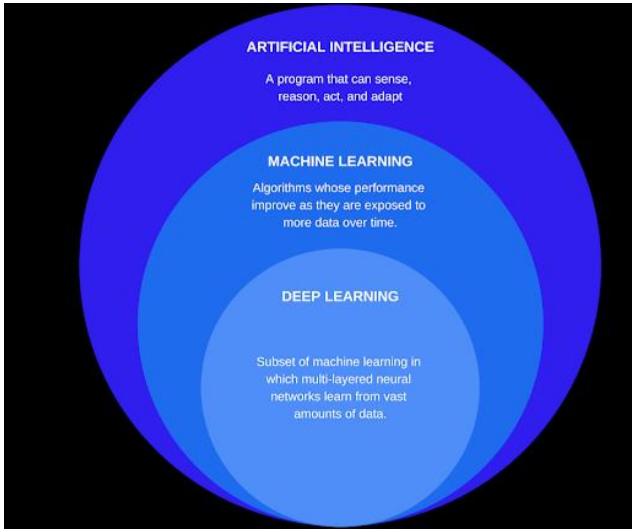

ARTIFICIAL INTELLIGENCE

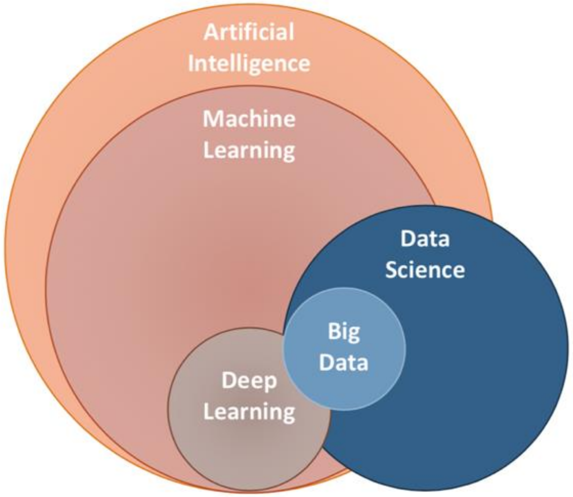
A program that can sense, reason, act, and adapt

MACHINE LEARNING

Algorithms whose performance improve as they are exposed to more data over time.

DEEP LEARNING

Subset of machine learning in which multi-layered neural networks learn from vast amounts of data.


Artificial Intelligence

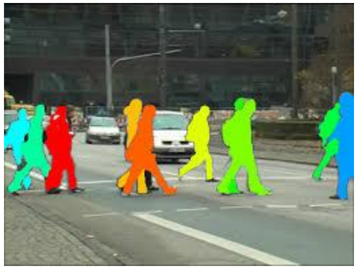
Machine Learning

Data Science

Big Data

Deep Learning

Learning **from Data**
(Machine Learning)



Recomended lectures

- ***Machine learning: Trends, perspectives, and prospects***. M. I. Jordan and T. M. Mitchell (Science, 2015)

- **AI Career Pathways: *Put Yourself on the Right Track.*** Workera. 2019

- ***Bots, Machine Learning, Servicios Cognitivos Realidad y perspectivas de la Inteligencia Artificial en España***, 2018. PWC & Microsoft.

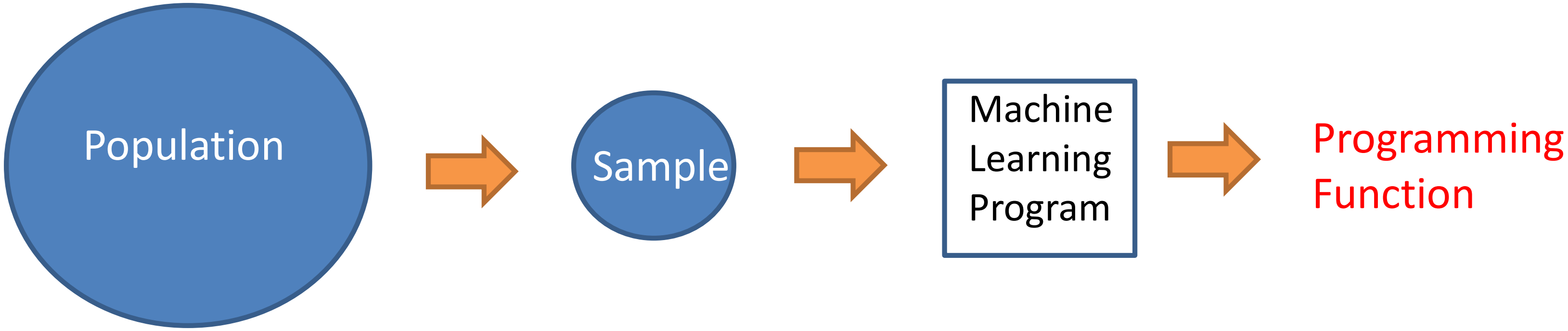
A simple example of ML

- Let suppose we know sales data of flats in a city, and we also know the flats distance to the city center: $(s_i, d_i), i = 1, ..N$

- Question: How to select the function that best predicts the cost given a distance?

  - Try to guess what decision we should make and what tasks we have to solve ...

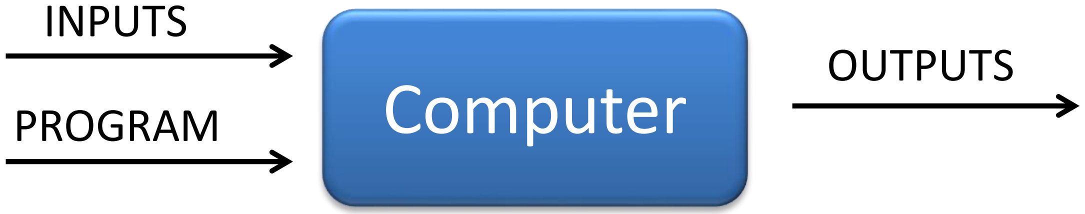
A learning task sketch



- Hypothesis:
  - The only available information is the sample
  - The function is computed from the sample using a Learning program
  - The function computes a prediction
- Main question: How to choose a sample and a Machine Learning Program to find an accurate prediction-function on the whole population?

Machine Learning

TRADITIONAL PROGRAMMING



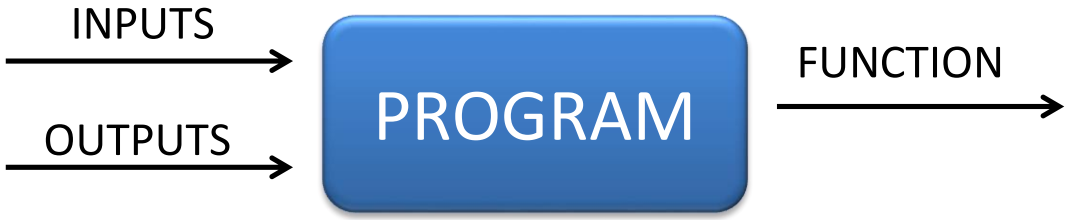
INPUTS

PROGRAM

Computer

OUTPUTS

MACHINE LEARNING



INPUTS

OUTPUTS

PROGRAM

FUNCTION

Computer learns a calculation to solve a task

Suitable for computer but difficult for human being

Machine Learning definitions

- Arthur Samuel : "the field of study that gives computers the ability to learn without being explicitly programmed." This is an older, informal definition.

- Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

- Example: playing checkers.

- E = the experience of playing many games of checkers
- T = the task of playing checkers.
- P = the probability that the program will win the next game.

- In general, any machine learning problem can be assigned to one of two broad classes:
  - Supervised learning : learn a function from labels and sample data
  - Unsupervised learning: learn a function from sample data

Machine learning: Trends, perspectives, and prospects. M. I. Jordan and T. M. Mitchell (Science, 2015)

Three main ML Paradigms

- **Supervised learning** : sample data + teacher (label) (static learning)

- **Reinforcement Learning**: sample data + rewards (not label) (dynamic learning)

- **Unsupervised learning**: only data (guessing)

Supervised Learning

How difficult is to define a tree?

A brown trunk moving upwards and branching with leaves . . .

Are these trees ?

Defining is Hard; Recognizing is Easy

Hard to give a complete mathematical definition of a tree.
Even a 3 year old can tell a tree from a non-tree.
The 3 year old has learned from data.

A pattern exists. We don’t know it. We have data to learn it.

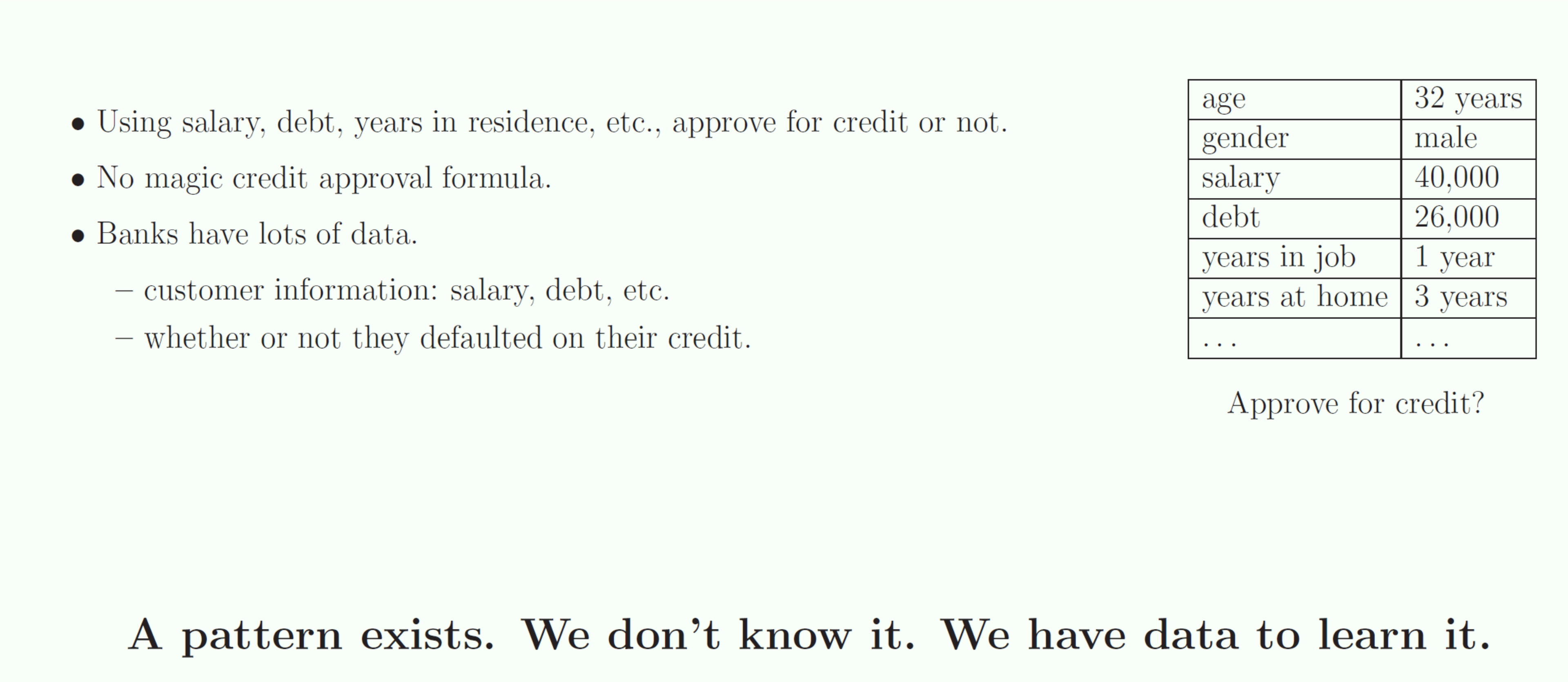
Visual Patterns

Identify handwritten digits in ZIP codes

Detecting faces in an image

A pattern exists. We don't know it. We have data to learn it.

Credit Approval



- Using salary, debt, years in residence, etc., approve for credit or not.
- No magic credit approval formula.
- Banks have lots of data.
  - customer information: salary, debt, etc.
  - whether or not they defaulted on their credit.

age	32 years
gender	male
salary	40,000
debt	26,000
years in job	1 year
years at home	3 years
. . .	. . .

Approve for credit?

A pattern exists. We don't know it. We have data to learn it.

More examples

- Speech and speaker recognition
- Object and Face recognition
- Document translation and composition
- Recommendation systems
- Stock Market Prediction
- Automatic Prognosis and Diagnostic in Medicine
- Etc, etc

Two main supervised paradigms

- Both represent PREDICTIONS !!

- Regression: the output is a real number (**continuous variable**)
  - Predict the height of a person from a data sample:
    - Features: weight; (weight, feet length); (weight, feet length, shoulders width), etc
  - Predict the temperature for the next day from a previous record of temperatures

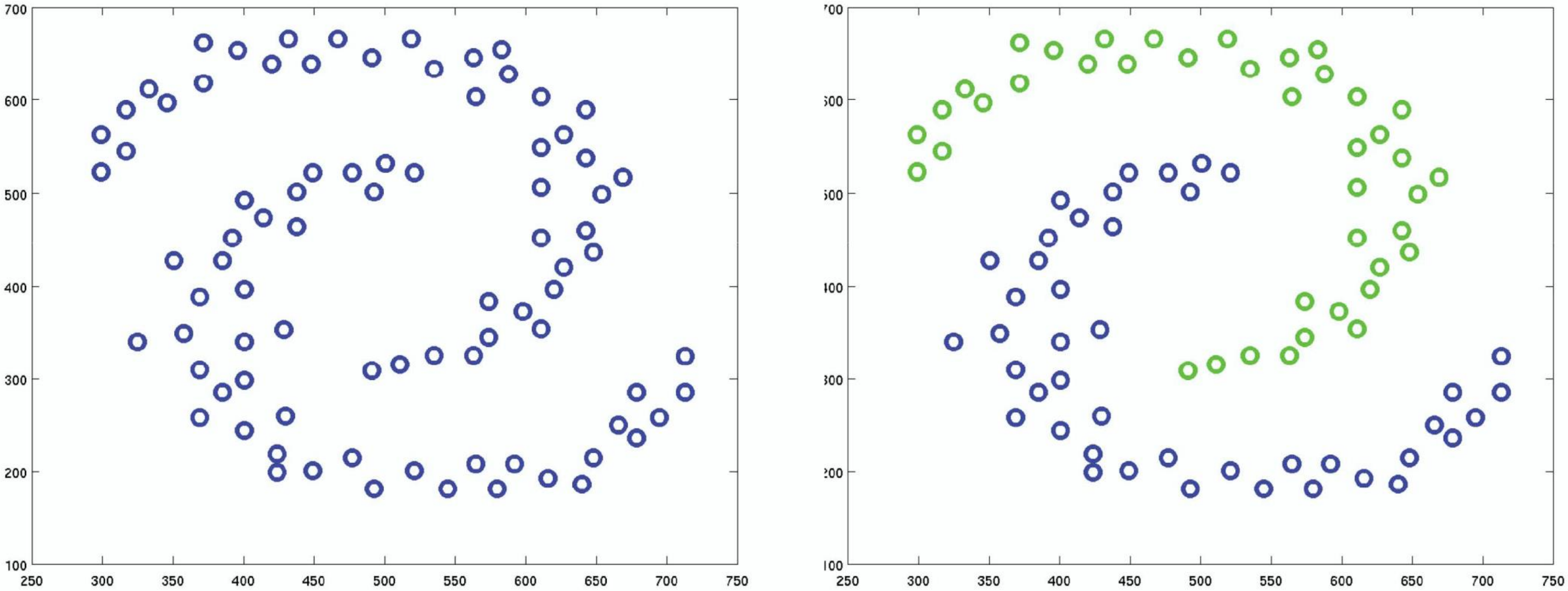
- Classification: the output is a class label (**discrete variable**)
  - Predict the weather for tomorow : (sunny, cloudy, windy)
  - Predict wether an image contains a face: (Yes,No), (0,1), (1,-1), etc
  - Predict wether an email is SPAM or not: (Yes, No)

Unsupervised Learning

- Data modeling to discover what is inside:
  - Geometric structure: **clustering**
  - Dependences discovering: **patterns**
  - **Dimensionality reduction**: relevant features
  - Etc

- Detangle sampled data in more independent features that facilitate downstream processes

Clustering

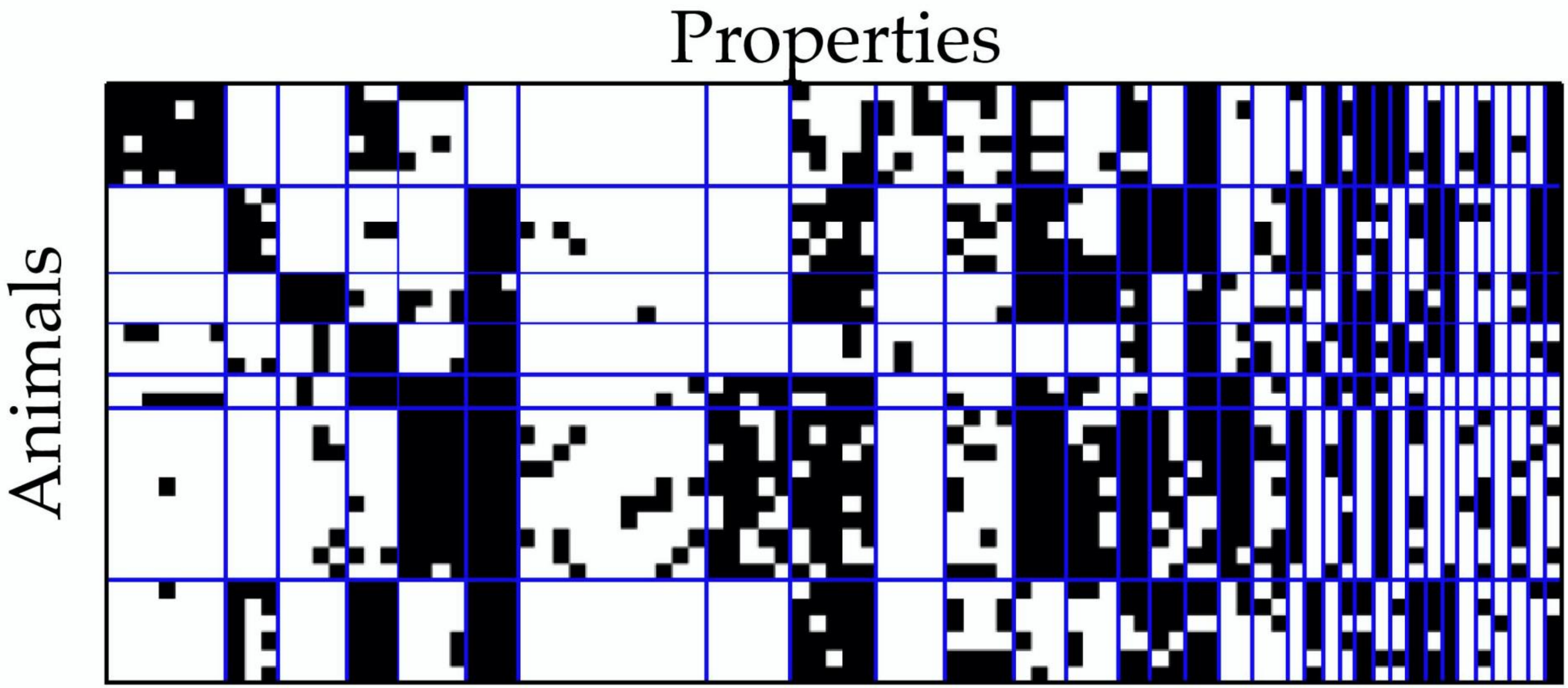

700
600
500
400
300
200
100
250
300
350
400
450
500
550
600
650
700
750
700
600
500
400
300
200
100
250
300
350
400
450
500
550
600
650
700
750

Example: Characteristics of Animals

- 50 animals
- 80 binary features
- Find interpretable groupings

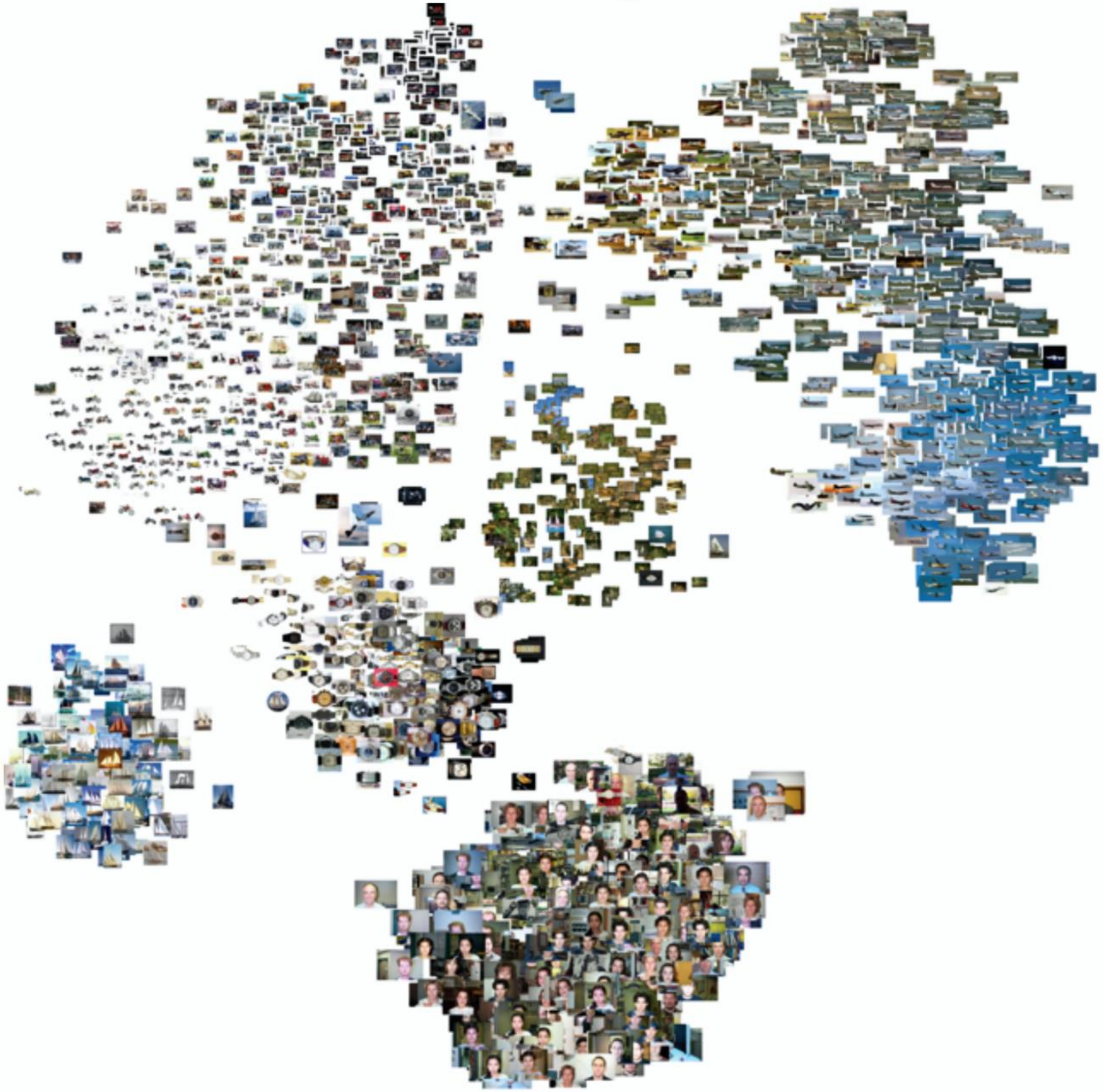
O1	killer whale, blue whale, humpback, seal, walrus, dolphin
O2	antelope, horse, giraffe, zebra, deer
O3	monkey, gorilla, chimp
O4	hippo, elephant, rhino
O5	grizzly bear, polar bear

F1	flippers, strain teeth, swims, arctic, coastal, ocean, water
F2	hooves, long neck, horns
F3	hands, bipedal, jungle, tree
F4	bulbous body shape, slow, inactive
F5	meat teeth, eats meat, hunter, fierce
F6	walks, quadrapedal, ground



(J. Tenenbaum)

Dimensionality Reduction



Review

A) Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, **what is T**?

1. The probability of it correctly predicting a future date's weather.
2. The weather prediction task.
3. None of these.
4. The process of the algorithm examining a large amount of historical weather data.

B) Suppose you are working on weather prediction, and use a learning algorithm to predict tomorrow's temperature (in degrees Centigrade/Fahrenheit). **Would you treat this as a classification or a regression problem?**

C) Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). **Would you treat this as a classification or a regression problem?**

D) Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. **Which of the following would you apply supervised learning to?** (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

1) Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".

2) Given historical data of children's ages and heights, predict children's height as a function of their age.

3) Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.

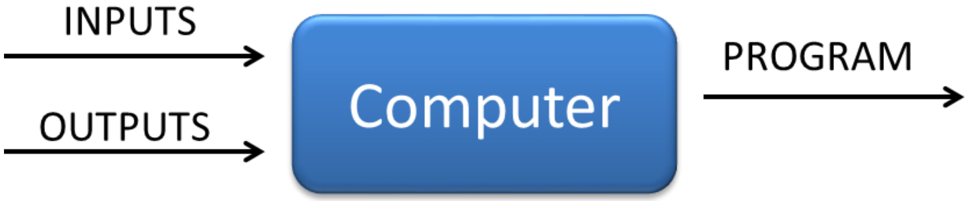
4) Given 50 articles written by male authors, and 50 articles written by female authors, learn to predict the gender of a new manuscript's author (when the identity of this author is unknown).

E) **Which of these is a reasonable definition of machine learning?**

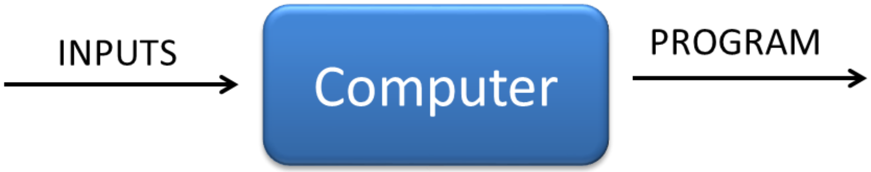
1. Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.
2. Machine learning learns from labeled data.
3. Machine learning is the science of programming computers.
4. Machine learning is the field of allowing robots to act intelligently.

What we already know

SUPERVISED MACHINE LEARNING



UNSUPERVISED MACHINE LEARNING



"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Try to identify the main elements….

Identify handwritten digits in ZIP codes

Detecting faces in an image

Regression or classification ?
Supervised or unsupervised ?

Try to identify the Key Elements

Let us assume you work for a company that is interested in discover a rule to fix the price to pay for the mango harvest in a large extension crop.

- The company is only interested in paying for the tasty mangos that can be sold in the next days.
- Unfortunately you knowledge about mangos crop is very limited or null
- But, you can visit the farm and take a sample of fruit to identify the mangos properties (photo, color, size, texture, etc) associated with tasty/non-tasty mangos.

What to do?

Can you identify the principal elements of the problem?

How set-up a dataset that can be helpful?

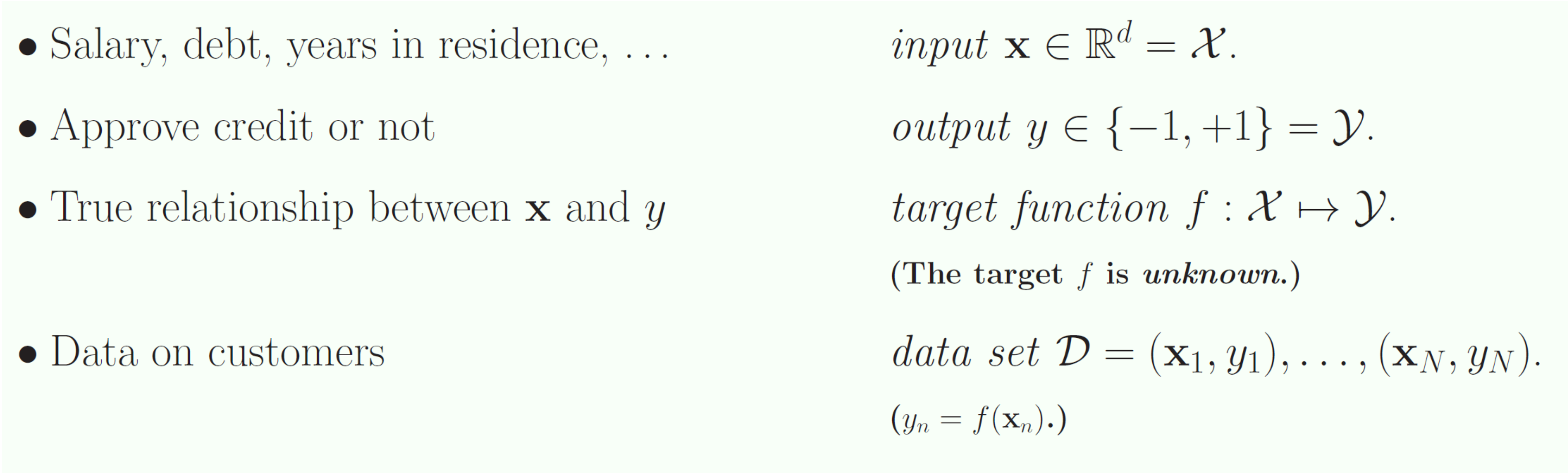
Let’s formalize the approach

1. What is the available information?
   - Where the data come from? $\mathcal{X}, \quad f: \mathcal{X} \rightarrow \mathcal{Y}$
     - What features to use?
   - Is the data collection relevant? Why?: $\mathcal{P}$
     - Should we assume some hypothesis ? independent identically distributed (i.i.d.)

2. How to set-up a model ?
   - What representation use?:
     - Which class of function are we going to use? $\mathcal{H}$
     - How to characterize each element h$\in\mathcal{H}$? h

3. How to search inside $\mathcal{H}$?
   - What criteria to use to guarantee learning? ERM, SRM, MDL
     - The function to optimize: (Loss function)
   - What algorithm to use to find the best function of $\mathcal{H}$? $\mathcal{A}$

Elements of the Credit Approval problem



• Salary, debt, years in residence, . . .	*input* $\mathbf{x} \in \mathbb{R}^d = \mathcal{X}$.
• Approve credit or not	*output* $y \in \{-1, +1\} = \mathcal{Y}$.
• True relationship between $\mathbf{x}$ and y	*target function* $f : \mathcal{X} \mapsto \mathcal{Y}$. (**The target f is *unknown*.**)
• Data on customers	*data set* $\mathcal{D} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. ($y_n = f(\mathbf{x}_n)$.)

We have identified some of the main elements of a learning task:

Input: feature vector
Output: class of label
Target function: unknown
Data Sample: i.i.d $\mathbf{x}_i$
Training sample: labeled data

When it is understood data set means training sample

More Examples

1. Use socio-economic data of families living in different areas of a city joint to the data of the last sales to predict the price of a house/flat.

2. Medicine: in cancer diagnosis the mitosis detection is relevant problem. Collect cells image to learn a binary classifier to predict mitotic cells.

3. Build a classifier for wine quality from physicochemical data of different type of wines and vineyard.

4. Build a classifier to recognize la presence of an object/face in an image.

5. Build a classifier to recognize a generic object(categories): dog, cat, tree, pedestrian, etc

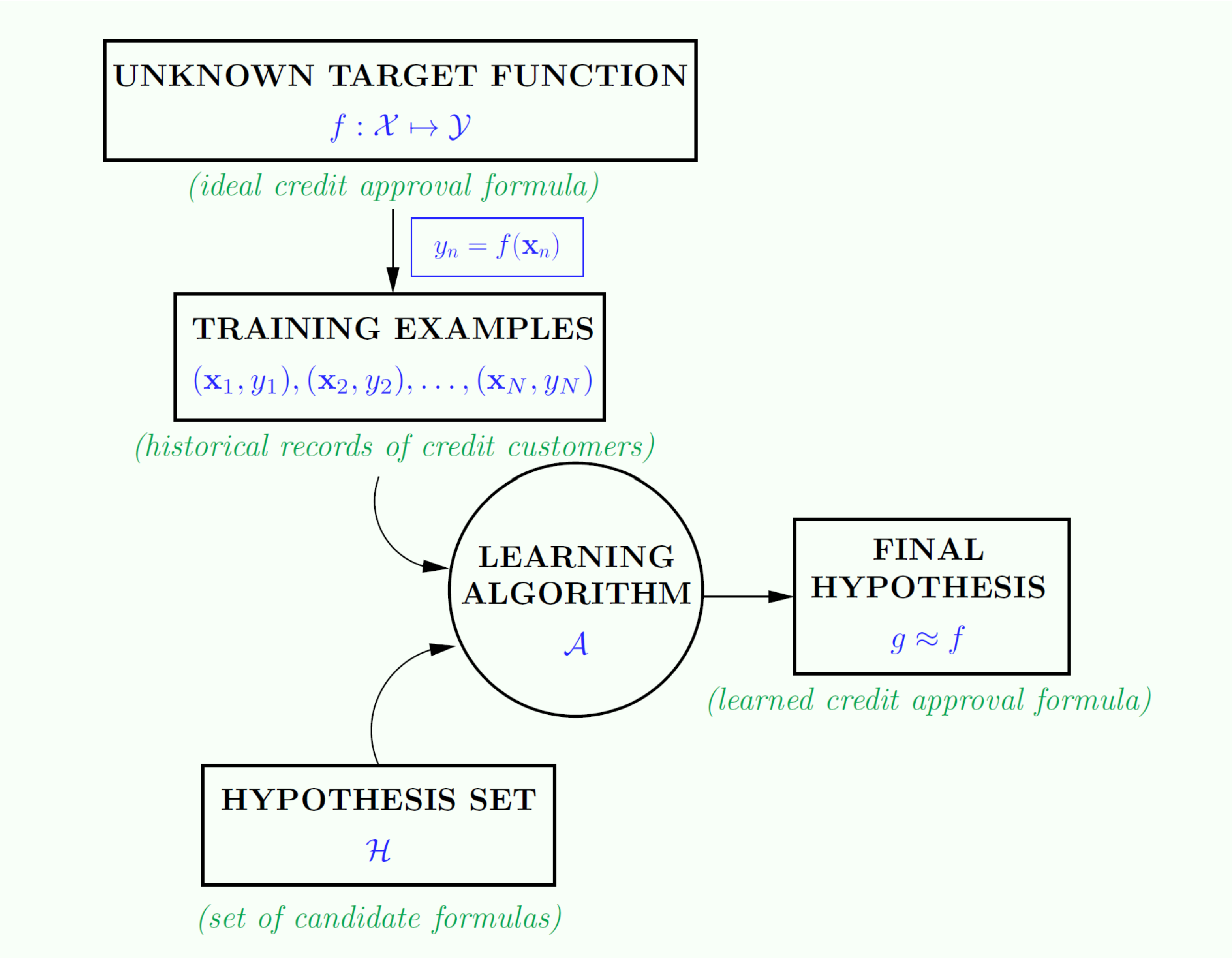
6. Classify documents according to its topic

7. Voice recognition

8. Automatic translation

9. etc, etc

Diagram of the initial Learning Setup



UNKNOWN TARGET FUNCTION

$f : \mathcal{X} \mapsto \mathcal{Y}$

(ideal credit approval formula)

$y_n = f(\mathbf{x}_n)$

TRAINING EXAMPLES

$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$

(historical records of credit customers)

LEARNING ALGORITHM

$\mathcal{A}$

FINAL HYPOTHESIS

$g \approx f$

(learned credit approval formula)

HYPOTHESIS SET

$\mathcal{H}$

(set of candidate formulas)

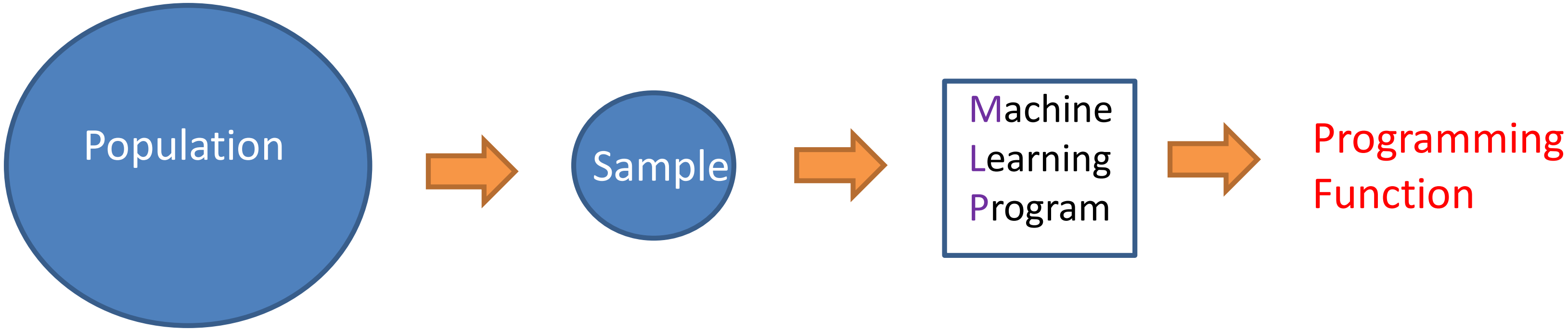
Detailed Summary

- A feature space/ Domain set/ Set of feature: $\mathcal{X}$
- A label set: $\mathcal{Y}$ (discrete: $\{(0,1),(-1,1),(1,2,3,..,M)\}$ o continuos: $\mathbb{R}$)
- Unknown target/label function: sample→ label ($f: \mathcal{X} \rightarrow \mathcal{Y}$ The unknown true function)
- Training examples: set of labeled samples , $\mathcal{D} = \{(x_i, y_i)), i = 1, \cdots, N\}$
  - independent and identically distributed (i.i.d.) samples from $(\mathcal{X}, \mathcal{Y})$

Modelling

- Hypothesis Set: $\mathcal{H}$ (Set of candidate function where the approximated solution is chosen from)
  - Solution hypothesis: a function $g \in \mathcal{H}$
- Learning optimization criteria ($\mathcal{L}$): Criteria used to choice the function g (ERM,SRM, MDL)
- **Loss Function**: $L(h) \in \mathbb{R}^+, h \in \mathcal{H}$ (Criteria used to assign a value of merit to each function)
- Learning Algorithm: $\mathcal{A}$ (Algorithm that determines, **from samples**, the best global hypothesis.)

A learning task sketch



Main question: How to choose a sample and learn from it an accurate prediction-function in the whole population?

- Sample: i.i.d from the population
- Machine Learning Problem: equivalent to fix $(\mathcal{H}, \mathcal{A}, L)$

Review

- What we already know:
  - The formal elements of a learning-from-data approach
  - The elements to be fixed before starting
  - The learning-from-data goal

Learning vs Design: What Learning is not !

- Some approaches only use data to fix some parameters of a well specified problem: this is design !

- Example: Let assume we want to build a model to recognizing coins from its size and mass: { (size_i,mass_i), i=1,..,N}

- Design: We collect information on the size and mass of each coin type and the number of coins in use. We build a physical model for mass and size, taking into account the variations by the use and the measured errors. Finally, we build a probability distribution on (size,mass) that we use to classify.

- Learning: We collect labeled data from each type of coin. The learning algorithm searches for a hypothesis that classifies the data well. To classify a new coin we use the learned hypothesis

- The information about *f* is key to adopt one or another approach.

Approaches of learning

- Machine Learning (computer science):
  - Main focus is on accurated prediction from large scale problems (generalization is important!)
  - Algorithm efficiency is an issue
  - Very dependent on the advances in optimization and regularization techniques.
  - Cons: Overfitting is always a possibility

- Statistical Learning: (statistical goals)
  - Main focus is inference (explaining the data) using probability distributions
  - Good results only under the assumed hypothesis
  - Very poor attention to very large scale problems

- Data Mining (statistical & computer science):
  - The main focus is extract dependences between variables in large databases. That is large inference
  - Shares many tools with M.L.
  - Algorithms and hardware with high level of scalability are important

- Bayesian Learning (probabilistic)
  - A full probabilistic approach based on a-priori distributions as prior knowledge
  - Overfitting is not in general an issue
  - Much more complex mathematically and computationally
  - Very poor attention to algorithm and computational issues