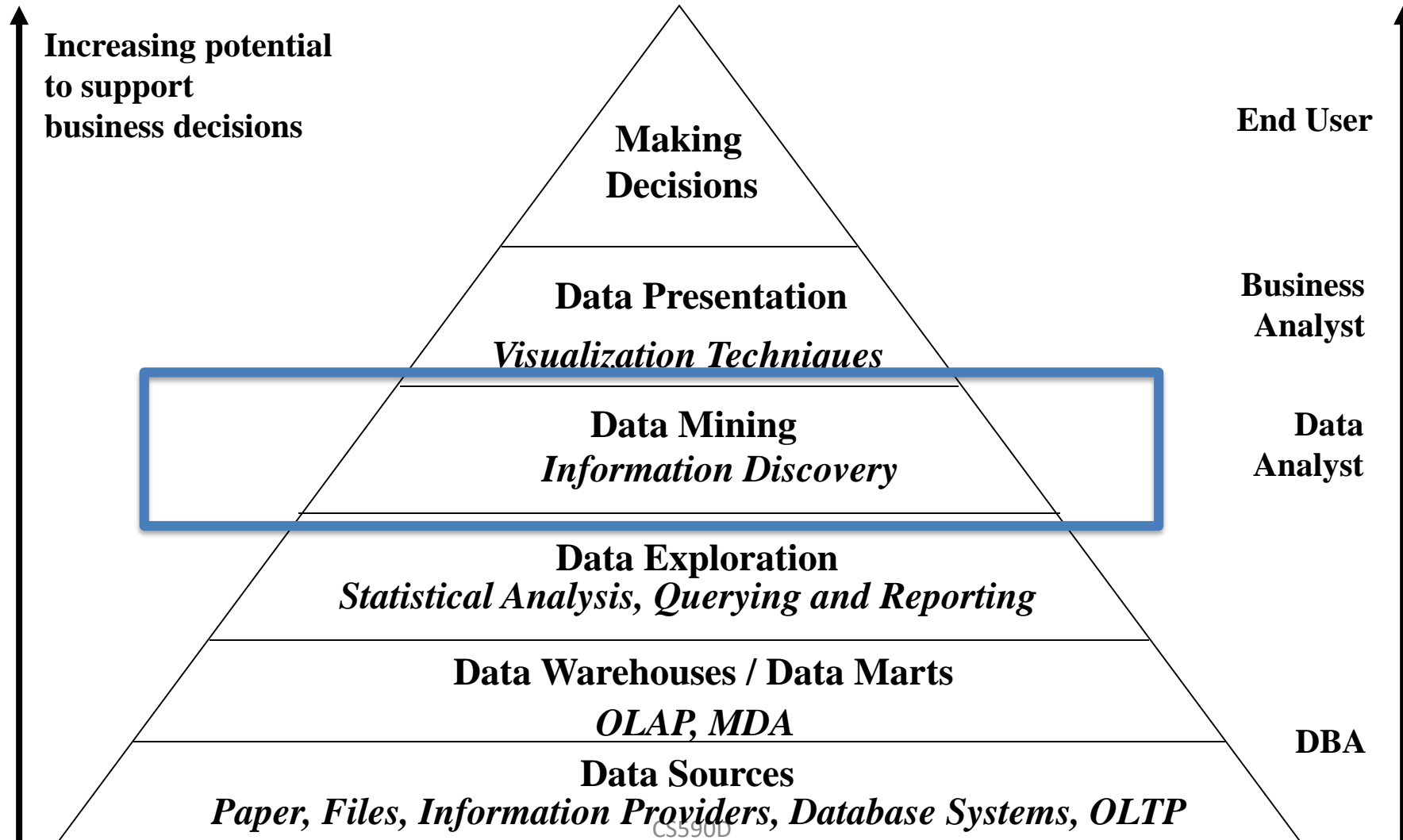


Data Analytics

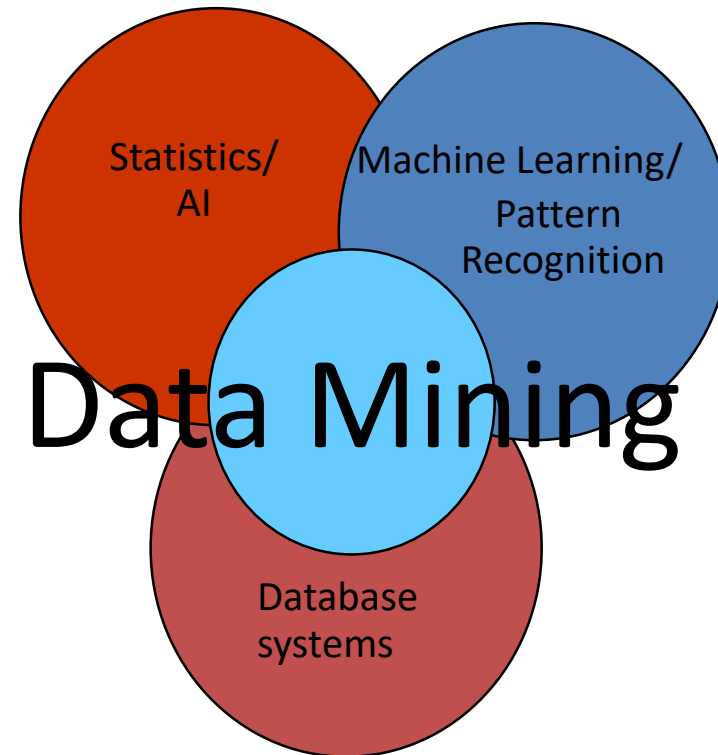
(many alternative names: Data Mining, Decision Support Systems, Machine Learning, Knowledge Discovery)

Data Mining and Business Intelligence



Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional (computational) techniques may be unsuitable due to**
 - **Enormity of data**
 - **High dimensionality of data (millions of attributes)**
 - **Heterogeneous, distributed nature of data**



Why Mining Data? Commercial Viewpoint

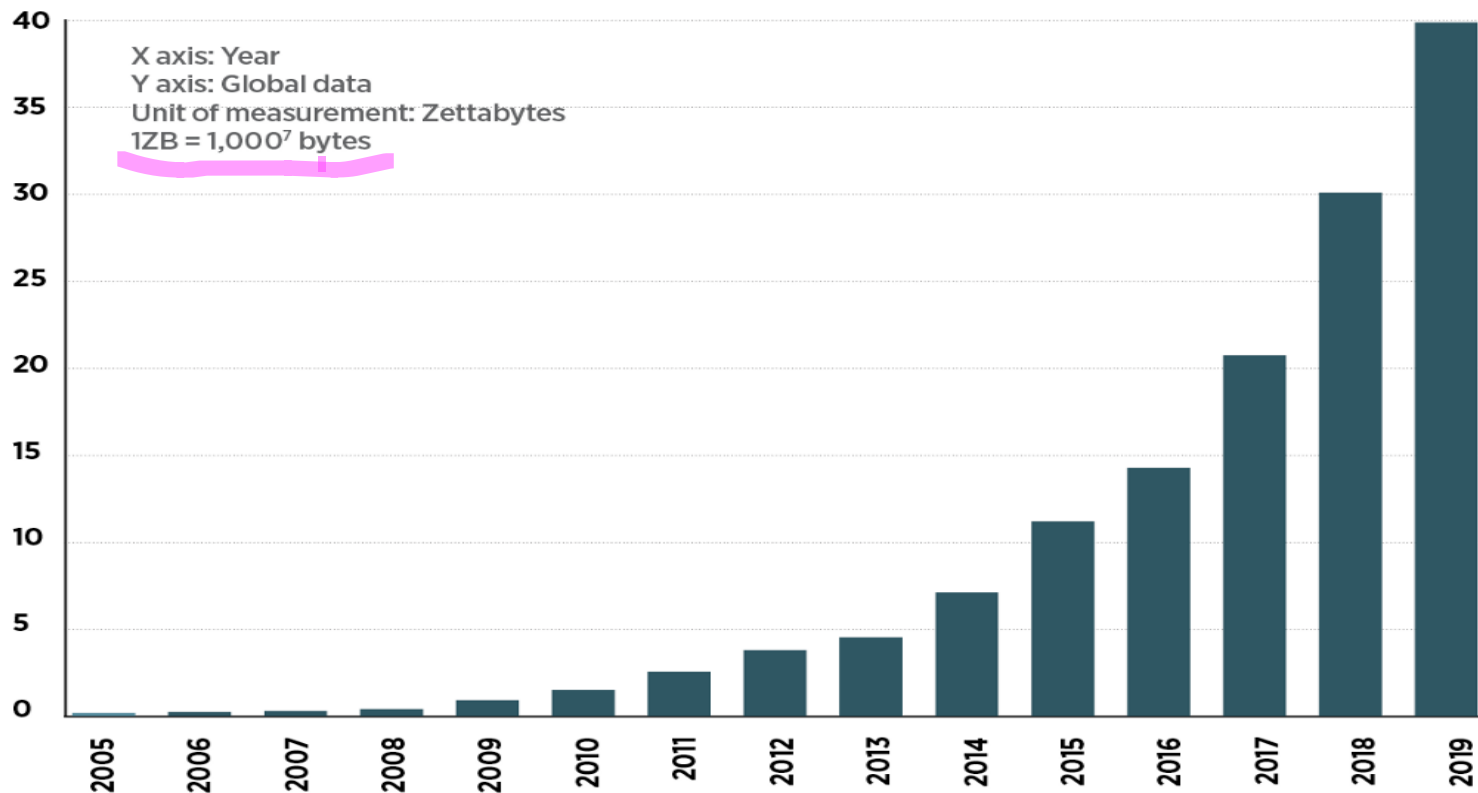
- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - purchases at department/grocery stores
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- **Competitive Pressure** is Strong
 - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)



Mining Large Data Sets - Motivation

- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all

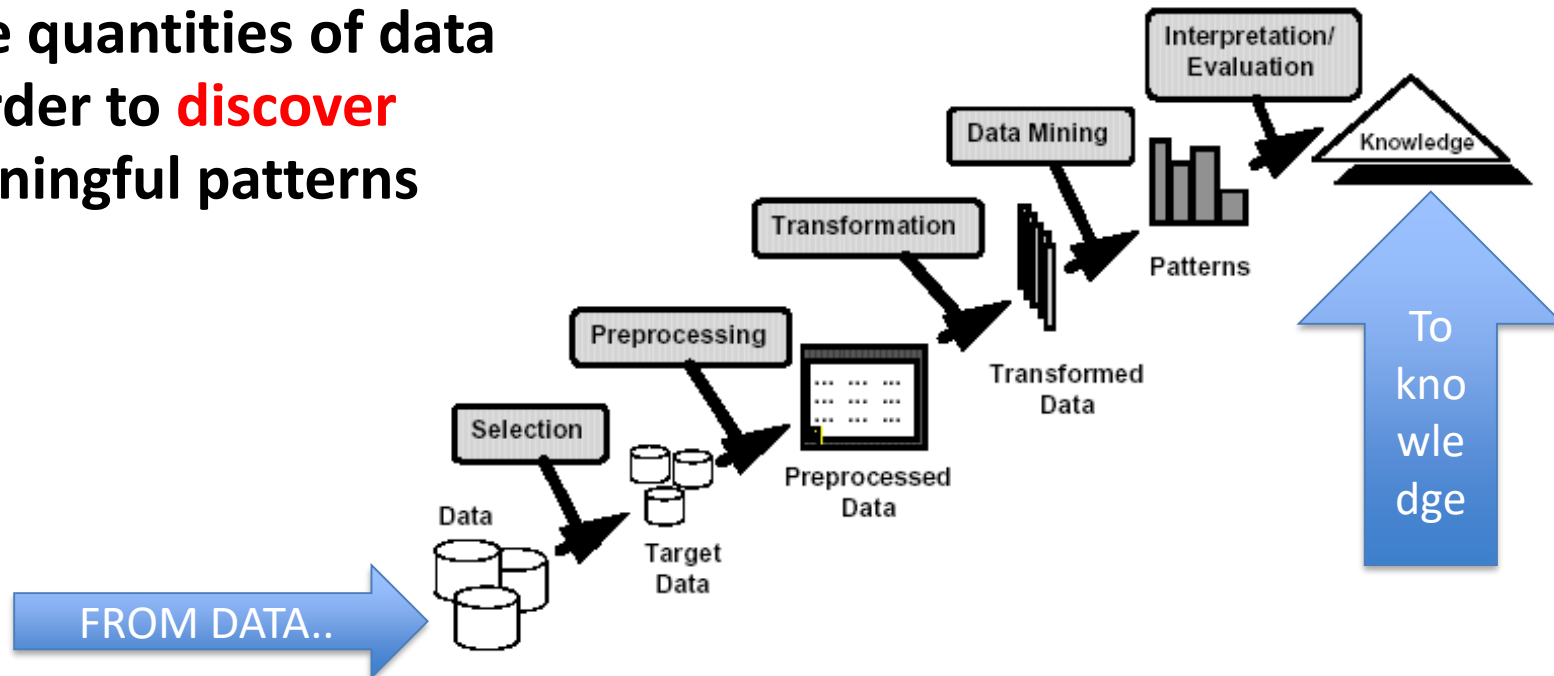
DATA GROWTH



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

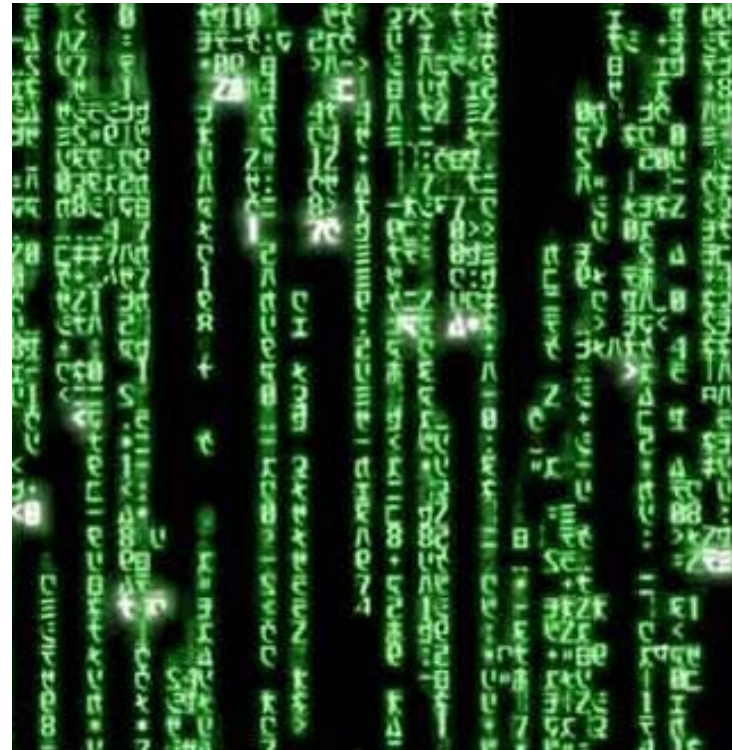


A blue ribbon graphic with a 3D effect, featuring a dark blue shadow on the left side. The ribbon is horizontal and contains white text.

Data, Information & Knowledge

Data

- Data **are** raw facts and figures that on their own have no meaning
- These can be any alphanumeric characters i.e. text, numbers, symbols



Data Examples

- Yes, Yes, No, Yes, No, Yes, No, Yes
- 42, 63, 96, 74, 56, 86
- 111192, 111234

- None of the above data sets have **any meaning** until they are given a **CONTEXT** and **PROCESSED** into a usable form

Data → Information

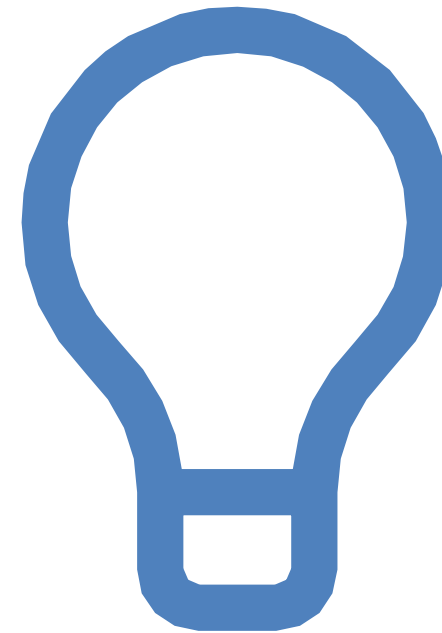
- To achieve its aims the organisation will need to **process** data into information.
- Data needs to be turned into **meaningful information** and presented in its most useful format
- Data must be processed in a **context** in order to give it **meaning**

Information

- Data that has been processed within a **context** to give it meaning

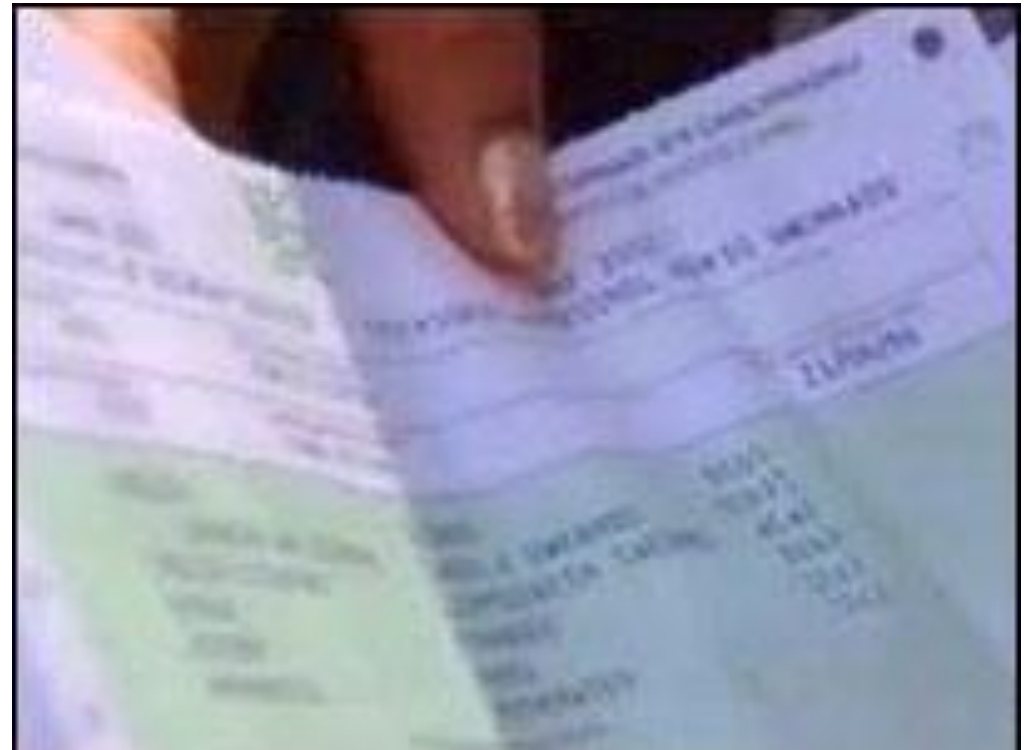
OR

- Data that has been processed into a **form** that gives it meaning

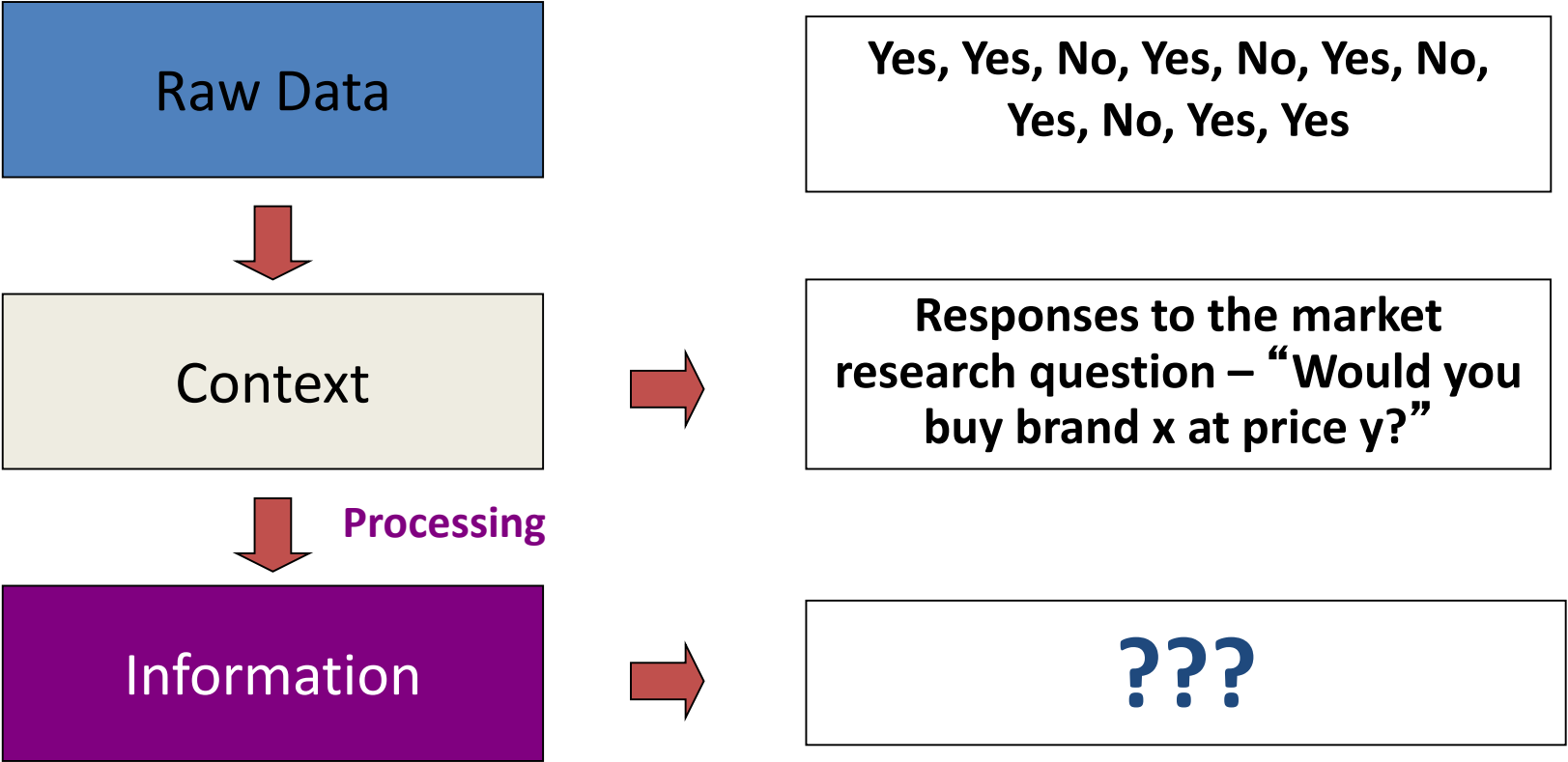


Examples

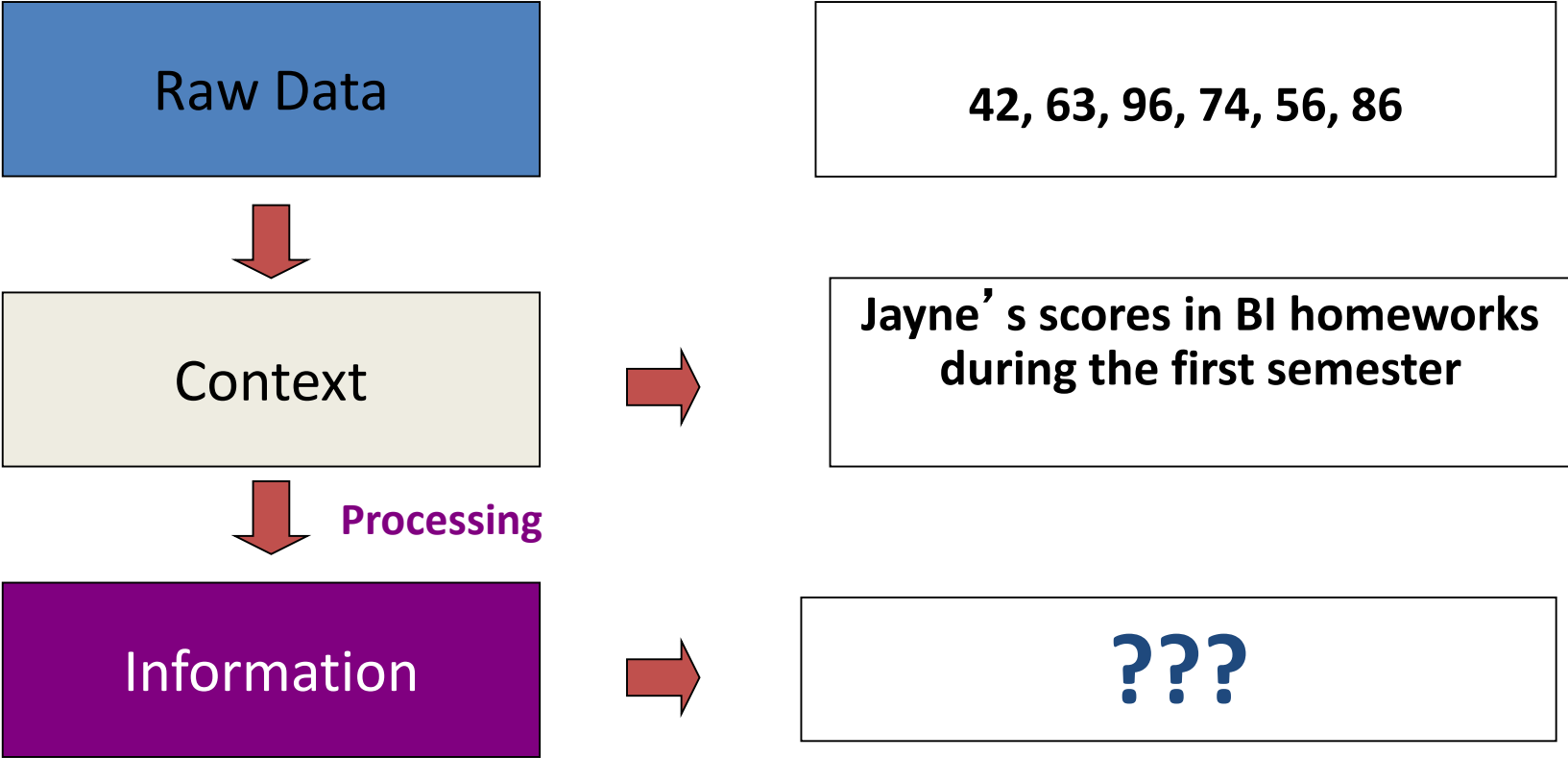
- In the next 3 examples we explain how the data could be processed to give it meaning
- What information can then be derived from the data?
- Can YOU answer?



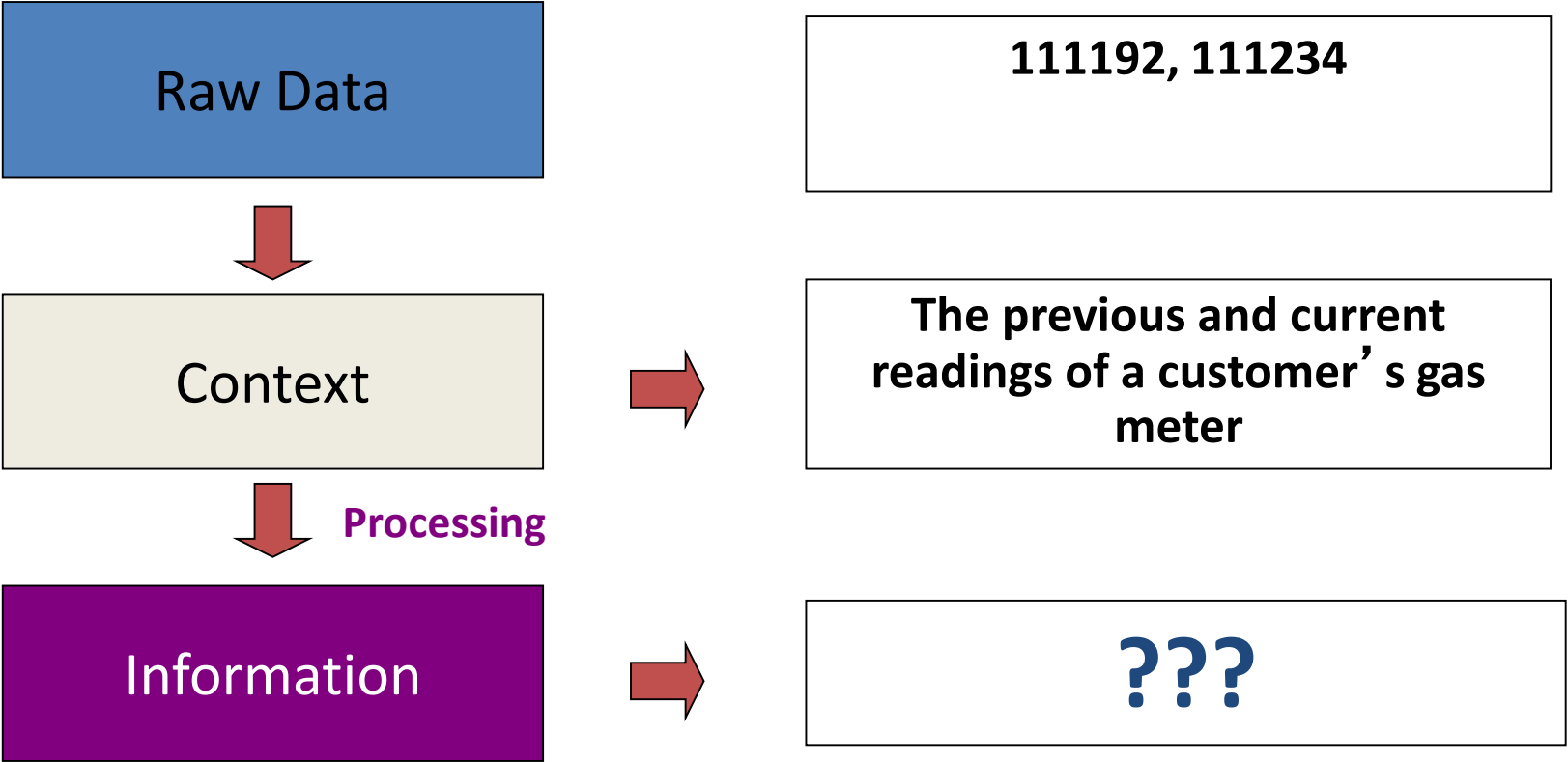
Example 1



Example 2



Example 3



Knowledge

- Knowledge is the understanding of rules needed to interpret information

“...the capability of understanding the relationship between pieces of information and what to actually do with the information”

Debbie Jones – www.teach-ict.com

Knowledge Examples

- Using the 3 previous examples:
 - A Marketing Manager could use this information to decide whether or not to raise or lower price of an item **(help settling pricing policies)**
 - Jayne's teacher could analyse the results to determine whether it would be worth her re-sitting a module **(help defining personalized training policies)**
 - Looking at the pattern of the customer's previous gas bills may identify that the figure is abnormally low and they are fiddling the gas meter!!! **(help in fraud detection)**

Data Mining Applications

- Some application domains (briefly discussed here)
 - Data Mining for Financial data analysis
 - Data Mining for Retail and Telecommunication Industries
 - Data Mining in Science and Engineering
 - Data Mining for Intrusion Detection and Prevention
 - Data Mining and Recommender Systems

3 questions

- **What kind of data?** (what do we have available or could be available)
- **What kind of INFORMATION** (knowledge) we would like to extract? (e.g., “clusters” of customers with similar purchase behavior)
- **What for** (what kind of predictive/ prescriptive analysis)? (e.g., addressing specific campaigns targeted for these customers)

Financial/Marketing
applications

- Predicting the impacts of customer engagement for a particular direct marketing promotion in a retail environment using historical promotional engagement **data** such as
 - customer information,
 - their location,
 - their responses to a promotional campaign or
 - how actively they have been engaging with websites or apps
- Identifying and preventing fraudulent transactions for banks by monitoring of customer transactions and flagging transactions which deviate from a standard customer behavior, identified for each customer of the bank from **data** such as:
 - transaction history and
 - the geographical locations of those transactions

Healthcare

- Preventing hospital-acquired infections by predicting the likelihood of patients susceptible to central-line associated bloodstream infections
- Using machine learning to predict the likelihood that patients will develop a chronic disease
- Assessing the risk of a patient not showing up for a scheduled appointment using predictive models
- **Which data:** Electronic Medical Record (EMR) data, along with hospital's internal data warehouse records on historical cases

Transportation

- Predictive Maintenance: Using vehicle sensor data (for cars or trucks), we can potentially help customers develop a predictive analytics solution, which can take this data to predict which components might fail or not perform as required.
- Dynamic Pricing: transportation businesses might be able to optimize the end-product costs based on real-time changes in operating factors such as fuel costs, security-related delays in shipments, and external factors, such as weather.

Retail Industry

24

Customer retention: Analysis of customer loyalty

- Use customer loyalty card information to **register sequences of purchases** of particular customers and predict future purchases
- Use sequential pattern mining to predict **changes** in customer consumption or loyalty
- Suggest **adjustments on the pricing** and variety of goods

Data Mining for Recommender Systems

25

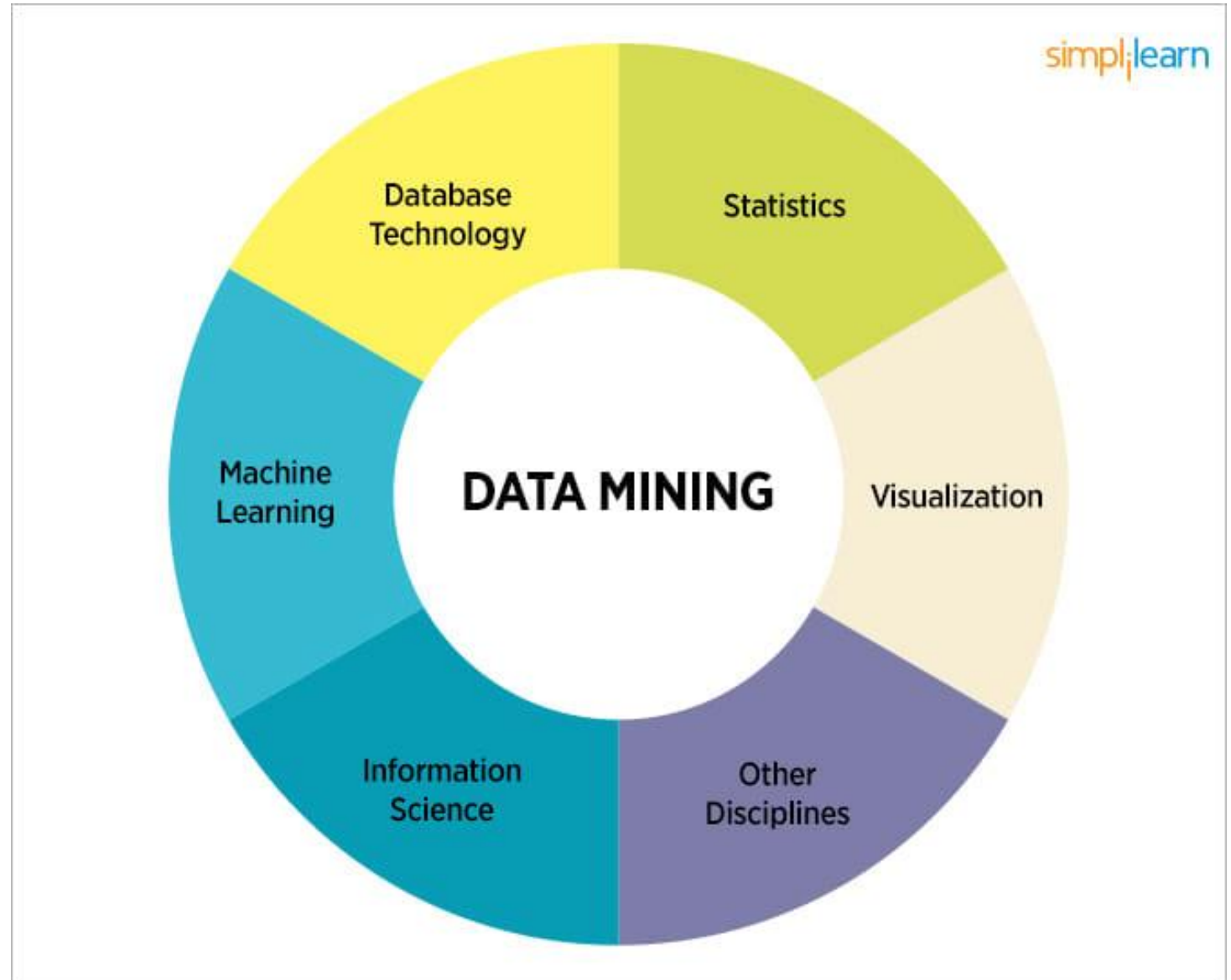
- Recommender systems: Personalization, making product recommendations that are likely to be of interest to a user. A very popular application!
 - What kind of data? User transactions, sales data, data (and opinions) on products
 - What kind of information we can extract?
 - User profiles
 - Product similarity, opinions on products
 - What the user is likely to like
 - What kind of knowledge?
 - Will user x purchase item y?
 - Will a new product y receive the appreciation of buyers?
 - Why item y was (was not) successful?

Data Mining Tasks

- **Prediction Methods**
 - Use historical data to predict unknown or future values of some attributes.
 - Example: predict if user x will purchase product y (given his/her past purchases and preferences)
- **Prescriptive Methods**
 - Find human-interpretable **patterns** or **traits** in data, that (synthetically) describe the data, and **act** accordingly.
 - Example: discover strong relationships between specific demographic attributes and vote during elections, or discover common buyer behaviors (e.g. if customers buy items X and Y, they also often buy Z, so suggest place items X, Y and Z near to each other to increase the sales of Z - e.g. outfit recommendation in fashion market)
 - Example: learn driving strategies, learn to play chess (physical and virtual robots)

Data Mining vrs. Machine learning

- Data Mining is extracting information from data, includes descriptive, prescriptive, predictive analytics, and also data visualization
- Machine learning can be considered as a subset of Data Mining: ML are algorithms that find «patterns» in the data for prescriptive an predictive analytics



What is Machine Learning (in a nutshell)

A set of methodologies to find regularities in data



```
graph LR; A[A set of methodologies to find regularities in data] --> B[These findings are used to predict future outcomes or to prescribe optimal strategies]
```

These findings are used to **predict** future outcomes or to **prescribe** optimal strategies

What data are used to learn?

- Historical («*labelled*») data: data **collected in the past**, for which the outcome is known (example: patient histories where we know if a cardiovascular event has occurred or not; bank customer's histories for which we know if they have been defaulters or not)
- *Unlabelled* data: data with no labels, for example the sequence of purchases of a user on an e-commerce web site, or sequence of actions on flight actuators by a human pilot of an airplane

What is «labelled»? Usually the task is learning to predict the value of some variable (e.g., cardiovascular risk). Historical data provide examples of such values.

Example of predictive learning: Credit risk assessment

Customer ID	AGE	INCOME	EDUCATION	DEFAULT
ID1	27	30.000	YES	1
ID2	50	45.000	NO	0
ID3	60	46.000	YES	0
.....				
ID1348	32	55.000	YES	0

- Credit scoring is a fairly widespread practice in banking institutions, whose main objective is to discriminate between borrowers, based on their *credit worthiness*.
- Decision on whether granting credit to new customers is based on **past data on customers who experienced a default or not**
- Machine learning can help assessing **the risk of default** of new customers based on a «risk model» learned from **past data**

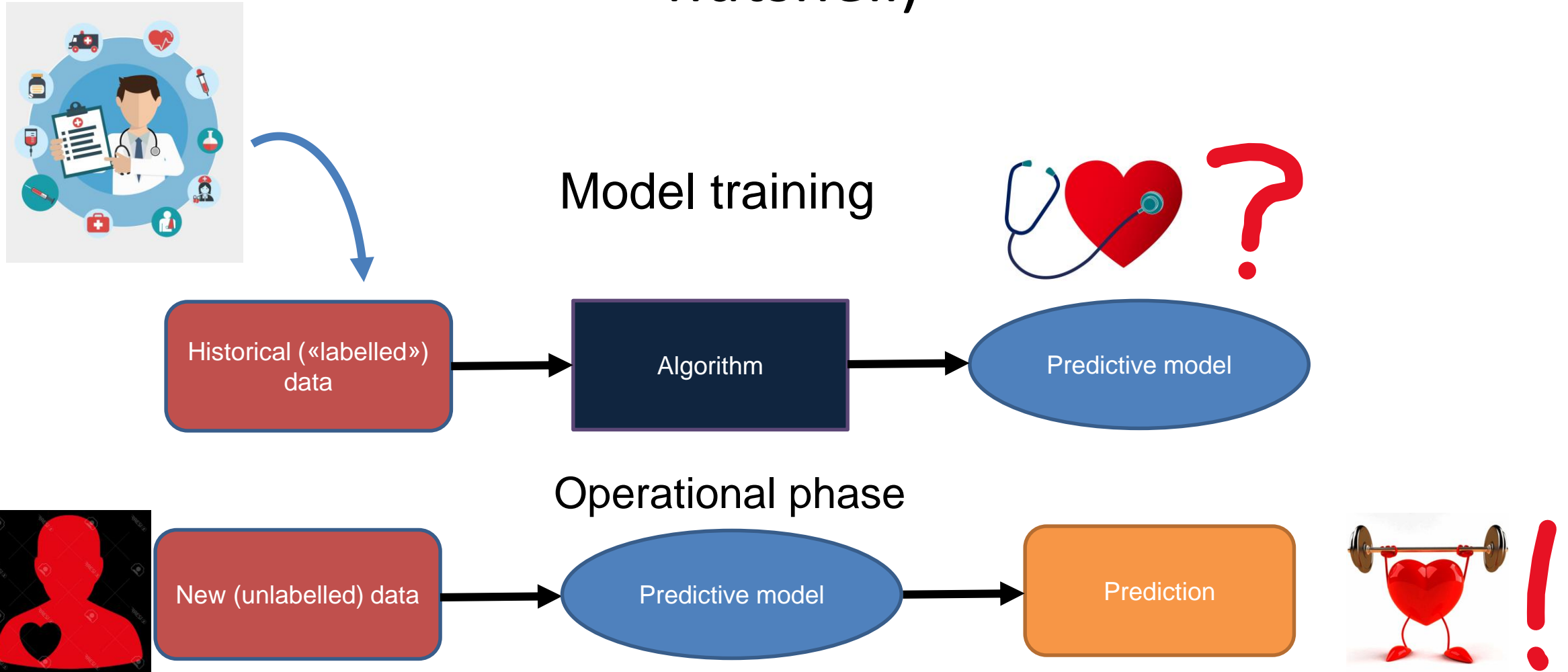
* Here data are «**labelled**», to mean that historical data include the label (value) of the **variable we want to predict**, «Default» in this example. Note that Default is a binary variable, but as we will see, ML algorithm can learn predicting either discrete or continuous variables.

Example of feature identification: cardiovascular risk assessment

# Diabetes_012 0 = no diabetes 1 = prediabetes 2 = diabetes	# HighBP 0 = no high BP 1 = high BP	# HighChol 0 = no high cholesterol 1 = high cholesterol	# CholCheck 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years	# BMI Body Mass Index
# Smoker Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes	# Stroke (Ever told) you had a stroke. 0 = no 1 = yes	# HeartDiseaseorAt... coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes	# PhysActivity physical activity in past 30 days - not including job 0 = no 1 = yes	# Fruits Consume Fruit 1 or more times per day 0 = no 1 = yes
# Veggies Consume Vegetables 1 or more times per day 0 = no 1 = yes	# HvyAlcoholConsum... Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes	# AnyHealthcare Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes	# NoDocbcCost Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes	# GenHlth Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
# MentHlth How often are you bothered by the following things: low thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days?	# PhysHlth Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days?	# DiffWalk Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes	# Sex 0 = female 1 = male	# Age 13-level age category (LAGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
# Education Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8	# Income Income scale (INCOME2 see codebook) scale 1-8 1 = less than \$10,000 5 = less than \$35,000 8 = \$75,000 or more			

- Electronic patient records are now widely available. They collect the «history» of patients, their clinical tests, treatments and diseases
- Doctors can be supported in deciding the best therapy, or in estimating a specific risk of complications (e.g., cardiovascular risk) by machine learning systems, based on the analysis of historical data of previous patients

Basic workflow of a *predictive* ML system (in a nutshell)



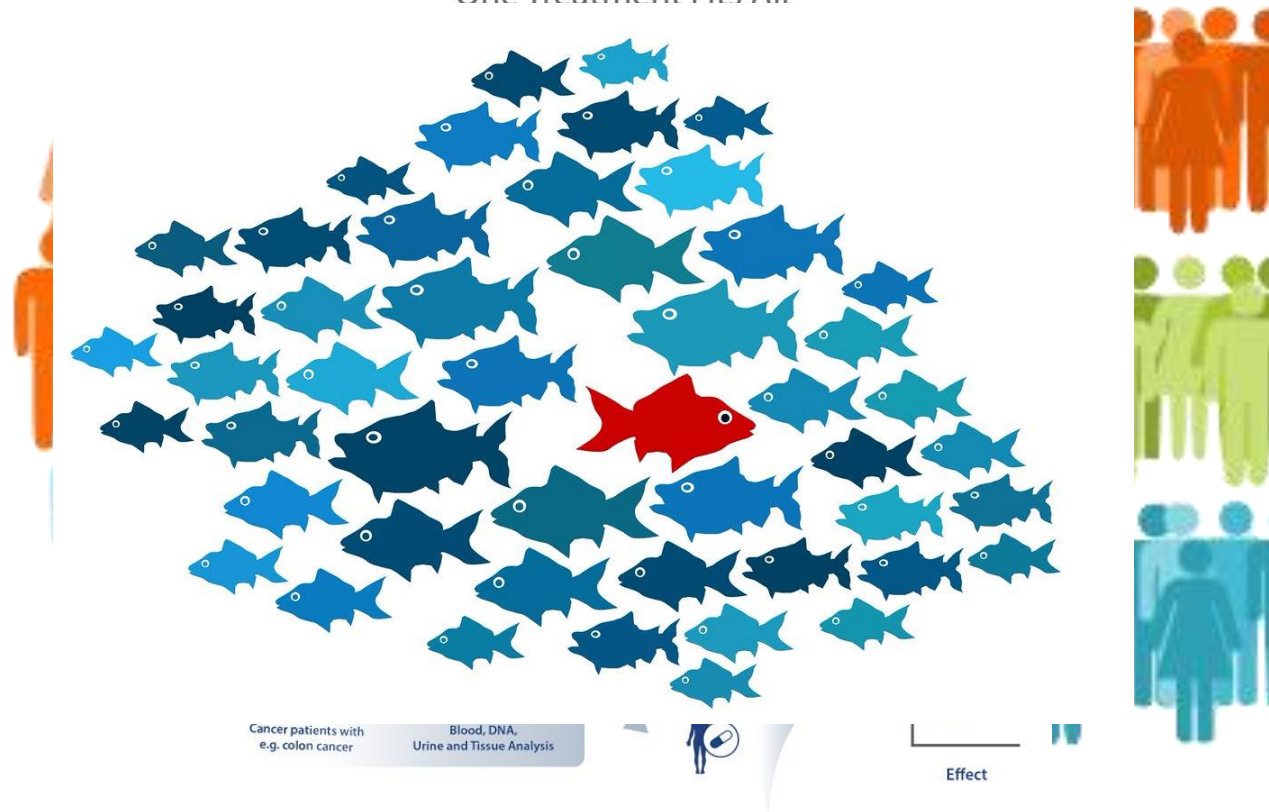
Note, **not all ML systems work in this way**. This is the most popular category of ML systems, named **Supervised Machine Learning**.

What is the task?

- Predictive:
 - Given previous historical «labelled» data, learn a model to predict future outcomes (e.g., see what happened to past credit applicants, or to past patients, and learn what may happen to new applicants or new patients)
 - Examples: predict patients' risk of a complication, predict future sales of a new product, users' satisfaction in a market campaign..
- **Prescriptive/Descriptive:**
 - **Given available data, or given an environment and some stimuli, prescribe «how to», e.g., best actions to be performed**
 - **Example: customer segmentation according to their profiles, best strategy to win a game, best way for a robot to execute a given task – e.g., drive a car – how to improve on-line sales by recommending the right items to customers**

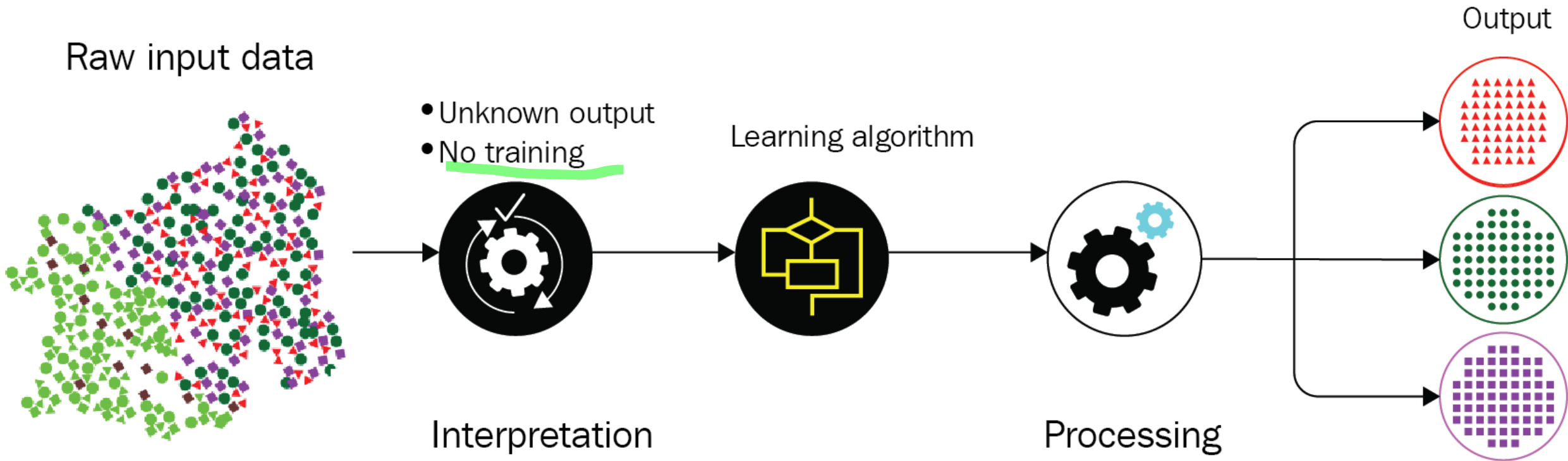
Example of prescriptive analytics: customer segmentation, precision therapies, anomaly detection

Current Medicine One Treatment Fits All



- Given data on customer profiles, cluster them into groups of «similar ones»
- Then, use these groups to identify best personalized marketing campaigns to optimize revenues

Example of clustering

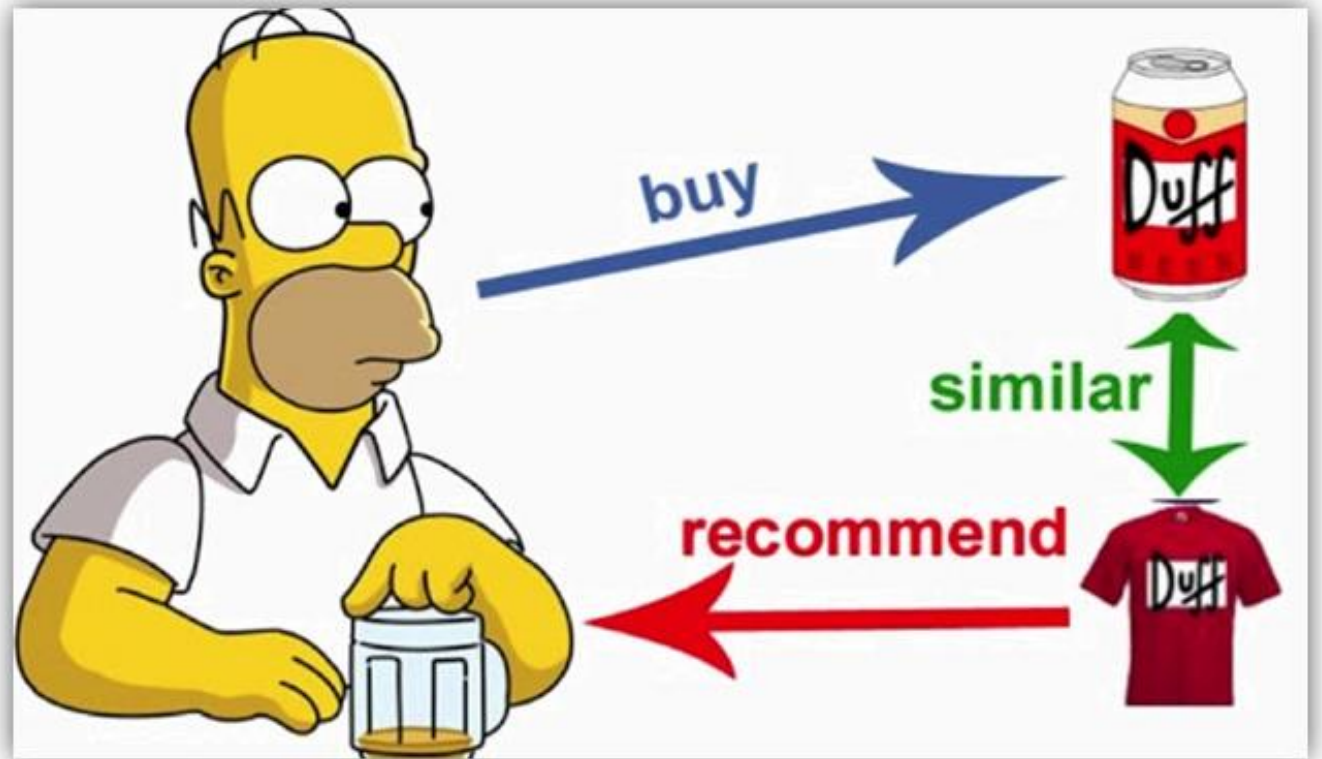


Note: system learns from unlabelled data, these ML models are called **Unsupervised Machine Learning** models

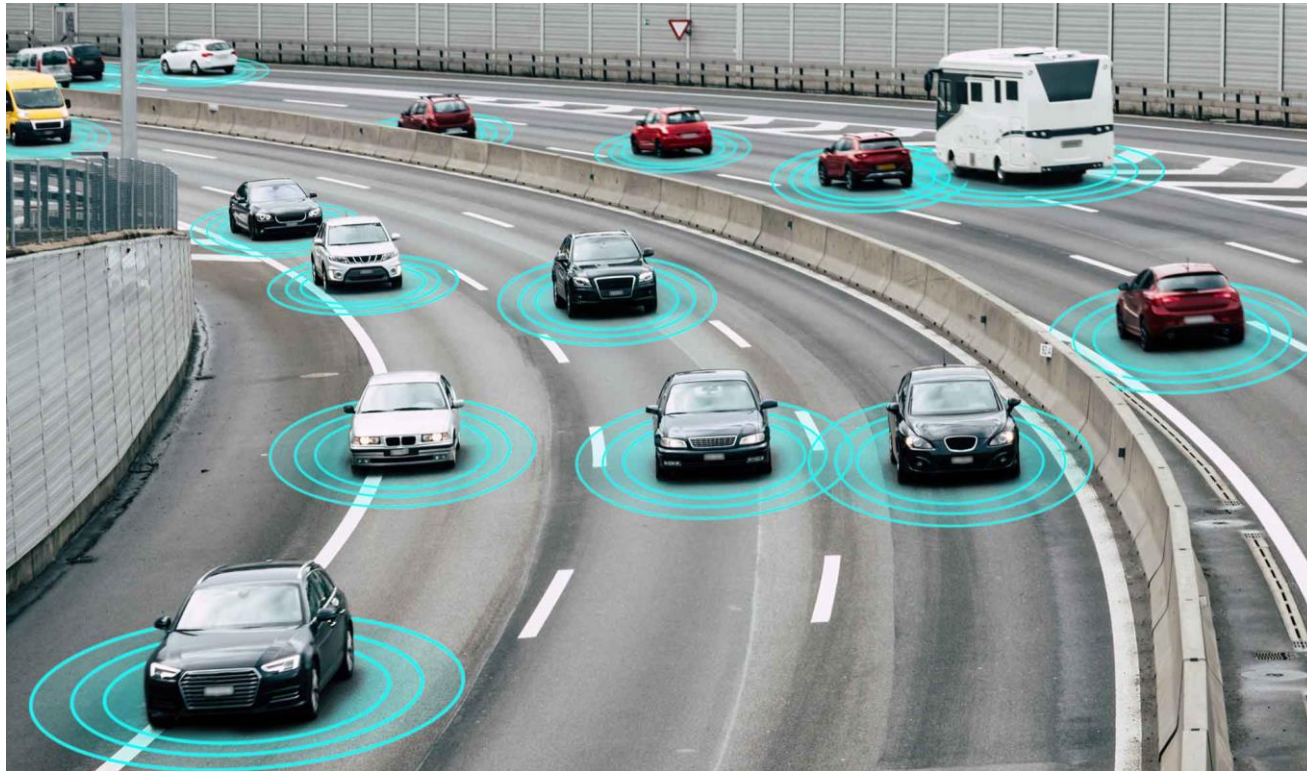
Example of prescriptive analytics

Recommender systems

- Observe a behavior and “recommend” items to buy, music to listen, people to follow on social,...

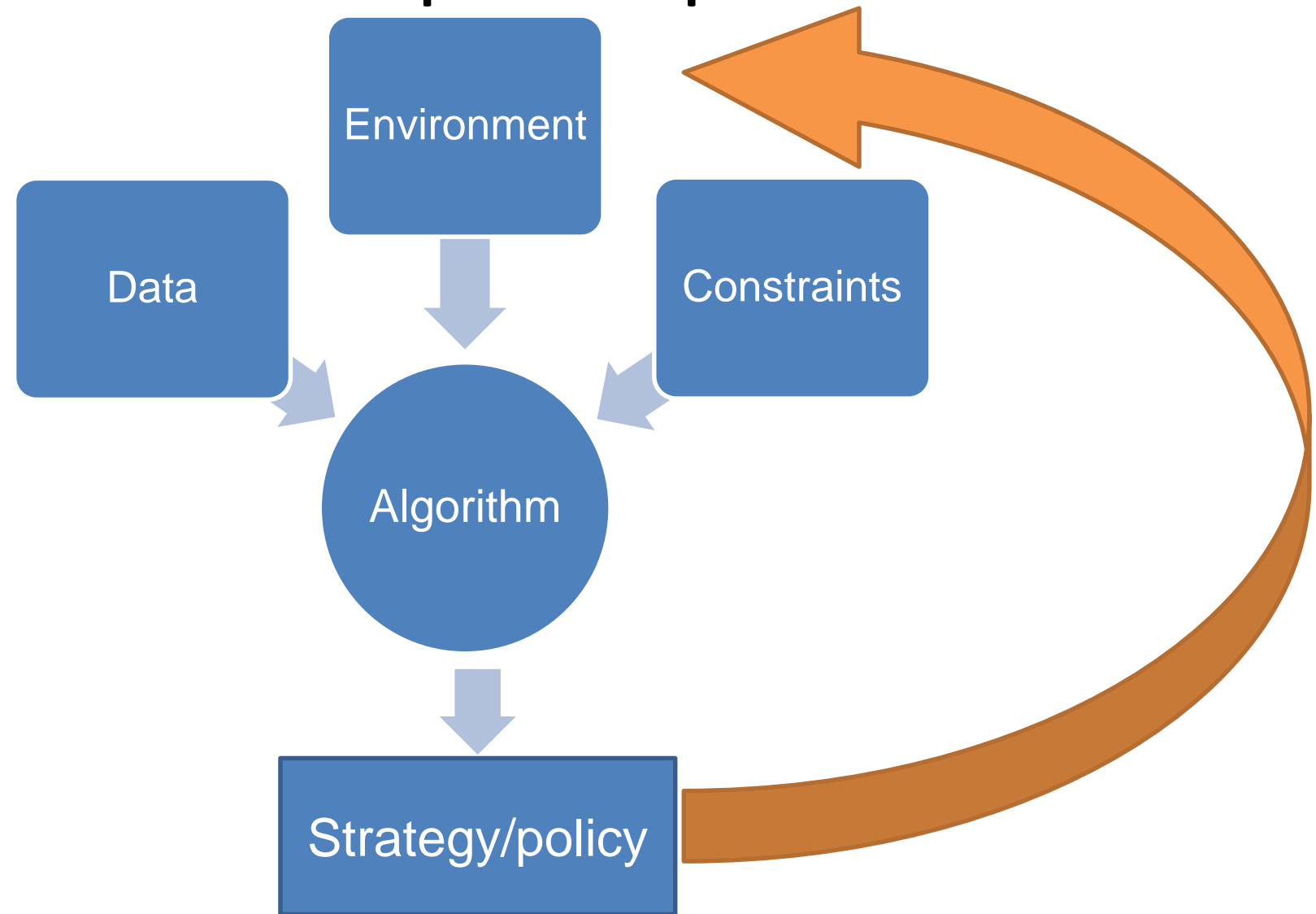


Example of prescriptive analytics: self driving cars



- Analyse driving behaviours of million «human» drivers
- Learn best strategy to react to the environment (driving strategies) in any condition

Workflow of prescriptive ML





Issues in Machine Learning

Issues in Machine Learning

“How can we program systems to automatically learn from «data» and to improve predictive/prescriptive capabilities with experience? “

Need to ponder on how human beings learn..

- **What** is learning?
- **What** can we learn?
- **What** is “experience”??
- **How** do we learn?
- **How** can we “improve”, and over what??

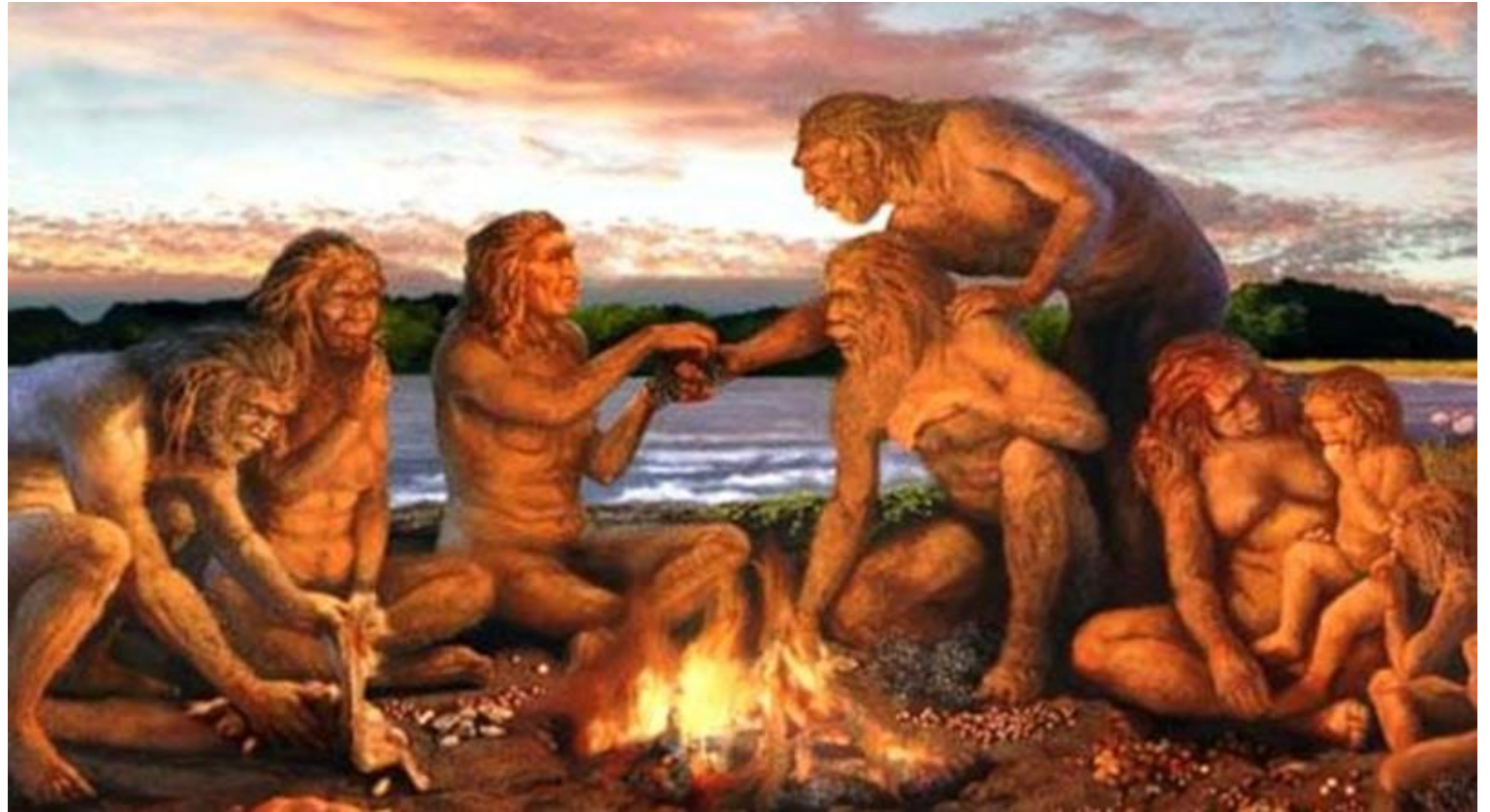
What is learning??





Fire burns!

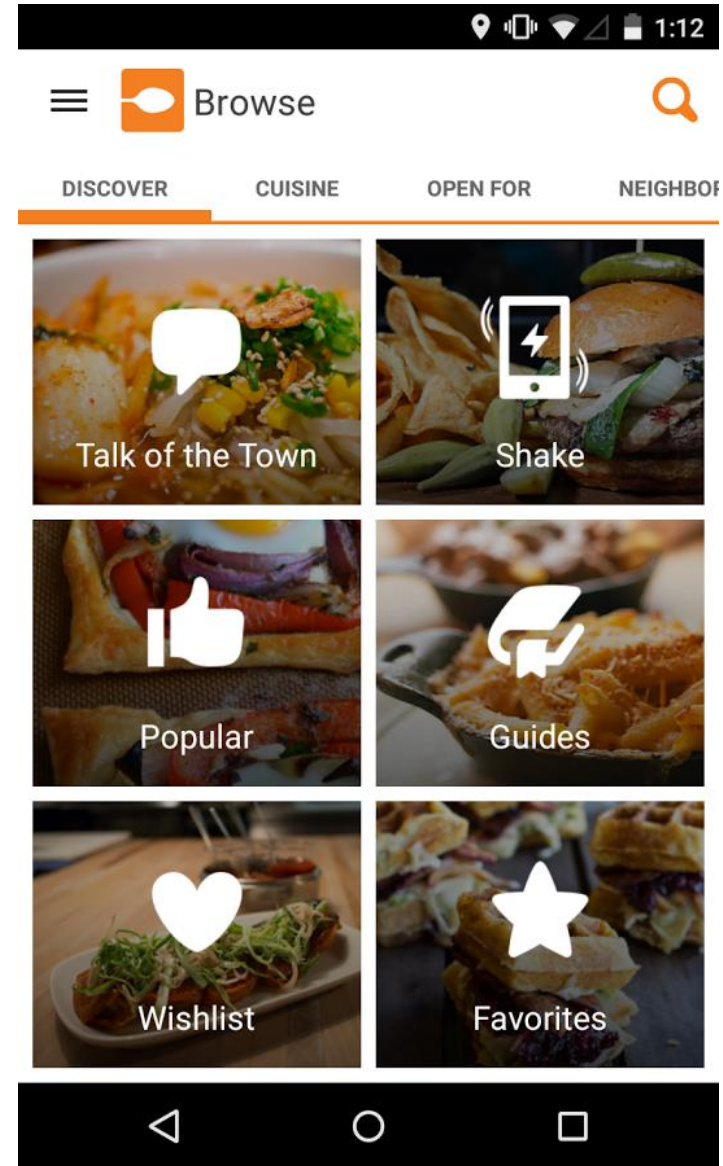
But we
eventually
learned
using it



You can
study (learn)
Machine
Learning



And then build
an app to
reccomend best
restaurants
based on
people's
preferences



So, what is learning (for humans)?

- **Make sense** of a **subject, event** or **feeling** by interpreting it into our own words or actions
- **Use** our newly acquired ability or knowledge - in conjunction with skills and understanding we already possess - **to do something useful** with the new knowledge or skill

What is learning?

COLLECT AVAILABLE DATA (*ingest*)

+

GAIN KNOWLEDGE (*understand*, interpret data and transform it into knowledge)

+

**USE NEW KNOWLEDGE
TO DO SOMETHING (*act*)**



But, **how** do we
learn??

How do humans learn?

- Someone tell us (teacher, or watching others)
- Try and test (learning by doing) as in the fire example

Basically, ML systems learn in one of these two ways



There is only one thing more painful than learning from experience, and that is not learning from experience.

Laurence J. Peter

Is there something humans cannot
learn??



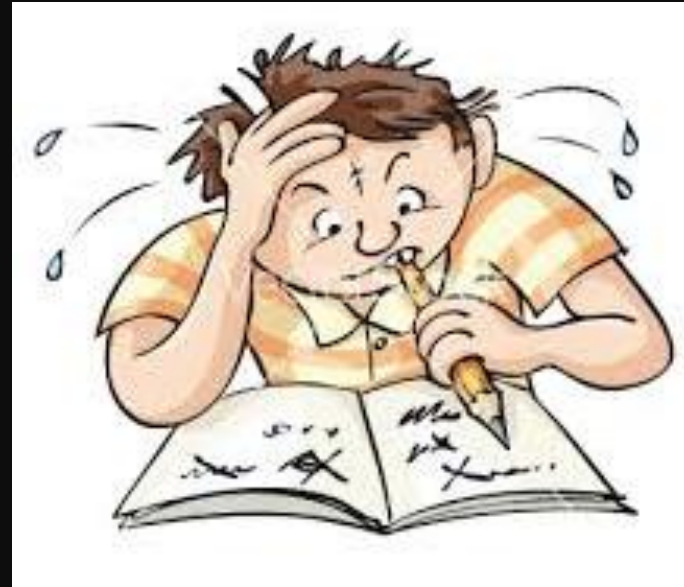
As a matter of facts, machines can learn to fly,
swim, run..

- Surprisingly, with rather different strategies ...

https://www.youtube.com/watch?v=4ZqdvYrZ3ro&feature=emb_imp_woyt

Besides things that humans cannot learn (but possibly machines can..), there are others that are either..

- Difficult to learn
- Difficult to teach



When is it difficult for humans to learn?

If there are **too many data**, humans cannot easily make sense of them (e.g. finding regularities in the human genome, learning to recognize one among millions of objects, market analysis and forecasts)



Stock market values
And quotes

When is it difficult
for humans to
learn?

If data **change too frequently**,
humans might be unable to
continuously adapt their
knowledge (e.g. personalized
recommendations, market
analysis forecast)



When is it difficult
for humans to learn?

If the environment is dangerous, “learning by doing” cannot be applied (e.g. rescue systems)



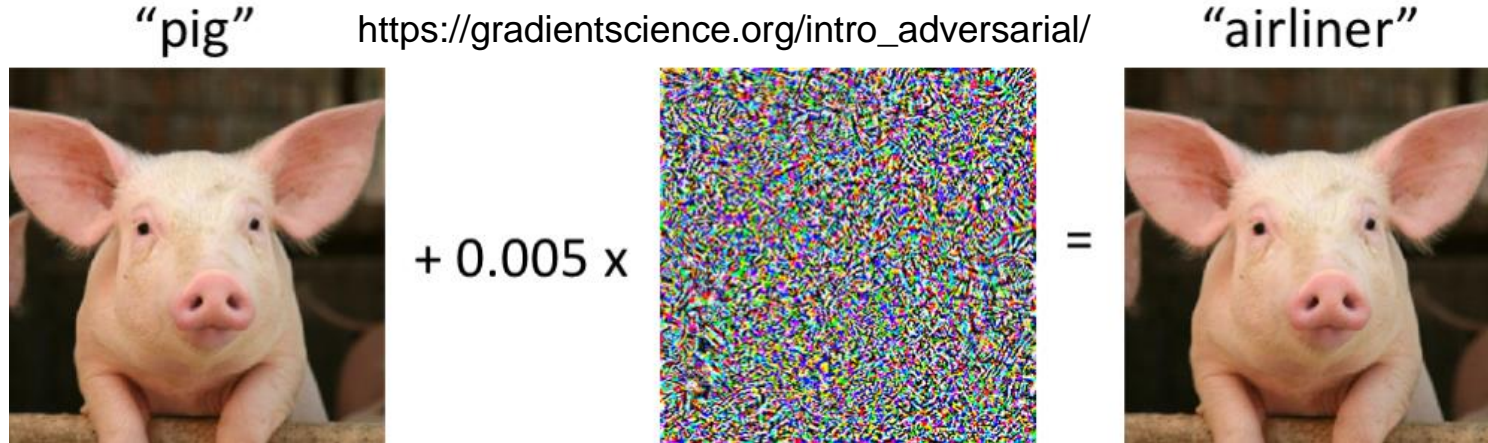
When is it difficult for humans to teach?

If there is not enough information or previous expertise to “understand and gain knowledge”

(we actually **do not understand** the image and speech recognition process by humans – it is not “teachable”)

Are machines better than humans?

- Yes in some applications and «to some extent» (e.g., games, precision surgery, image understanding..)



- However, using machines (in general) is not always advisable
- A useful question is: WHEN is the support of machine learning truly needed?

So when is
it
advisable
to use
Machine
Learning?

ML is used when:

- **No expertise**
 - Human expertise does not exist (navigating on Mars), or there is a danger
 - Humans are unable to explain their expertise (speech/image recognition)
- **Too many data**, data change frequently:
 - A solution changes in time (market data for market forecast)
 - A solution needs to be adapted to particular cases (personalized systems for a recommendation, diagnosis, etc.)

So when is it advisable to use Machine Learning?

- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (**knowledge engineering bottleneck**).
 - Expert systems (more frequently named *Decision Support Systems*)
- Develop systems that can automatically adapt and **customize** themselves to individual users.
 - Personalized news or mail filters
 - Personalized tutoring
 - Personalized therapies
 - Recommenders
- Discover new knowledge from large databases (**data mining**).
 - Customer preferences (learn from large samples of customers' shopping behaviours)
 - Medical text mining (electronic health records)
 - Social network mining (messages and friendship relations)
 - Emotion detection (from large datasets of people's images)

An interdisciplinary topic: many related disciplines!

Artificial Intelligence

Data Mining

Probability and Statistics

Information theory

Numerical optimization

Computational complexity theory

Control theory (adaptive)

Psychology (developmental, cognitive)

Neurobiology

Linguistics

Philosophy

ML is perhaps the most interdisciplinary of CS

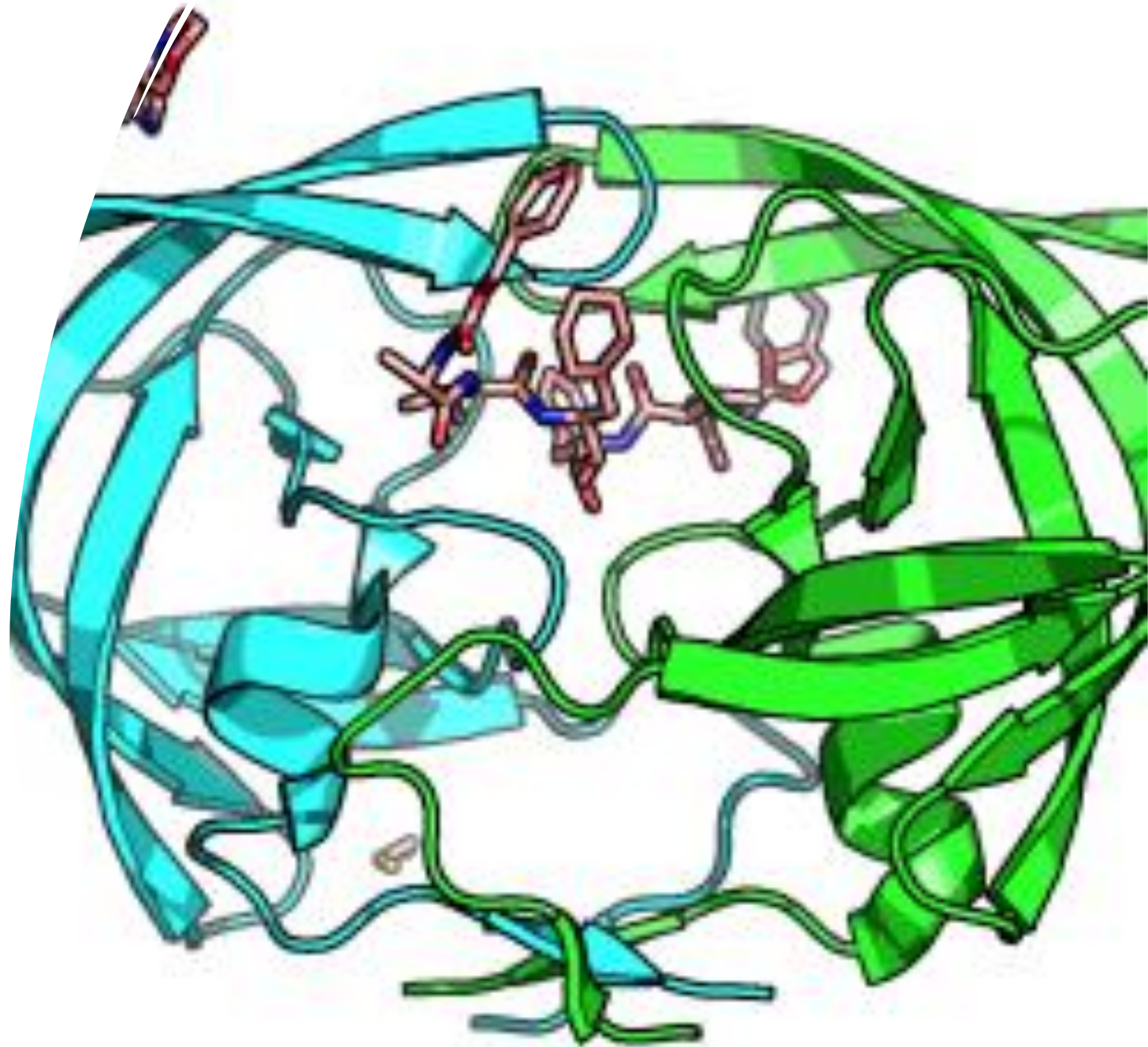
Some “real hot” ML applications

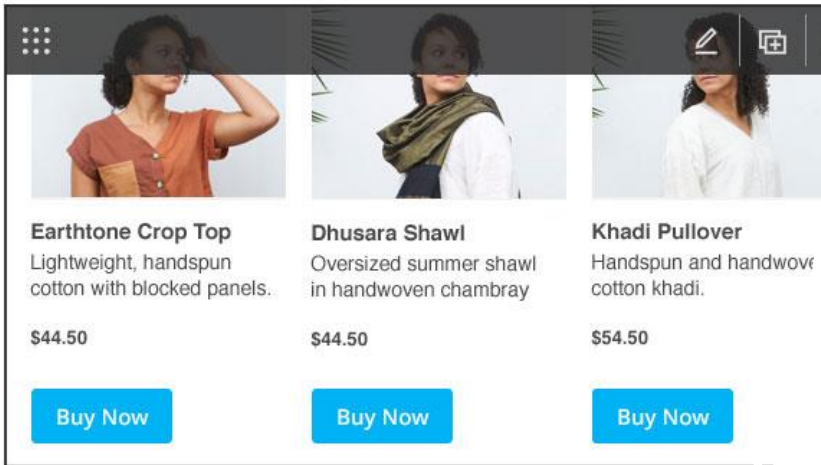
- It is really hard to find a problem where machine learning is not already applied -- machine learning is practically everywhere, in business applications and science!
- Let's see a list of (truly) “hot” applications...



Computational Biology & E- health

- Predicting diseases and complications from genomic data (metabolic, gene-disease relations, ..)
- Drug repurposing through the analysis of biological networks (e.g. interactions between proteins)
- Predicting epidemics through the analysis of human interaction data (e.g., population density, data on population movements, climatic data, etc.)

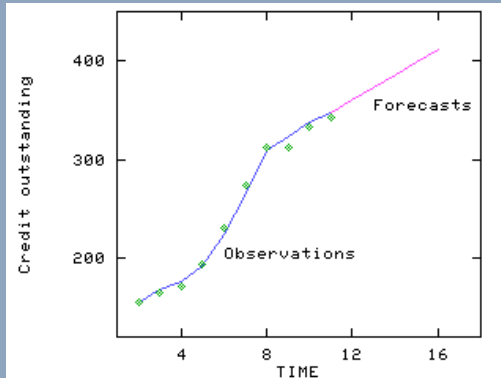




Web Search and Recommendation Engines

- Find relevant searches, predict which results are most relevant to us, return a ranked output (Google)
- Recommend similar products (e.g., Netflix, Amazon, etc.)

Finance



- Predict if an applicant is credit-worthy
- Detect credit card frauds
- Find promising trends on the stock market (*algorithmic trading*)

Text and Speech Recognition

- Handwritten digit and letter recognition at the post office
- Voice assistants (Siri)
- Language translation services

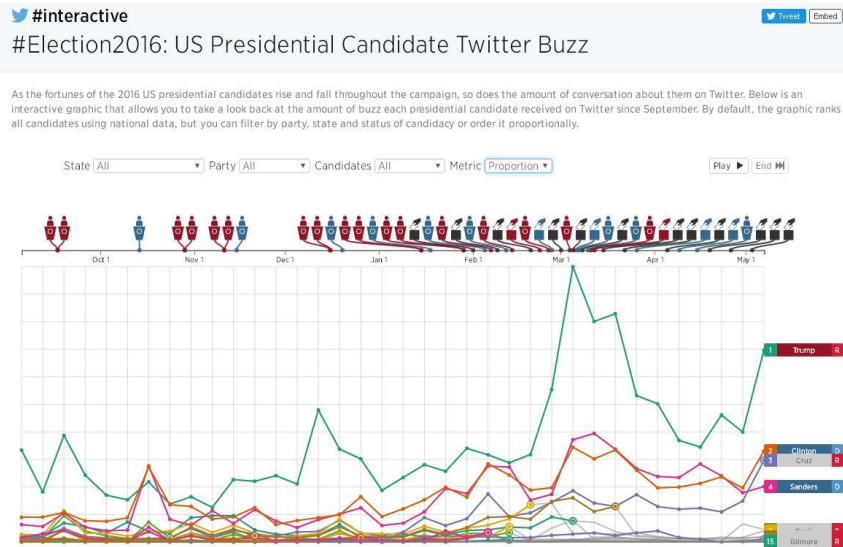




Image Understanding and Robotics

- Identification of relevant information (objects) in large amounts of Astronomy data
 - Robotics for industry, energy saving, and smart cities
 - Self-driving cars
-

Social Networks and Advertisement



- Social data mining:
 - data mining of personal information
 - Predict/analyze opinions, political choices, purchase behaviors

And the latest (based on OpenAI/LLM/GTP models)

- Generating code from NLP instructions (e.g., NL2SQL)
- Domain-general question answering
- Visual recommender systems: given data, recommend best charts to gain insight from data (DATA2Viz)
- Generating artistic images from NLP descriptions (NL2Images)
- Automated image captioning (IMAGE2NL)
- Summarization
- Automated problem solving
-exponential growth of applications

ML
Algorithm
types
(we will
shortly
analyze)

- Classification [Predictive]
- Clustering [Descriptive/prescriptive]
- Association Rule Discovery [Descriptive/prescriptive]
- Sequential Pattern Discovery [Descriptive/prescriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

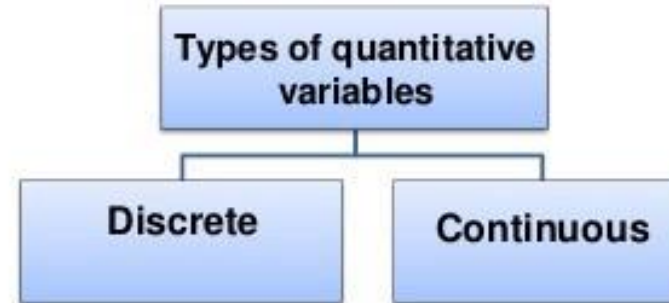
ML algorithm types

- **Classification and regression [Predictive]**
- Clustering [Descriptive/prescriptive]
- Association Rule Discovery [Descriptive/prescriptive]
- Sequential Pattern Discovery [Descriptive/prescriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

What is a classifier?

- Given records (instances) representing «data» of a domain (e.g., customers, products, patients, hotels..) and given a particular attribute (that we call also *feature*, or *variable*, or *descriptor*, or *field*..) describing these records (for example: being a returning customer, experiencing a failure, having a heart attack, being overbooked..),
- The algorithm learns, from past data, predicting the value of the attribute for the future (unseen instances)
- Attribute must be **categorical** (discrete, finite set of values), this is why is it called classifier

Categorical vrs continuous variables



A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child in an elementary school.

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

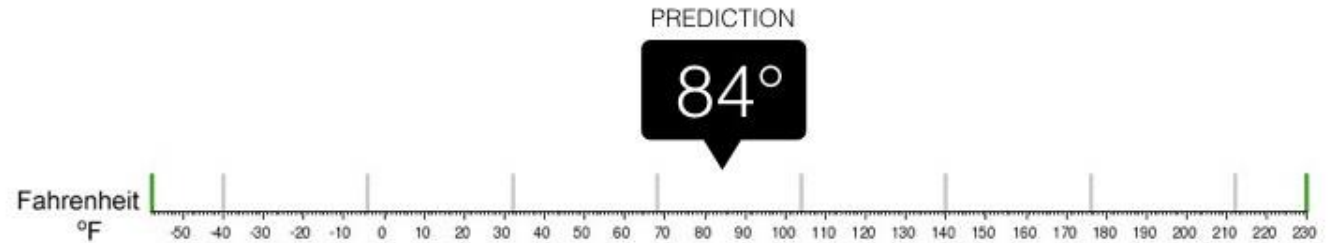
No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Classification vrs. Regression



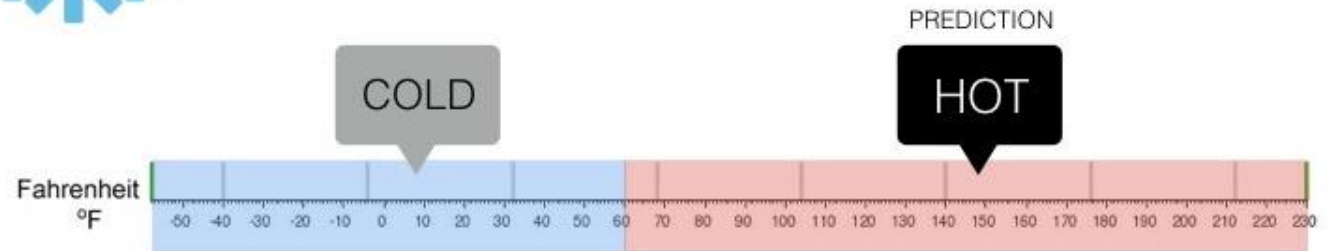
Regression

What is the temperature going to be tomorrow?

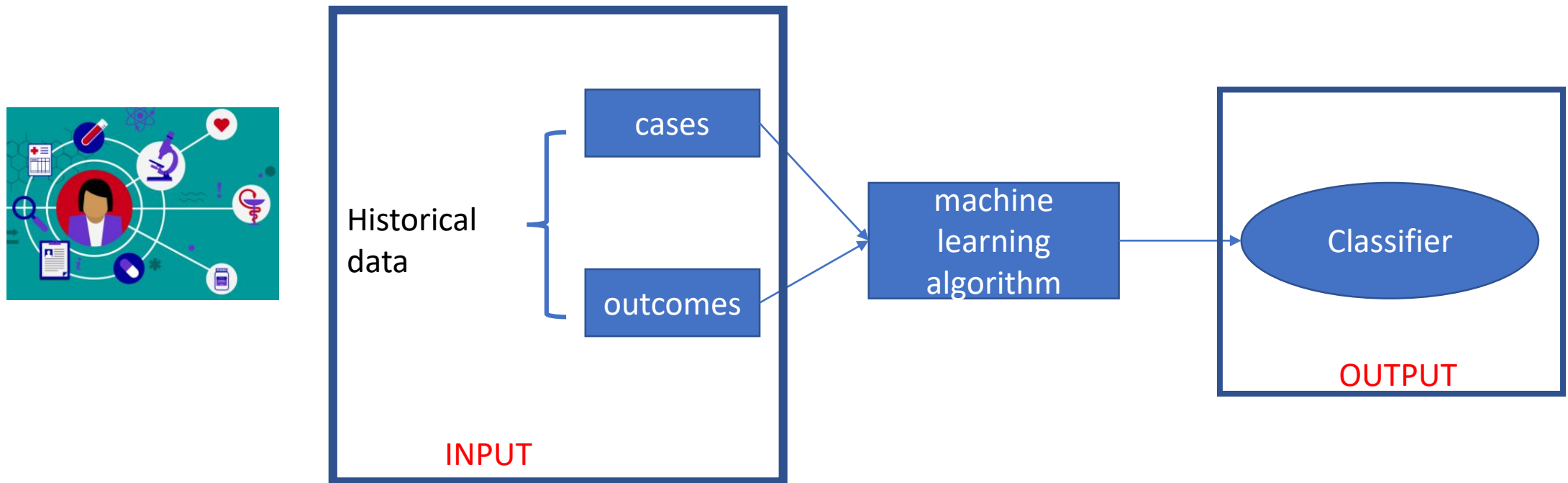


Classification

Will it be Cold or Hot tomorrow?



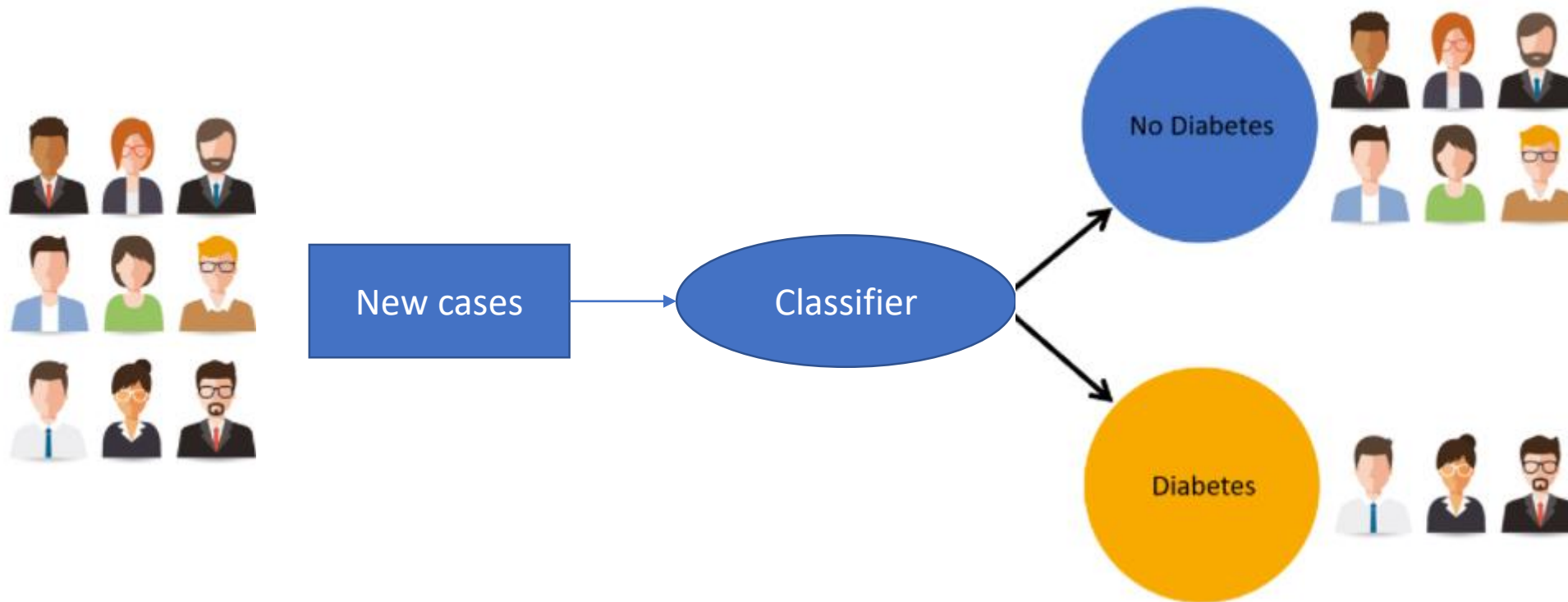
Classifiers: Training phase



Example:

- «cases» are clinical records and medical data of past patients; «outcomes» is the observed efficacy of a therapy, or the onset of a complication (e.g., diabetes)
- Classifier: a model (a function) that predicts an outcome on future patients (e.g., the onset of a complication)

Prediction phase

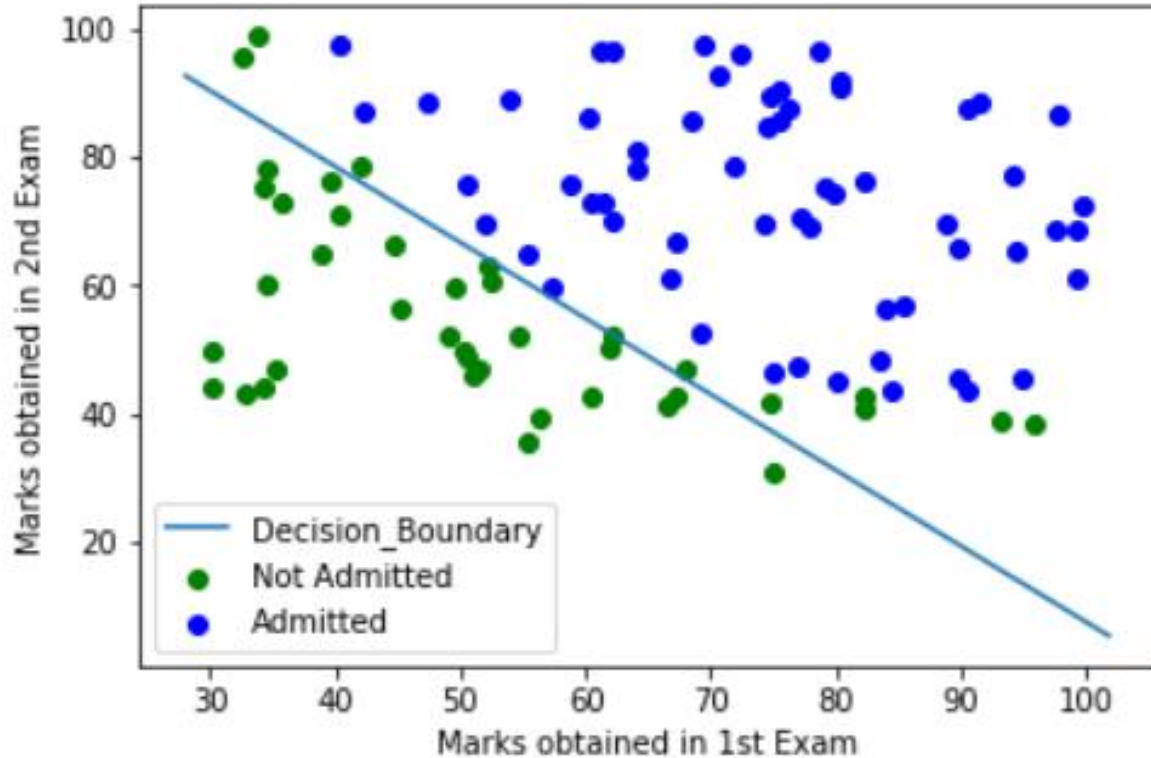


During the prediction phase, the predictive model (the classifier) becomes operational. It is provided with new data – e.g., data on new patients - and it uses the learned model to predict the future

Classifiers: a formal definition

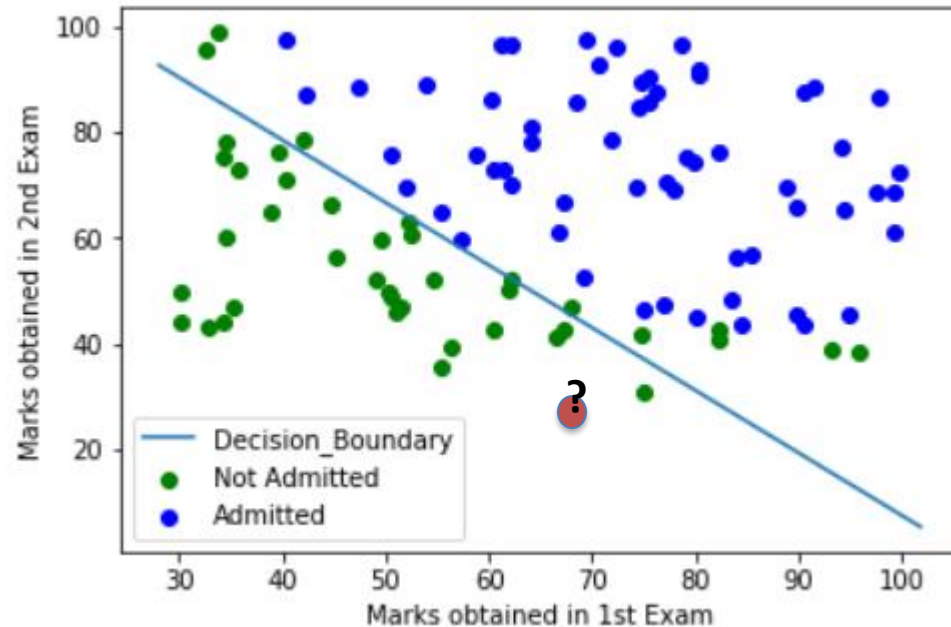
- Given a collection of records D (*training set*) in a given domain (e.g., customers of a bank)
 - Each record x_i in D contains a set of *attributes*, one of the attributes is named the *class* (= the attribute c whose value we would like to learn predicting, e.g., c =defaulter, with values yes/no). It must be a *discrete* variable (binary or multi-valued)
 - The training set represents historical data for which the value of c is KNOWN
- **Objective:** Learn a *model* $c(x)$ for the *class* attribute as a function of the values of other attributes.
- **Method:** the target is to learn a “mapping” function $c(x): x \rightarrow C$ such that $c(x)$ reproduces, for all records x in the historical dataset D , the same class values.
- **Formally:** given pairs (x_i, c_j) in $D, \forall x_i \in D, c(x_i) = c_j$
 - D is named learning (or training) set, and offers examples of correct classifications for a set of instances (the pairs (x_i, c_j))
 - 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.
 - If accuracy is sufficiently good on test data, the model becomes operational

Dataset of student's marks in the first and second exam



- Let's say we have records of student's results for the first second and third exam
- Binary class value: admitted or not admitted
- The table shows the distribution of historical data, blue are those who passed, green those who did not pass
- Objective of ML algorithm is to learn predicting from past data if the student will pass or not the third exam
- "learn a model" that will predict the future outcome for future students, given results of first and second exam

How is the prediction made?



- The «model» is some discriminative function $y=f(x)$
- In this example, $f(x)=mx+q$ is the blue line: it approximately separates the positive (admitted) and negative (not admitted) examples
- This line is used as decision boundary for future predictions
- When presented with a new data instance x (the red dot) the model uses the following classification rule:
- IF $f(x)<0$ THEN y =«not admitted» ELSE IF $f(x)>0$ THEN y =«admitted» (remember, $f(x)=0$ for all x lying on the blue line, $f(x)>0$ for all x lying above the blue line, $f(x)<0$ for all points lying below the blue line).

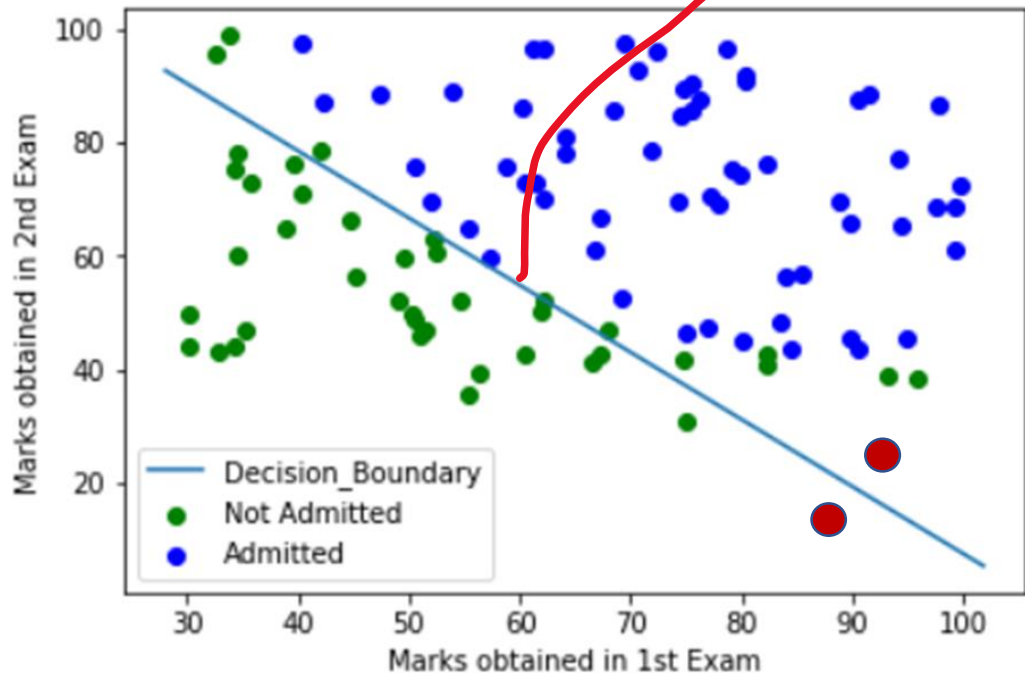
Learning a Classifier is in 3 phases

- 1. Training, or learning phase:** it analyses historical data where a classification (decision) has been taken (e.g. credit assignment, risk assessment, patient diagnosis..) and learn a classification MODEL
- 2. Test phase:** a portion of historical data not used during training is submitted to the model for classification. Model predictions are compared with «ground truth» to assess quality of the model. If quality not sufficient, data pre-processing and model learning algorithm are improved
- 3. Prediction (operational) phase:** the model becomes operational, new unseen data are submitted to the model for classification and decision support

What is a “Model”?

- A model is a function $y=f(x: a_1, a_2..a_n)$ where y is a value, and $\mathbf{x}: a_1, a_2..a_n$ is the set of *attributes* or features of instances (records) \mathbf{x} of a dataset (e.g. e.g., \mathbf{x} is a credit card applicants, or a patient, or a product, etc. represented by a set of attributes).
- Learning a model means that for **every combination of attribute values**, the system is able to compute the value of $c(x)$
- This ability mimics human ability of generalization: from few evidences we learn a strategy that can be applied also to cases that we have not seen before.
- A model can be any algebraic, logic or probabilistic function
- If the values of $f(x)$ are DISCRETE, then we have a **classifier**, if they are continuous, then we have a **regressor**.
- For example, if, given a record describing an applicant credit history, we want to learn a model that predicts if it is a possible defaulter, the model is a CLASSIFIER ($f(x)$ takes 2 discrete values, defaulter and not defaulter)
- If the model must predict a future stock market value given its past values and other market-related attributes, then $f(x)$ is a continuous value, and the model in a regressor.

What does it mean model learning?



- The «model» here is the linear separator $y=mx+b$ represented by the blue line, where m and b are initially **unknown** (they are called the «model parameters»)
- Parameters m and b of the line are *learned* based on the examples by «some» ML algorithm
- Once the line is learned, then, we can use the model for predictions, as explained before

Example 2: Predicting loan defaulters

Attribute Name	age	Job profile	Income	Emp Length	Loan Amount	Duration	Purpose	Housing	Loan History
Type	Numeric	Nominal	Numeric	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal
Values Example	33, 50, 46	Less, Moderately, Highly skilled	5000, 80000	1, 8, 15	1200, 3500, 11000	12, 24, 48 in months	Car loan, House Loan, Business Loan etc	Rent, Own, Free	Defaulter, Not Defaulter

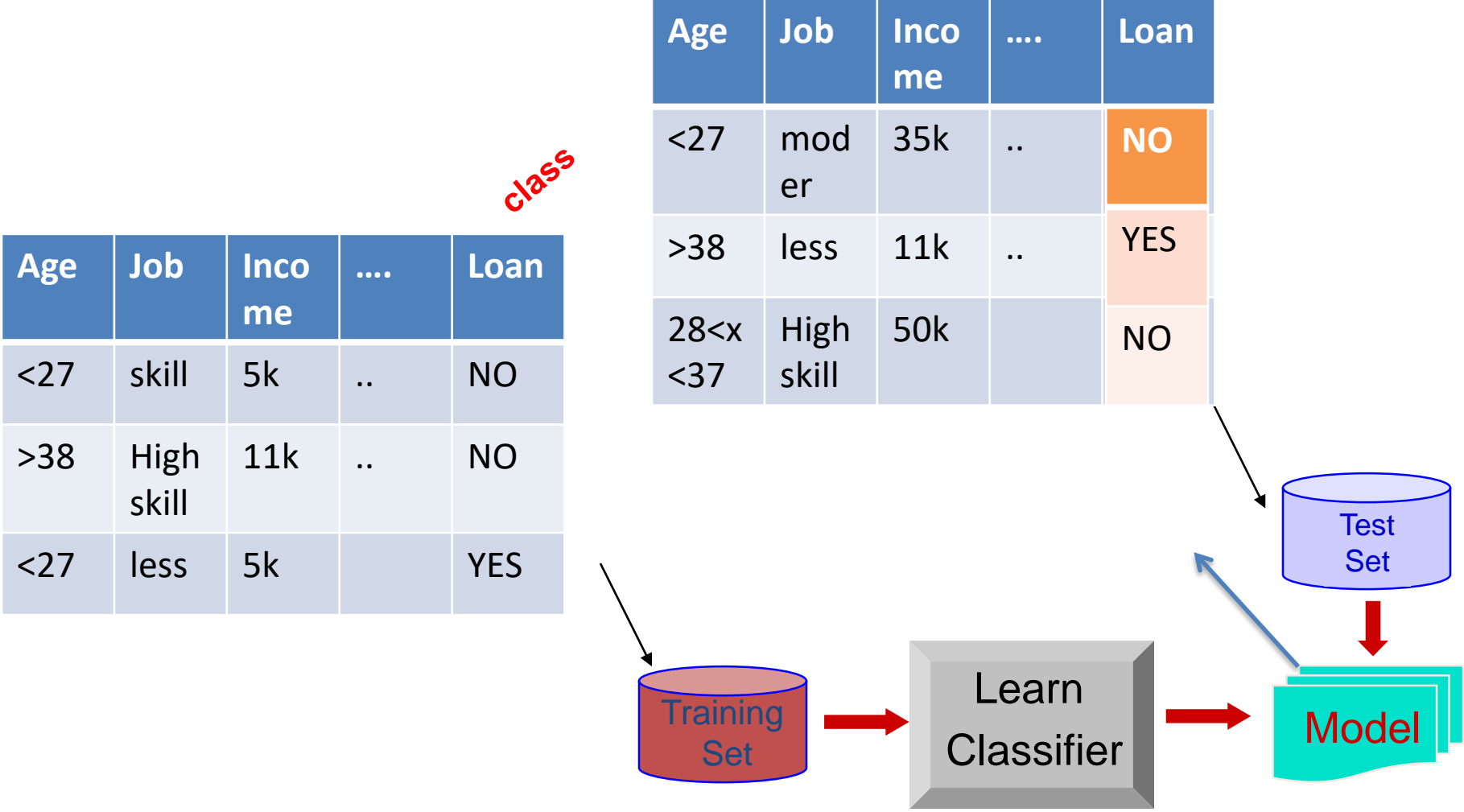
- Input data are «histories» of previous loan receivers.
- In the ETL step, «relevant» (for the task) attributes have been selected
- One of the attributes must be the one that we would like to learn predicting. This is named «class». It must be discrete (that is, it may take a **finite** set of values)
- In this example, the *class* is «**Loan History**», and has 2 values, defaulter or non-defaulter (it is a binary variable).
- **Objective:** given past histories, train a classifier algorithm to learn predicting the value of this class (defaulter or non-defaulter) – and then, use the classifier to predict if new customers (not in the record of past customers) will be defaulters or non-defaulters, and use this prediction to decide whether to grant credit or not.

Processing the DW

Attribute Name	Age	Job profile	Income	Emp Length	Loan Amount	Duration	Purpose	Housing	Loan History
Type	Nominal	Nominal	Numeric	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal
Values Example	<=27, 28<=X<=37, >=38	less, moderately, highly skilled	5000, 80000	<1, 1 to 4, 4 to 7 7+	1200, 3500, 11000	36, 60 in months	Car loan, House Loan, Business Loan etc	rent, own, other	Defaulter, Not Defaulter

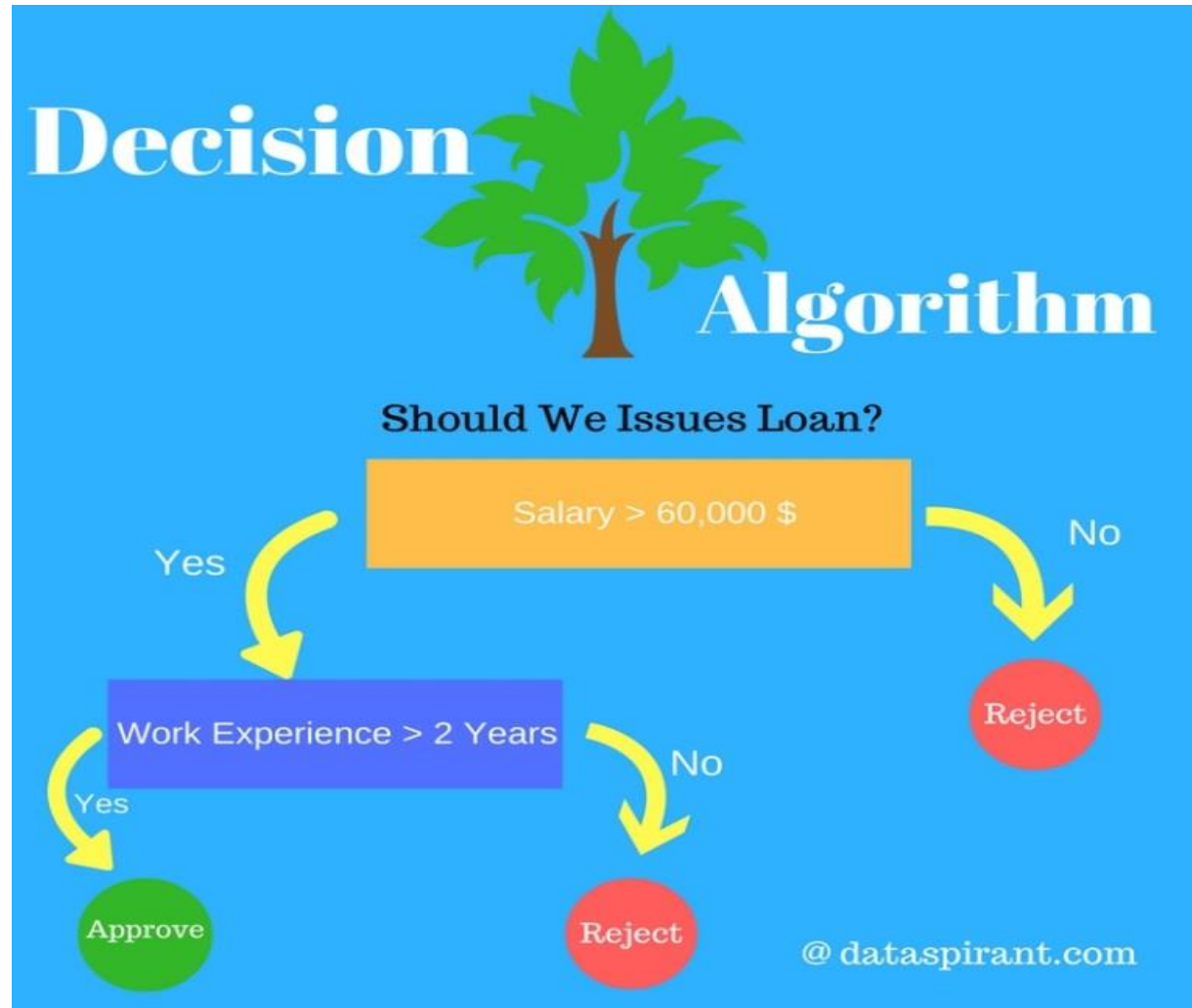
- If some attributes have too many values, it might be difficult to find regularities in the data.
- In this case, age and employment length have been discretized (values are replaced with intervals)

Example of train/test phases



Note: actually the values of the "class" attribute are known also for the test set, but they are not provided to the system, in order to test its accuracy.

A very simple example of classifier (decision tree)

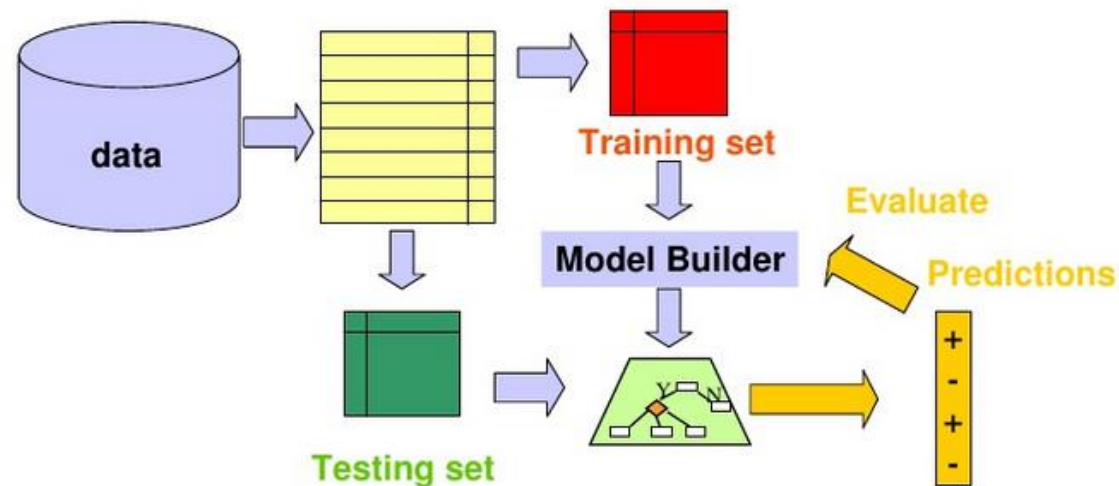


Example: loan application (with simpler attributes..)

- Problem: a bank has data on previous granted loans. For each customer, the following *attributes* (all with finite *DISCRETE* values) are available:
 - **Age** (young middle old)
 - **Has_job** (true, false)
 - **Own_house** (true, false)
 - **Credit_rating**: (fair, good, excellent)
 - **Loan** (yes, no) (whether the loan was granted or not)
- The last attribute is called “class”. It represents the (binary, in this case) variable y for which we wish to learn a predictive model:
$$loan=f(age,has_job, own_house,credit_rating)$$
- **Task**: learn a model to help deciding whether to grant a loan or not to new customers based on history of past granted loans.
- When the model is learned, given a new customer and his/her attributes, the model predict the class *loan* (with values *YES* or *NO*), i.e. it suggests if the loan should be granted or not
- Clearly the example is very simple: in the reality, many more attributes (even order of millions!) are considered to build a model

Example: Loan Application (2)

- How do we learn the model?
- Learn the model from a fragment of available data (training set) based on previous experience (e.g. loan histories where we know if the recipient could refund the loan or he/she was a defaulter)
- Evaluate the quality of predictions using another fragment of the available data (test set)
- If performance is good (e.g., error rate is $<2-3\%$) then use the model to help decisions on new customers



Example: Loan Application (3)

ID	Age	Has_Job	Own_House	Credit_Rating	Defaulter
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

A simple training set

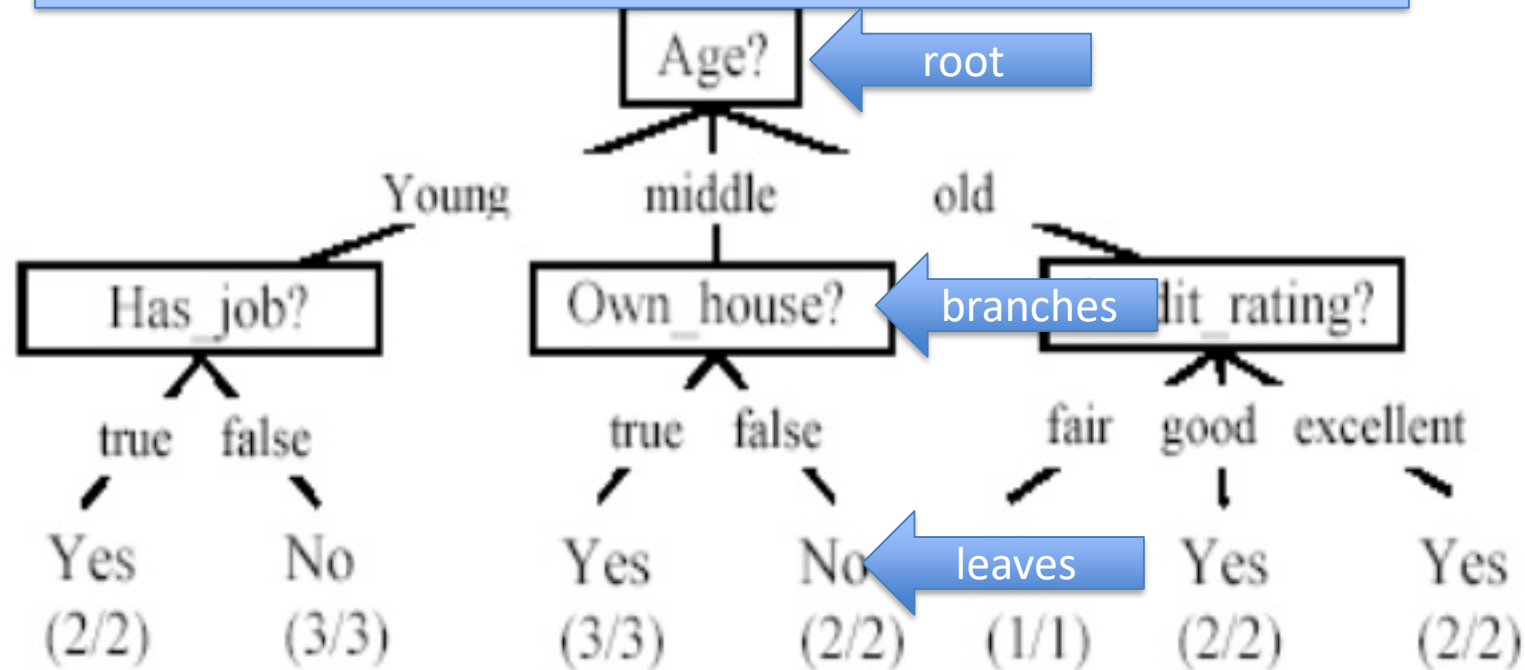
Example: Loan Application (4)

- So we have data on past loans. Each record represents a previously processed customer request, with 4 attributes and one class.
- Both the class and the attributes have symbolic discrete (not real-valued) values
- What kind of prediction function we can learn?
- As we said, functions can be algebraic (e.g. polynomial, exponential ..) logic (e.g. rules) or probabilistic (a probability is learned for each class value)
- We here introduce Decision Trees, a type of **logic function (it learns RULES)**

A decision tree from the loan data

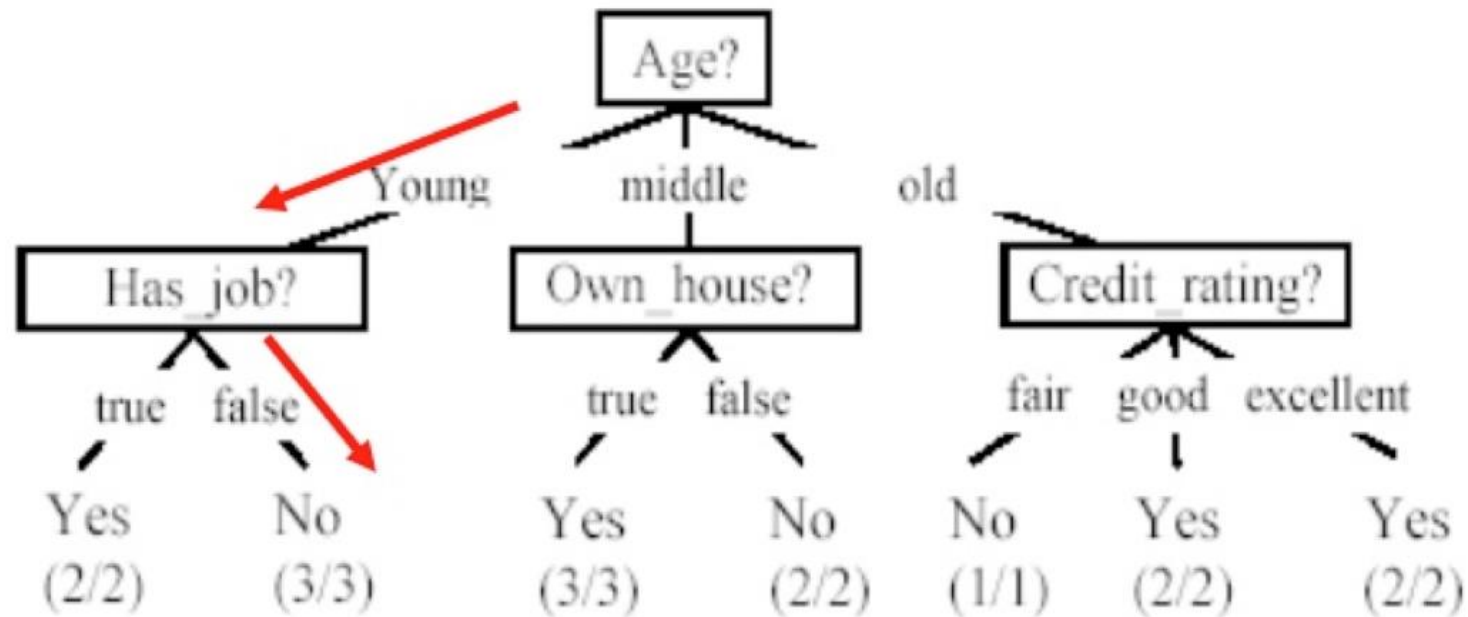
Decision nodes and leaf nodes (classes)

A tree has a root, branches and leaves



Root and branch **nodes** are TEST on attribute VALUES; leaf **nodes** are DECISIONS on class values; branch **edges** are labeled with attribute values.

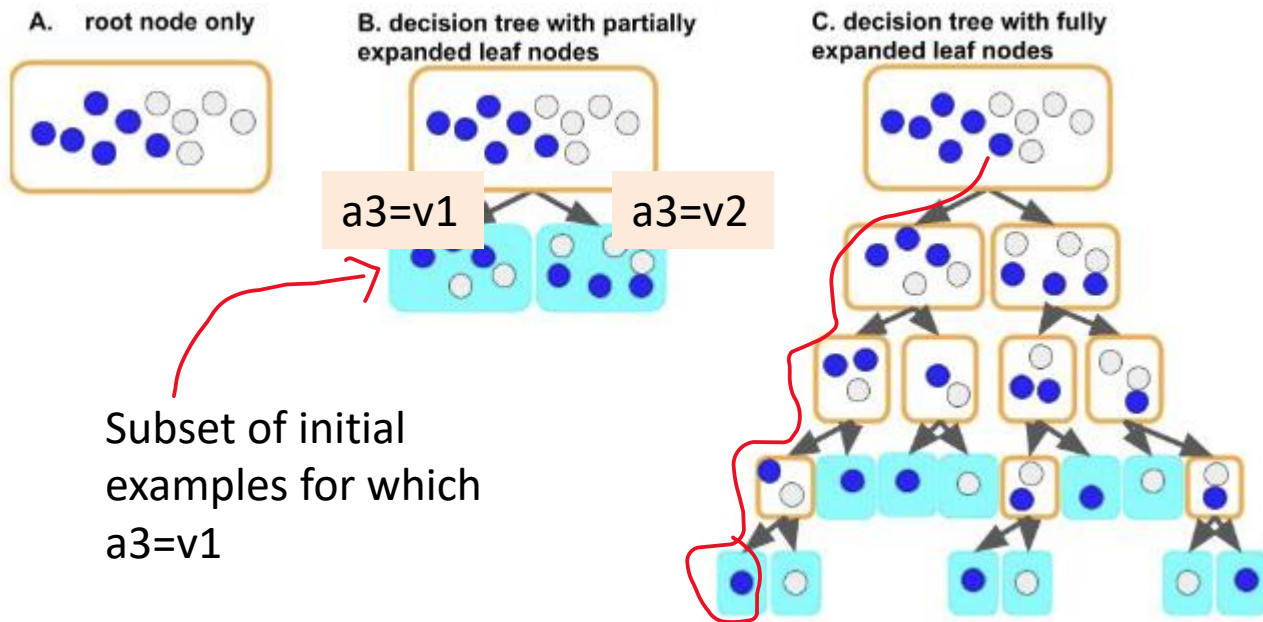
A decision tree from the loan data (2)



Every path from the root to a leaf can be interpreted as a RULE, e.g.,:

IF Age=Young AND Has_job=false THEN loan=NO

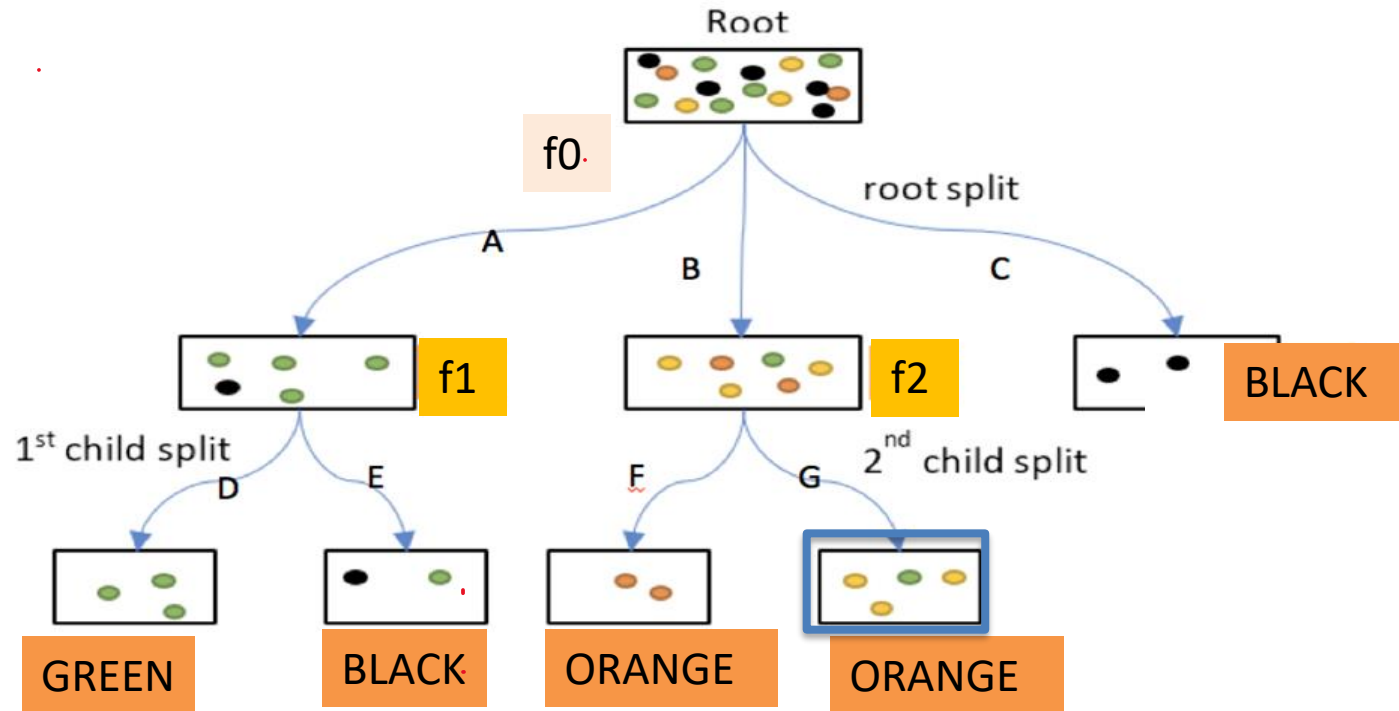
How is a DT generated?



A high-level example

- Blue and grey dots are the «labelled» examples in the training set, described by a set of attributes (let's call them a_1, a_2, a_3, \dots)
- The objective is to split our data in subgroups based on the values of attributes such that all the members of a group belong to the same class («blue» or «gray»)
- At each step, the algorithm search for the attribute that allows a split of the data which is «closest» to the objective (let's say, in step B, that we split the data according to attribute a_3 , whose value are, e.g., v_1 and v_2)
- When all the examples in a subset have the same label, then, output a decision
- The red path corresponds to a rule:
IF $a_3=v_1$ AND $a_1=v_4$ AND... THEN class=blue

Support and confidence of rules



Rule R: IF f0=B AND f2=G
THEN ORANGE

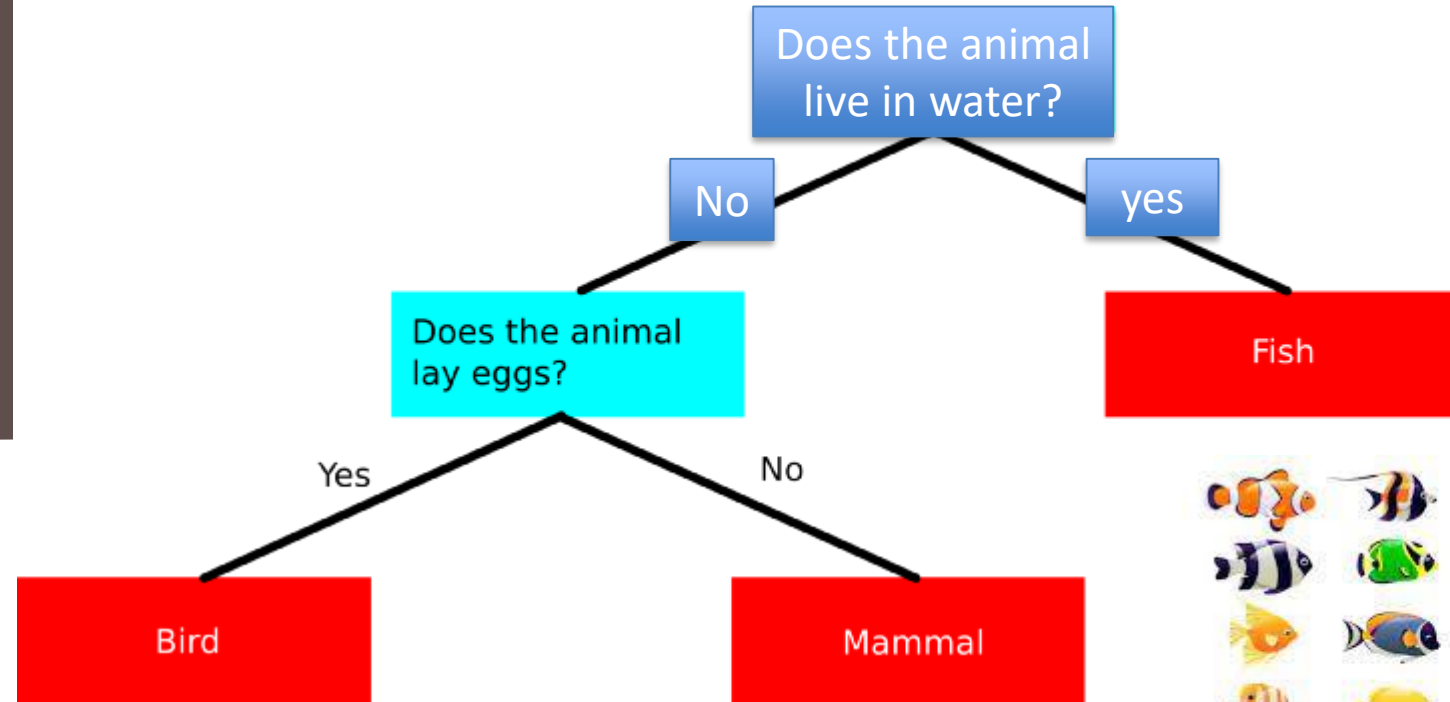
$$S(R) = \frac{3}{15}, C(R) = \frac{3}{4}$$

104

$S(R)$ the support of a rule is the ratio between the examples in the dataset that satisfy the rule, and the total number of available examples

The confidence $C(R)$ of a rule is the ratio between the examples that satisfy both the conditions (the IF part) of a rule AND the consequence (the THEN part) and the number of examples that satisfy only the condition. If the numerator is lower than the denominator, it means that the rule is not 100% correct, given the data.

Another example



A whale lives in water but it is a mammal! So out of 21 examples 1 does not match the rule!

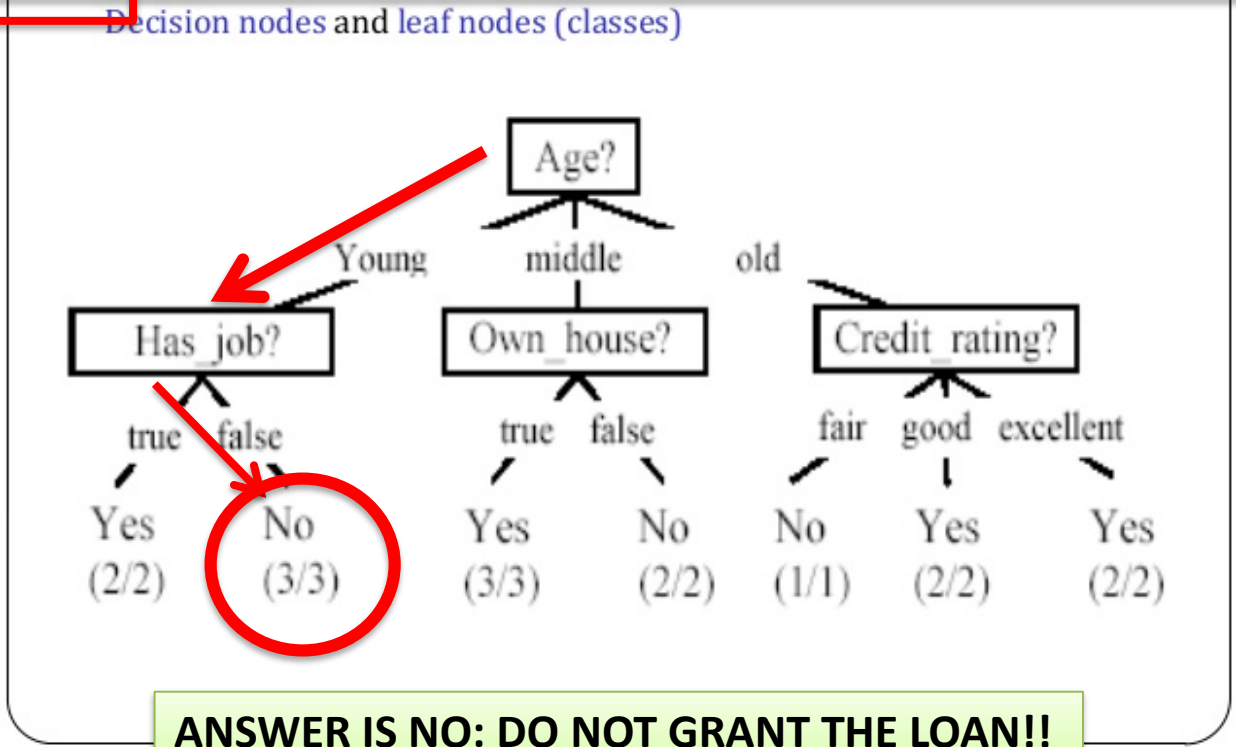
Once a model is learned, it can be used for new predictions

A new loan request

Age	Has_Job	Own_house	Credit-Rating
young	false	false	good

?

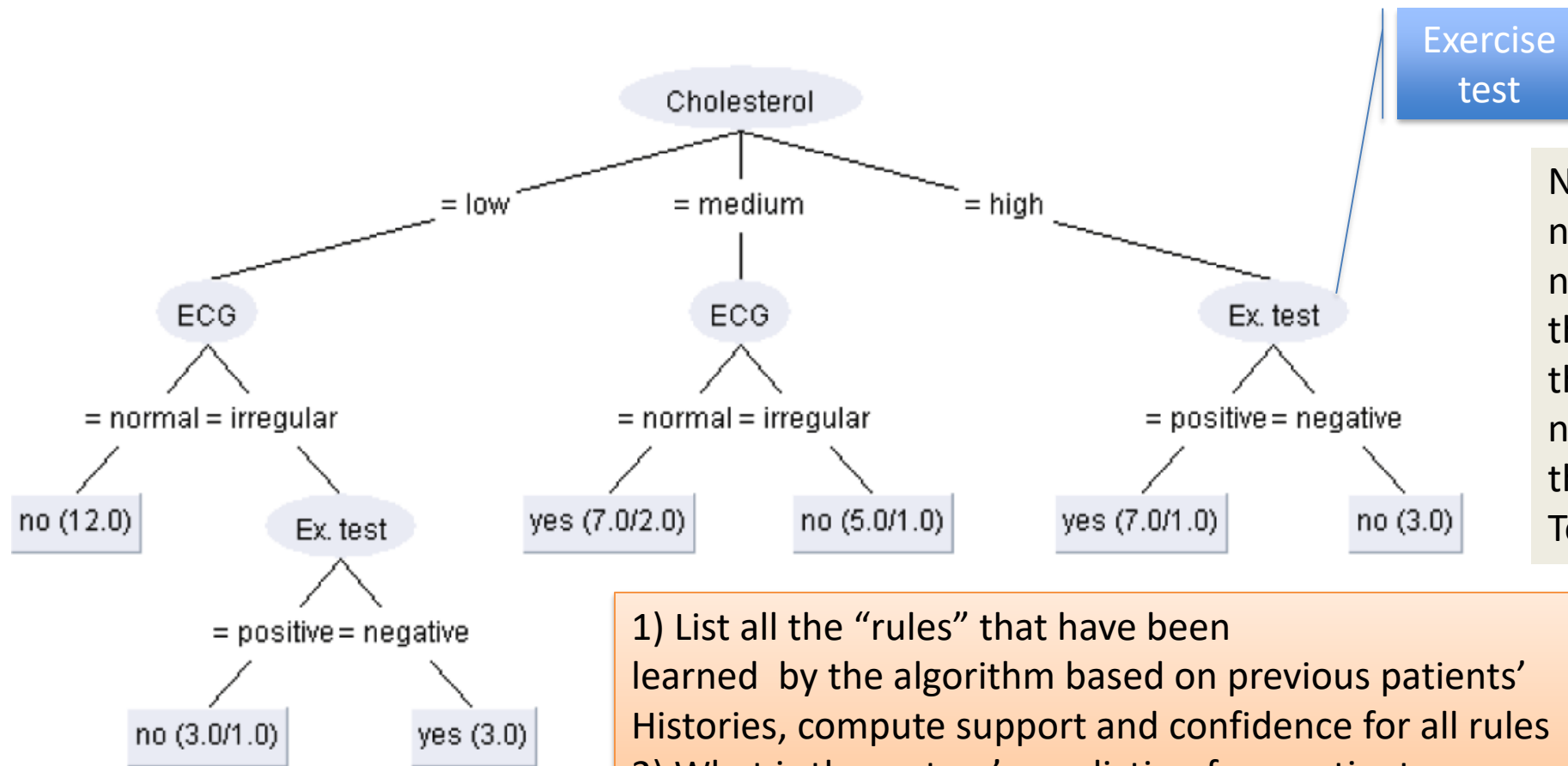
Decision nodes and leaf nodes (classes)



How do we actually learn decision trees from data?

- Several ML algorithms with code are freely available
- However, “Not your business” learning how they work and how to use them (although in Watson you will run a few examples)
- Decision trees have an advantage: they provide RULES in output. **Expert users can inspect generated rules and select those that seem more reliable and/or convincing**
- In general, in business it is always better to select data mining algorithms that provide an EXPLANATION of their learned model. Deep (state of the art) ML algorithms are black boxes!!
- Interpretability is a recent research area in machine learning
- Fairness and bias are other relevant issues

In class exercise: can you “read” this Dtree?



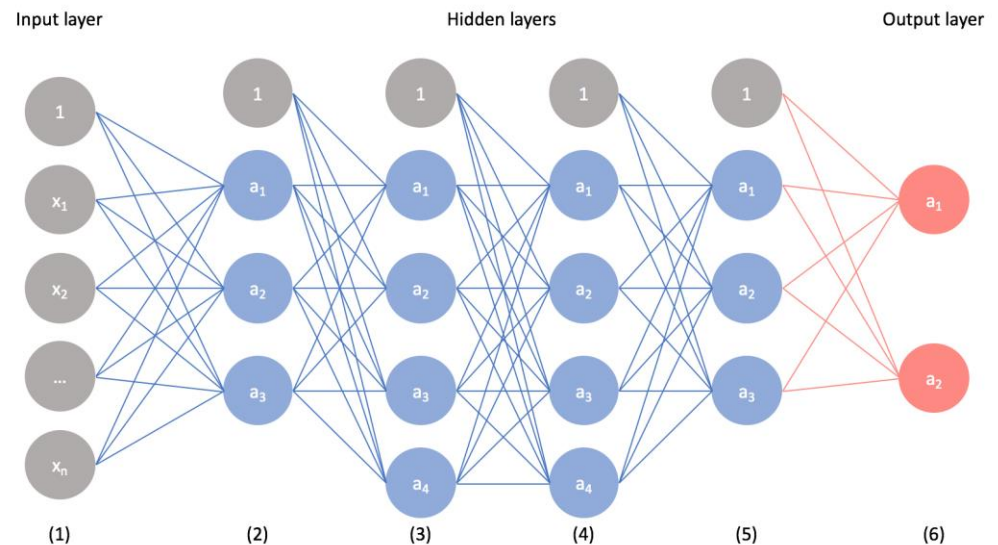
Note: the first number is the numerator of the support the second the numerator of the confidence
Total 45 examples

- 1) List all the “rules” that have been learned by the algorithm based on previous patients’ Histories, compute support and confidence for all rules
- 2) What is the system’s prediction for a patient with high cholesterol and negative exercise test?

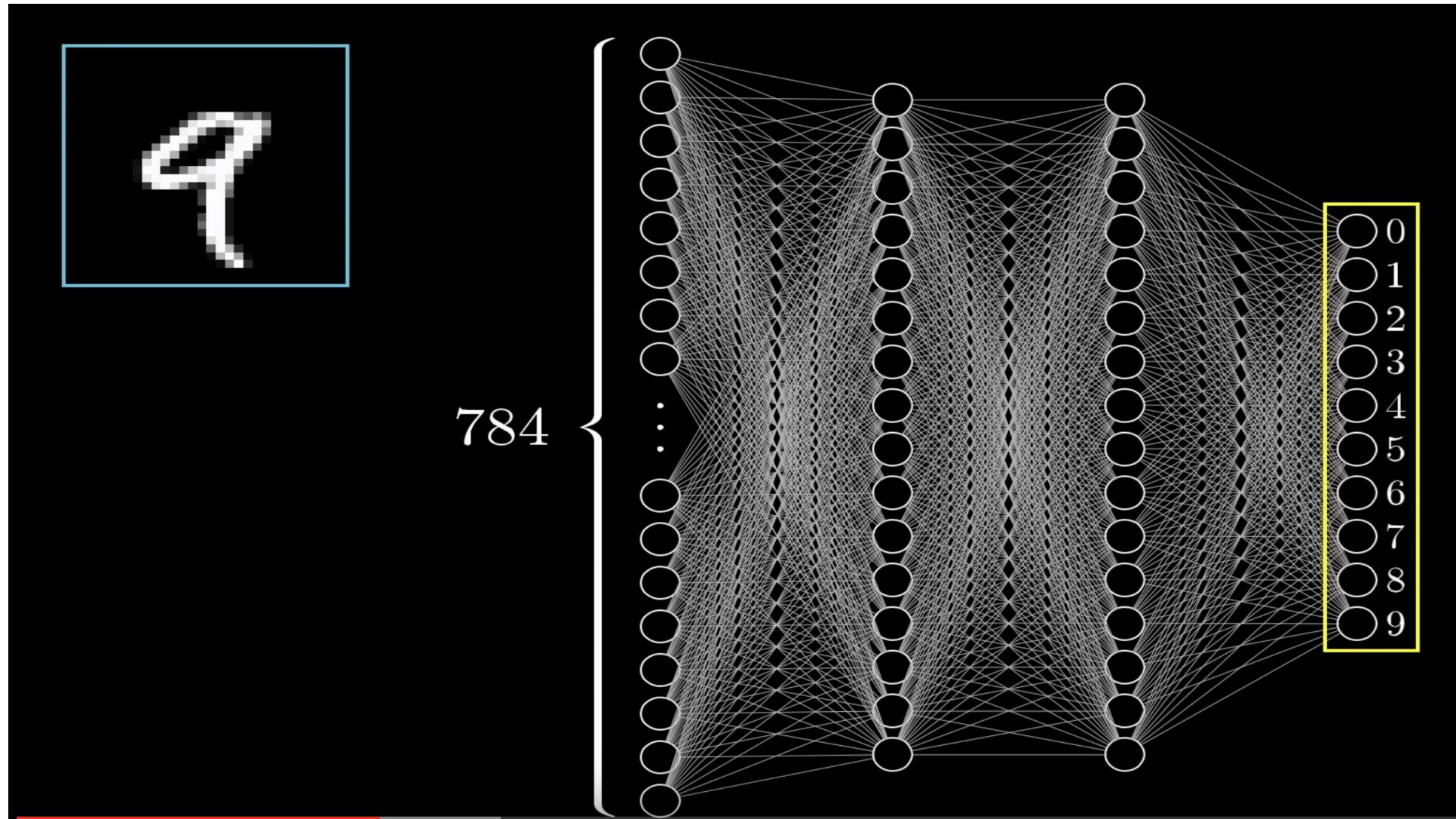
“class” is *risk_of_diabetes*, with two values: yes or no

Classification algorithms

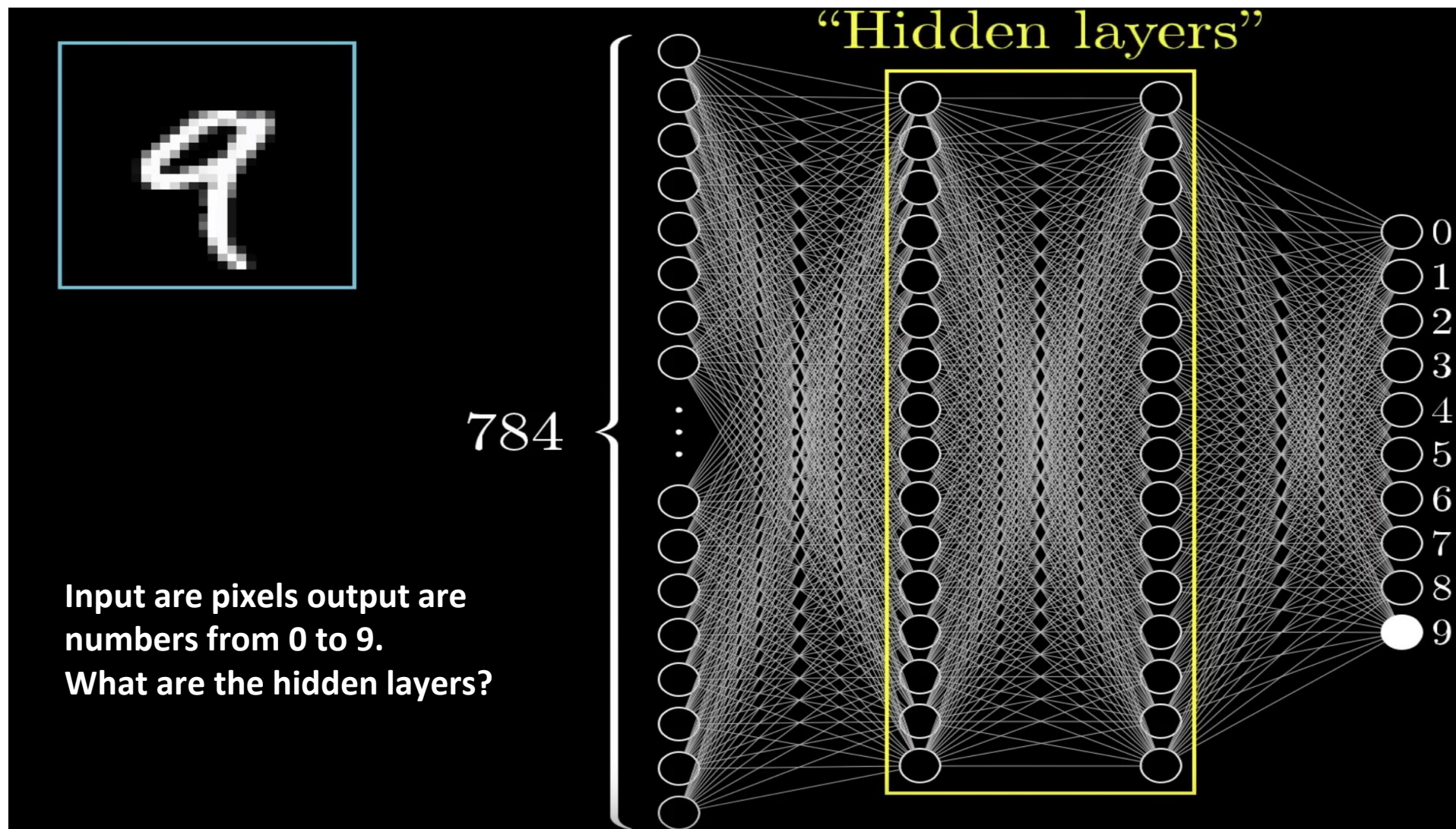
- We have shown decision trees because they are easy to interpret
- State-of-the-art classification algorithms are much more complex, often not based on logics. Rather, they learn algebraic functions
- Example: Neural networks



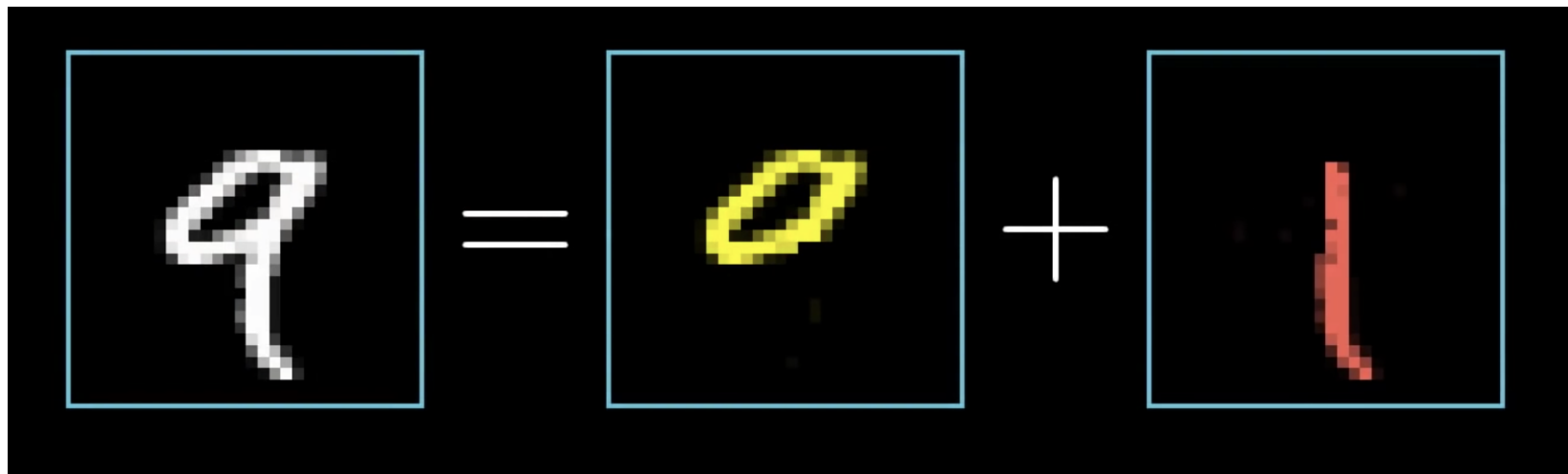
A visual intuition of how neural networks work (character recognition)



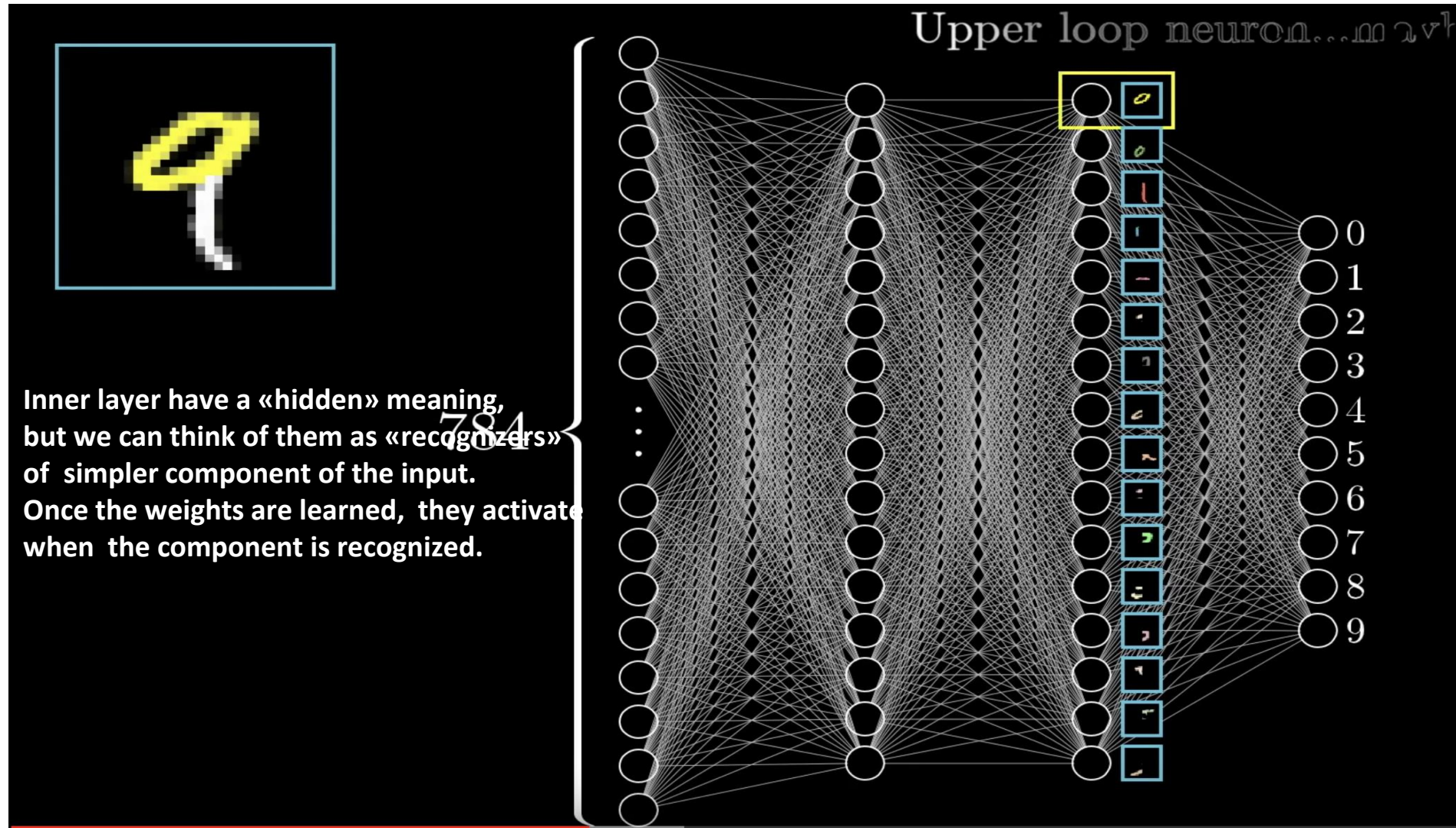
A visual intuition of neural networks (2)



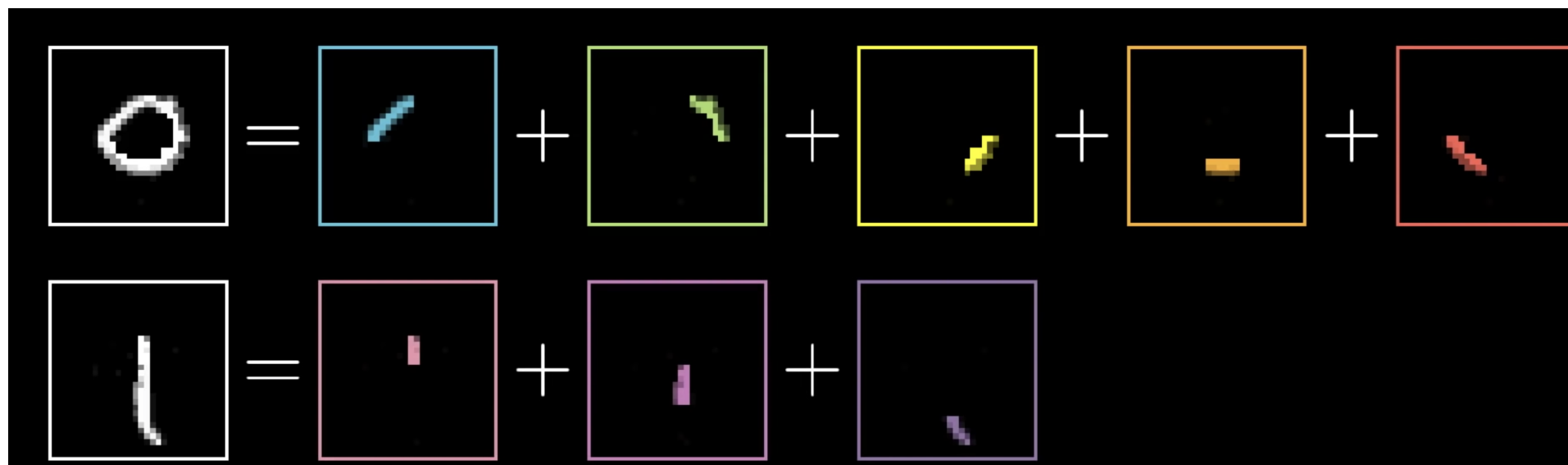
A visual intuition of neural networks (3)



An intuition of neural networks (4)



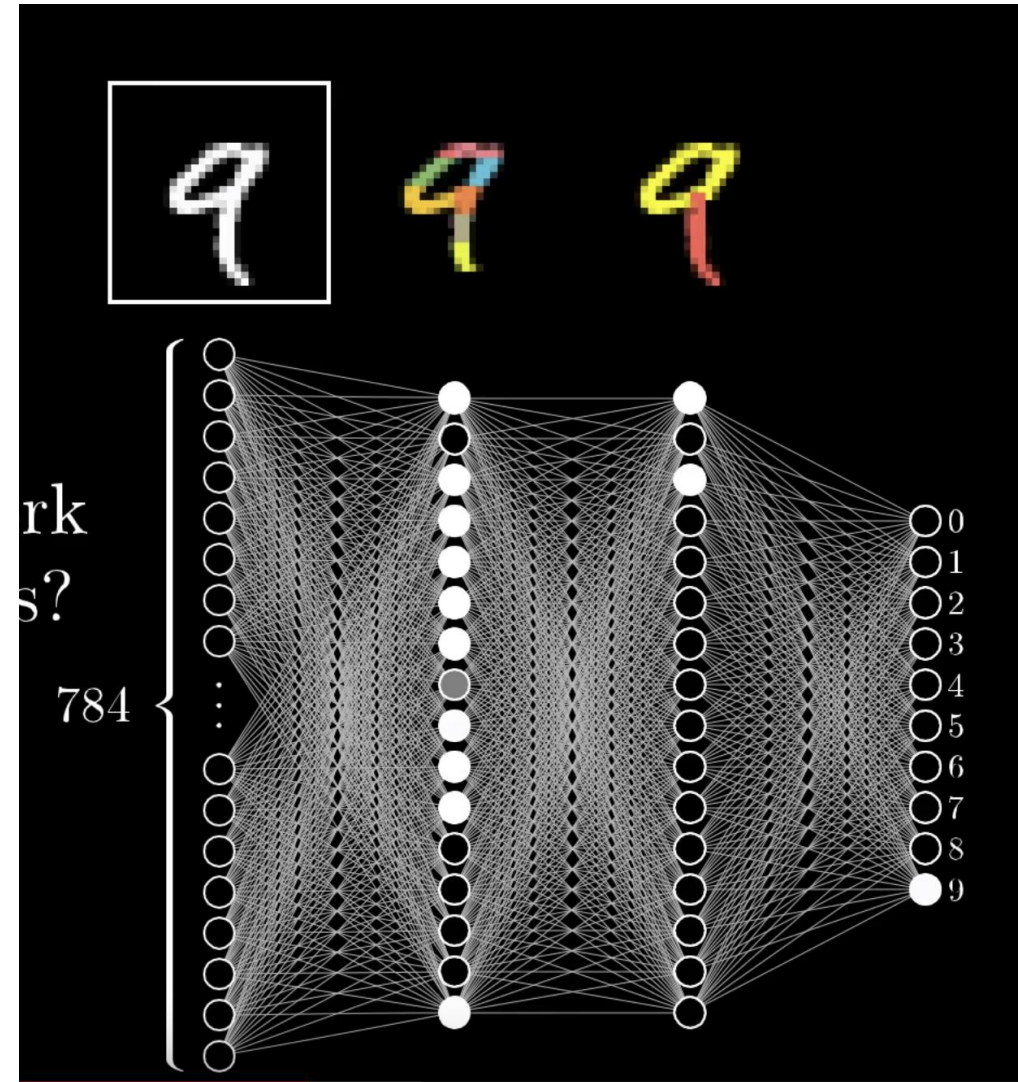
An intuition of neural networks (5)



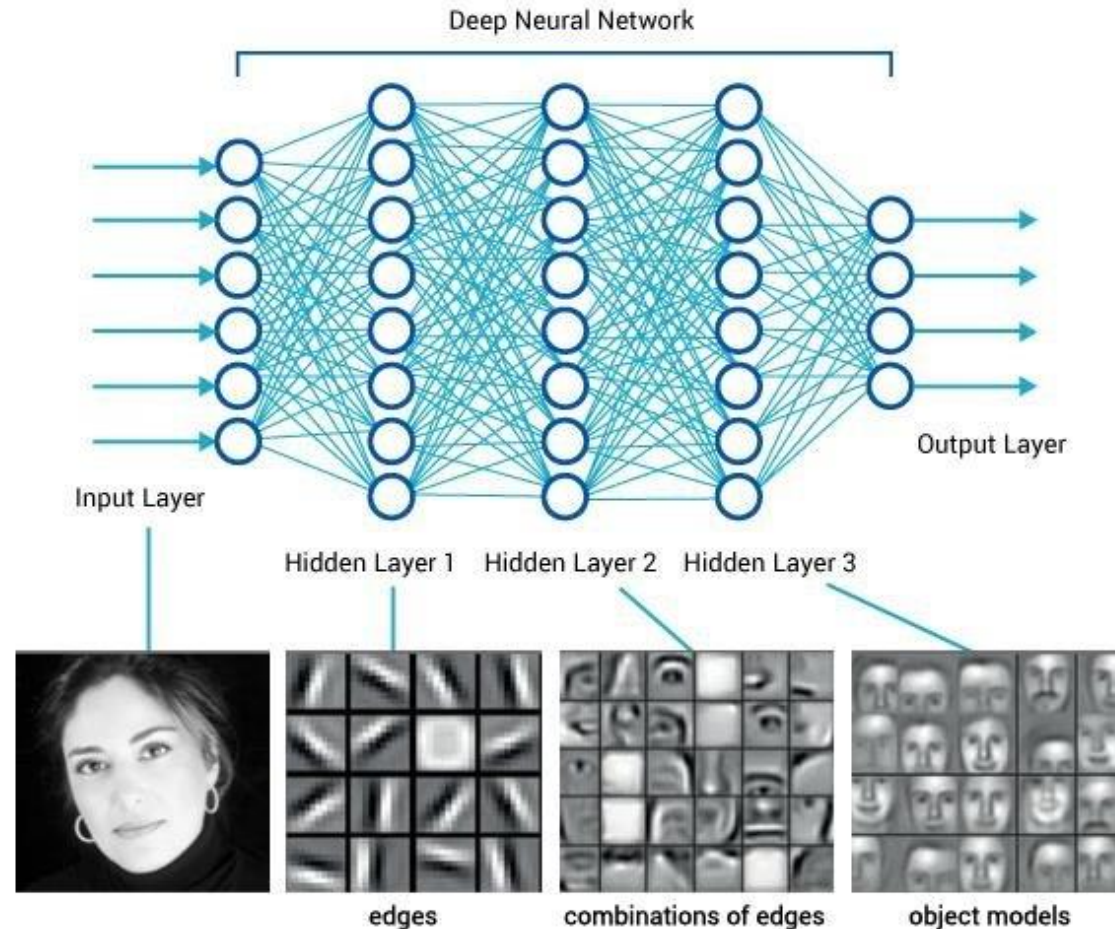
A visual intuition of neural networks (6)

Layer 2 recognizes smallest component, layer 3 more complex component.

HOWEVER this is an intuitive explanation. **We can't actually tell the network does this..**

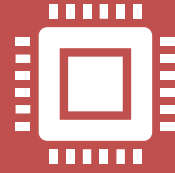


Another visual intuition: face recognition



HOWEVER this is an intuitive explanation. We can't actually tell the network does this..

What kind
of models
do these
algorithm
learn?

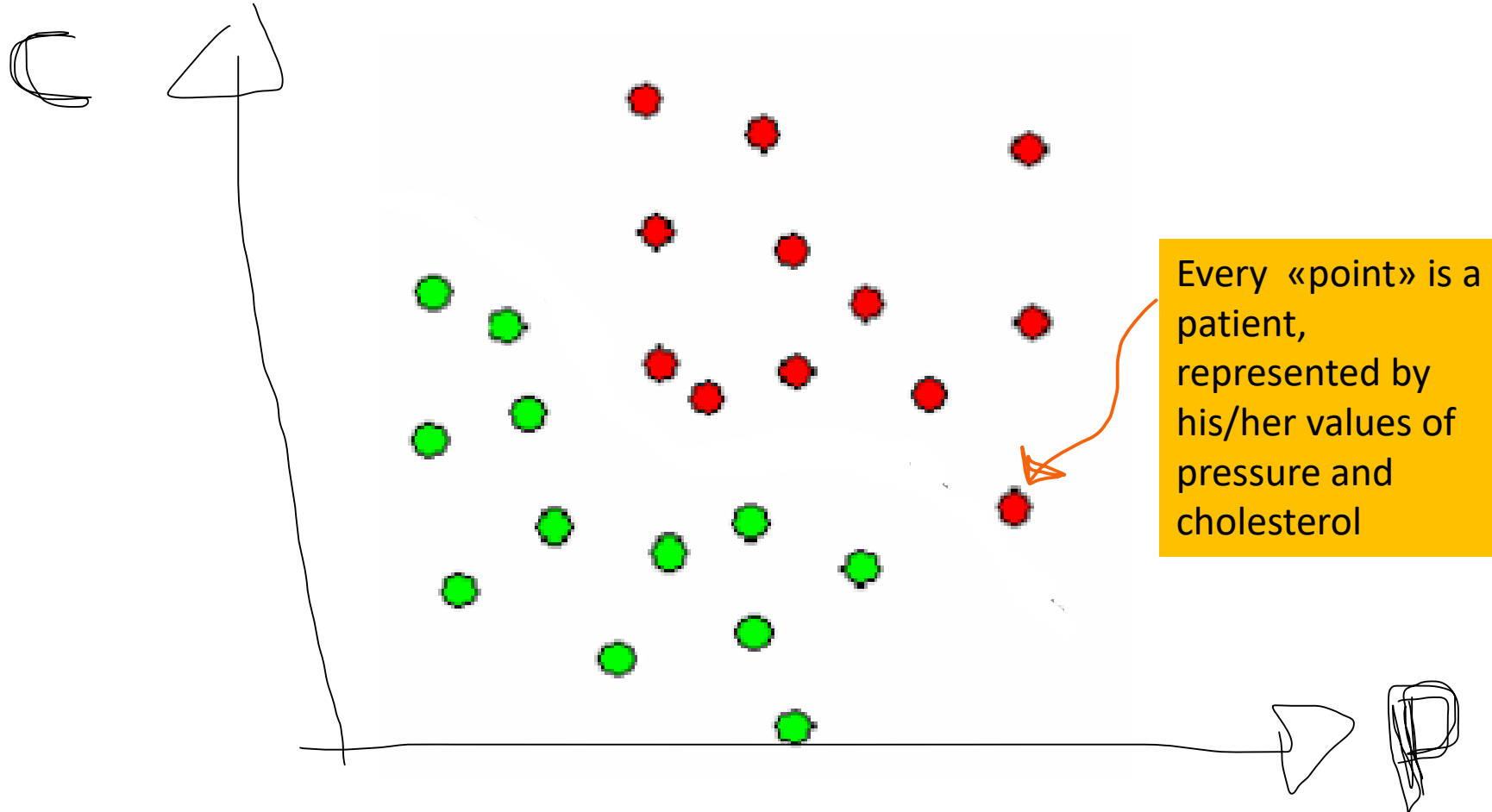


Several off-the shelf and deep algorithms learn a model that is based on an algebraic function



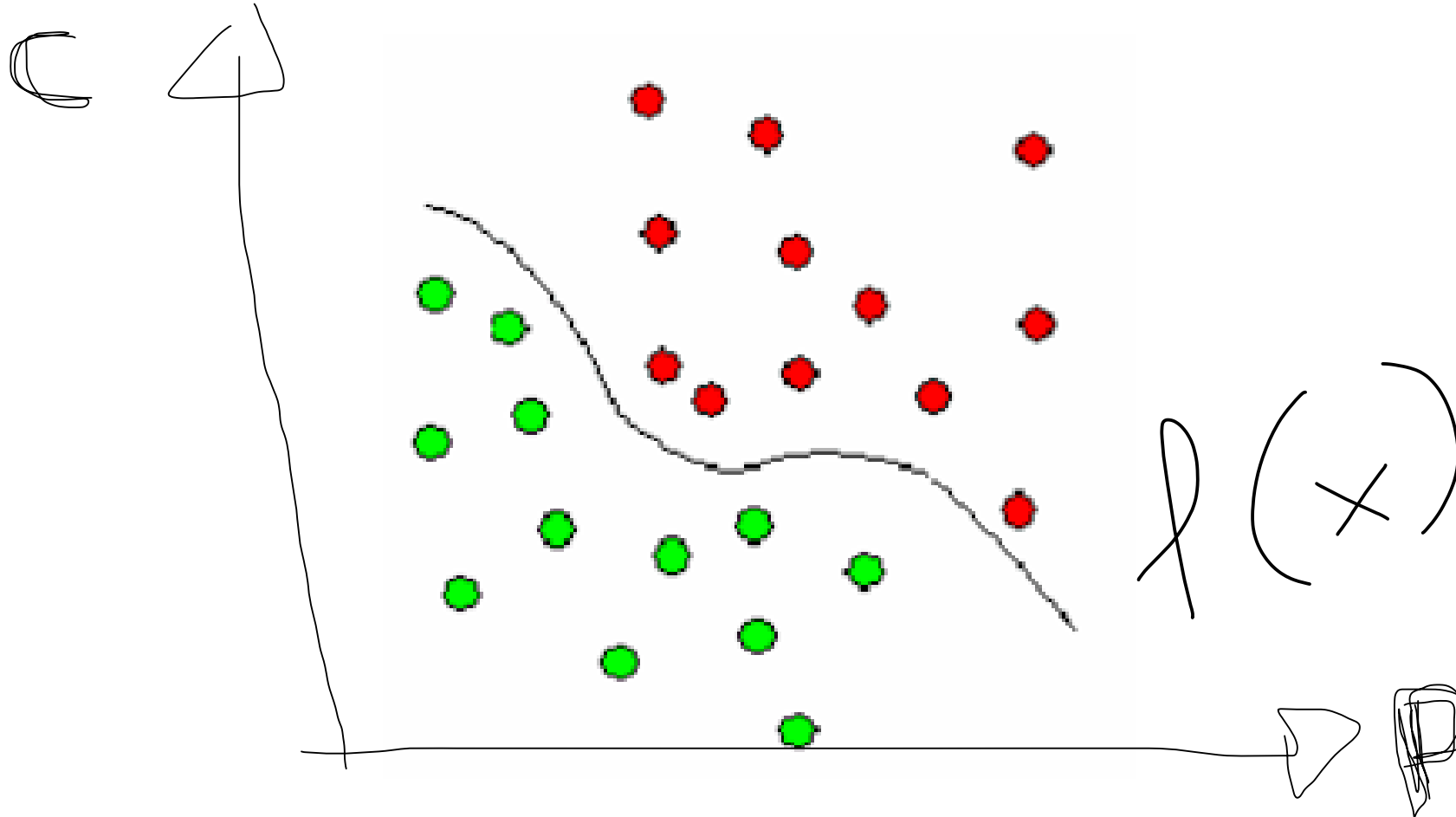
Models are often more complex that «lines» (as for the student grade's example

Learning fase



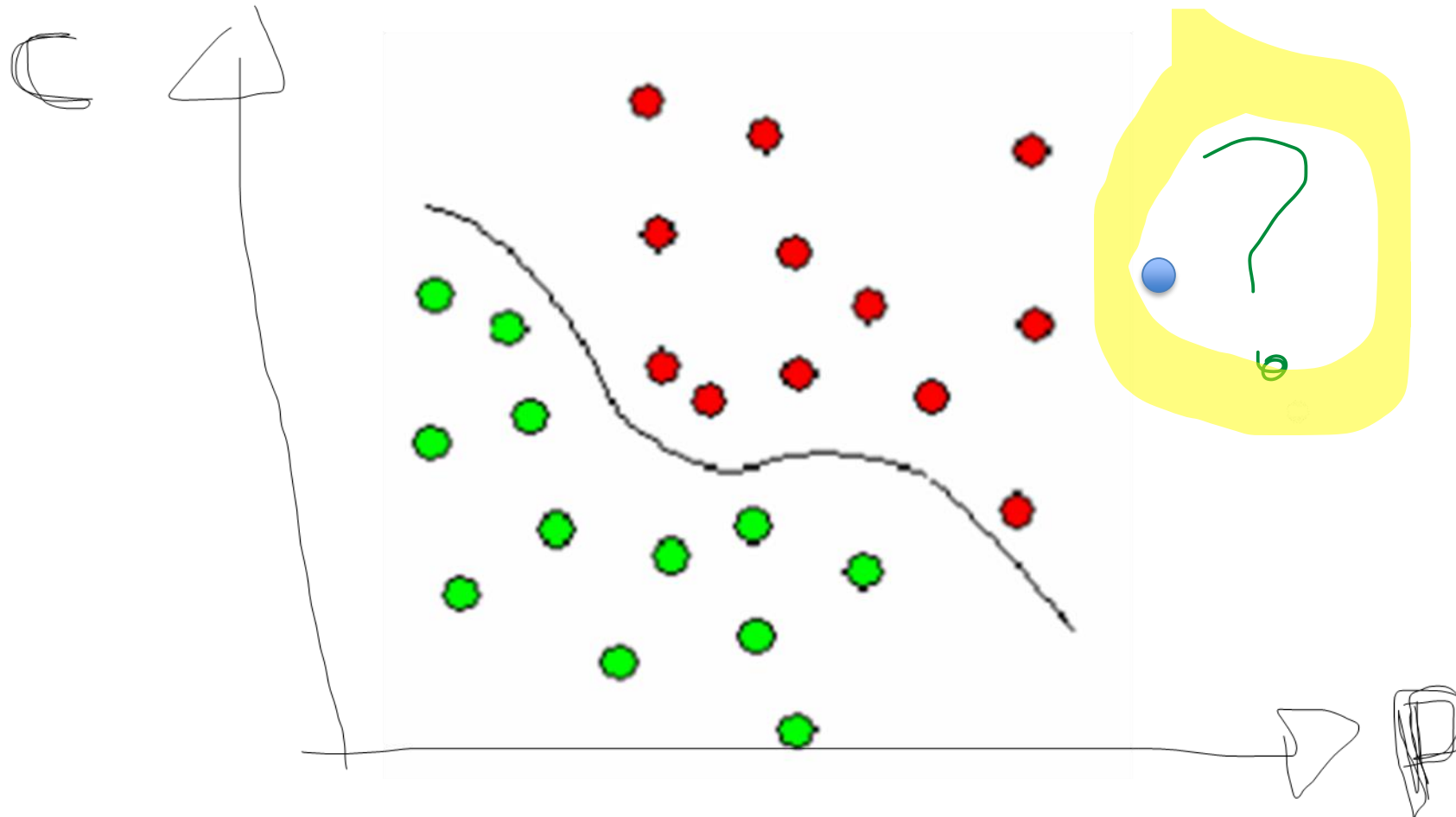
Suppose we are given examples of patients with only two attributes, pressure and cholesterol. Red are patients that ended up having cardiovascular complications, green are patients who didn't

Learning fase



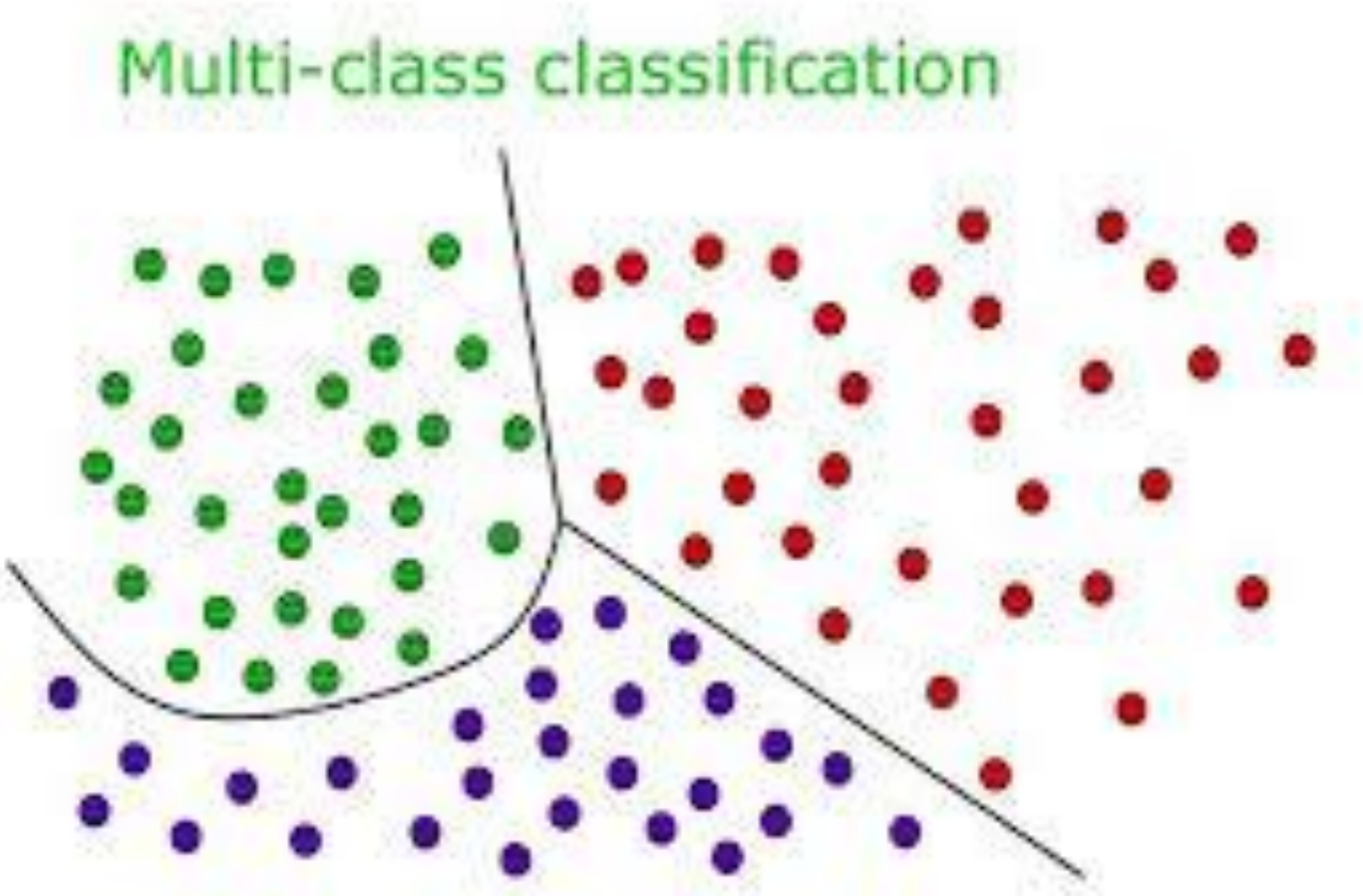
The MODEL that the system tries to learn from examples is a «separating» line (linear or more complex) $f(x)$ (separating hyperplane if multiple attributes)

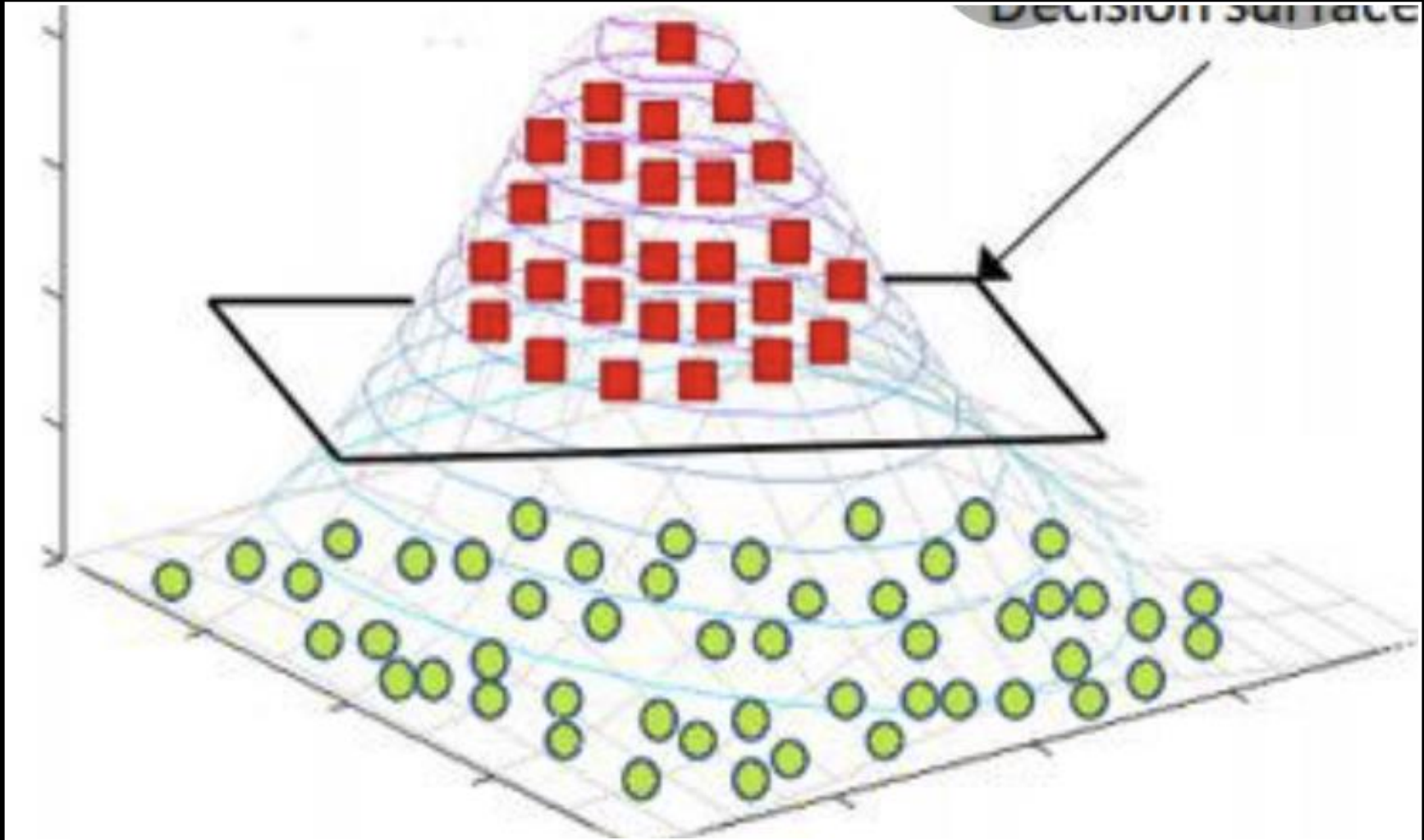
Prediction phase



Given a new «unseen» patient, the model will predict «at risk of cardiovascular problems» if his/her P and C values place the patient above the learned curve, which is called DECISION BOUNDARY

This can be extended to multiple classes





And to multiple attributes

Classification: application examples

- **Direct Marketing**
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - Use the data for a similar product introduced before (e.g., targeted with a traditional approach, like phone calls).
 - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
 - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - Use this information as input attributes to learn a classifier model.

Classification: applications

- **Fraud Detection**
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When 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** (e.g., a tree) for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

Classification: applications

- **Customer Attrition/Churn:**
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he/she calls, what time-of-the day he/she calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty (class is loyal /no-loyal).

Data Mining Tasks...

- Classification [Predictive]
- **Clustering** [Descriptive/prescriptive]
- Association Rule Discovery [Descriptive/prescriptive]
- Sequential Pattern Discovery [Descriptive/prescriptive]
- Regression [Predictive]
- Deviation (Anomaly) Detection [Predictive]

Clustering Definition

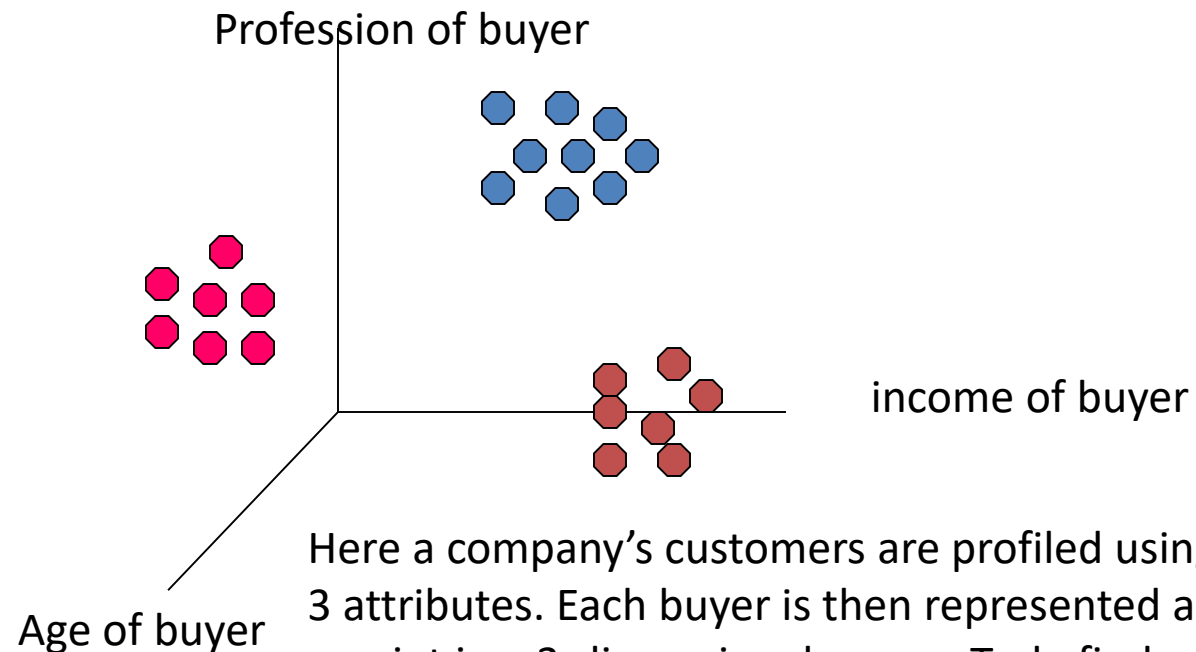
- Given a set of records (data items), each having a set of attributes, and a similarity measure among them, find clusters (set of records) such that
 - Members in one cluster are more similar to one another.
 - Members in separate clusters are less similar to one another.
- Similarity Measures:
 - Euclidean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

Illustrating Clustering

x Objective: create clusters (groups of records) such that:

Intracluster distances
are minimized

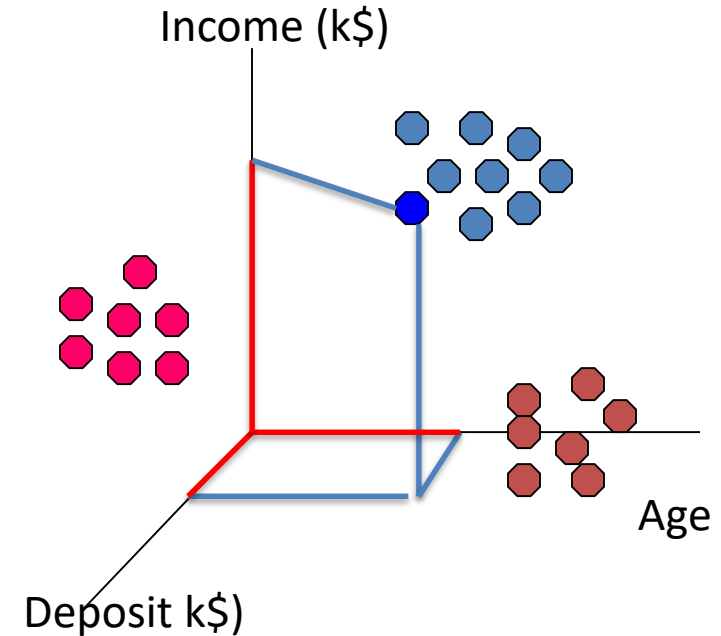
Intercluster distances
are maximized



Here a company's customers are profiled using 3 attributes. Each buyer is then represented as a point in a 3-dimensional space. Task: find groups of customers which are "similar".

Another example

- Suppose records in the database are a bank's customers, described by three attributes, e.g., *age*, *income*, *deposit*.
- Since attributes are numeric, we can represent each record in a 3-D space where each dimension is an attribute and each coordinate of a point is the value of the attribute for the considered record



Customer-333(Age=50,Income=5k\$,Deposit=3k\$)

What if we
have more
than 3
attributes?

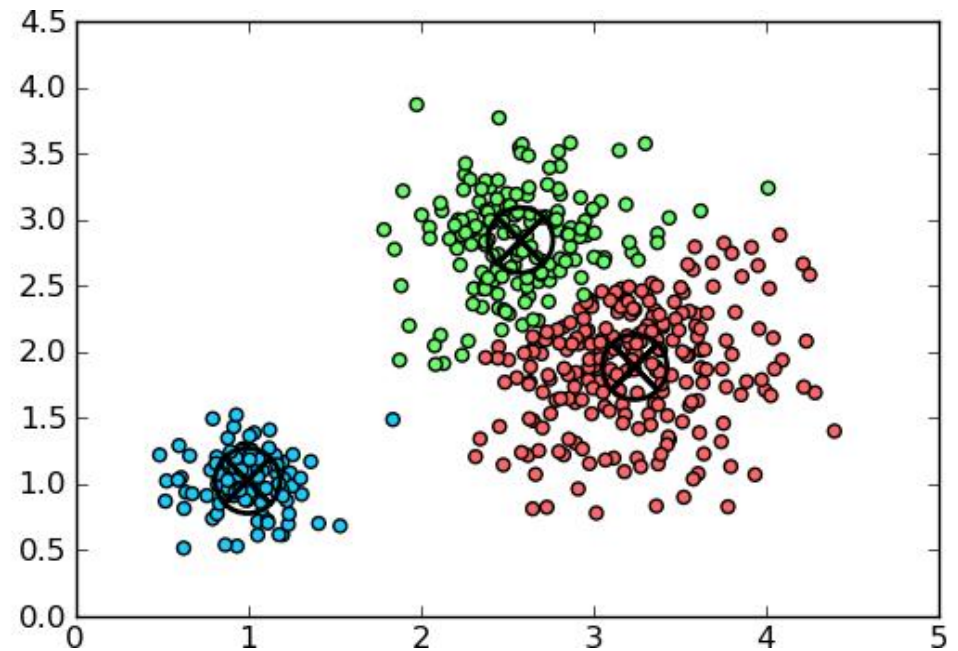
In real life, we can only perceive (and draw) 3 dimensions, but in math, no problem..

Algorithms can handle hyper-spaces (with as many dimensions as needed) ..

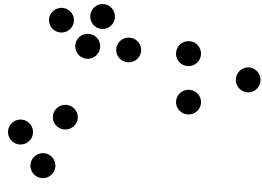
Why is clustering helpful?

- **No training is required:** we do not learn a class (e.g., loan yes/loan no) we learn clusters (*groups of similar ones*)
- **At prediction time,** the algorithm is given a new record (e.g., a new bank customer) and the algorithm is able to “associate” it with one cluster, according to the customer’s profile.
- Now we can “predict” his/her future behaviour based on the past behaviour of his/her similar ones. E.g., we can propose investments that others in the cluster have subscribed.

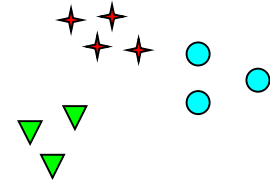
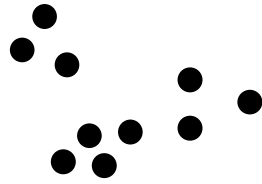
However,
clustering is
complex.
Many solutions
are possible..



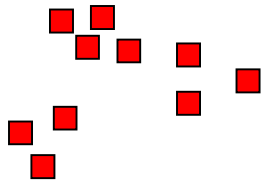
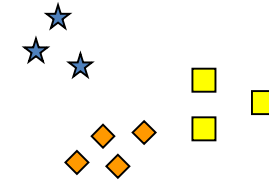
Notion of Cluster can be Ambiguous!



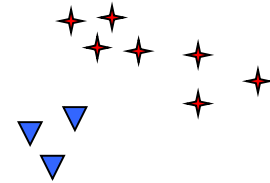
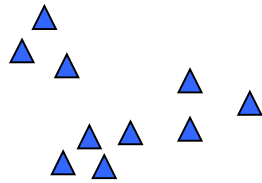
How many clusters?



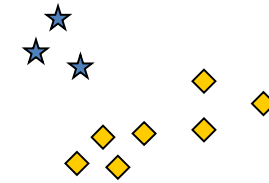
Six Clusters



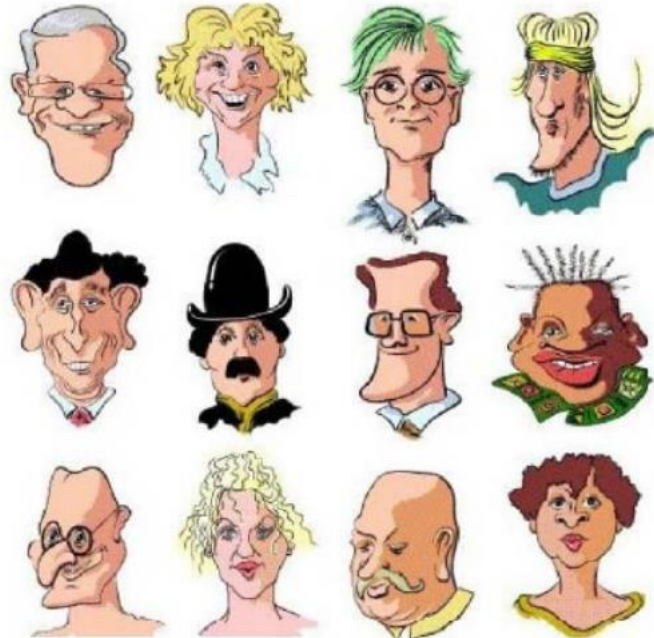
Two Clusters



Four Clusters



Example: group people according to their traits



- Far more complex than supervised learning
- What is “similar”?
- How many groups?


Very ill-formed problem:
Many ways of grouping, e.g.
Bold/hairy, female/male,
hair color, age..
Depends on similarity function




Are they similar?



Clustering results depend on the attributes that we select to measure similarity!

Machine Learning: Clustering



By color	
By shape	
By size	
	etc...

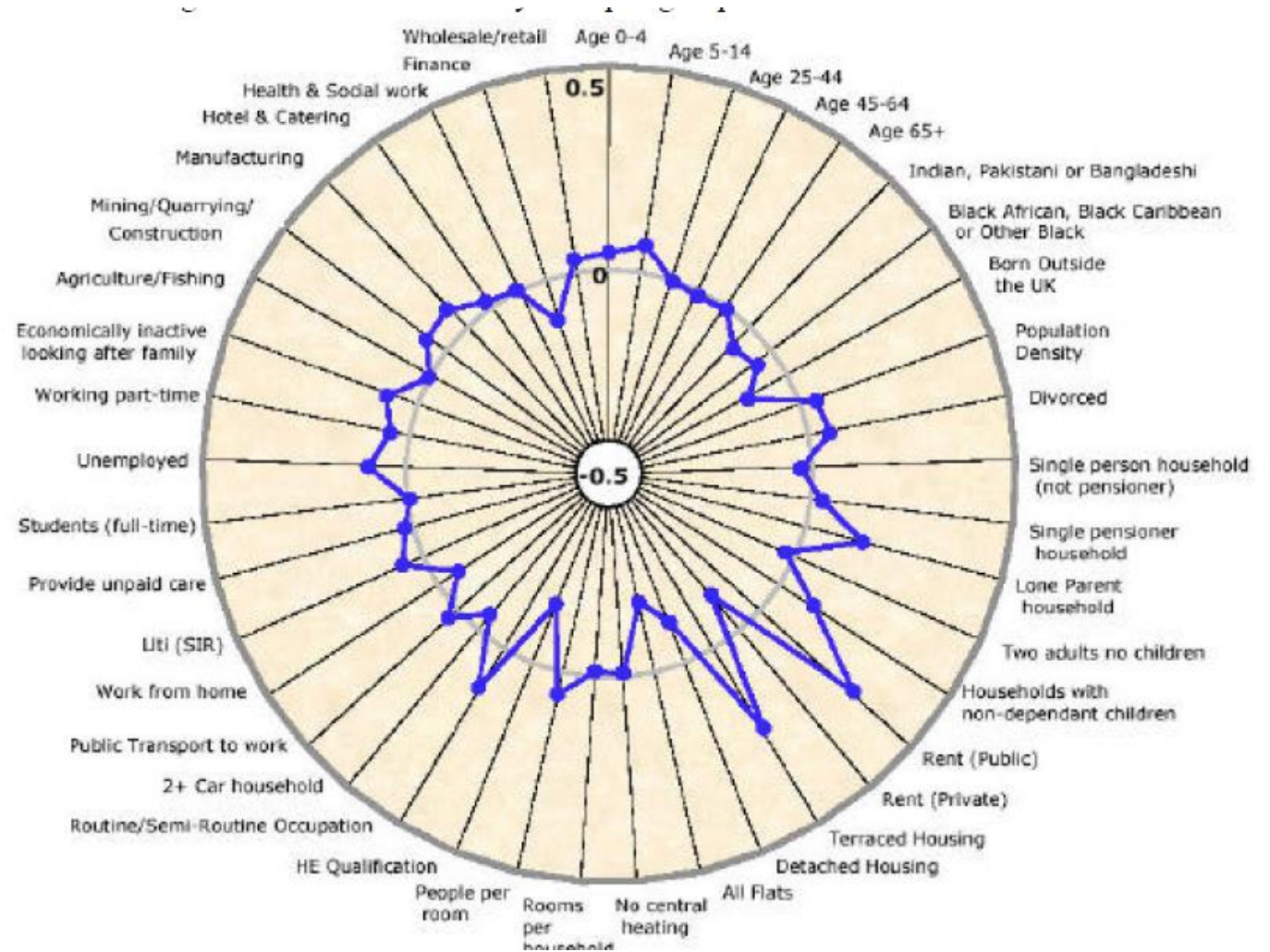
ENSTOA

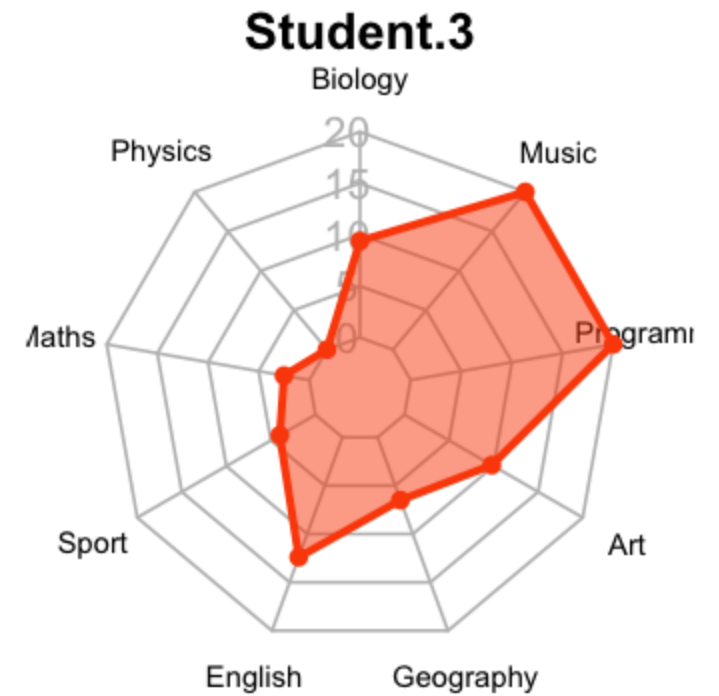
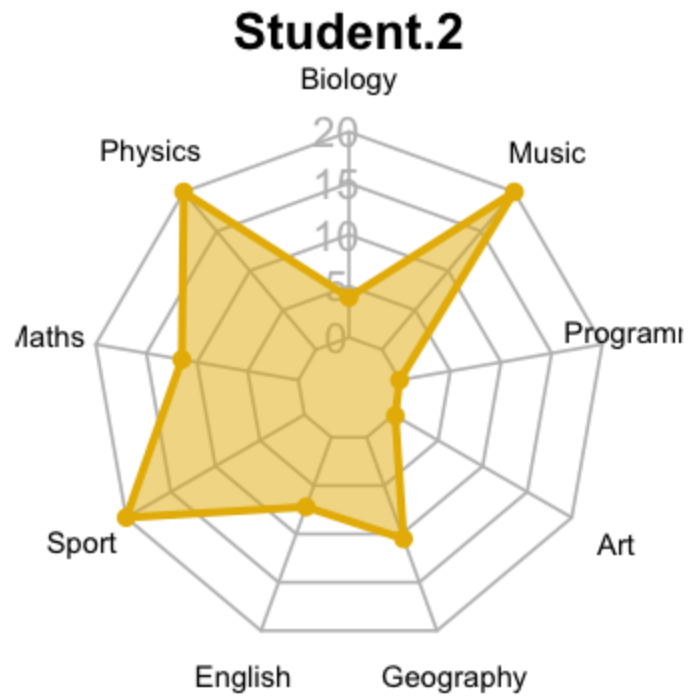
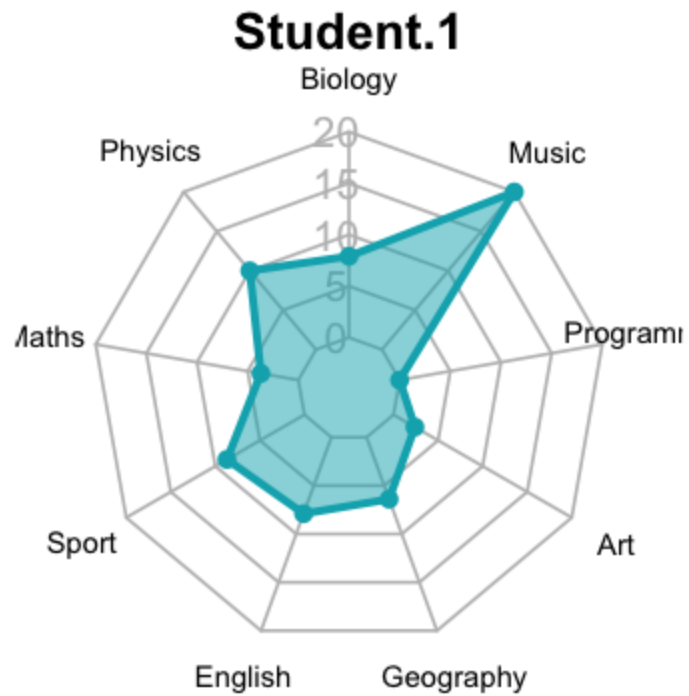
How to describe clusters?

- Clustering is different from classification. In classification, we are GIVEN a set of categories (e.g., defaulters and non-defaulters, or high-risk, mid-risk, low-risk) and the task is to predict the right category. In clustering, the system creates groups of similar-ones, but there is no class label.
- How can we understand the result of a clustering algorithm?
- Especially if data instances are described by very many attributes, it might be difficult to interpret the results

Radar plots

- A radar plot function for visualising clusters of people.
- People is described by 41 attributes (those shown outside the circle) – blue plot describes one specific detected cluster
- This is the plot of “blue collars”
- Blue line tells how much of a specific feature (e.g., divorced, use of public transportation to work, etc) the members of this cluster have.
- To derive this visualization, all attributes have been normalized between +0.5 and -0.5





Visualizing more than one cluster (students clusters by performance in courses). Here, 1,2.. Are the cluster names. Each cluster includes a subset of students of the original sample

Clustering applications

Cluster analysis for the following application:

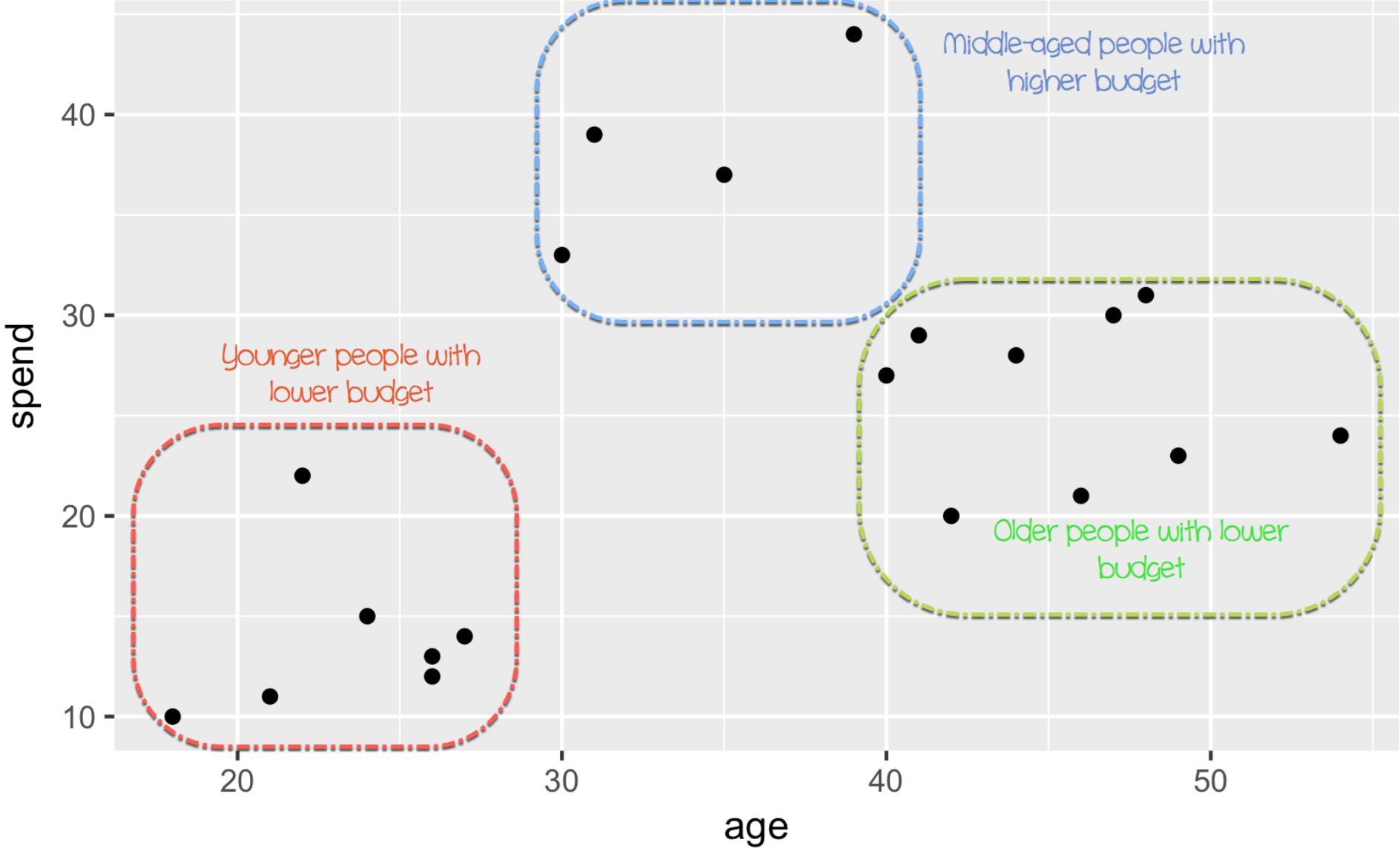
- Customer segmentation: Looks for similarity between groups of customers
- Stock Market clustering: Group stock based on performances
- Reduce dimensionality of a dataset by grouping observations with similar values

Clustering analysis is meaningful as well as actionable for business.

Clustering: Application 1

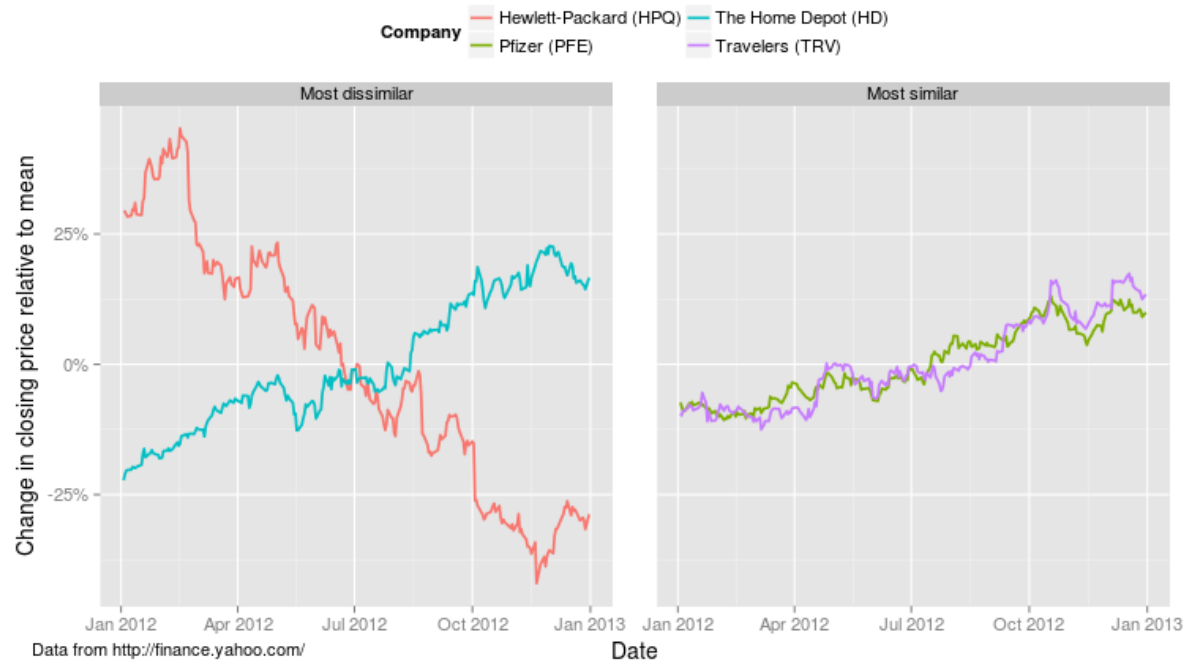
- **Market Segmentation:**
 - Goal: subdivide a market into distinct subsets of "similar" customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

MARKET SEGMENTATION EXAMPLE (simplified)



Clustering: Application 2

- **Stock market data:**
 - Goal: To find groups of stocks that are similar to each other based on their up/down movements.



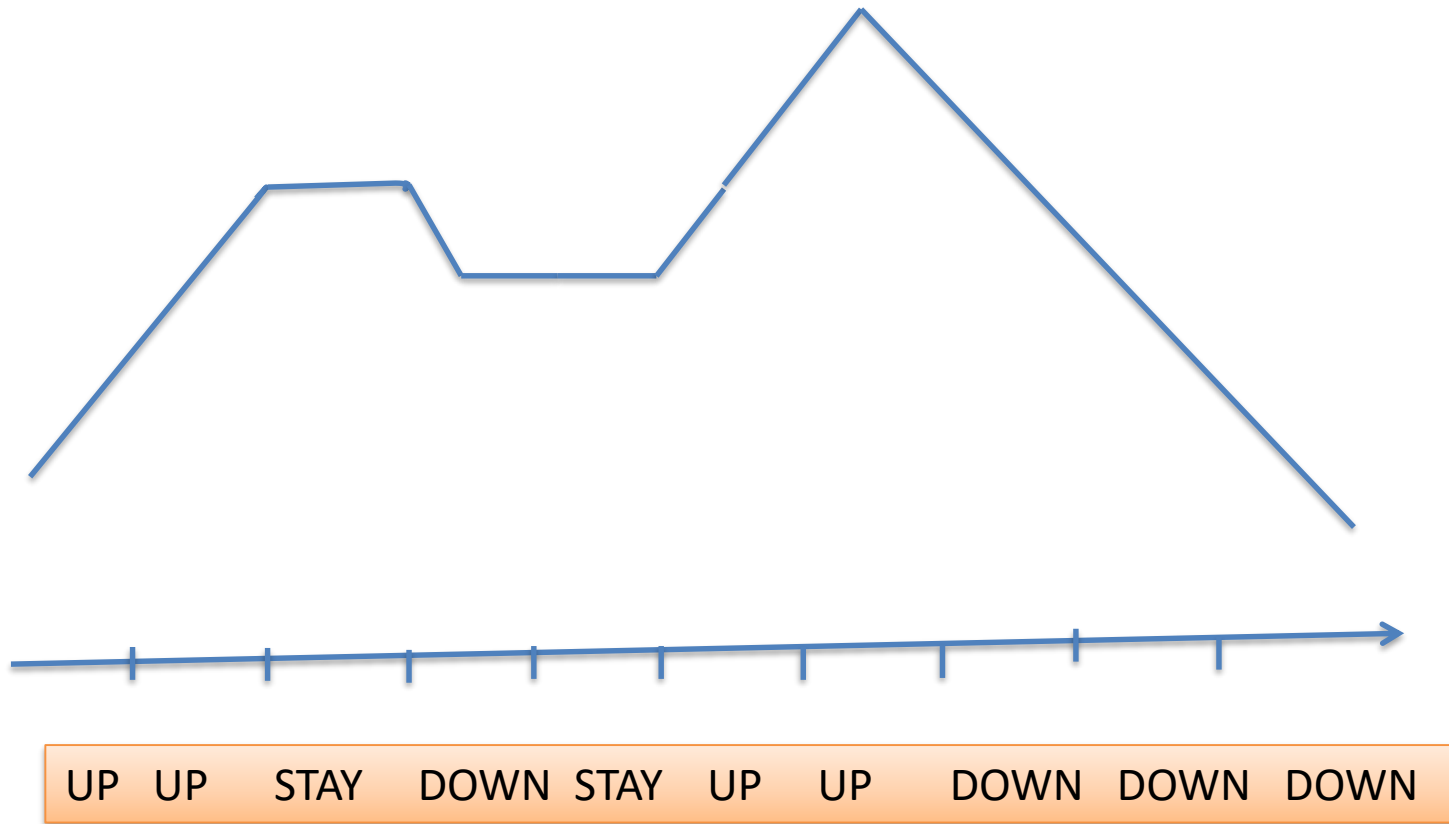
Clustering of S&P 500 Stock Data

- Observe Stock Movements in real time.
- Discretize into: Stock- $\{UP/DOWN/STAY\}^*$
- Similarity Measure: Two stocks are more similar if their sequences (up/down/stay) are frequently similar in the same days.

	<i>Discovered Clusters</i>	<i>Industry Group</i>
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, Oracle-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-Mac-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

e.g., a stock movement is transformed in a discrete sequence, like: ***up up up stable down down stable stable down....etc.***

Example of transformation:



Clustering stock market companies

- Another application is clustering stock market companies
- Criteria (attributes) for computing similarity can be: return of asset, rate of return on equity, earning per share, operating profit margin,

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- **Association Rule Discovery [Descriptive]**
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation (anomaly) Detection [Predictive]

Association Rule Discovery: Definition

- Given a set of records representing transactions – e.g., each contains some number of *items* from a given collection , for example products purchased in a supermarket);
 - Produce **dependency rules** which will predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
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}

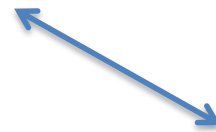
{Diaper, Milk} --> {Beer}

Rules read as follows:

- 1) those who buy milk, also buy coke
- 2) those who buy diapers and milk, also buy beer

Digression: How to represent transactions with tables in a database?

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



TID	Bread	Coke	Milk	Diaper	Beer
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1

Association rules

- IF (item₁ AND item 2 AND...) THEN (item_y AND item_x...ANDitem_z)
- IF (**ANTECEDENT**) THEN (**CONSEQUENT**)
- Reads: “IF a transaction includes all the items in the antecedent, then *it is likely* to include also all the items in the consequent”
- Like for Decision trees, association rules have support and confidence, that determine the strenght (or likelihood) of the rule, given the data from which it has been extracted
- Differently from Decision Trees, association rules are extracted from data which are not annotated with a class label (it is an untrained method to learn from data)
- A well known algorithm is APRIORI

Association Rule Discovery: Application 1

- **Marketing and Sales Promotion:**

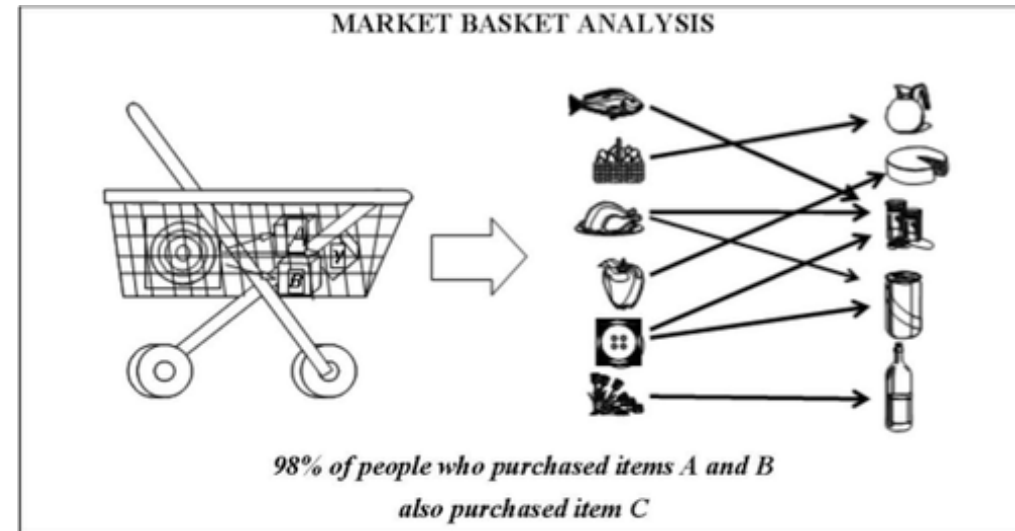
- Let the rule discovered be:

{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 other 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! (consider all rules with badgel in antecedent and Potato chips in consequent. Which other products are mentioned in these rules?)

Association Rule Discovery: Application 2

- **Market basket analysis .**
 - Goal: To identify items that are bought together by sufficiently many customers.
 - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
 - A classic rule --
 - If a customer buys diaper and milk, then he is very likely to buy beer.
 - So, don't be surprised if you find six-packs stacked next to diapers!



Association Rule Discovery: Application 3

- Inventory Management:
 - Goal: A consumer appliance repair company wants to **anticipate the nature of repairs** on its consumer products and keep the service vehicles equipped with right parts to reduce number of visits to consumer households.
 - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the **co-occurrence patterns**.
 - So that we know that **if one part fails, another part is likely to fail soon**

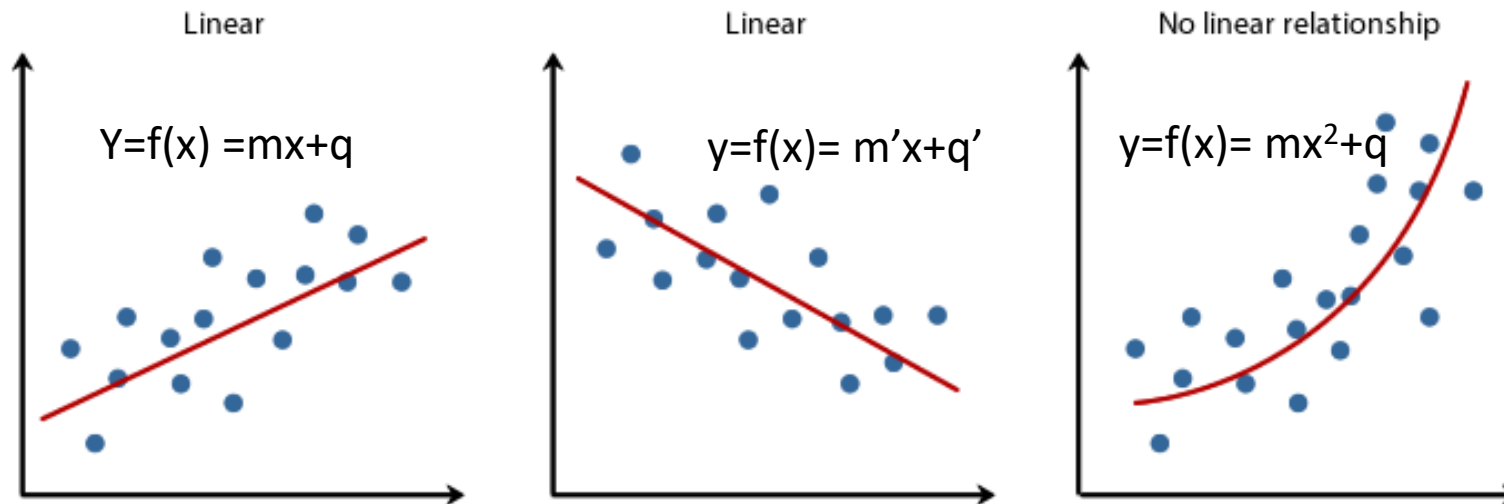
Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- **Regression** [Predictive]
- Deviation (anomaly) Detection [Predictive]

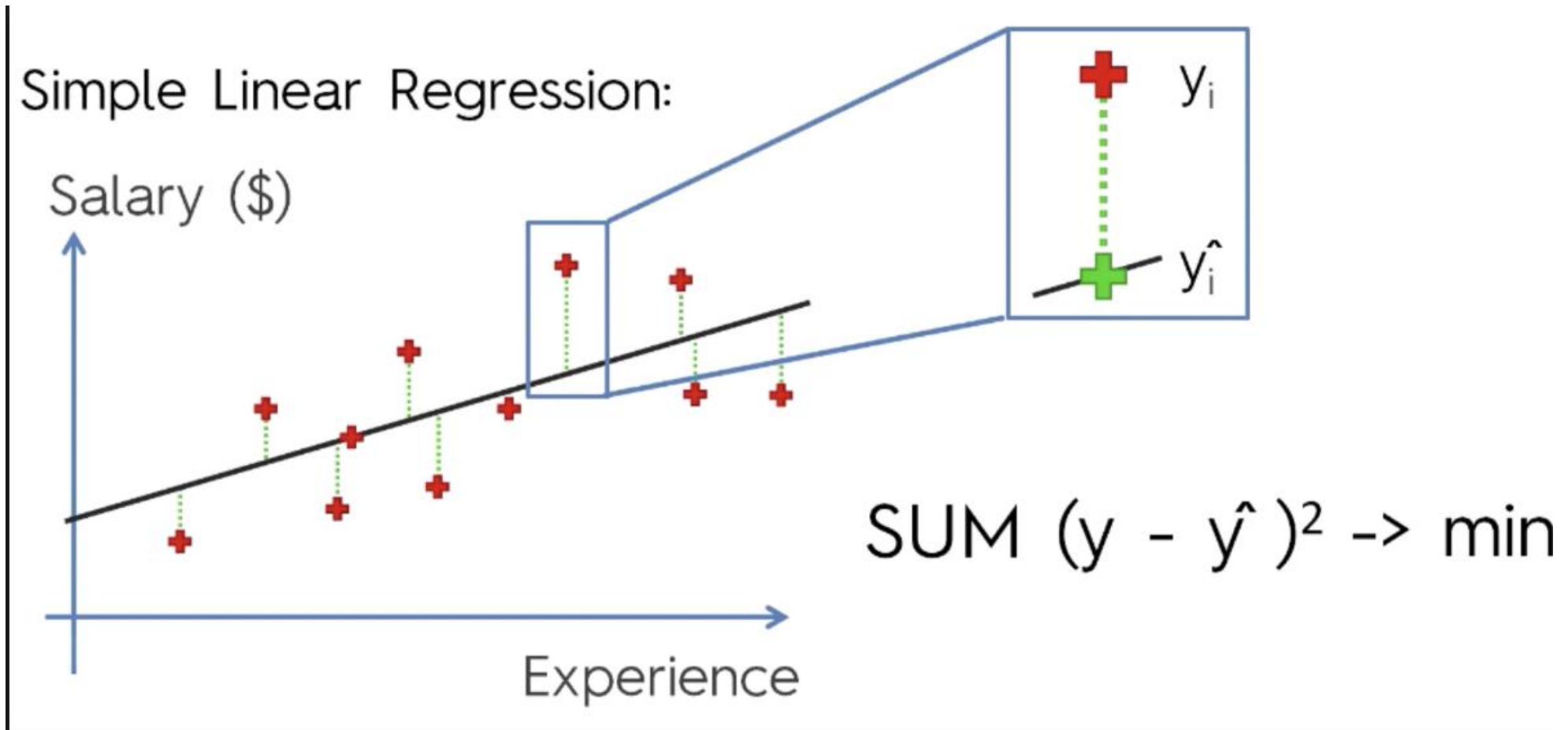
Will skip this

Regressions

- **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 and neural network fields.
- Example: we here have simple two-valued records (x is the independent variable (attribute), y the **dependent** that we would like to predict), represented as points in a bi-dimensional space
- Objective: given a *training set of historical data* where for each input x the output value is given, find a function $y=f(x)$ (the red line) which minimizes the error of the predicted values wrt the measured points



Example: predicting salary when given years of experience



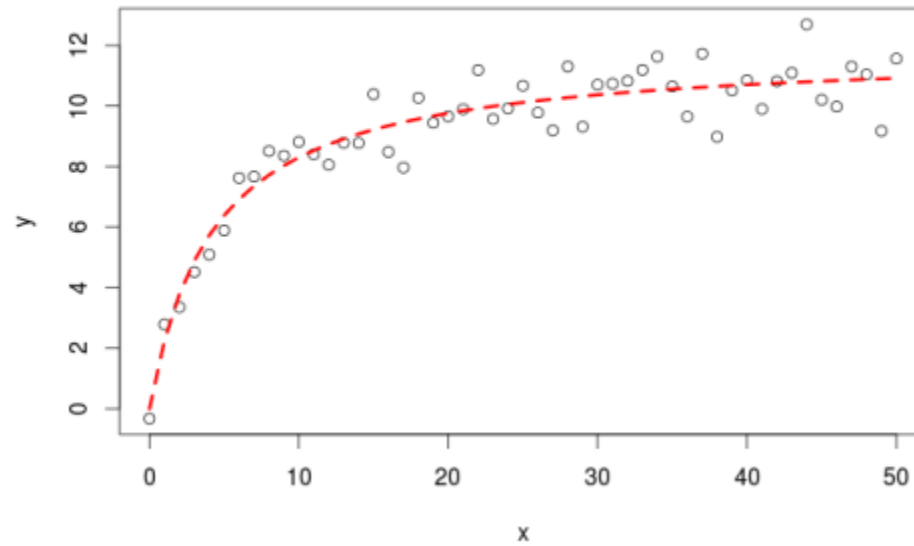
Regressions vrs classifiers

- Different from classification: now the learned function $y=f(\mathbf{x})$ is not used to SEPARATE the (hyper)plane into decision areas (and use it to predict the *category label* of a new point, eg a patient, or a credit applicant)
- Rather, here the function is used to predict a **future value of some attribute** (e.g. a stock market value, a probability of risk, the temperature of nextcoming day..)
- « \mathbf{x} » is in general not a single variable, rather, it is a set of variables (the «attributes» of a record)
- $y=f(x_1, x_2, \dots, x_n)$, example: $y=\text{current-Uber-taxi-ride-price}$, $f(\mathbf{x})=f(\text{traffic-severity, available-riders, time-of-the-day, city-type, city-area, driver-popularity, \dots})$
- $f(\mathbf{x})$ can be any algebraic function, for example a polynomial combination of the attribute values.

Example

- x_1 =traffic-severity, x_2 =available-riders, x_3 =time-of-the-day, x_4 =city-type, x_5 =city-area, x_6 =driver-popularity
- y =current-Uber-taxi-ride-price
- The system objective is to learn some function $y=f(x_1,x_2,x_3,x_4,x_5,x_6)$ that approximates the example data in the training set
- Example: $y=w_1x_1+w_2(x_2)^2+w_3x_3+w_4(x_4)^3+w_5x_5+w_6x_6$

Example of non-linear regression



Regressors vrs Classifier

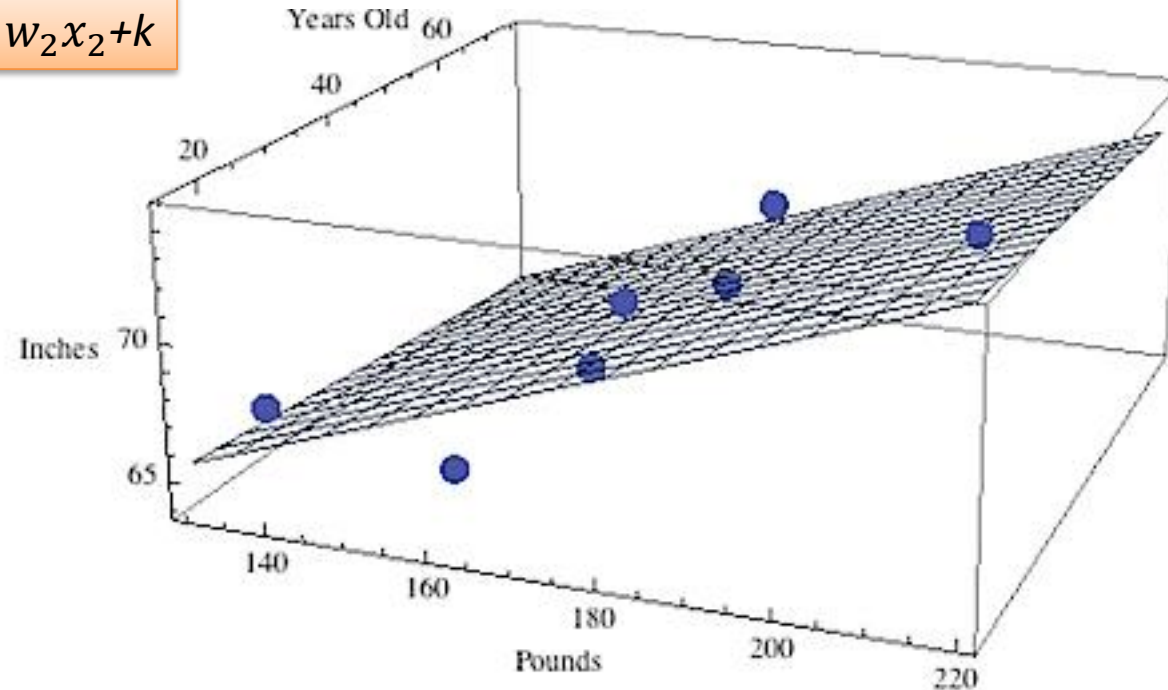
- To summarize:
 - We use classifier if we want to predict one among a set of finite category values (risk/no risk, sport politics entertainment,...), given the attribute values of some entity (patient, document..)
 - We use a regressor if we want to predict the VALUE of a continuous function (temperature, stock market price, ..) given the attribute values of some entity

What is the learning task with regression?

- Example: linear regression with two attributes x_1 and x_2 and one class y (target)
- $y = w_1x_1 + w_2x_2 + k$
- w_1 , w_2 and k are **unknown** coefficient
- We are given a training set (a table with attributes x_1, x_2, y), therefore we know triples of values (x_1, x_2, y) for a number of “points”
- The learning system must learn the values of the coefficients
- It does so by searching for the values of w_1, w_2, k that minimize the prediction error over the examples in the training set

Example

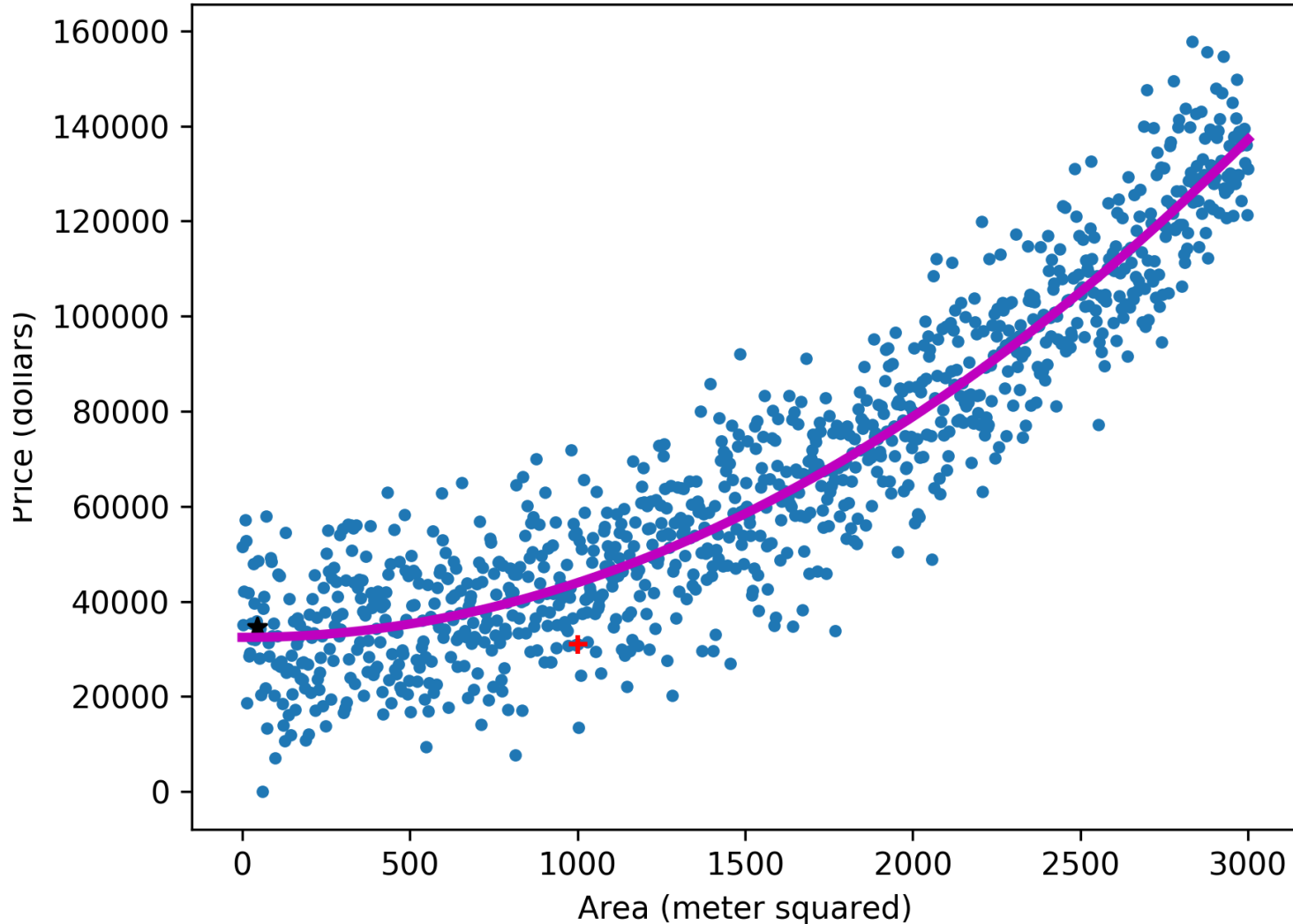
$$y = w_1x_1 + w_2x_2 + k$$



Say you want learn predicting the variable y =**weight** (pounds) given **age** (x_1) and **height** (x_2) (in inches). The algorithm is given a “learning set” of real people, where we know weight (the “ y ”), age (x_1) and height (x_2). These are the **blue** points. We want to “learn” a linear equation (the equation of a semiplane) such that when given any pair of values (x_1 x_2), we solve the equation and predict y . Learning a linear equation means learning the coefficients w_1 , w_2 and q .

Predicting house prices based on house descriptive attributes

House data of city Branalle



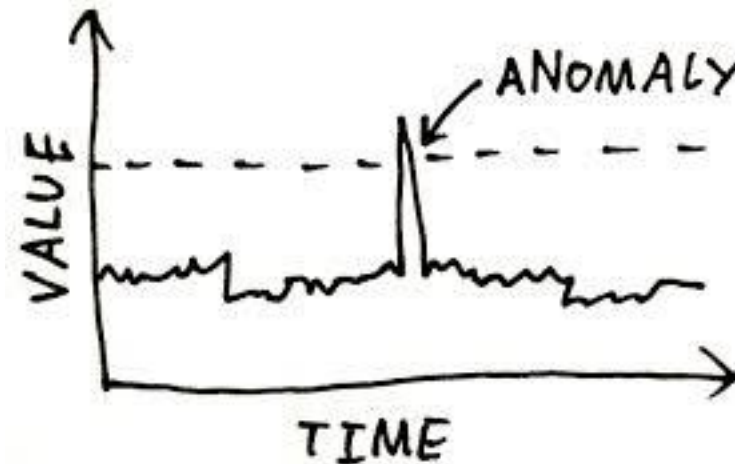
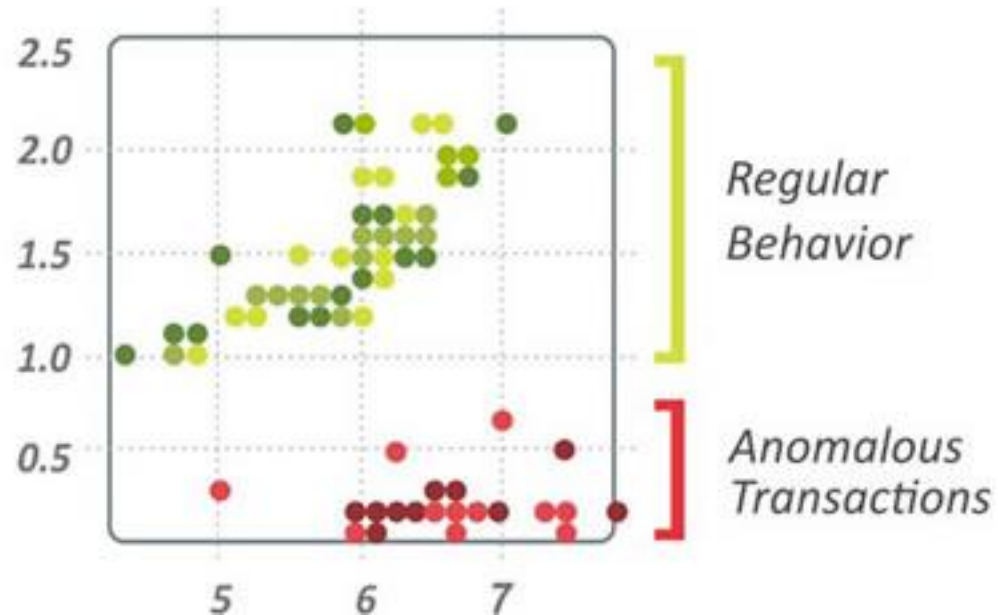
For simplicity and visibility only one attribute is shown here, i.e. dimension of the house in sqm. For many attributes, the problem has a solution in a multi-dimensional space

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- **Deviation (anomaly) Detection [Predictive]**

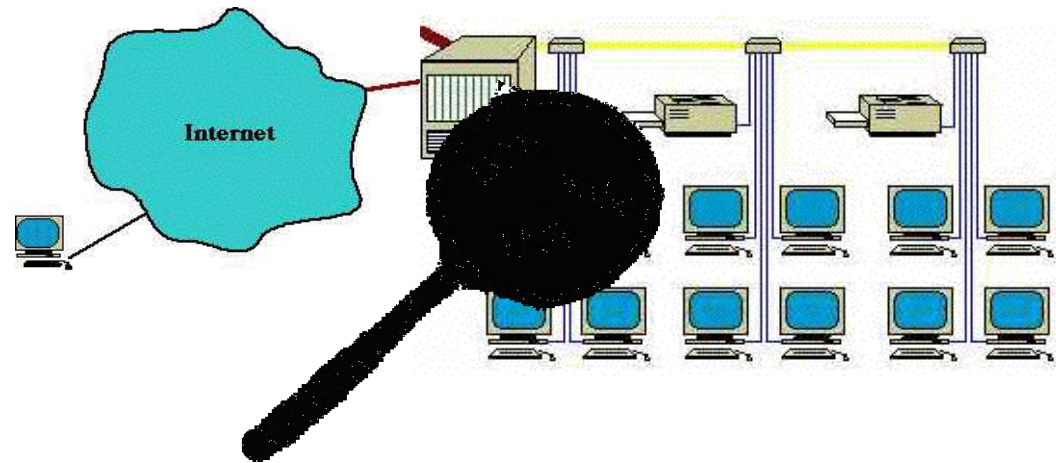
Deviation (anomaly) detection

- Detect significant deviations from normal behavior
- Input: “normal” behaviour (or, examples of previous abnormal data)
- The system learns a model of “normality” and then it detects significant deviations from normality
- Often difficult to find examples of abnormality (fake news, credit card frauds) so we train on normal cases

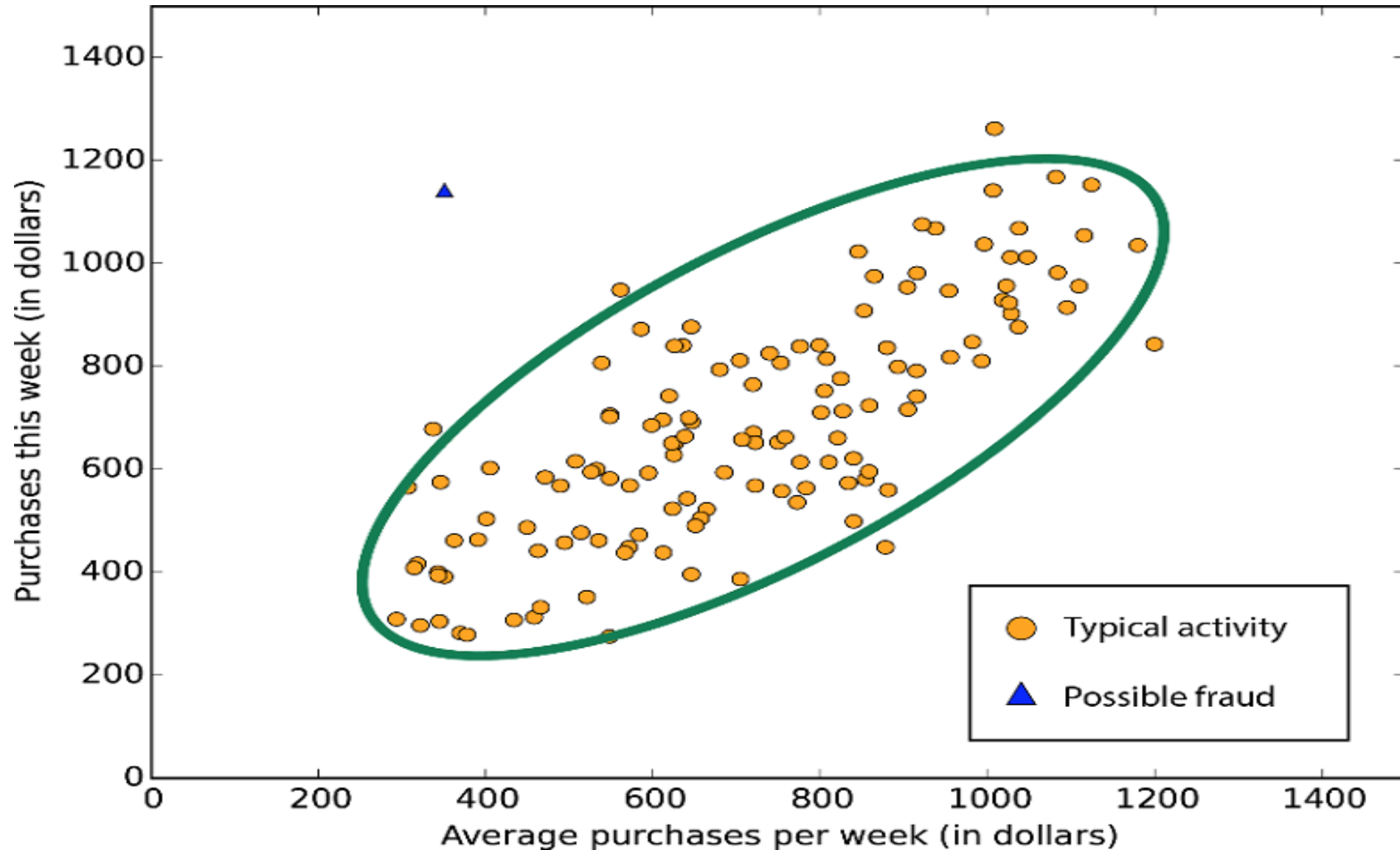


Deviation/Anomaly Detection

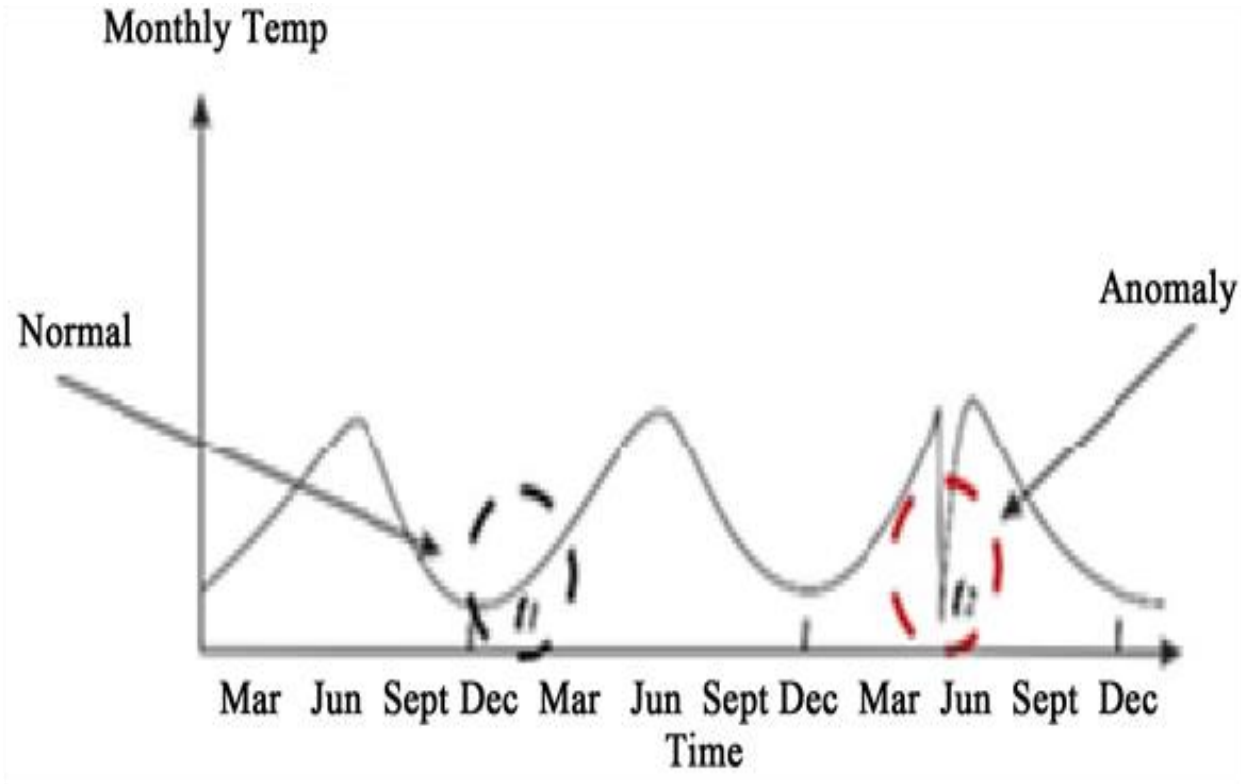
- Applications:
 - Credit Card Fraud Detection
 - Network Intrusion Detection
 - Industrial damage detection



Type of anomalies: point anomaly



Contextual Anomaly



- A negative spike in balance occurs in an anomalous period
- Note: the same spike is “normal” e.g., in January
- Here the “context” is the period of occurrence
- Other examples of contextual anomalies: behavioural anomaly detection (depends on individuals, daytime, season..)

Example

A Real Fraud Example

My credit card statement—**Can you see the fraud?**



May 22	1:14 PM	FOOD	Monaco Café	\$127.38
May 22	7:32 PM	WINE	Wine Bistro	\$28.00
...				
June 14	2:05 PM	MISC	Mobil Mart	<u>\$75.00</u>
June 14	2:06 PM	MISC	Mobil Mart	<u>\$75.00</u>
June 15	11:48 AM	MISC	Mobil Mart	<u>\$75.00</u>
June 15	11:49 AM	MISC	Mobil Mart	<u>\$75.00</u>
May 28	6:31 PM	WINE	Acton Shop	\$31.00
May 29	8:39 PM	FOOD	Crossroads	\$128.14
June 16	11:48 AM	MISC	Mobil Mart	<u>\$75.00</u>
June 16	11:49 AM	MISC	Mobil Mart	<u>\$75.00</u>

All same \$75 amount?

Monaco?

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- A number of “theory-related” challenges (algorithms)

Read and report use cases of machine learning/predictive systems in these domains:

- DATA MINING FOR FINANCIAL APPLICATIONS
- Customer Segmentation Using Clustering and Data Mining Techniques
- Fraud Analytics using prescriptive, predictive and social network techniques