

Bachelor's degree in Bioinformatics

Introduction to Data Mining and Machine Learning

Prof. Renato Bruni

bruni@dis.uniroma1.it

*Department of Computer, Control, and Management Engineering (DIAG)
“Sapienza” University of Rome*

What is Machine Learning ?

Some definitions

1959: Arthur Samuel “*programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning*”

1997: Tom Mitchell “Machine Learning is the study of computer algorithms that *improve automatically* through experience”

or, more precisely

“We say that a machine *learns* with respect to a particular task *T*, performance metric *P*, and type of experience *E*, if the system reliably *improves its performance P* at task *T*, following experience *E*”

An example: Spam Detection

An e-mail filter able to decide which mail should be classified as «spam» or «not spam» learning from your decisions on past emails

- **T (task)** classify mail as «spam» or «not spam»
- **P (performance measure)** the percentage of correctly classified mails
- **E (experience)** your e-mail classification as «spam» or «not spam»

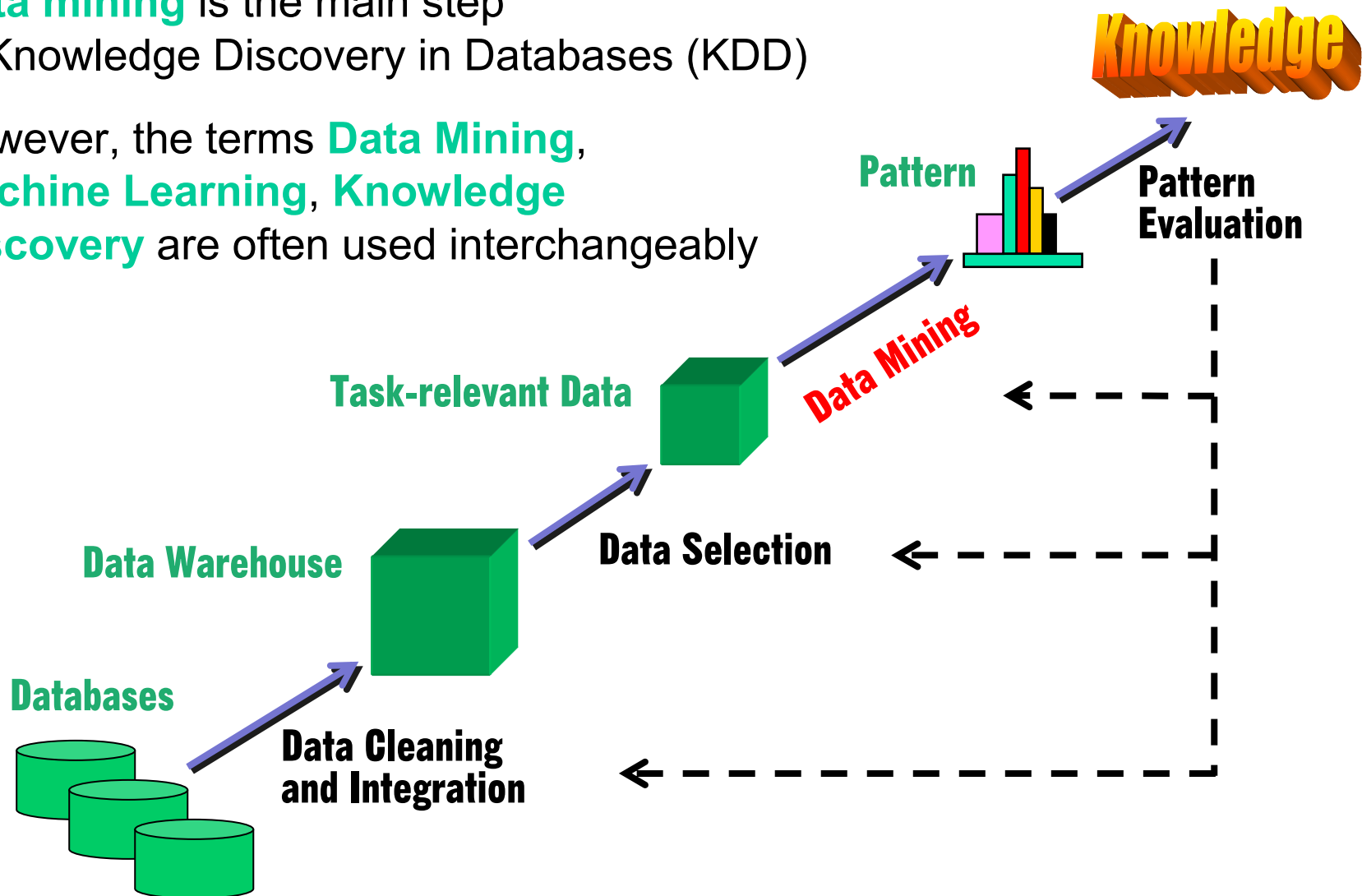
What is Data Mining ?

- **Information explosion** (a.k.a. data flood) is the rapid increase in the amount of data produced and stored
- **A circle**: technology improvements allow to use more data → using even more data becomes **necessary** → this requires further technological improvements
- We are drowning in data, but starving for **knowledge**!
Managing those data becomes more and more difficult. We need effective techniques, or we risk an information overload
- **Data mining**:
Extraction from large data sets of information that is **not obvious**, **not immediately available** and **potentially useful** (rules, regularities, patterns, etc. = knowledge) using automatic or semi-automatic methods

Where is Data Mining ?

Data mining is the main step of Knowledge Discovery in Databases (KDD)

However, the terms **Data Mining**, **Machine Learning**, **Knowledge Discovery** are often used interchangeably

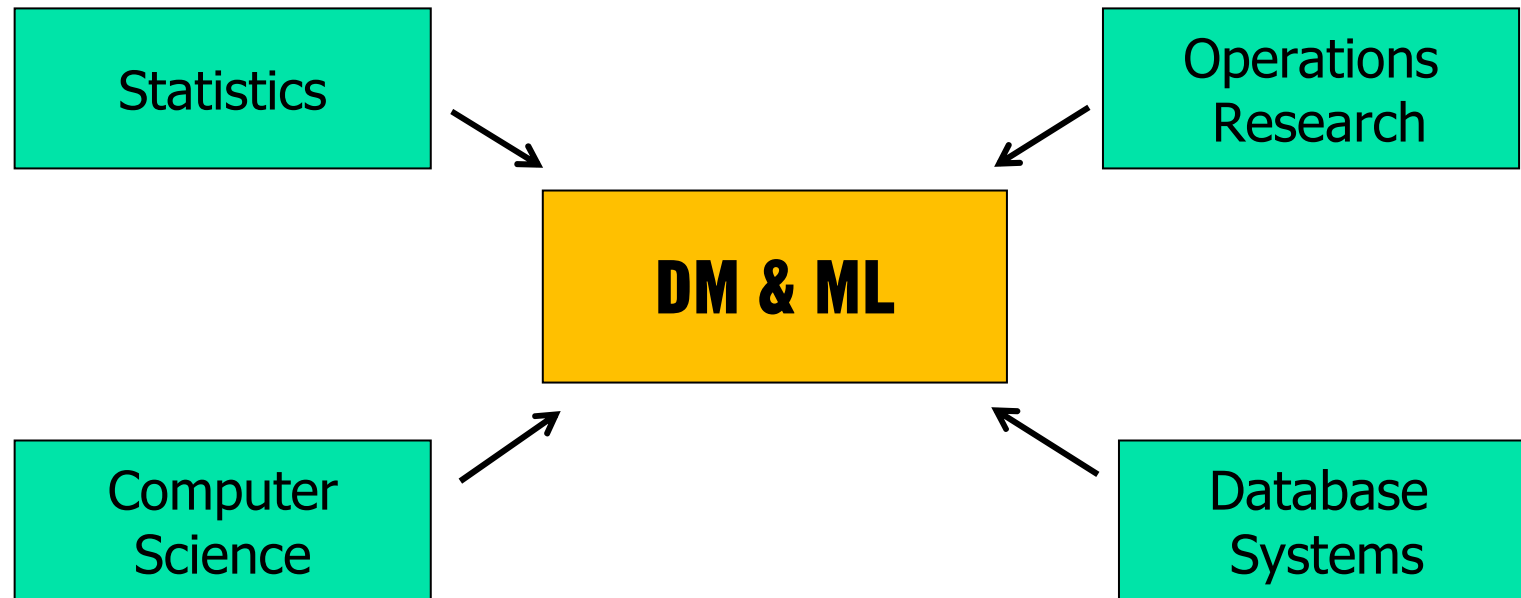


Who needs Data Mining ?

Obtaining **knowledge** and not just **data** is essential in many applications. Some examples:

- Database Analysis (Rules extraction, Associations)
- Market Analysis (Customer profiling, Marketing)
- Risk Analysis (Finance planning, Investments)
- Fraud Detection (Credit cards, Food adulteration)
- Decision Support (Resource management, Allocation)
- Medical Analysis (Diagnosis, Donors management)
- Text mining (Search engines, Anti spam)
- Analysis of Economical or Social Policies (Rule learning)
- ...

What do I need for ML and DM ?



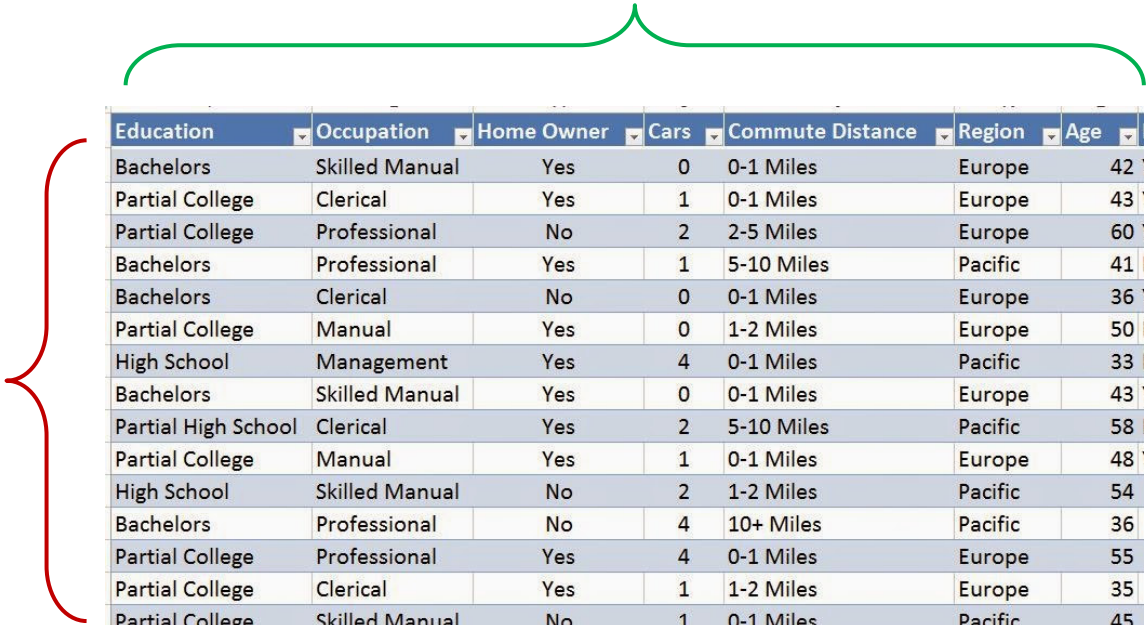
- Several different competences are required to do Machine Learning and Data Mining
- It is a very **interdisciplinary Area**
- For this reason, many things are called with **different names** in the different communities

What exactly is Data ?

- A collection of **objects**, each having some **attributes**
- Each object is usually stored in a record
- An attribute is a property or characteristic of an object
Examples: name, eye color, income, etc.

Attributes, a.k.a. fields, features, variables, columns, ...

Objects, a.k.a. records, tuples, instances, observations, points, samples, rows, ...



Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age
Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	42
Partial College	Clerical	Yes	1	0-1 Miles	Europe	43
Partial College	Professional	No	2	2-5 Miles	Europe	60
Bachelors	Professional	Yes	1	5-10 Miles	Pacific	41
Bachelors	Clerical	No	0	0-1 Miles	Europe	36
Partial College	Manual	Yes	0	1-2 Miles	Europe	50
High School	Management	Yes	4	0-1 Miles	Pacific	33
Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	43
Partial High School	Clerical	Yes	2	5-10 Miles	Pacific	58
Partial College	Manual	Yes	1	0-1 Miles	Europe	48
High School	Skilled Manual	No	2	1-2 Miles	Pacific	54
Bachelors	Professional	No	4	10+ Miles	Pacific	36
Partial College	Professional	Yes	4	0-1 Miles	Europe	55
Partial College	Clerical	Yes	1	1-2 Miles	Europe	35
Partial College	Skilled Manual	No	1	0-1 Miles	Pacific	45

Data Units or Records

A record scheme is a set of **fields** $R = \{f_1 \dots f_m\}$

A record instance is a set of **values** $r = \{v_1 \dots v_m\}$

Each field f_i has its **domain** D_i that is the set of all possible values

Example: fields can be *age*, *marital status*, corresponding values can be *18*, *single*, etc.

Fields can be: $\left\{ \begin{array}{ll} \blacksquare \text{ numerical or quantitative} & \left\{ \begin{array}{l} \blacksquare \text{ continuous: real-valued} \\ \blacksquare \text{ discrete: integer or binary} \end{array} \right. \\ \blacksquare \text{ categorical or qualitative} & \left\{ \begin{array}{l} \blacksquare \text{ ordered (e.g. first, second)} \\ \blacksquare \text{ not ordered (e.g. red, blue)} \end{array} \right. \end{array} \right.$

Fields can be re-encoded differently. For example, many procedures convert **each field** f_i into one or more **binary ones**, that we will call **binary attributes** $a_i^j \in \{0,1\}$

Different Tasks in Data Mining

Depending on the application, different **activities** may be required. However boundaries are not sharp at all

- **Classification**: learning a function or a criterion to map objects on a pre-defined set of classes
- **Regression**: learning a function or a criterion to assign each object a real value
- **Clustering**: identification of a partition of the set of objects to group together similar objects
- **Learning of Dependencies and Associations**: identification of significant relationships among data attributes
- **Rule Learning or Summarization**: identification of a compact description of a set or subset of data

Learning Paradigms

Supervised learning: the “correct answer” (**label**) on the instances is available (at least for some of them).

We learn from the labeled data (=correct answers) to predict labels (=new correct answers) for unseen instances

Unsupervised learning: no “correct answers” available.

We use the data but the corresponding output values are not known in advance. Example: one wants to find similarity classes and to assign instances to the correct class

Very often, labeled data are scarce, but unlabeled data are easy to collect. **Semi-supervised learning**: techniques that learn from small amount of labeled data and also from large amount of unlabeled data

Learning Process

In many learning tasks, data are partitioned into:

■ **Training set** (data+labels, or just data for unsupervised): used to learn

- incrementally (on-line learning): Data are obtained incrementally during the training process
- batch (off-line) learning: Data of the training set are available in advance before entering the training process

■ **Validation set** (data): after the learning phase, we may need other data to tune parameters etc.

■ **Test set** (data): used for doing what we must do (if we know also the labels, we can compute the accuracy)

We deal with **large data sets** and possibly small training sets (e.g. rare events, not controllable events). Labeled data may be costly.

CLASSIFICATION EX.: FRAUD DETECTION

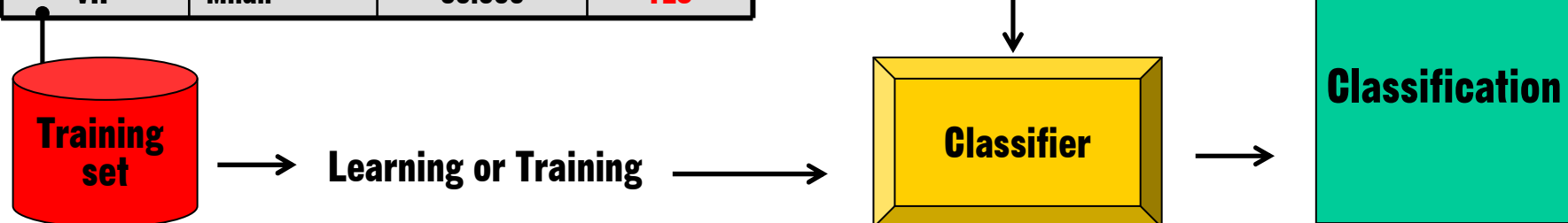
- Given a **training set** partitioned in classes, **predict** the **class** of new data, i.e., learn a classifier

Numerical or categorical

Class

customer type	Town	Income	Fraud
A	Rome	25.000	NO
B	Milan	15.000	YES
X	Florence	18.000	NO
VIP	Rome	45.000	NO
B	Neaples	20.000	NO
A	Bologna	16.000	YES
A	Turin	50.000	NO
X	Venice	28.000	NO
VIP	Milan	30.000	YES

Customer type	Town	Income	Fraud
X	Milan	30.000	?
A	Turin	22.000	?
VIP	Florence	18.000	?
A	Rome	14.000	?
B	Milan	55.000	?
X	Bari	26.000	?
A	Lecco		



REGRESSION EXAMPLE: PREDICT SALES

Independent variables (predictors)

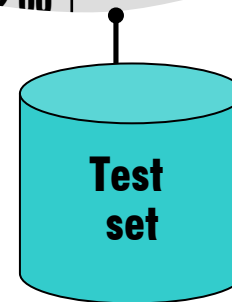
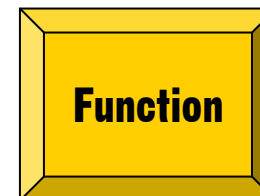
Dependent variable (numerical)

Cost	Price	Usage	Sales
5,00	11,50	Frequent	154
6,00	12,80	Rare	21
15,50	25,50	Frequent	234
15,50	33,95	Occasional	44
1,00	1,50	Frequent	79
13,50	20,50	Occasional	355
8,50	12,90	Frequent	988
19,00	35,90	Frequent	57
12,90	26,90	Rare	3

Cost	Price	Usage	Sales
10,00	19,90	Frequent	?
5,50	11,00	Occasional	?
14,50	25,90	Occasional	?
63,00	128,00	Rare	?
2,50	4,90	Frequent	?
24,00	49,90	Occasional	?
12,00	22,00	Frequent	?



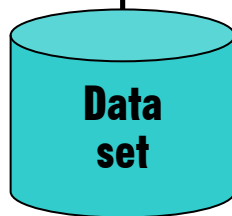
Definition of the model (linear, etc.) → Parameters learning



CLUSTERING EX.: MARKET SEGMENTATION

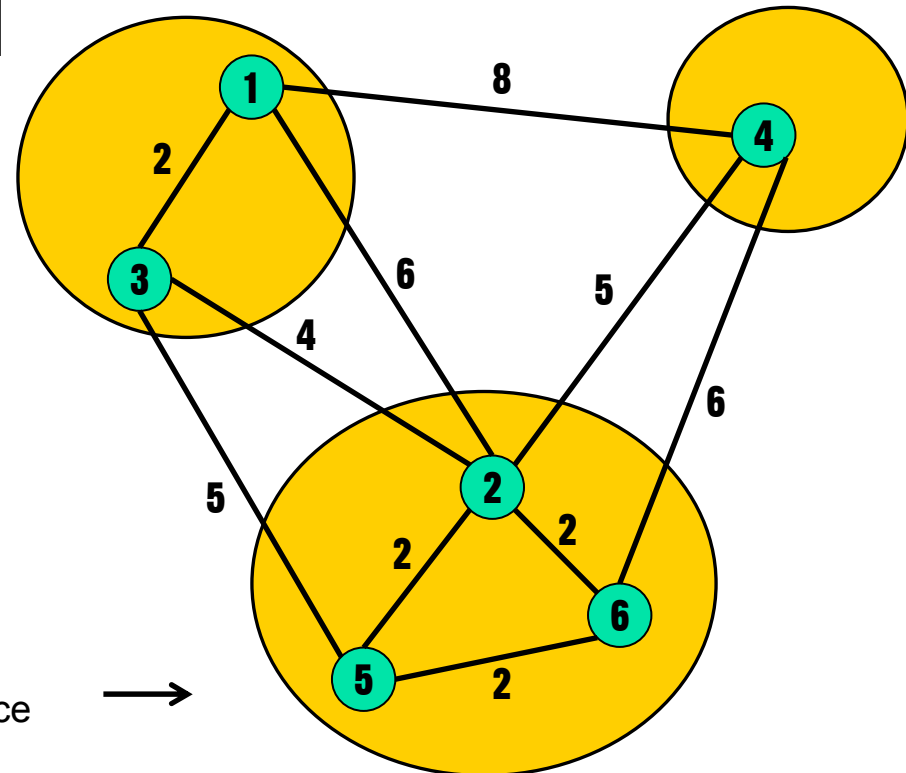
ID Cust.	Town	Income	Marital status	Revenue
1	Milan	21.470	unmarried	2.500
2	Rome	12.500	unmarried	400
3	Turin	63.600	Divorced	250
4	Neaples	21.900	married	12.000
5	Milan	20.300	married	645
6	Rome	40.500		

Given the data, partition all customers in $k=3$ groups that should be treated differently



Definition of a distance criterion
and computation of distances

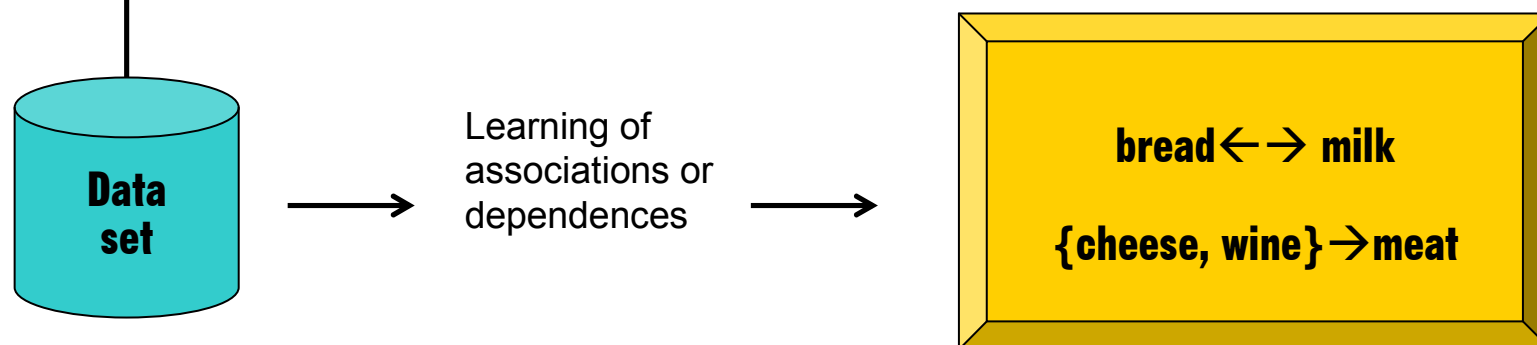
Partition in k groups minimizing intra-group
distance or maximizing group-to-group distance



ASSOCIATION EXAMPLE: FOOD SHOPPING

ID	Oggetti Acquistati
1	bread, milk, eggs
2	vegetables, cookies, juice, pasta
3	meat, cheese, cookies, wine
4	bread, cheese, milk
5	bread, wine, meat, vegetables, milk, cheese
6	pasta, juice, eggs

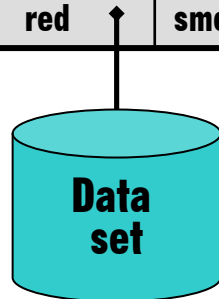
Each record contains a variable number of objects from a list of foods. Find dependences or associations in the records, so as to predict what a customer still has to buy and help him\her (the more we sell, the better)



RULE EXTRACTION: POIS. MUSHROOMS

Categorical	Categorical	Numerical	Numerical
Colour	Skin	Diameter	Heigh
red	granular	13	5
white	smooth	4	7
grey	granular	10	8
grey	smooth	6	12
red	granular	10	10
white	granular	5	9
grey	smooth	6	10
white	granular	3	6
red	smooth	10	16

Given the description of many poisonous mushrooms, find a compact description (an intensive description) of this set



Learning of the properties
of the records

**(NOT white AND granular) OR
(heigh/diameter > 1.5)**

“If you torture the data long enough, it will always confess”

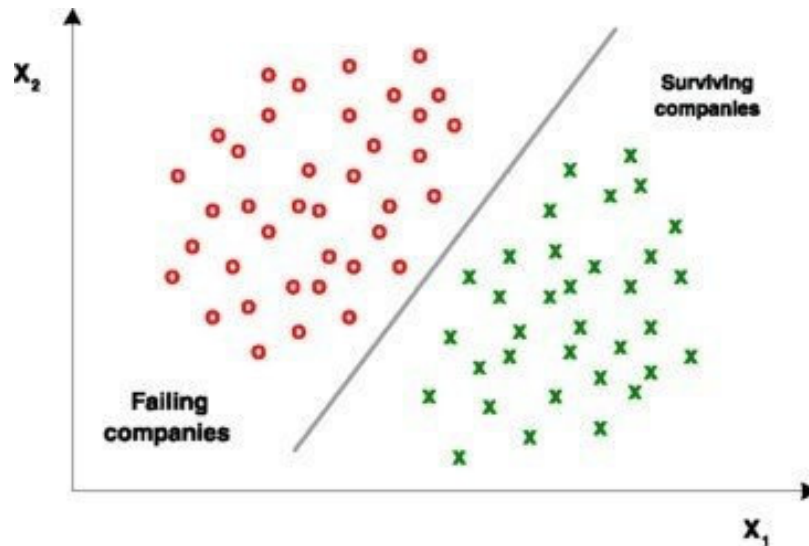
HOW TO OBTAIN THE RESULTS?

- There exist many approaches, each approach has several variants, and algorithms can also be designed by mixing approaches
- The background of researchers will often make the choice
- In general, **there is not** a “best technique”: no single algorithm is currently able to provide the best performance on all datasets
- This seems to be **inevitable**: if you chose a “best technique”, one can make a dataset composed of the records wrongly labeled by this “best technique” and make it the “worst technique” (no free lunch theorem)
- Therefore, **Ensemble techniques**: use many weak learners and combine their outputs to obtain both accuracy and robustness

Part 2: Neural Networks

Linearly separable data

Easy situations

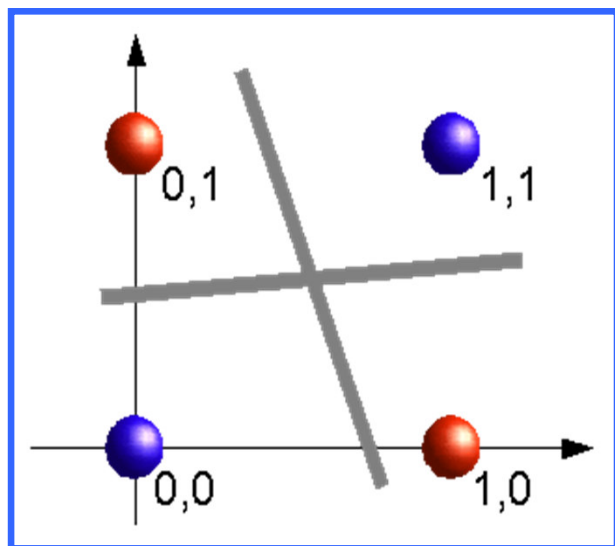
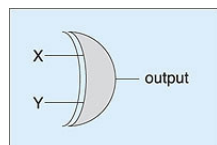


Not linearly separable

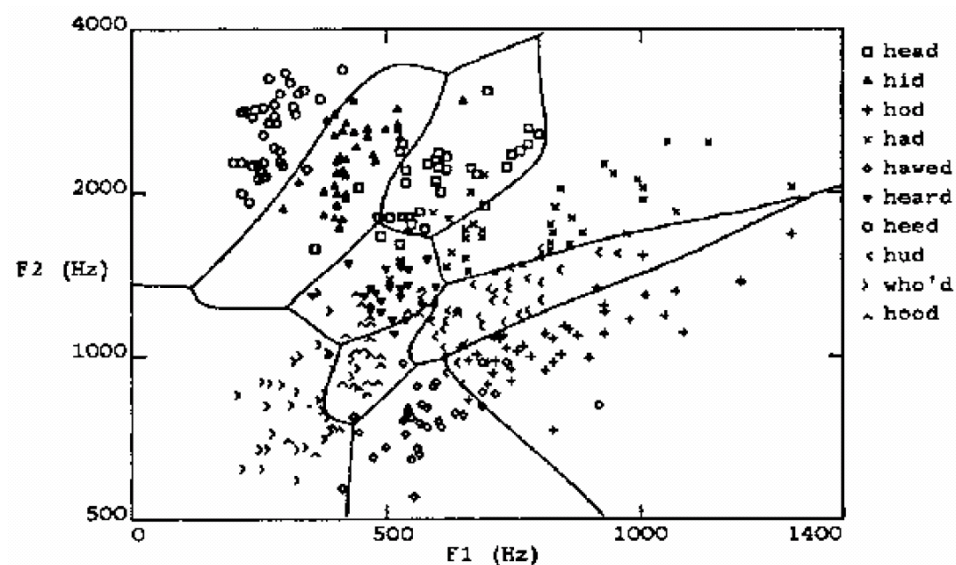
$f: X \rightarrow Y$

- f might be non-linear function
- X (vector of) continuous and/or discrete vars
- Y (vector of) continuous and/or discrete vars

The XOR gate

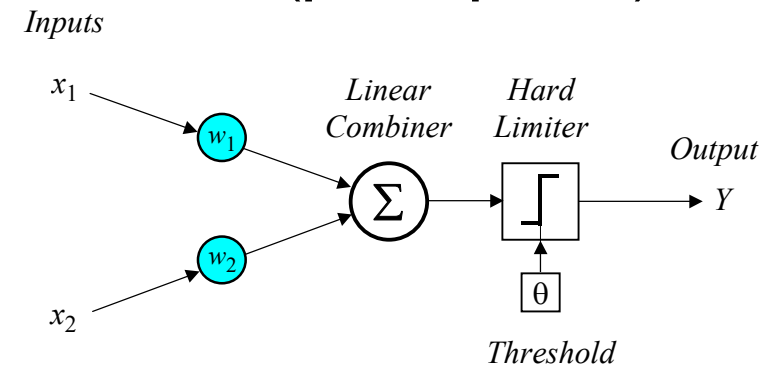
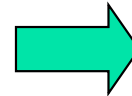
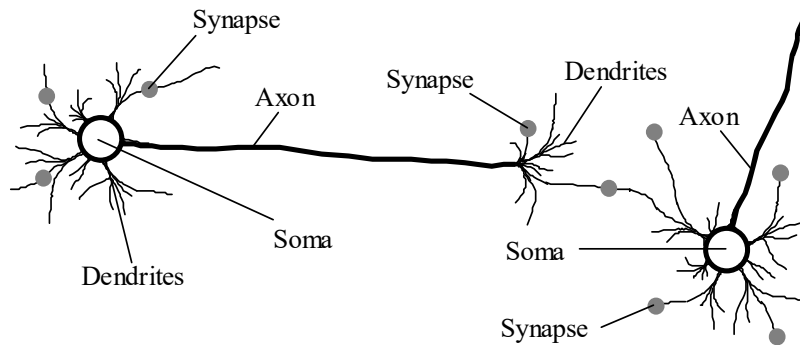


Speech recognition



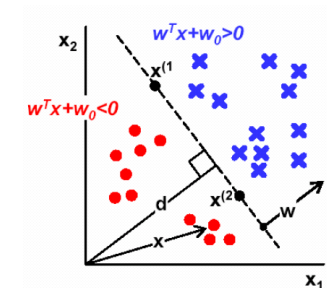
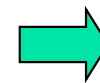
Perceptron

■ From biological neuron to artificial neuron (perceptron)



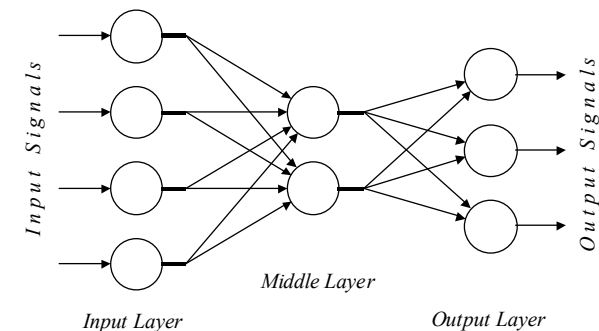
■ Activation function

$$X = \sum_{i=1}^n x_i w_i \quad y = \begin{cases} +1, & \text{if } X \geq \omega_0 \\ -1, & \text{if } X < \omega_0 \end{cases}$$

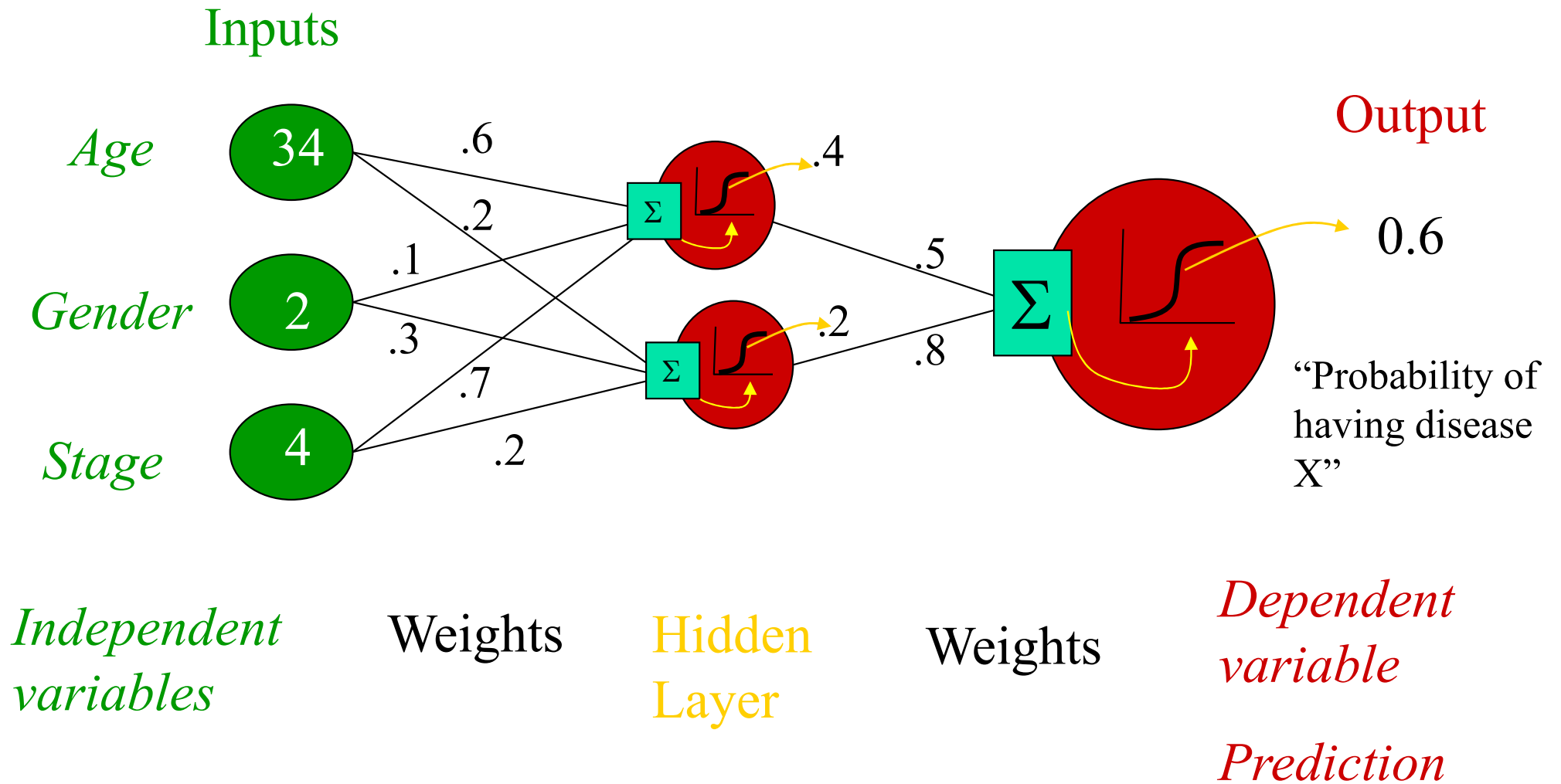


■ Artificial neuron networks

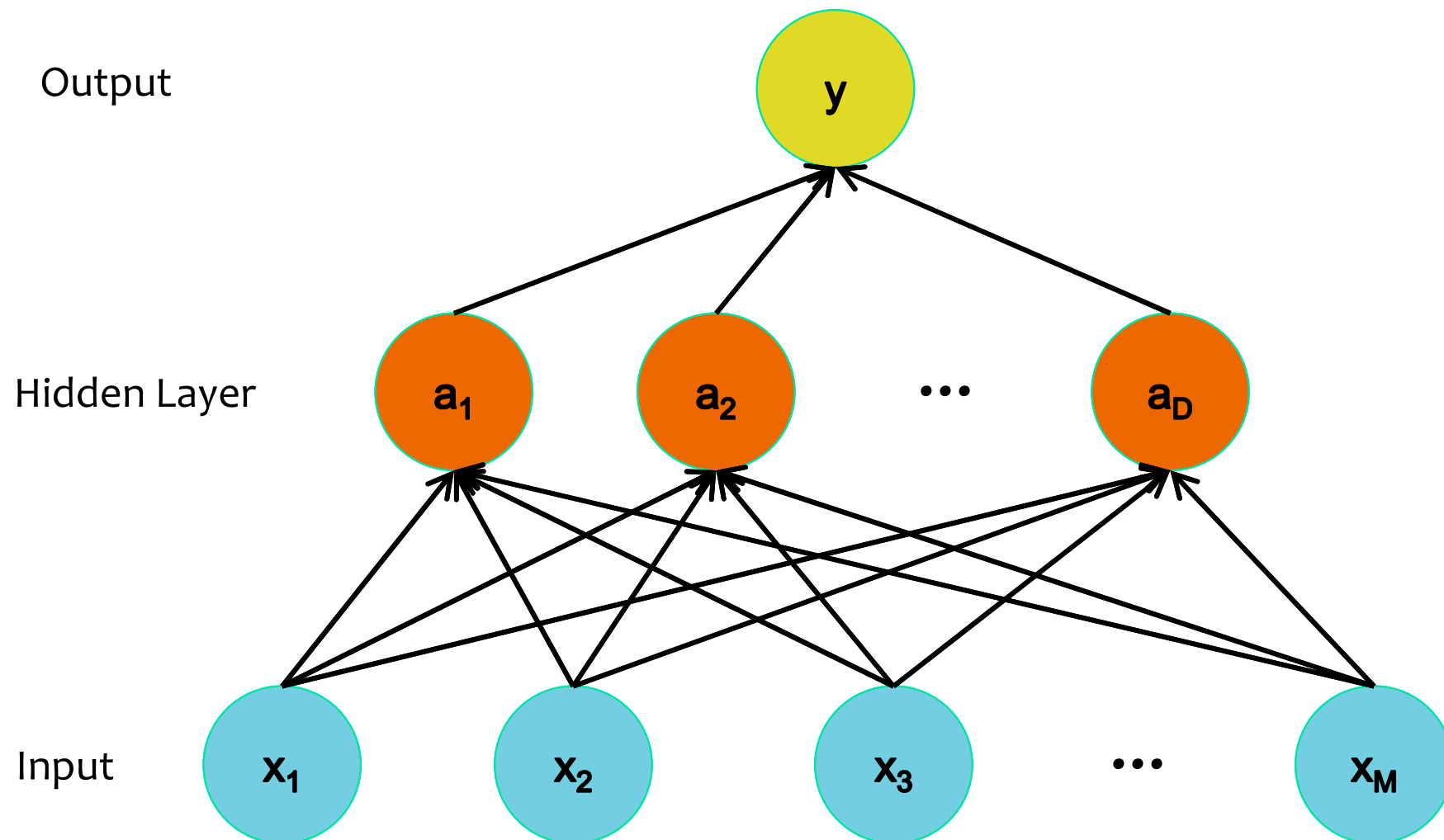
- supervised learning
- gradient descent



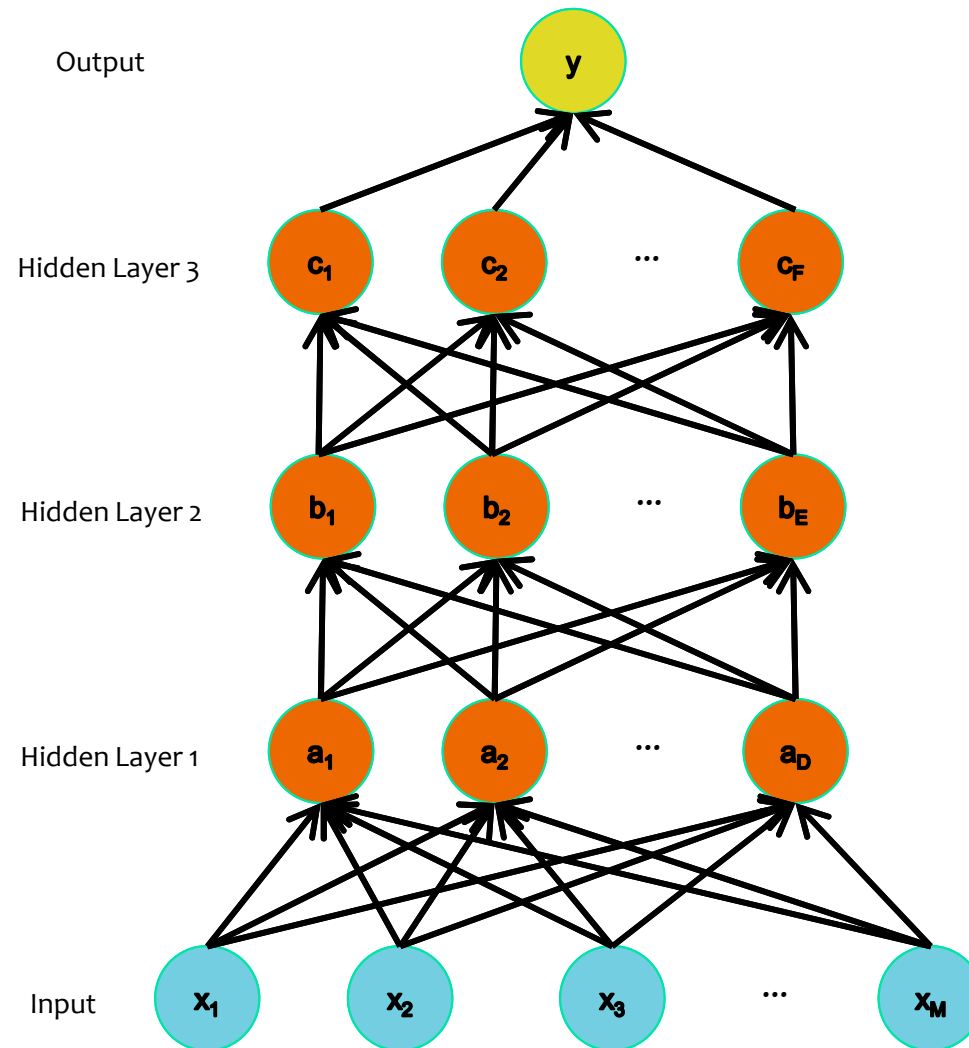
Neural network model



Building a neural network



Deep networks



Different levels of abstraction

- We don't know the “right” levels of abstraction
- So let the model figure it out!

Feature representation



3rd layer
“Objects”



2nd layer
“Object parts”

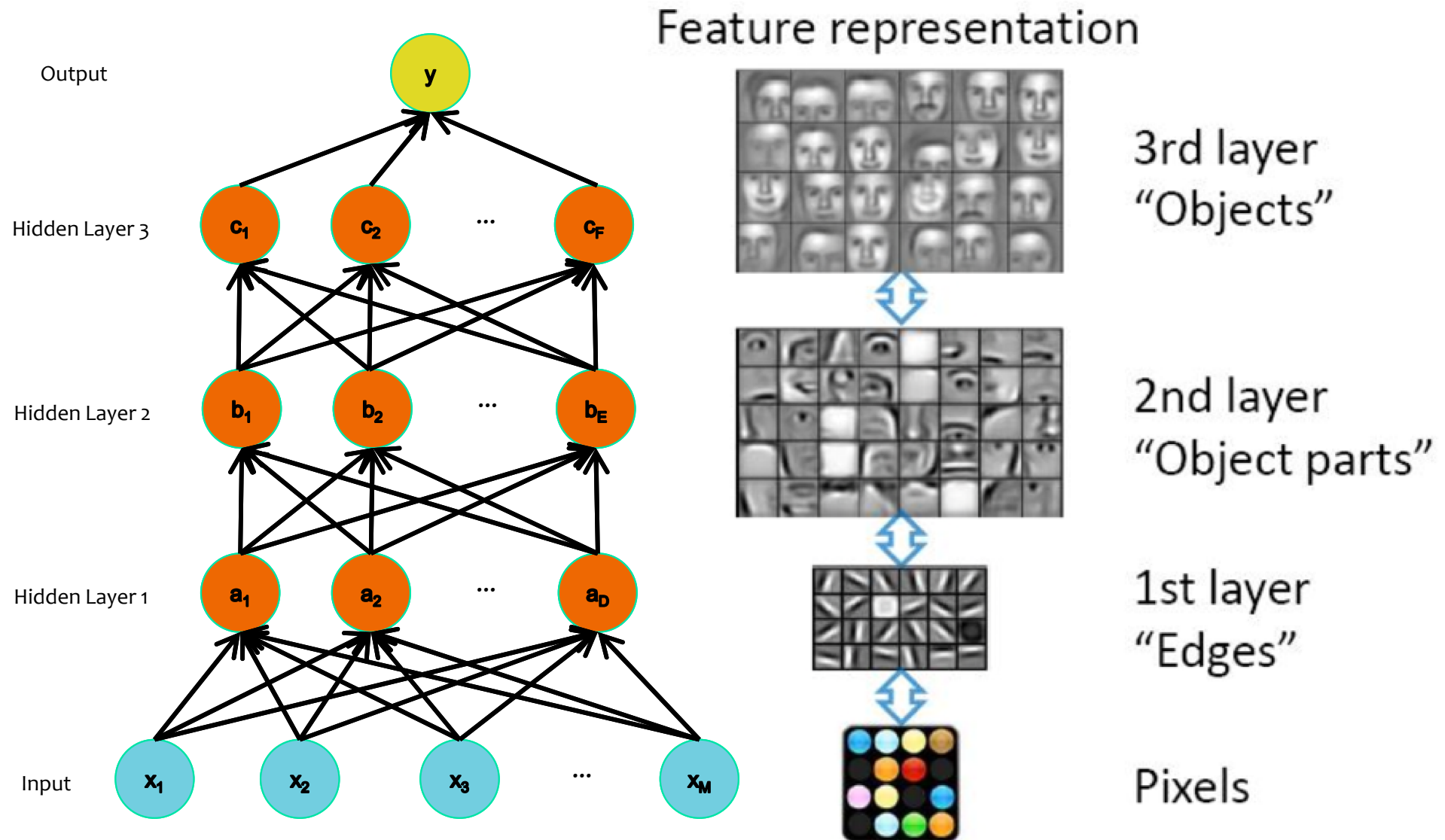


1st layer
“Edges”



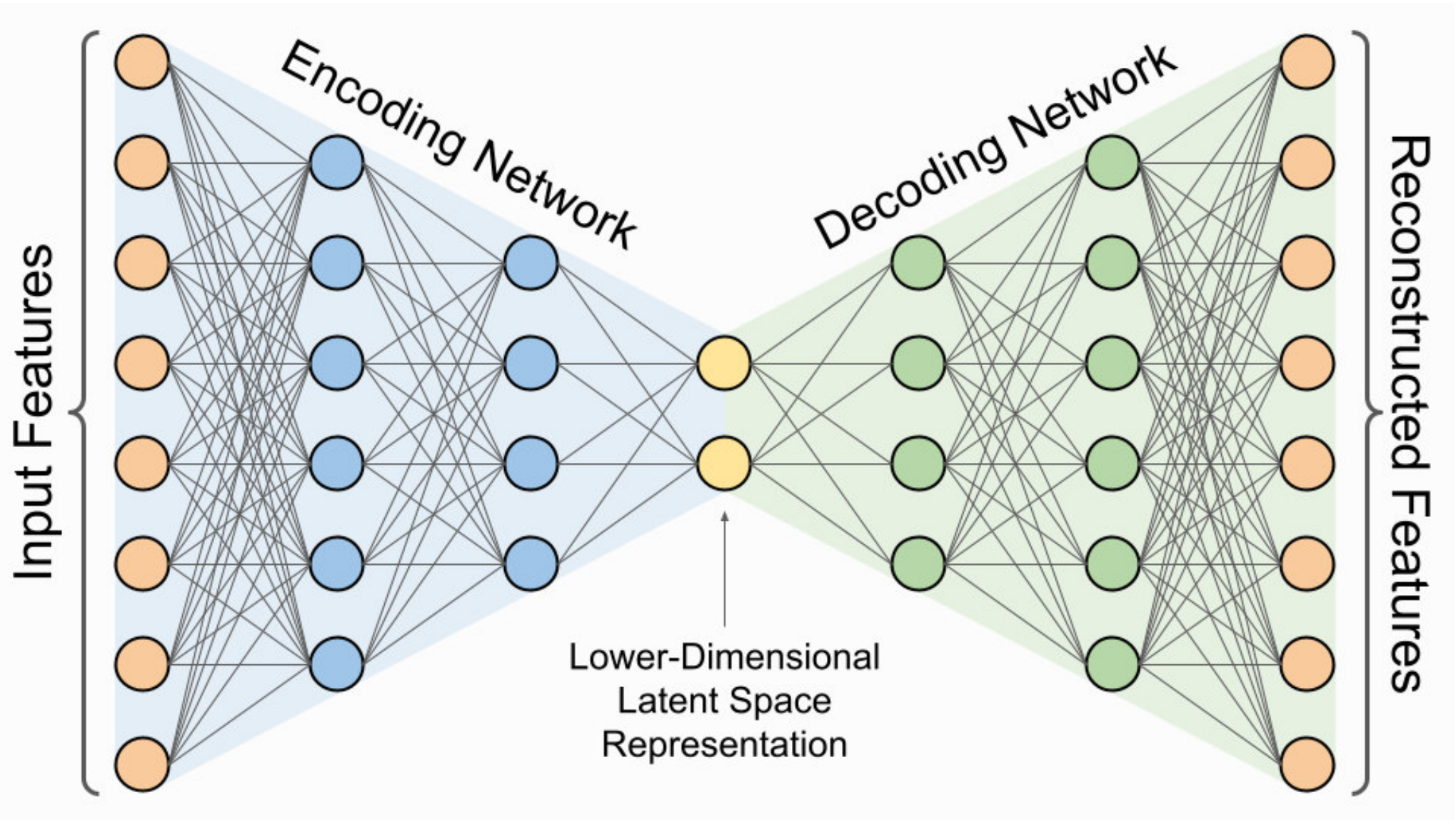
Pixels

Features handled by different levels



Example from Honglak Lee (NIPS 2010)

Autoencoders



Autoencoders

Are good for

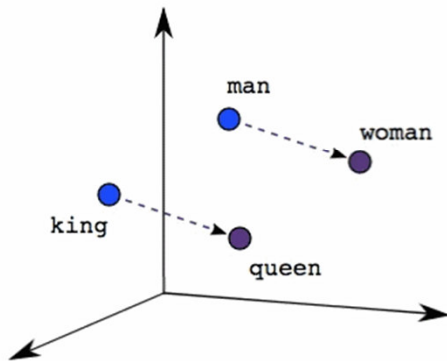
- **Extraction of relevant features**
- **Noise removal**
- **Anomalies detection**
- **Fraud detection**
- **Any data transformation that should keep the essential structure**
(e.g. colorization, magnification, etc)

This network architecture is the starting point for many modern powerful AI systems, including transformers, for example GTP

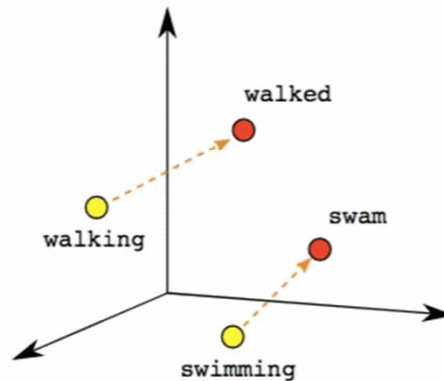
Word embedding

If we want Natural Language Processing in Large Language Models

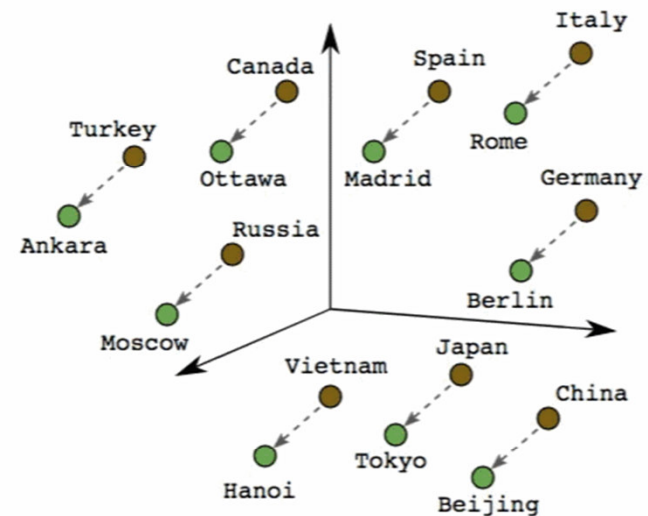
- Words must be encoded in numbers
- Many possible older techniques: one-hot encoding, bag of words, N-gram, TF-IDF. They cannot take context into account
- Aim: represent words in dense vectors of a given dimension, so that similar words are close to each other in the embedding space



Male-Female



Verb Tense



Country-Capital

How do we obtain the embedding ?

We use an **embedding layer in a neural network**

Algorithm word2vec: we consider sequences of n words (e.g. 3) and

1. Given the extreme, we try to guess the middle, or
2. Given the middle, we try to guess the extreme

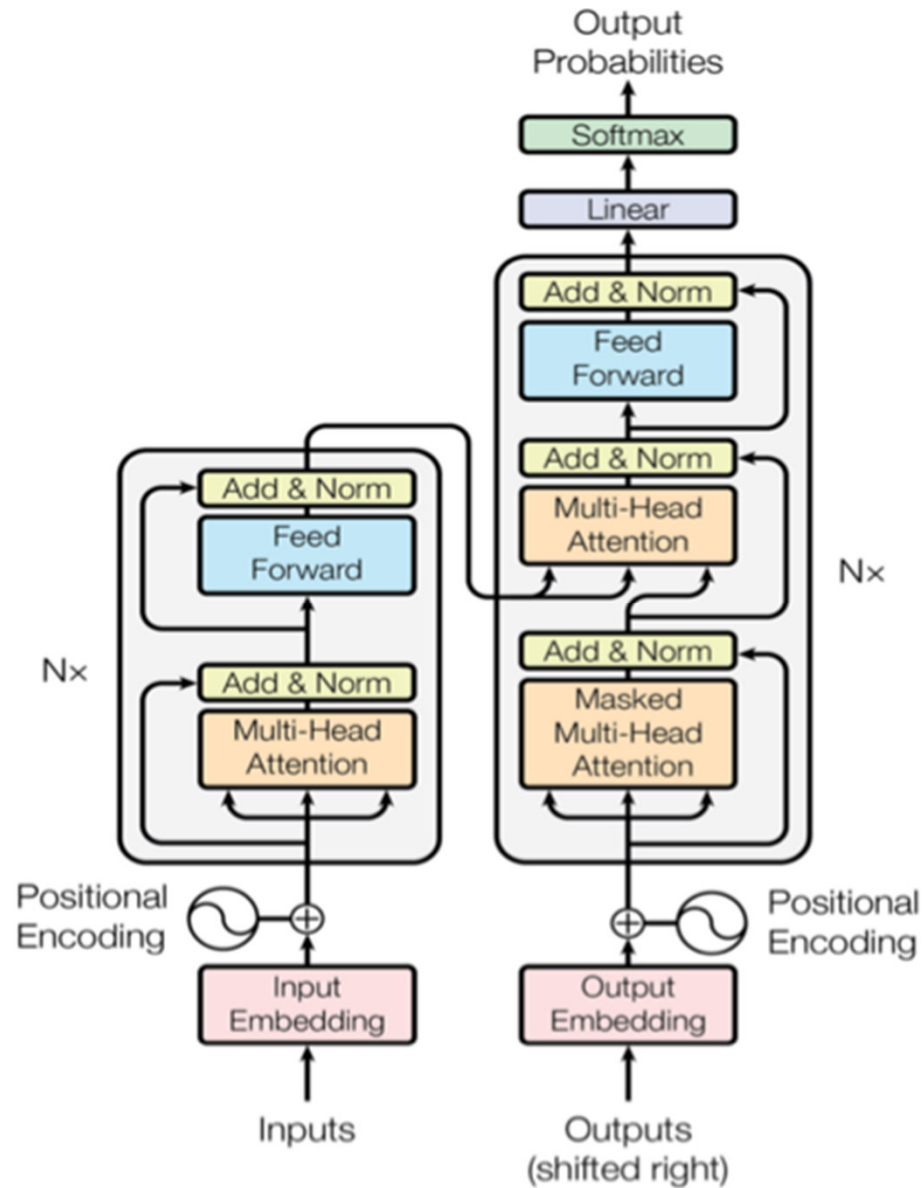
After training, the weights become the embeddings

Algorithm GloVe: not just local dependencies of the words, but also the global context

Algorithm fastText: uses the subwords (pieces of words)

Algorithm ELMo: each word has a representation that is a function of the entire input sentence using LSTM (long short term memory)

Transformers



Transformers

Important advantages

- **Attention mechanism**
- **Able to deal with multi modal input: numbers, text, images, voice**
- **Has embedding layers**
- **Has encoders and/or decoders**

However

- **Very computationally demanding**
- **Must be pre-trained**
- **Transformer hallucination**

Transformers evolutionary tree

