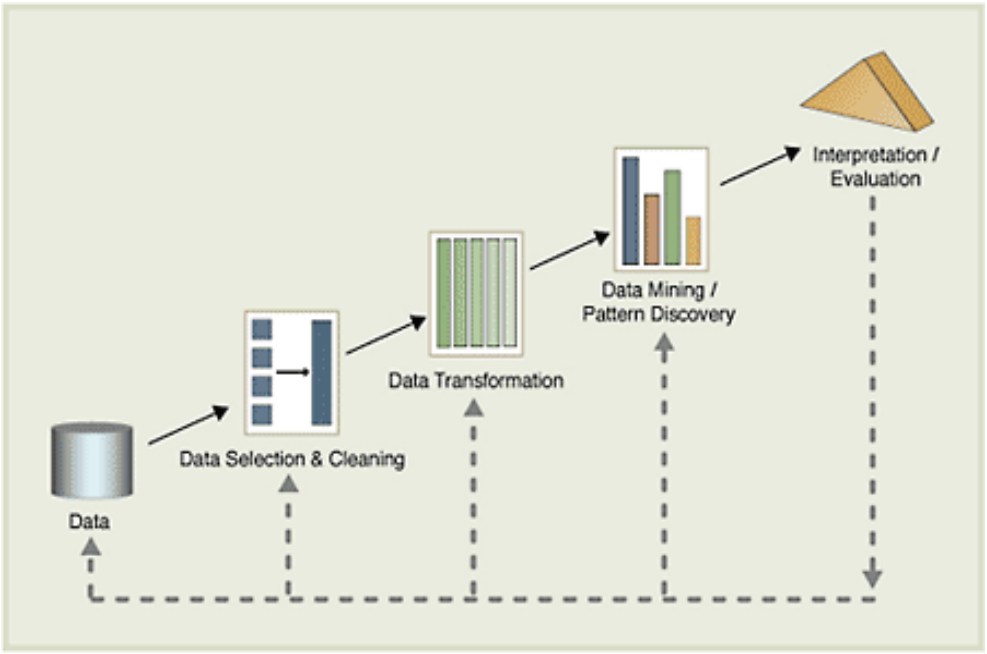


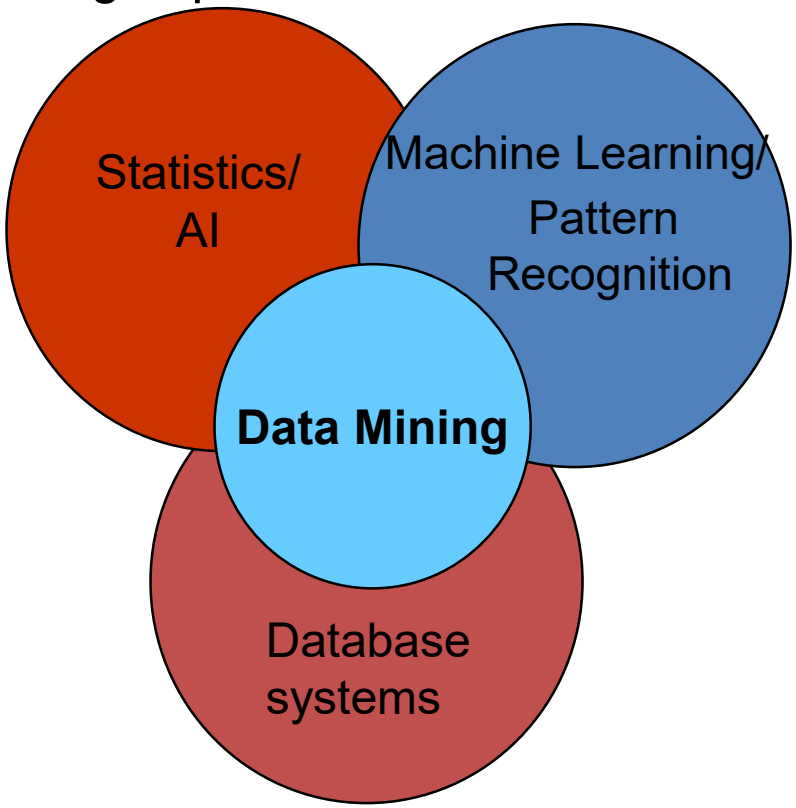
What is Data Mining?

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

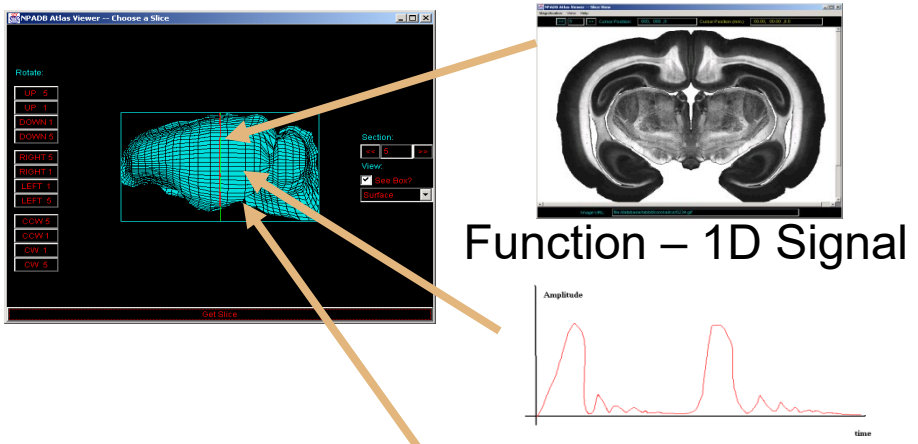




Data Mining: On What Kind of Data?

- 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: C2#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
YBL080C	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

Vassilis S. Kontogiannis

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.

What is Machine Learning?

"The subfield of computer science that "gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)

"***Machine learning (ML)*** *is concerned with the design and development of algorithms and techniques that* ***allow computers to "learn"****. The* ***major focus*** *of ML research is to* ***extract information from data automatically****, by computational and statistical methods. It is thus closely related to data mining and statistics*".
(Svensson and Söderberg, 2008)

What is Machine Learning?

Definition by Tom Mitchell (1998):

Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E

A well-defined learning task is given by $<P, T, E>$

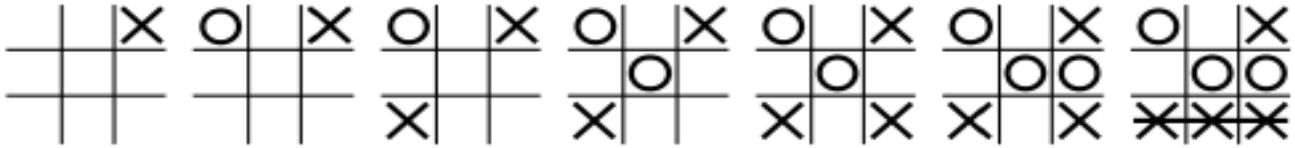
Examples:

T: Playing Chess (or Go)
P: Percent games won against an opponent
E: Playing games against itself

T: Classify emails as legitimate or spam
P: Percentage of emails labeled correctly
E: Repository of emails, some with human-specified labels

Example: tic-tac-toe

- How to program the computer to play tic-tac-toe?



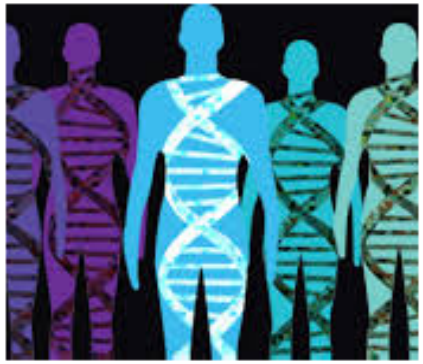
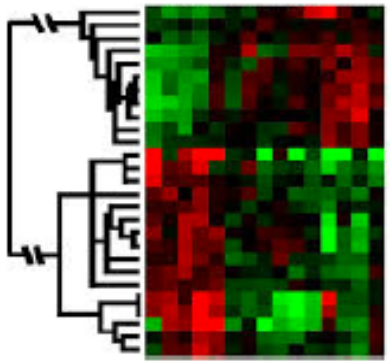
- Option A: The programmer writes explicit rules, e.g. 'if the opponent has two in a row, and the third is free, stop it by placing your mark there', etc (lots of work, difficult, not at all scalable!)
- Option B: Go through the game tree, choose optimally (for non-trivial games, must be combined with some heuristics to restrict tree size)
- Option C: Let the computer try out various strategies by playing against itself and others, and noting which strategies lead to winning and which to losing (='machine learning')

When Do We Use Machine Learning?

ML is used when:

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)



Learning isn't always useful:

- There is no need to "learn" to calculate payroll

Why "Learn" ?

- *"The goal of machine learning is to make a computer learn just like a baby — it should get better at tasks with experience."*

- A machine learning system can be used to
  - *Automate a process*
  - *Automate decision making*
  - *Extract knowledge from data*
  - *Predict future event*
  - *Adapt systems dynamically to enable better user experiences*
  - ...

Write code to explicitly do the above tasks

- How do we build a machine learning system?

Write code to make the computer ***learn*** how to do the tasks

When We Need Machine Learning

Tasks involving big data

- Genomics
- Internet search
- Anomaly detection

Tasks requiring customization

- Email filters
- Personalized medicine
- Image inpainting

Tasks for which it is challenging to specify our knowledge

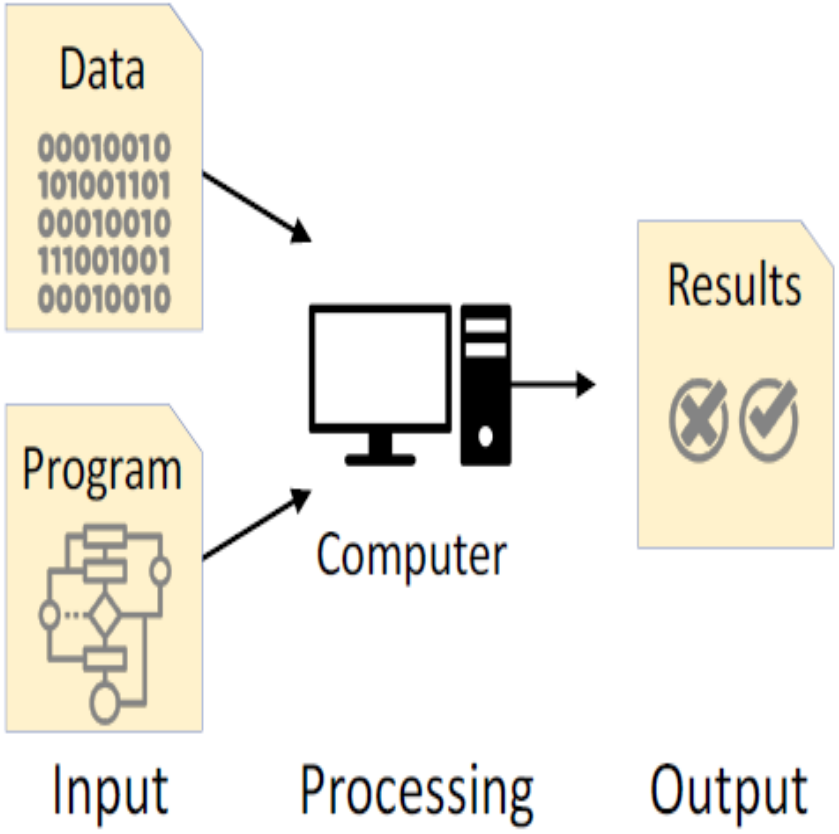
- Facial recognition
- Understanding speech
- Medical diagnosis

Tasks for which we don't have human expertise

- Space exploration
- Undersea manipulation
- Cellular robotics

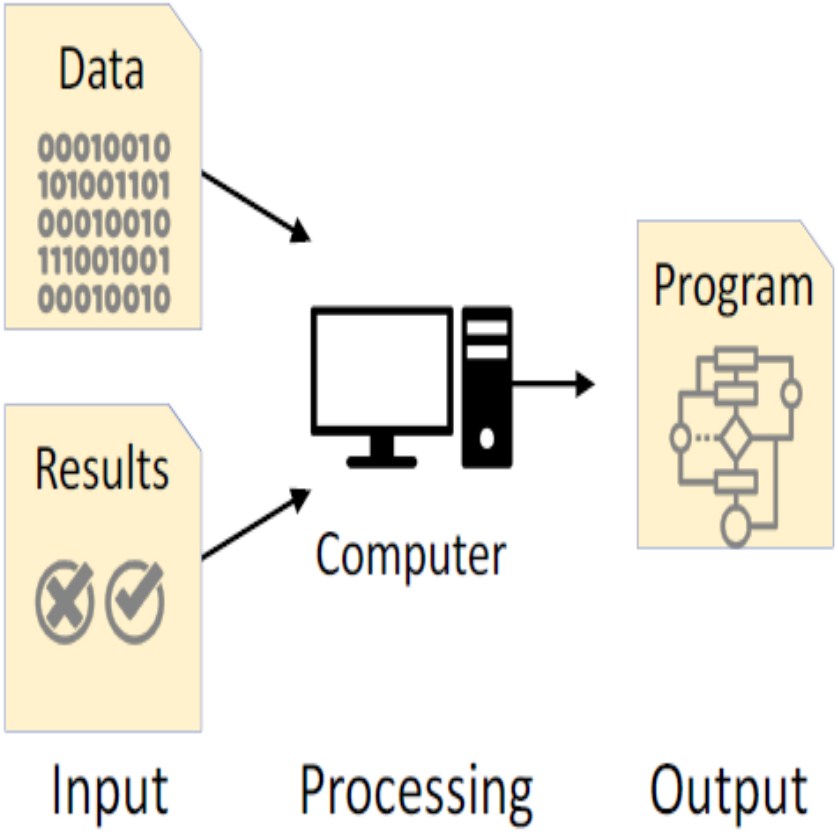
Traditional Programming

Works well when we know how to specify the program



Machine Learning

Needed when we don't know how to specify the program



History of Machine Learning - I

- 1940s, Human reasoning / logic first studied as a formal subject within mathematics (Claude Shannon, Kurt Godel et al).
- 1950s, The Turing Test is proposed: a test for true machine intelligence, expected to be passed by year 2000. Various game-playing programs built. 1956, Dartmouth conference coins the phrase artificial intelligence. 1959, Arthur Samuel wrote a program that learnt to play draughts (checkers if you are American).
- 1960s, A.I. funding increased (mainly military). Famous quote: Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved."
- 1970s, A.I. winter. Funding dries up as people realise it is hard. Limited computing power and dead-end frameworks.
- 1980s, Revival through bio-inspired algorithms: Neural networks, Genetic Algorithms. A.I. promises the world – lots of commercial investment – mostly fails. Rule based expert systems used in medical / legal professions.

History of Machine Learning - II

- 1990s, AI diverges into separate fields: Machine Learning, Computer Vision, Automated Reasoning, Planning systems, Natural Language processing… Machine Learning begins to overlap with statistics / probability theory.
- 2000s, ML merging with statistics continues. Other subfields continue in parallel. First commercial-strength applications: Google, Amazon, computer games, route-finding, credit card fraud detection, etc... Tools adopted as standard by other fields e.g. biology.
- 2010s, deep neural networks have led to significant performance improvement in speech recognition, reinforcement learning, image classification, machine translation, etc..
- Future?

Some links on machine learning history:
https://en.wikipedia.org/wiki/Timeline_of_machine_learning
https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/

Challenge

Recognize dogs in images

What a human sees

DOG

NOT A DOG

What a computer sees

```
01101111 01101110
01100101 01110011
00100000 01100001
01101110 01100100
00100000 01111010
01100101 01110010
01101111 01110011
```

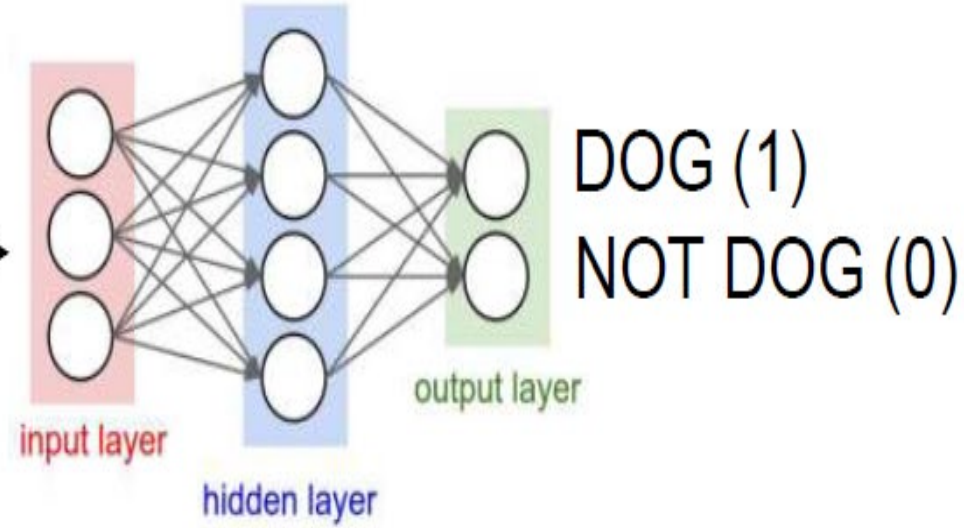
???

```
01101111 01101110
01100101 01110011
00100000 01100001
01101110 01100100
00100000 01111010
01100101 01110010
01101111 01110011
```

???

What do we need?

Training phase

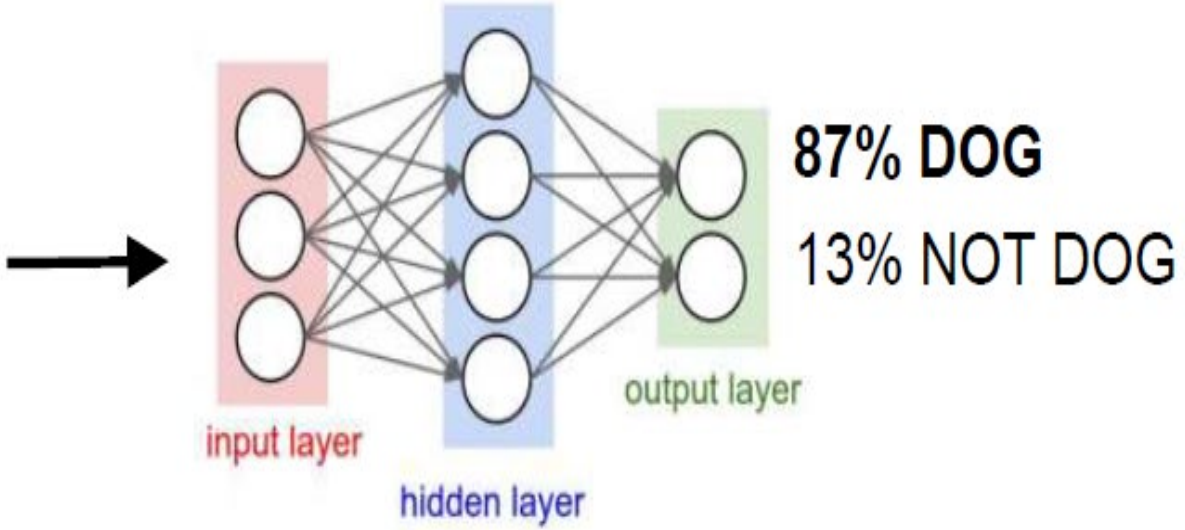


Labeled training set (dog/not dog)
> 1000 images

Untrained Neural Network

Prediction phase

Unlabeled image



Trained Neural Network

Machine learning in A.I.

Speech Recognition

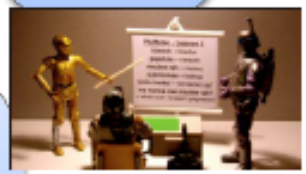
Robotics

Speech Synthesis

Machine Learning

Data Mining, Analysis, Engineering

Natural Language Processing



Computer Vision

Text Mining

- All of these are subfields of Artificial Intelligence (A.I.).
- Machine learning plays a significant role in A.I.

Machine learning Concept/Requirement

Machine Learning: Methodology

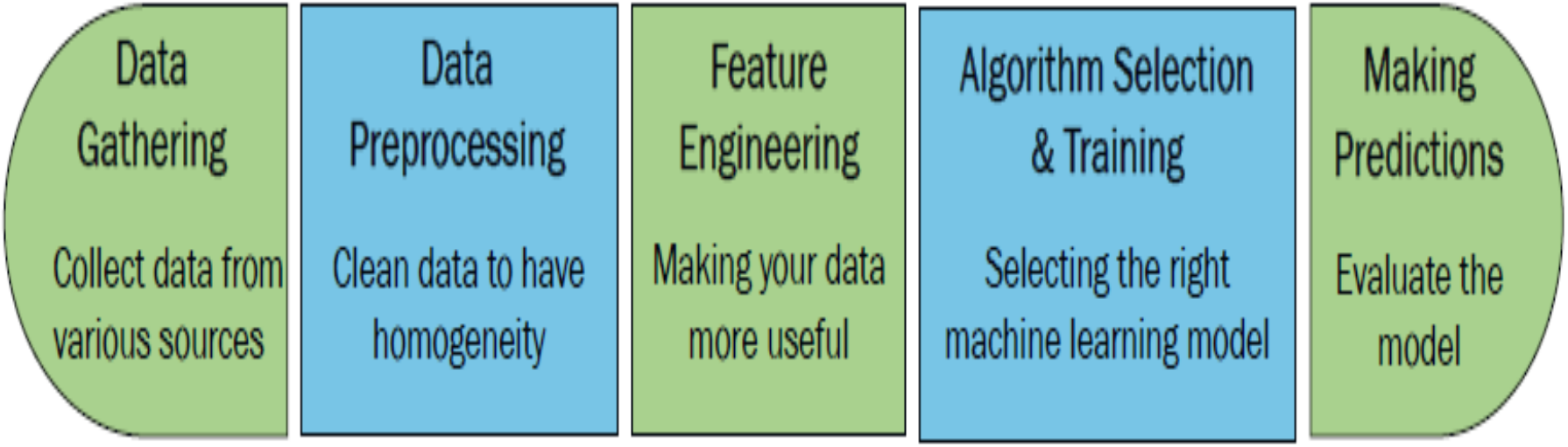
- Basic idea:
  - To represent experiences with **data**.
  - To covert a task to a **parametric model**.
  - To convert the learning quality to an **objective function**.
  - To determine the model through optimising an objective function.

- Machine learning research builds on optimisaton theory, linear algebra, probability theory…

Maths Knowledge Overview

- Linear Algebra:
  - Concepts: vector, matrix, etc.
  - Operations: transpose, sum, multiplication, trace, inverse, etc.
- Calculus:
  - Derivative, partial derivative, gradient, etc.

Steps to Solve a ML/DM Problem



Data Gathering

Might depend on human work

- Manual labelling for supervised learning.
- Domain knowledge. Maybe even experts.

The more the better: Some algorithms need large amounts of data to be useful (e.g., neural networks).

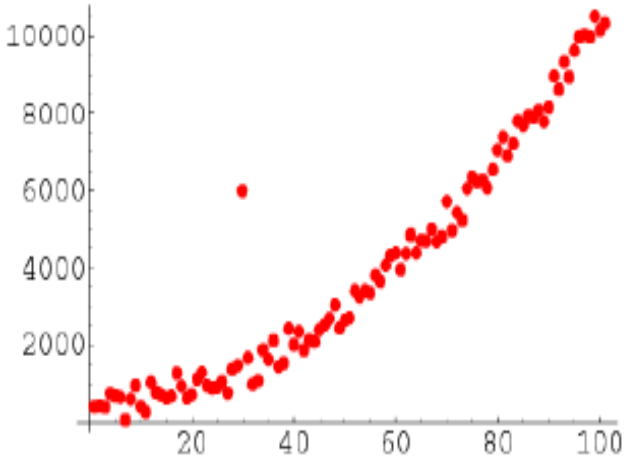
The quantity and quality of data dictate the model accuracy

Data Pre-processing

Is there anything **wrong** with the data?

- Missing values
- Outliers
- Bad encoding (for text)
- Wrongly-labeled examples
- Biased data
  - Do I have many more samples of one class than the rest?

Need to fix/remove data?



Feature Engineering

What is a feature?

A feature is an individual measurable property of a phenomenon being observed

Extract more information from existing data, not adding "new" data by itself

- Making it more useful
- With good features, most algorithms can learn faster

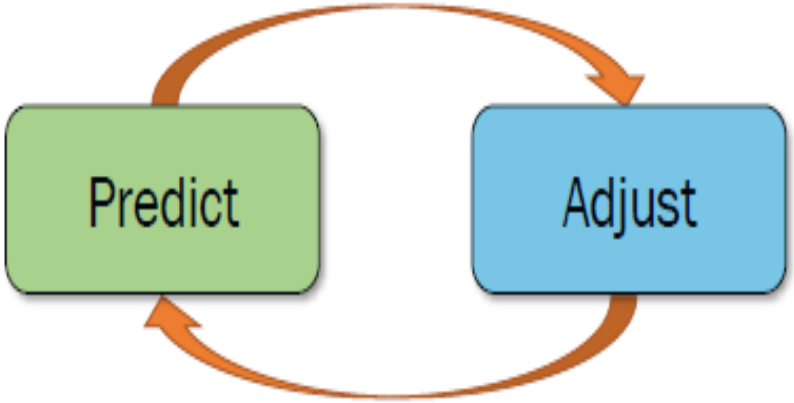
It can be an "art"

- Requires thought and knowledge of the data

Algorithm Selection & Training

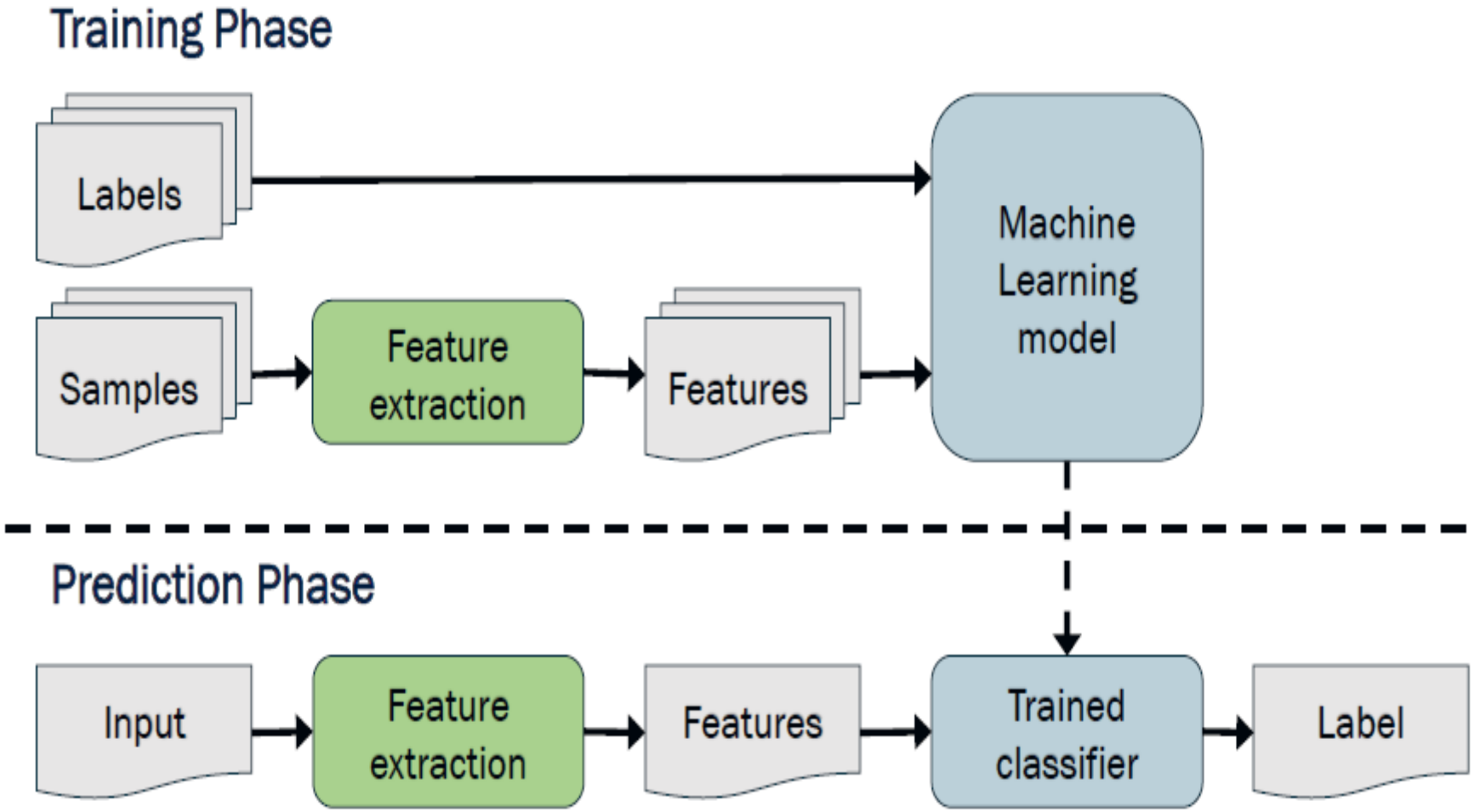
Goal of training: making the correct prediction as often as possible

- Incremental improvement:

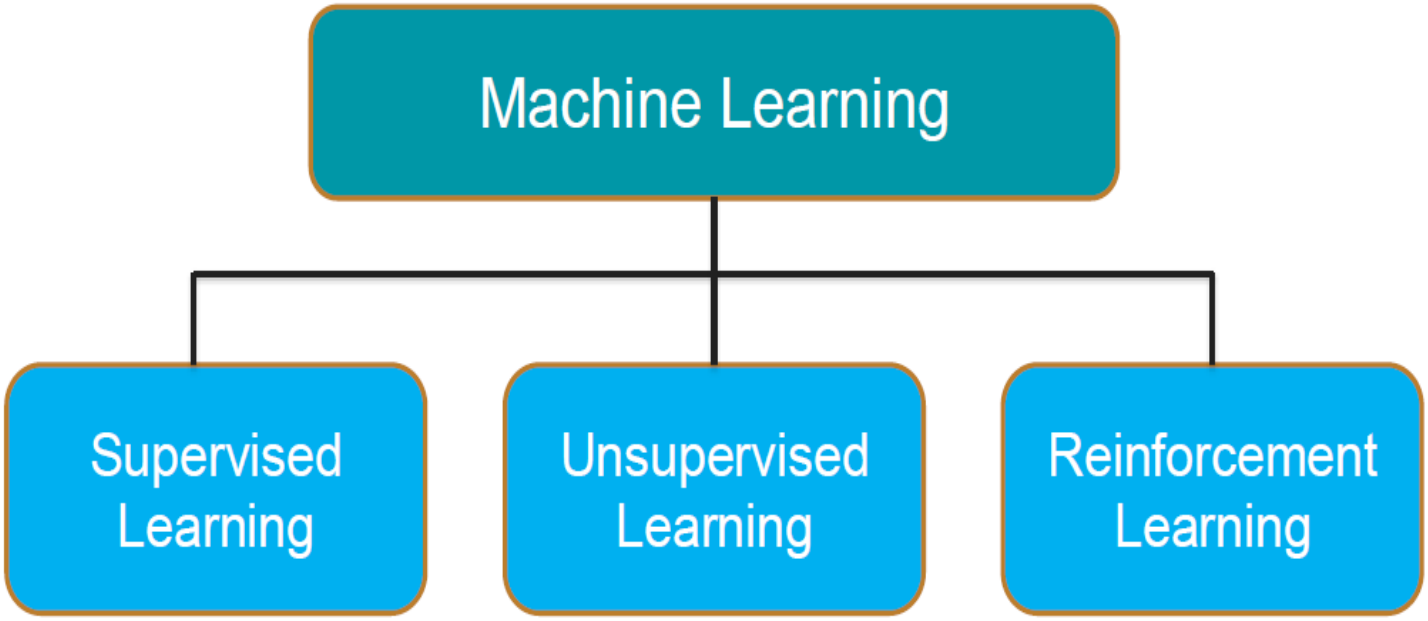


- Use of metrics for **evaluating** performance and comparing solutions
- **Hyperparameter tuning**: more an art than a science

Making Predictions



Types of Learning



Supervised learning

- Given: training data + desired outputs (labels)

Unsupervised learning

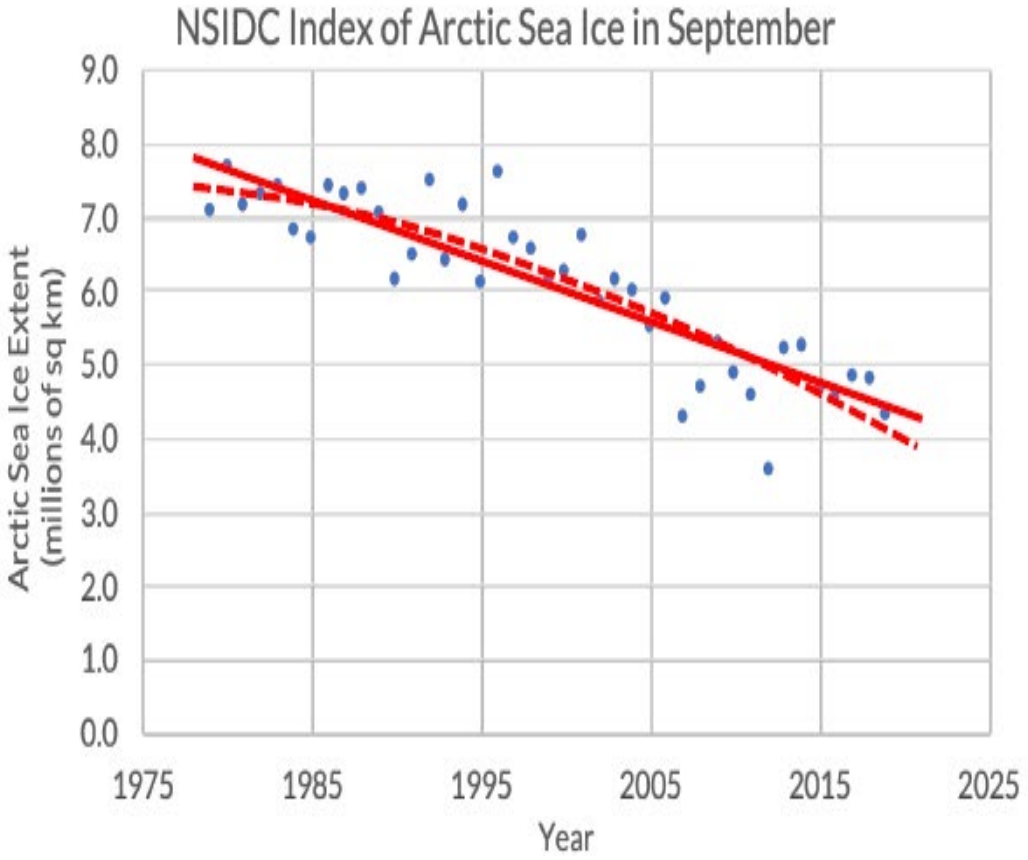
- Given: training data (without desired outputs)

Reinforcement learning

- Rewards from sequence of actions

Supervised Learning: Regression

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is numeric == regression



Supervised Learning: Regression Examples

The target output is a continuous number (or a set of such numbers).

- Finance: x=current market conditions and other possible side information, y=tomorrow's stock market price
- Social Media: x=videos the viewer is watching on YouTube, y=viewer's age
- Robotics: x=control signals sent to motors, y=the 3D location of a robot arm end effector
- Medical Health: x=a number of clinical measurements, y=the amount of prostate specific antigen in the body
- Environment: x=weather data, time, door sensors, etc., y=the temperature at any location inside a building

… … … this list can never end, applications of regression are vast and extremely active!

Example: Stock price prediction

- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

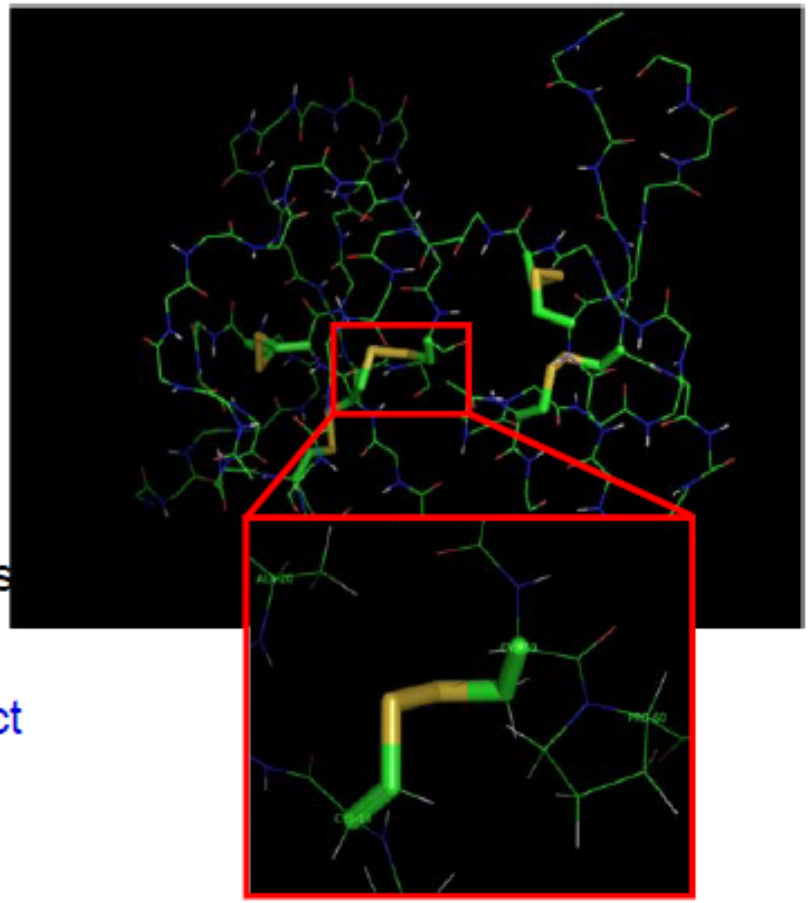
Example: Computational biology

x

AVITGACERDLQCG
KGTCCAVSLWIKSV
RVCTPVGTSGEDCH
PASHKIPFSGQRMH
HTCPCAPNLACVQT
SPKKFKCLSK

Protein Structure and Disulfide Bridges

y



Regression task: given sequence predict 3D structure

Protein: 1IMT

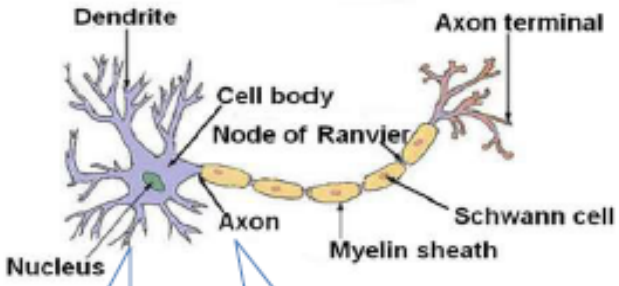
Neuron & Neural Networks

Pedestrian

Car

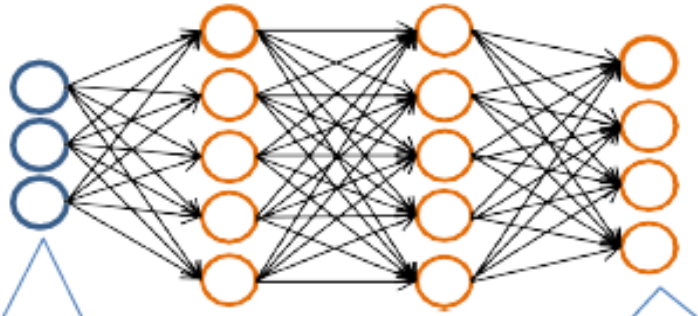
Motorcycle

Truck



Input

Output



Input (Image Pixel)

Output (Judgement)

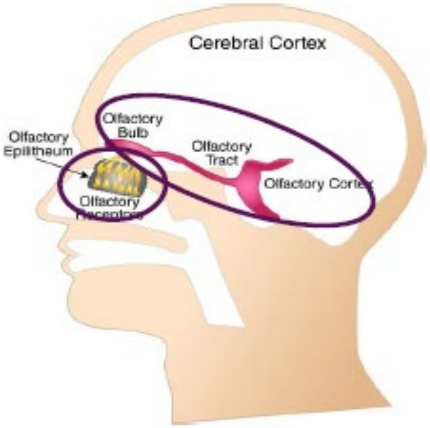
$h_\Theta(x) \in \mathbb{R}^4$

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

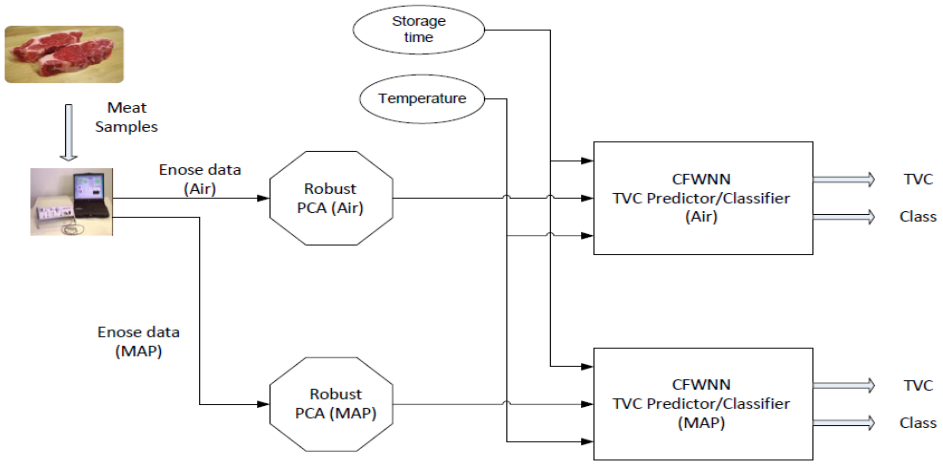
e-Nose: advanced sensing in complex environments

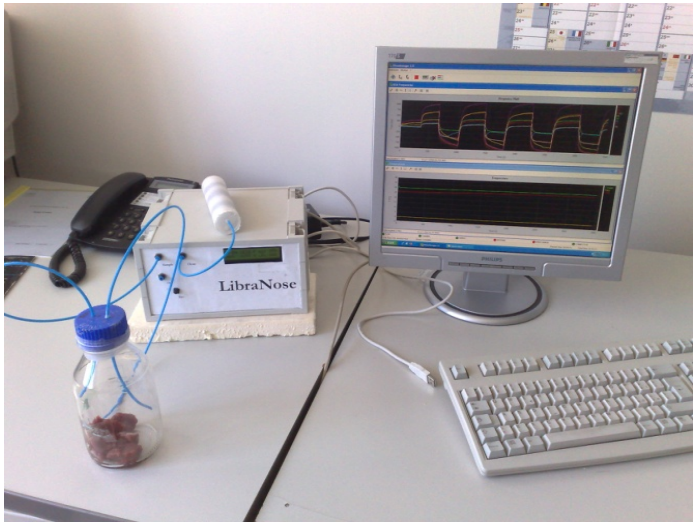


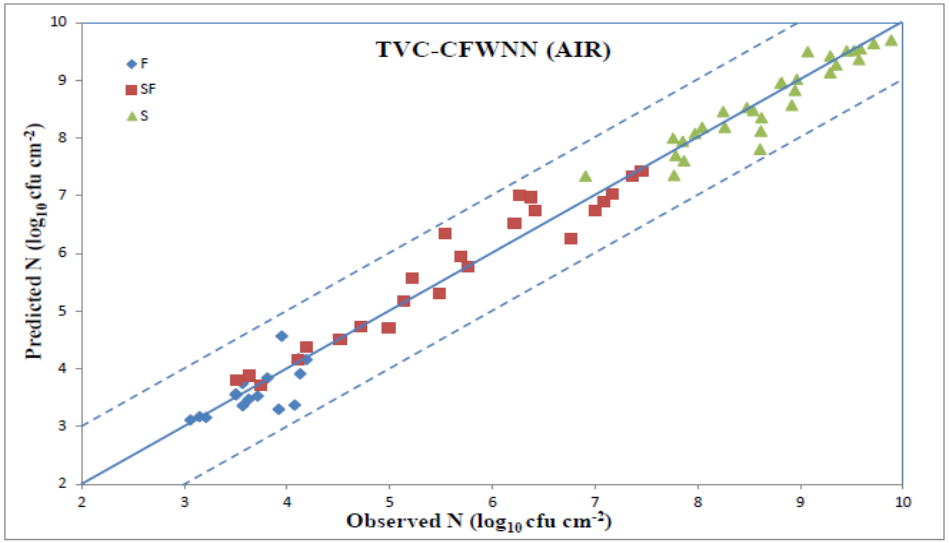
Each and every part of the enose is similar to human nose.

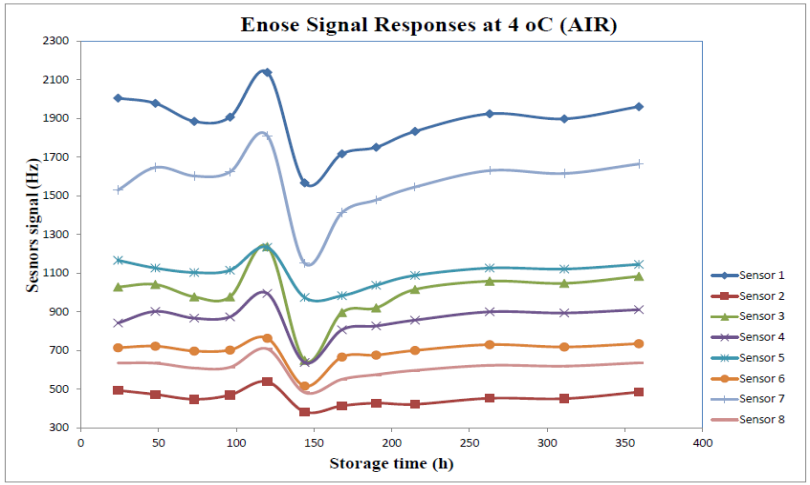
Biological Nose	E-Nose
Inhaling	Pump
Mucus	Filter
Olfactory epithelium	Sensors
Binding with proteins	Interaction
Enzymatic proteins	Reaction
Cell membrane depolarized	Signal
Nerve impulses	Circuitry and neural network

Proposed detection (meat spoilage) system using e-Nose (V. Kontogiannis)









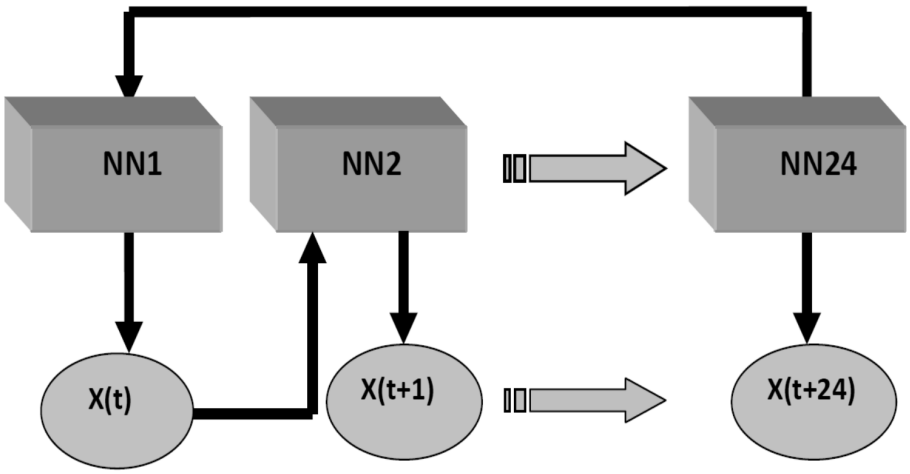
- ***Classify beef samples to one of three quality classes (Fresh, Semi-fresh and spoiled), based on their e-nose signal information***
- ***Predict Total Viable Count (TVC) on meat surface***

NEURAL NETWORK FOR LOAD FORECASTING IN SMART GRID

To achieve optimization of power configuration and energy saving, a large amount of new technologies are applied to the power system in the smart grid. Load forecasting is must as far as planning and operation of a power system is concerned.

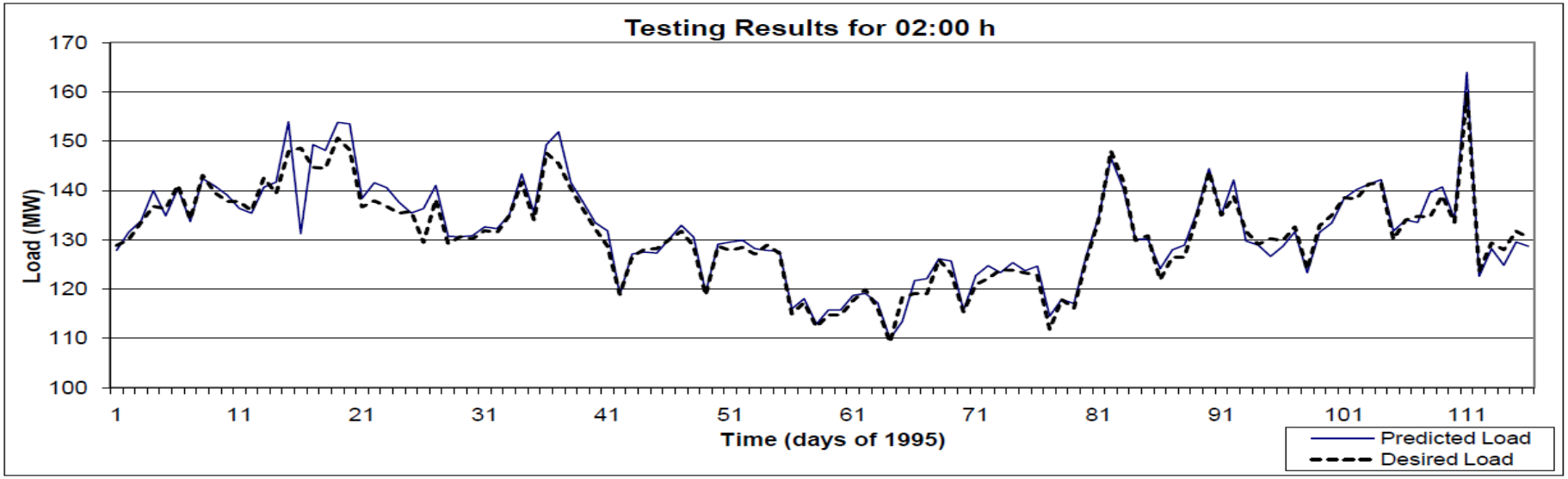
To give the exact information about the power purchasing and generation, high accuracy of load forecasting is required.

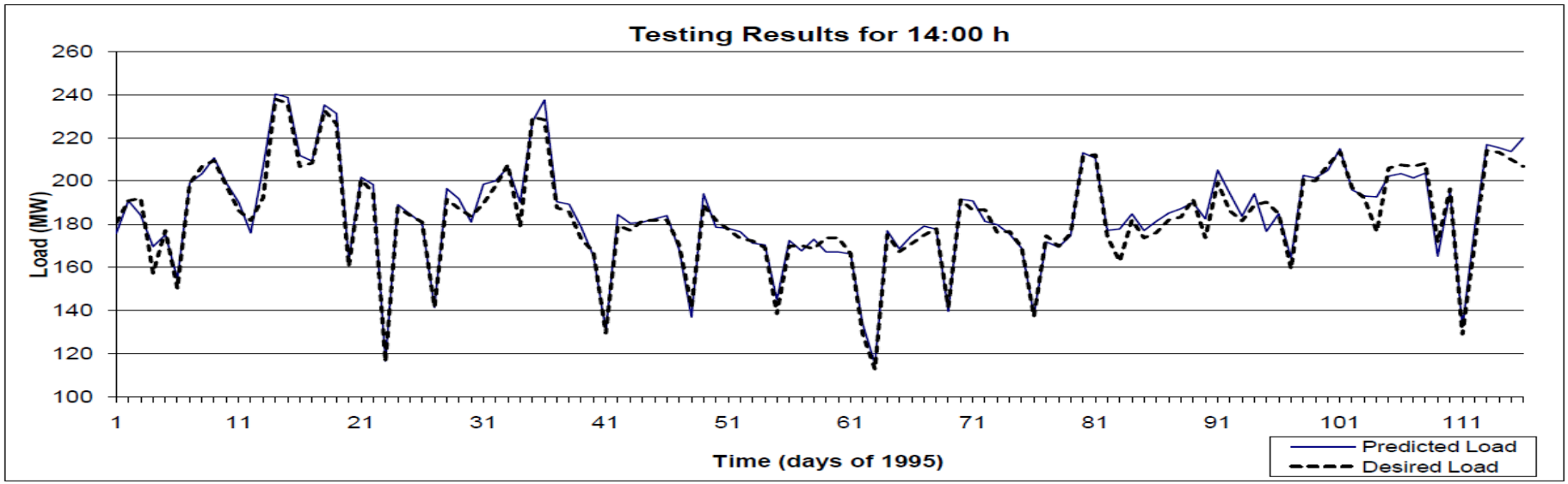
Factors such as season differences, climate changes, weekends and holidays, disasters and political reasons, operation scenarios of the power plants and faults occurring on the network lead to changes of the load demand and generations. Owing to the transcendent characteristics, **NNs is one of the most competent methods to do the practical works like load forecasting.**



An intelligent-based system has been implemented as a STLF models on the power system of the island of Crete. Public Power Corporation of Greece is the main provider of the electric power in Crete.

Energy Load Forecasting (research by V. Kontogiannis)

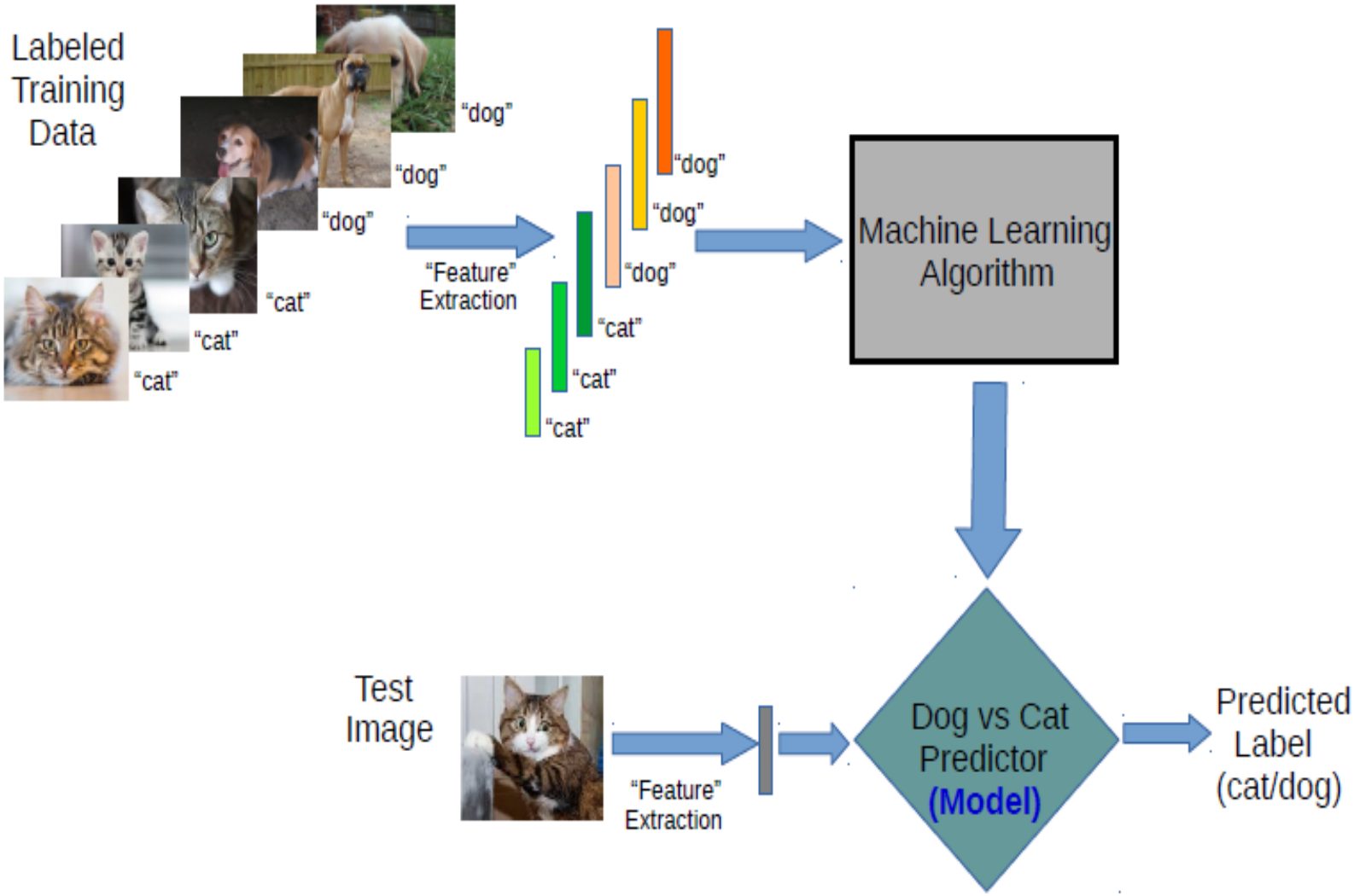




Supervised Learning: Classification

- 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.

Supervised Learning Procedure (Classification)



Supervised Learning: Classification Examples

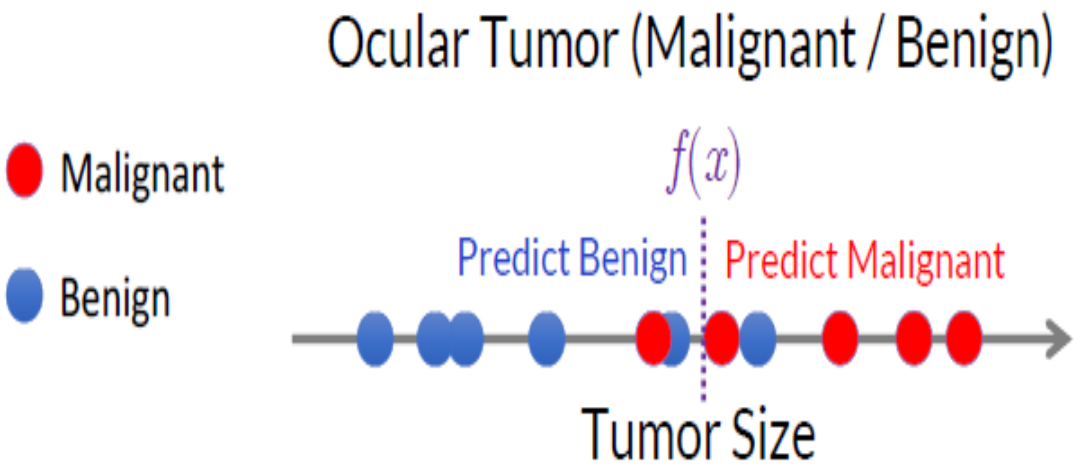
The target output is a category label.

- Medical diagnosis: x=patient data, y=positive/negative of some pathology
- Optical character recognition: x=pixel values and writing curves, y='A', 'B', 'C', …
- Image analysis: x=image pixel features, y=scene/objects contained in image
- Weather: x=current & previous conditions per location, y=tomorrow's weather

… … … this list can never end, applications of classification are vast and extremely active!

Supervised Learning: Classification

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification

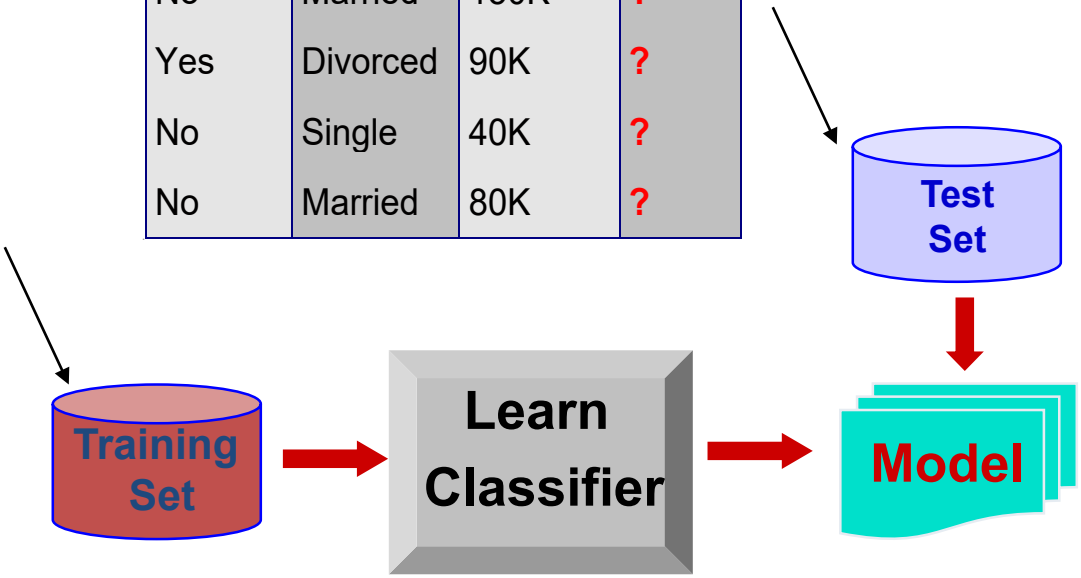


Classification Applications

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	?

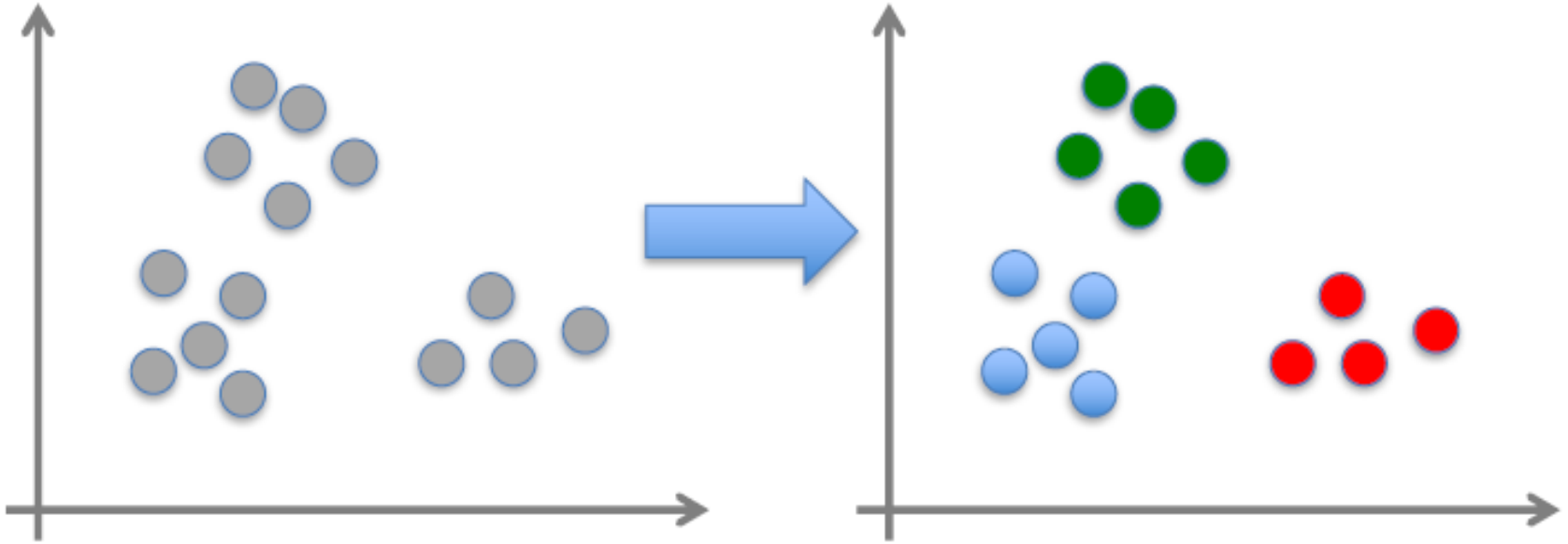


Classification: Applications

- Application area: Fraud Detection
- Goal: Recognize fraudulent cases in credit card transactions
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes
    - When and where does a customer buy? What does he buy?
    - How often he pays on time? etc.
  - Label past transactions as *fraud* or *fair* transactions
    This forms the *class attribute*
  - Learn a model for the class of the transaction
  - Use this model to detect fraud by observing credit card transactions on an account

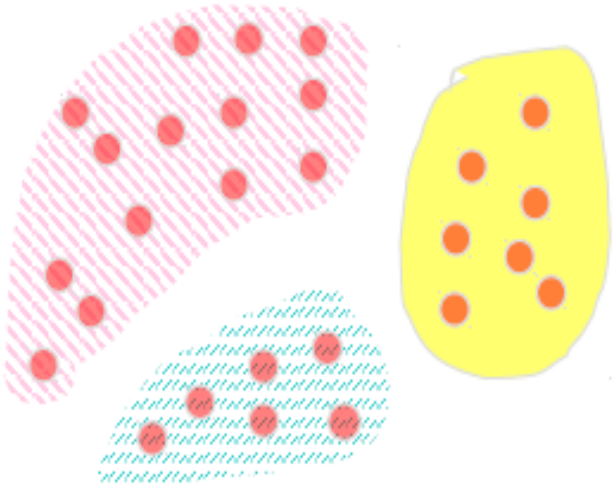
Unsupervised Learning

- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
  - E.g., clustering

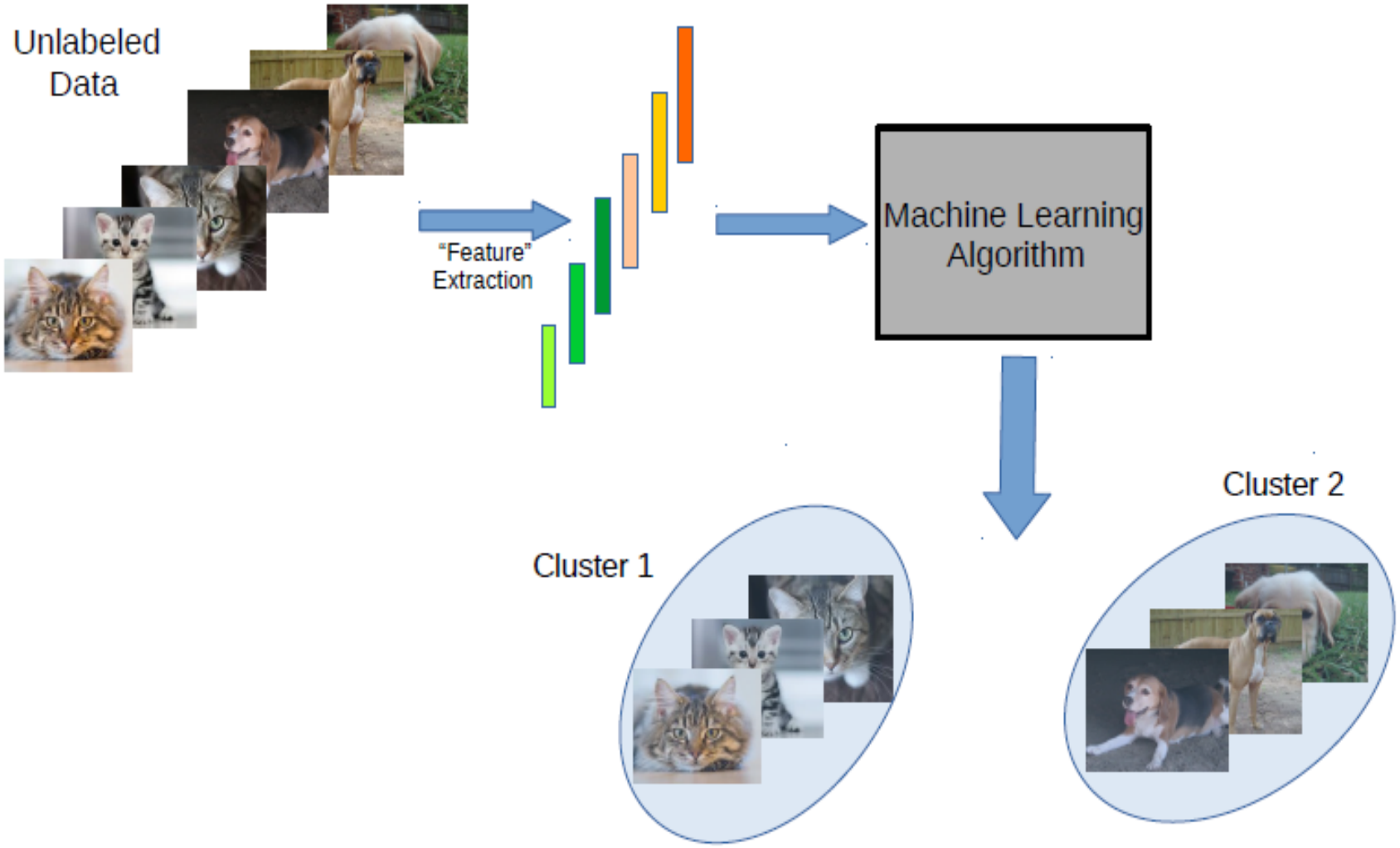


Unsupervised Learning: Clustering

- Given a set of data points, and a similarity measure among them, find clusters such that
  - Data points in one cluster are similar to one another
  - Data points in separate clusters are different from each other
- Result
  - a descriptive grouping of data points

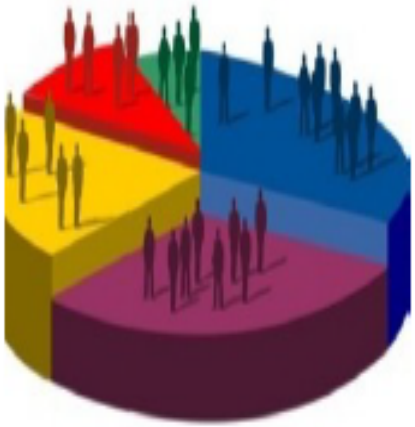


Unsupervised Learning Procedure (Clustering)

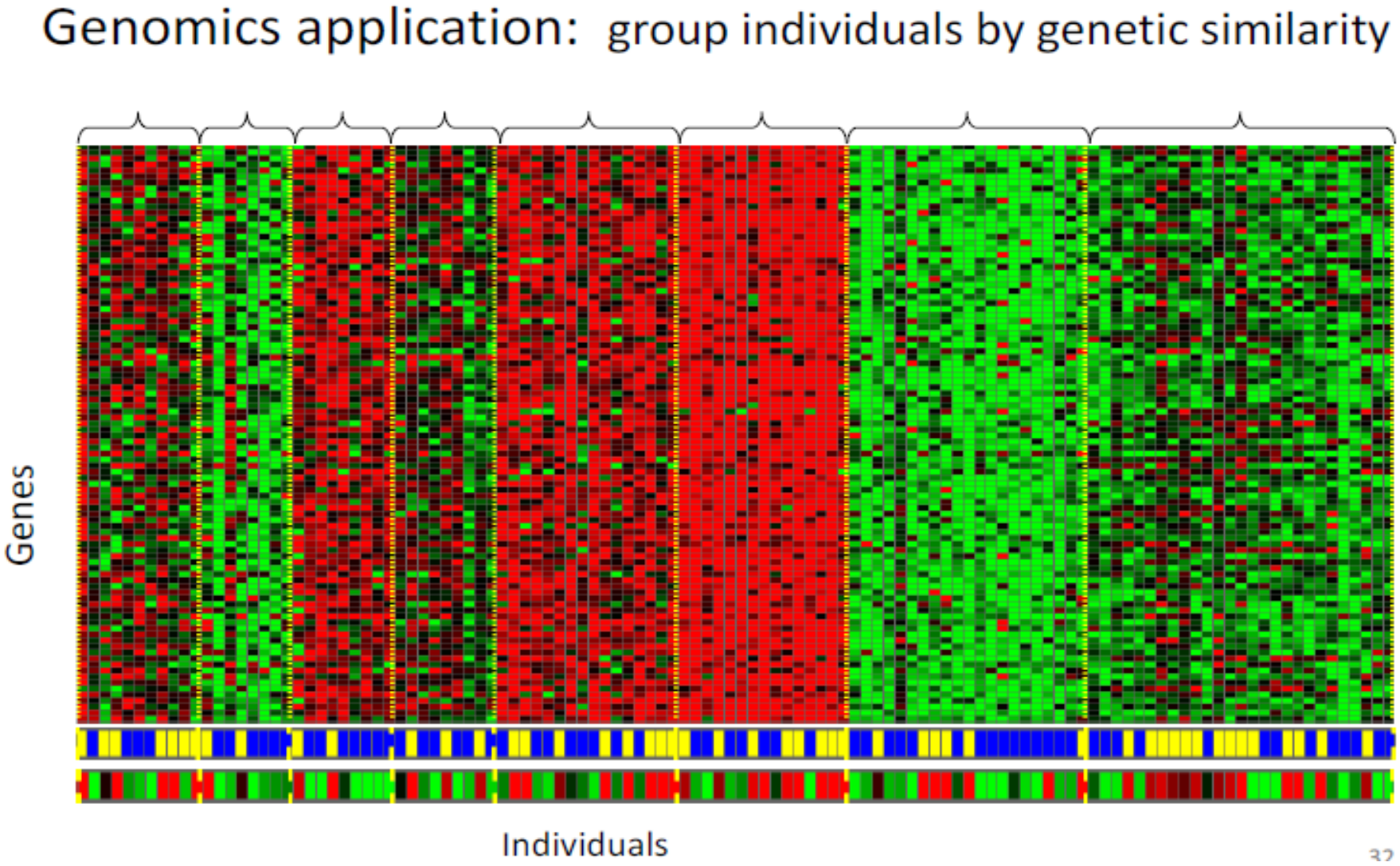


Clustering: Applications

- Application area: Market segmentation
- Goal: Subdivide a market into distinct subsets of customers
  - where any subset may be conceived as a marketing target to be reached with a distinct marketing mix
- Approach:
  - Collect information about customers
  - Find clusters of similar customers
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters



Unsupervised Learning: Clustering Example



Clustering: Applications

Application area: Document Clustering

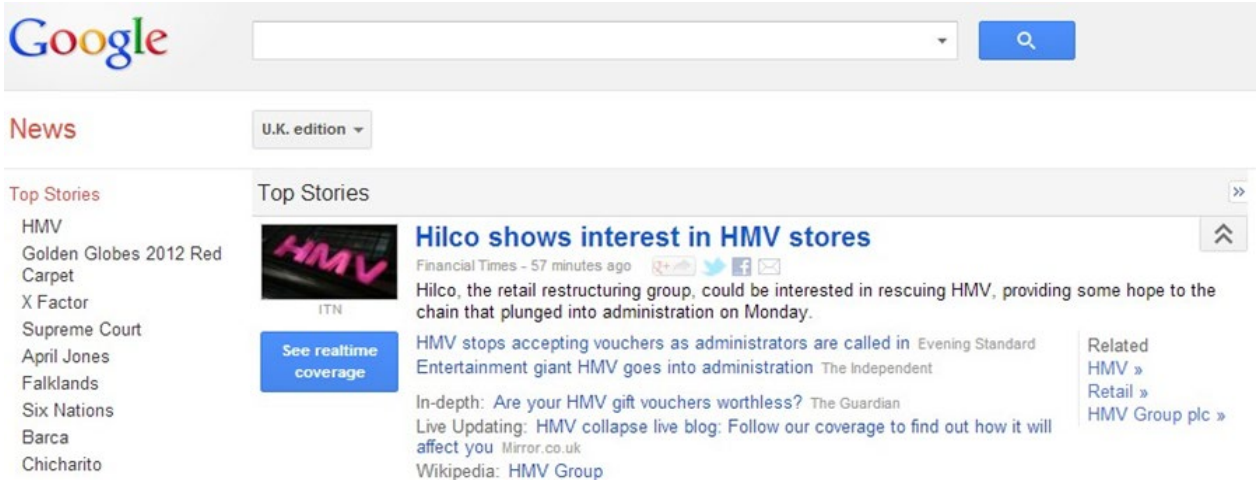
Goal: Find groups of documents that are similar to each other based on the important terms appearing in them

Approach

–Identify frequently occurring terms in each document

–Define a similarity measure based on the frequencies of different terms

Application Example: Grouping of stories in Google News



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}

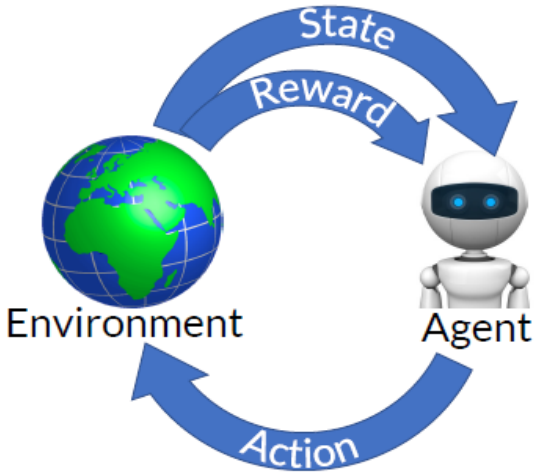
Association Rule Discovery: Applications

- Application area: Marketing and Sales Promotion
- Example rule discovered:

  {Bagels, Coke} --> {Potato Chips}
- Insights:
  - promote bagels to boost potato chips sales
  - if selling bagels is discontinued, this will affect potato chips sales
  - coke should be sold together with bagels to boost potato chips sales

Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Examples:
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand



Reinforcement Learning

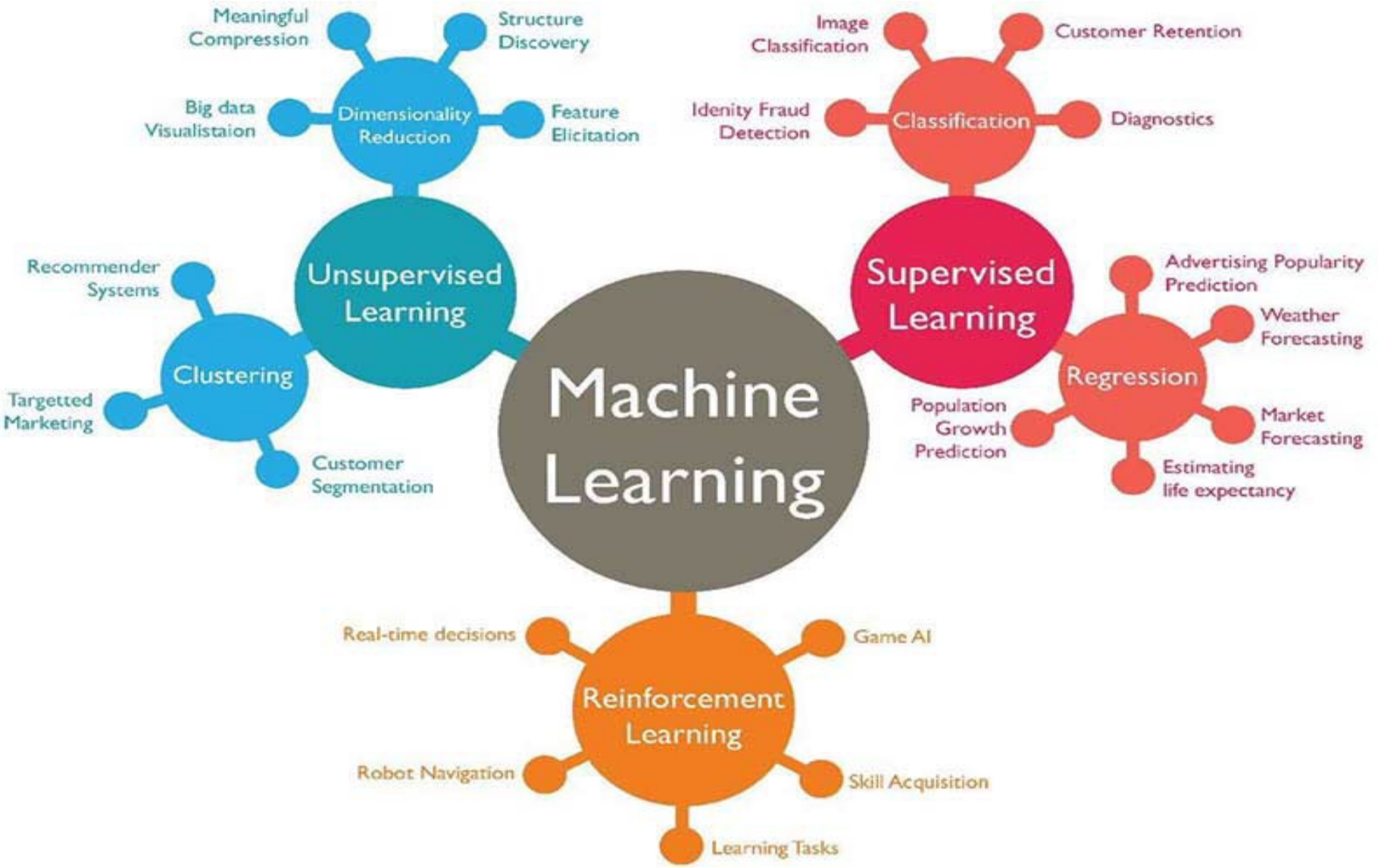
- Learn policy from user demonstrations

Stanford Autonomous Helicopter

http://heli.stanford.edu/

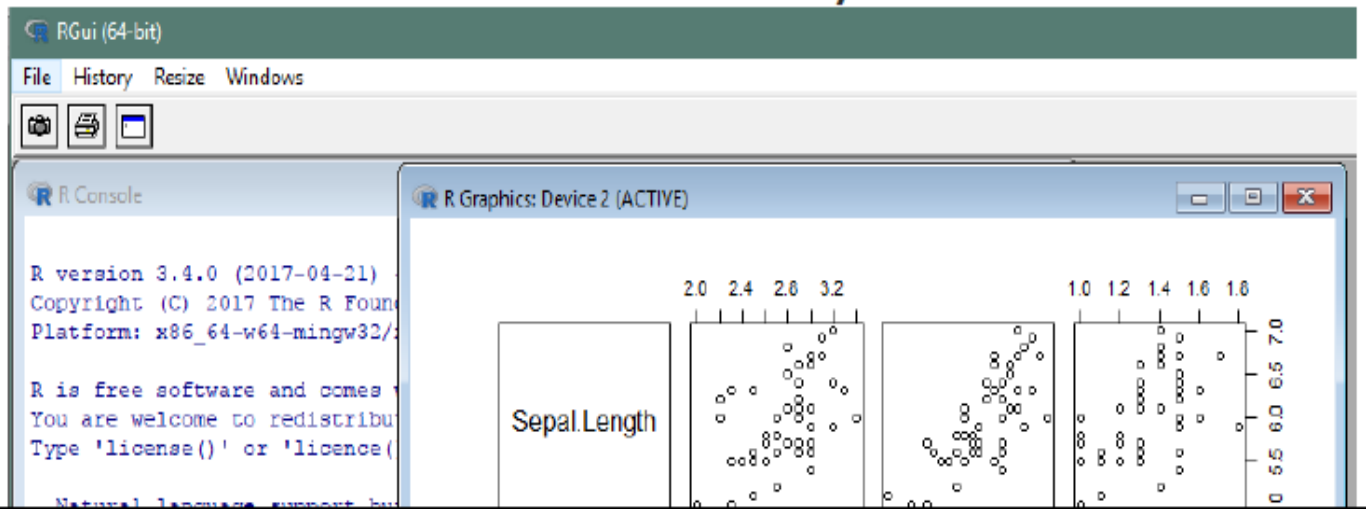
https://www.youtube.com/watch?v=VCdxqn0fcnE

Applications Overview



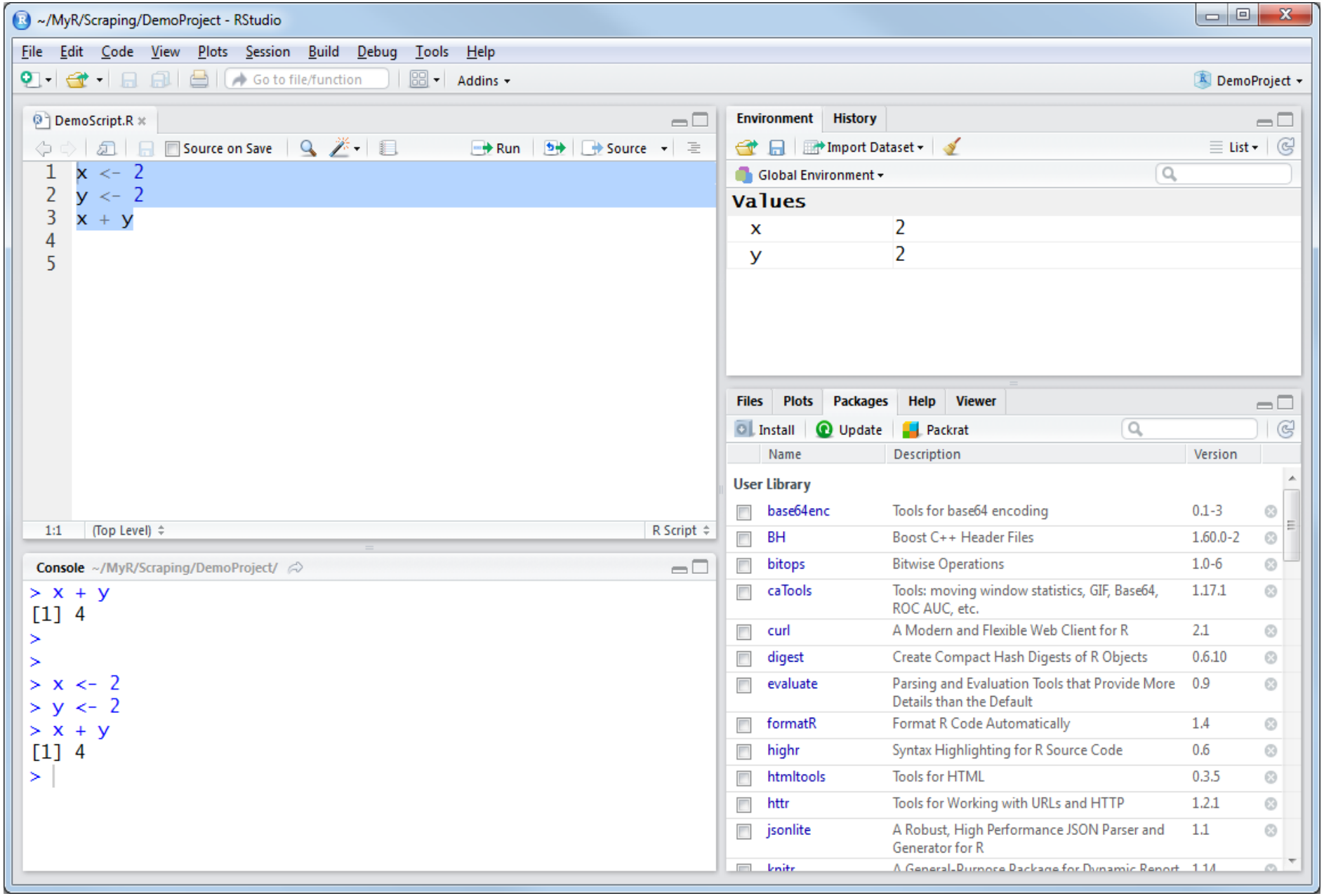
INTRODUCTION TO R

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



ikipedia

29


~/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

For this module (5DATA002W machine Learning & Data Mining), the main programming tool will be the R language.

R is a free-distributed software and can be downloaded from: *https://cran.r-project.org/*. Versions for Windows, Mac and Linux are available. If you wish to download/install in your laptop make sure that you download any version > 4.0. 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://www.rstudio.com/products/rstudio/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.

R language Computers at Labs are fully equipped with both R and RStudio tools. RStudio is also available on AppsAnywhere (from university website).

https://support.ecs.westminster.ac.uk/w/index.php/AppsAnywhere#Lockdown

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