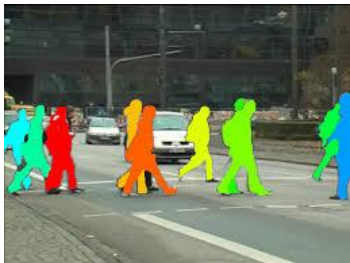# Learning **from Data** **(Machine Learning)**

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.

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
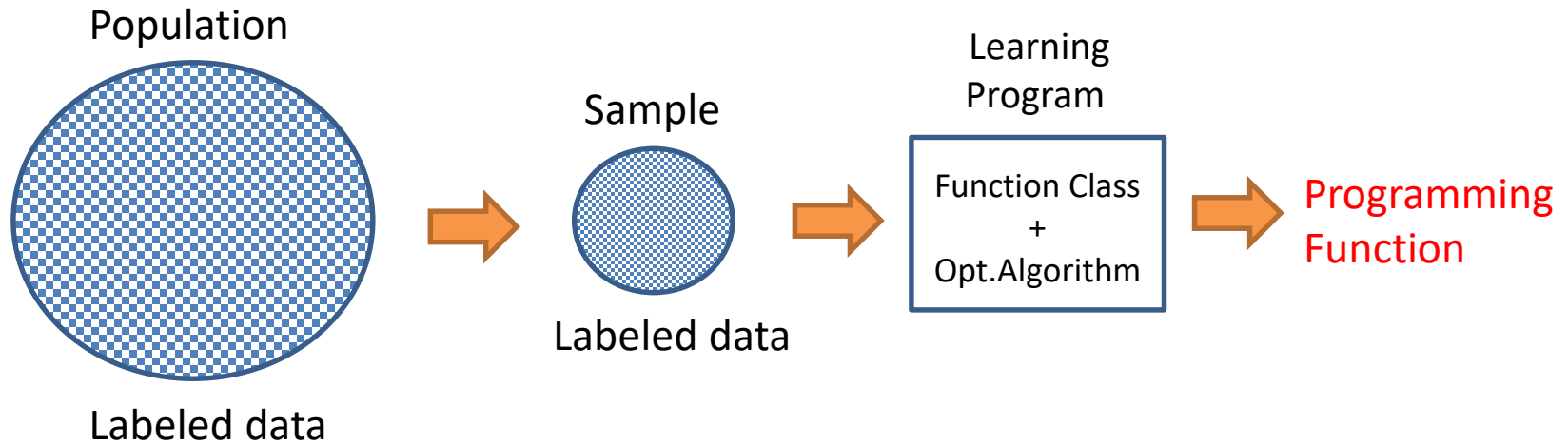 - Very poor attention to algorithm and computational issues

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: $(s_i, d_i), i = 1, \dots N$

- 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 …

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

- *Available information* about the problem *is the key* to adopting one or the other approach

A simple learning sketch

Population

Sample

Learning Program

Function Class
+
Opt.Algorithm

Programming Function

Labeled data

Labeled data

- 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?

Try to identify the function class and opt. algorithm for the previous example

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 (**representation**)

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
- And a very large etc,...

Main supervised tasks

- All 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/categorical variable**)
    - Predict the weather for tomorow : (sunny, cloudy, windy)
    - Predict if an image contains a face: (Yes,No), (0,1), (1,-1), etc
    - Predict if an email is SPAM or not: (Yes, No),), (0,1), (1,-1), etc

- Probabilistic Classification: the output is a probability vector over the labels
    - Predict real values
    - Solve the same problems as Classification

Let's check

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.
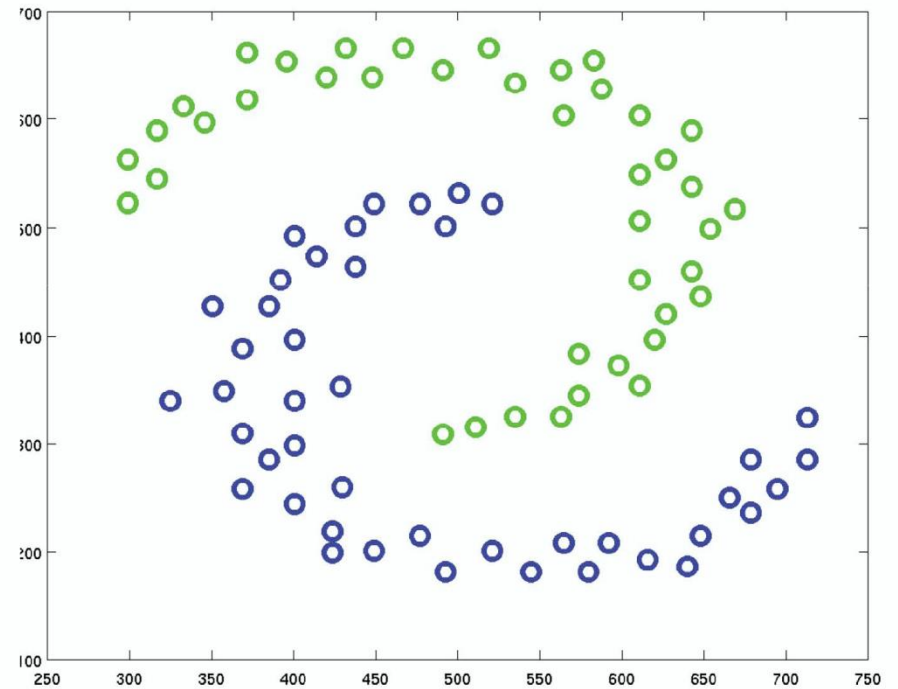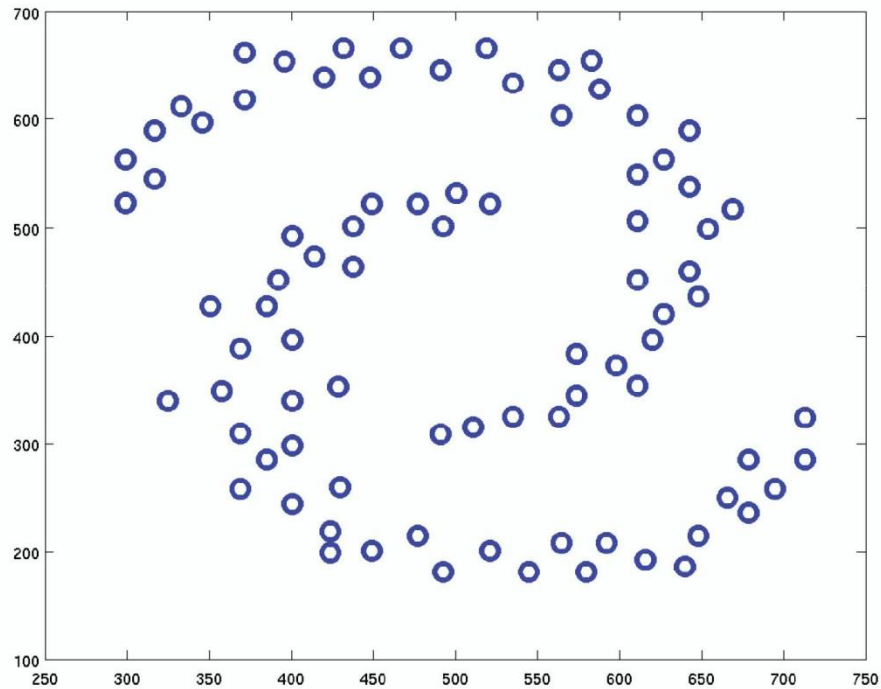
Unsupervised Learning

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

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

Properties

Animals
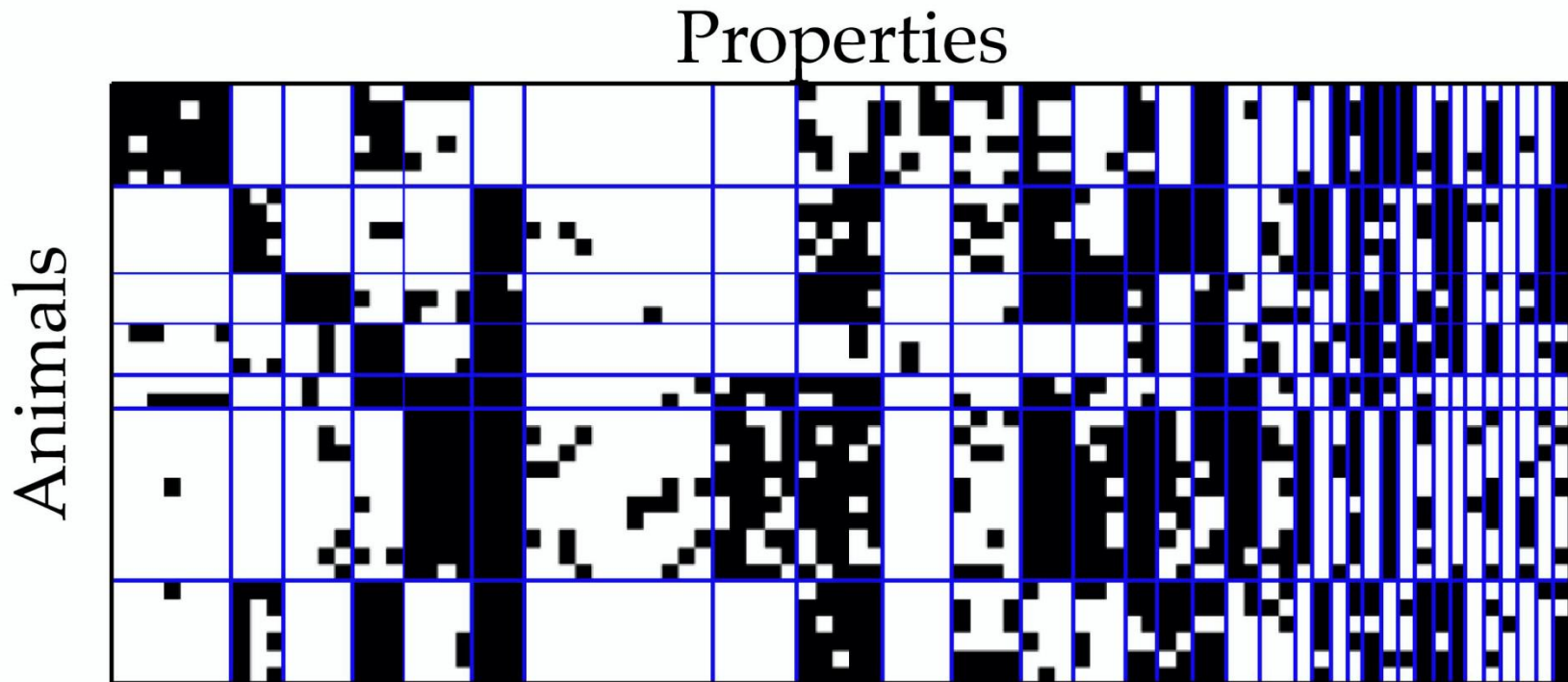
(J. Tenenbaum)

Dimensionality Reduction

Dimensionality Reduction

What we already know

SUPERVISED MACHINE LEARNING

INPUTS →

OUTPUTS →

Computer

→ PROGRAM

UNSUPERVISED MACHINE LEARNING

INPUTS →

Computer

→ PROGRAM

"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."

Let's start formalizing the approach

1. **What is the available information**?
   - Where the data come from? : $\mathcal{P}(\mathcal{D})$
   - What features to use? $\quad X$
   - Any sampling condition?: independent identically distributed (i.i.d.)

2. **What is the prediction task** ? $\quad f : X \rightarrow Y$ *(Labels)*

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

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

Let's try to identify the elements....

Identify handwritten digits in ZIP codes

Detecting faces in an image

Regression or classification ?
Supervised or unsupervised ?

Elements of the Credit Approval problem

- Salary, debt, years in residence, ... $\qquad$ $input\ \mathbf{x} \in \mathbb{R}^d = \mathcal{X}.$

- Approve credit or not $\qquad$ $output\ y \in \{-1, +1\} = \mathcal{Y}.$

- True relationship between $\mathbf{x}$ and y $\qquad$ $target\ function\ f : \mathcal{X} \mapsto \mathcal{Y}.$

  (The target f is *unknown*.)

- Data on customers $\qquad$ $data\ set\ \mathcal{D} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N).$

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

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

Input: feature vector
Output: class or 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

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?

Learning in action

- Let's assume X, Y and D are given by the learning task.

- Start with a set of candidate hypotheses H which you think are likely to represent f

$$H = \{h_1, h_2, \ldots, \}$$

  is called the hypothesis set or model.

- Select a hypothesis g from H. The way we do this is called a *learning algorithm*.

- Use g for new samples. We hope $g \approx f$.

The target f is fixed but unknown.

> **We choose H and the *learning algorithm* A**

This is a very general setup (eg. H could be the set of all possible hypotheses)

Diagram of the initial Learning Setup
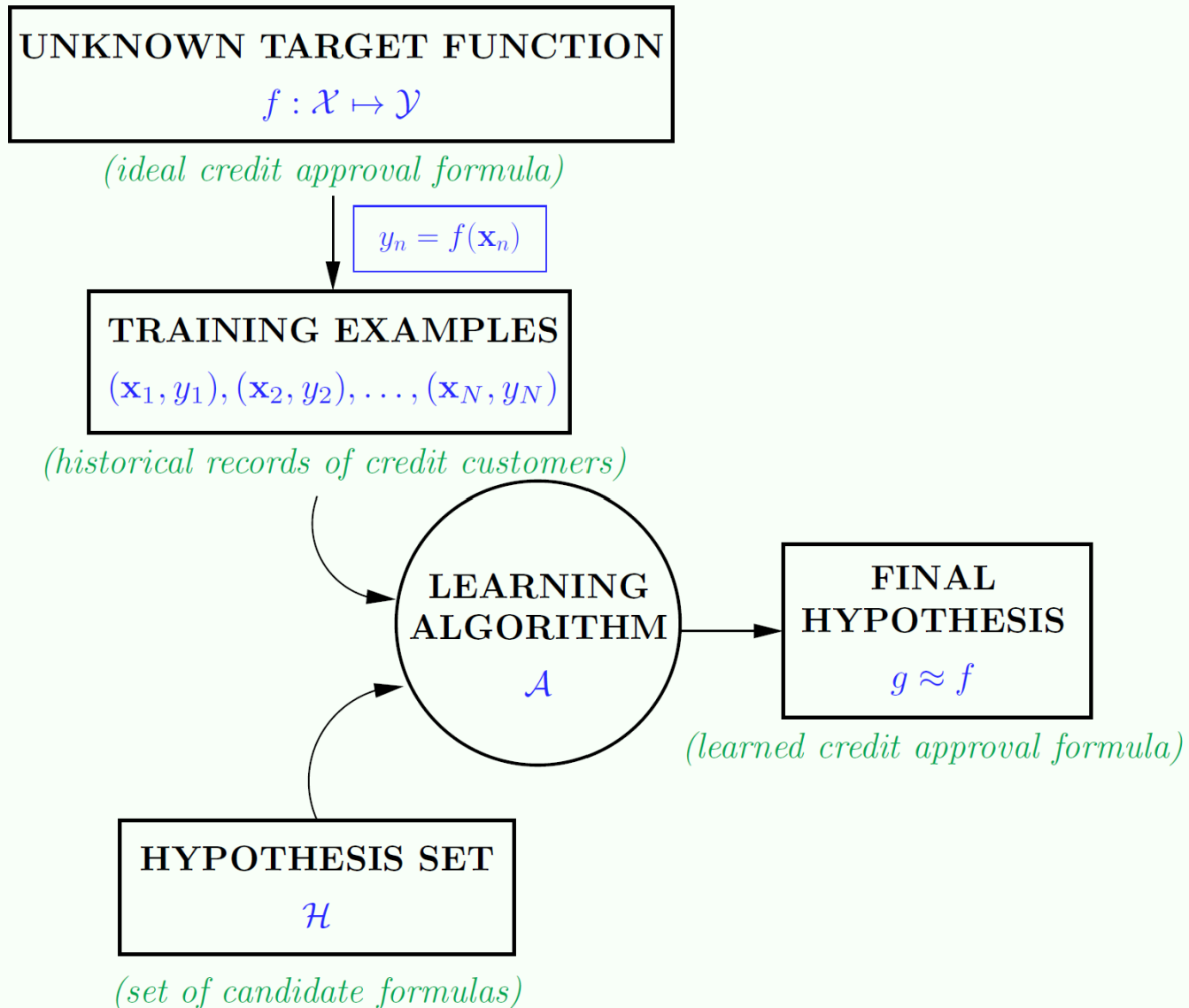
We already should know

– The formal elements of a learning-from-data approach

– The elements to be fixed before starting

– The learning-from-data goal