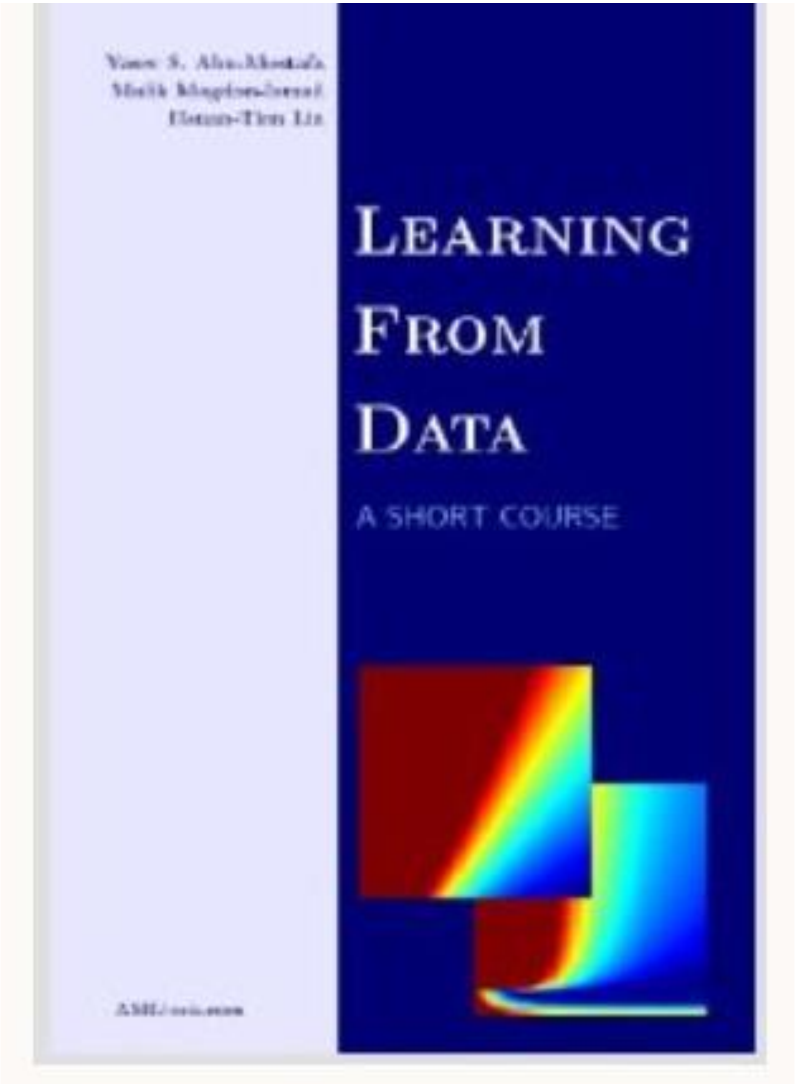
Machine Learning from Data

Lecture 1: Spring 2021

Resources

- Textbook (Yaser S. Abu-Mostafa, Malik Magdon-Ismail)
- Website
- Homework Submission: Submitty



Yaser S. Abu-Mostafa
Malik Magdon-Ismail
Hsuan-Tien Lin

LEARNING FROM DATA

A SHORT COURSE

AMLbook.com

Topics Covered in the Course

I

- What is Learning?
- Can we Do it? ✓
- How to Do it? ← Models
- How to do it well? Regularization
- General Principles of Learning
- Advanced techniques
- Other Learning Paradigms

Today’s lecture

- Motivation
- Learning Vs. Defining
- Formalize Learning
- Set-up a Machine Learning Problem

Application of Machine Learning
Manufacturing
Healthcare
Insurance
Customer Service
Transportation
E-commerce
Automobile

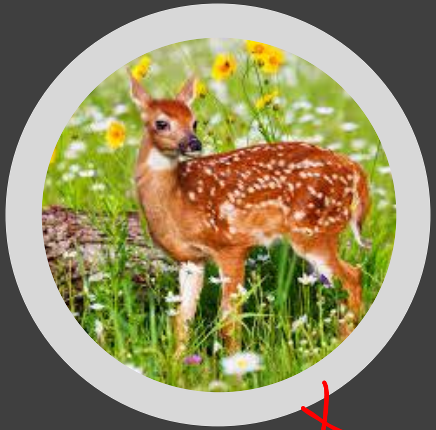
Machine Learning Everywhere

pic source: eduCBA

What is ~~Machine~~ Learning in General

- Ask a 5-year-old, is this a dog?
- Most likely, the answer is Yes

✓


✗

✗

✓

Are these Dogs?

- It is easy for humans to identify.
- Has anyone ever defined dogs for us?
- We have learned from data.

Can we define a dog?

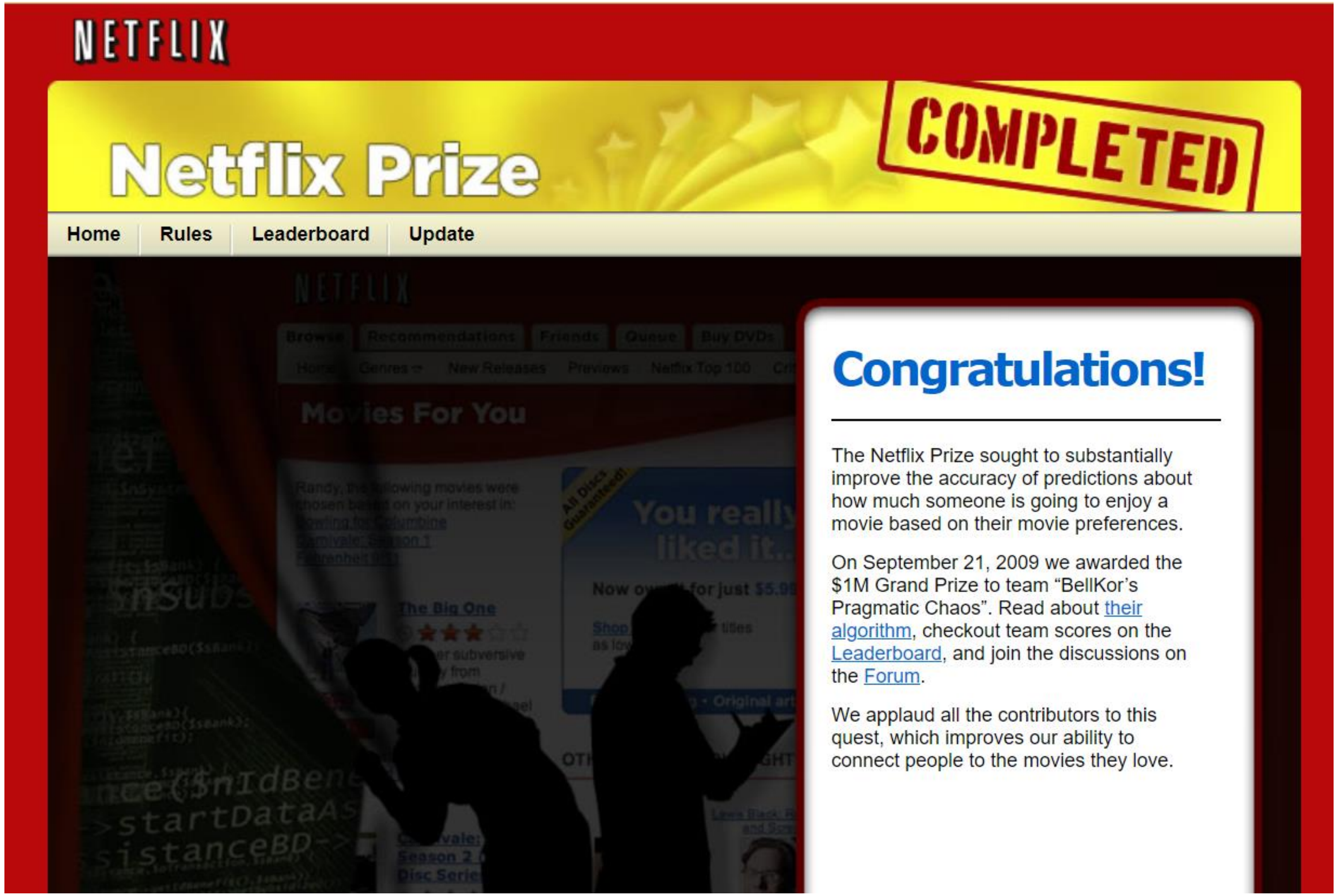
- Let us try.
  - Something that has 4 legs.
  - Runs with a certain speed
  - Something about facial features.

Color

Learning: “Which ones are dogs?”

- Defining is hard. ←
- Recognizing is Easy. ✓
- It is hard to give a mathematical definition of a Dog.
- A 5-year-old can tell the difference (they learned from Data).
- Learning from Data is used when we do not have an analytic solution.
  - We have data to construct an analytic solution.

The Netflix Problem



Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

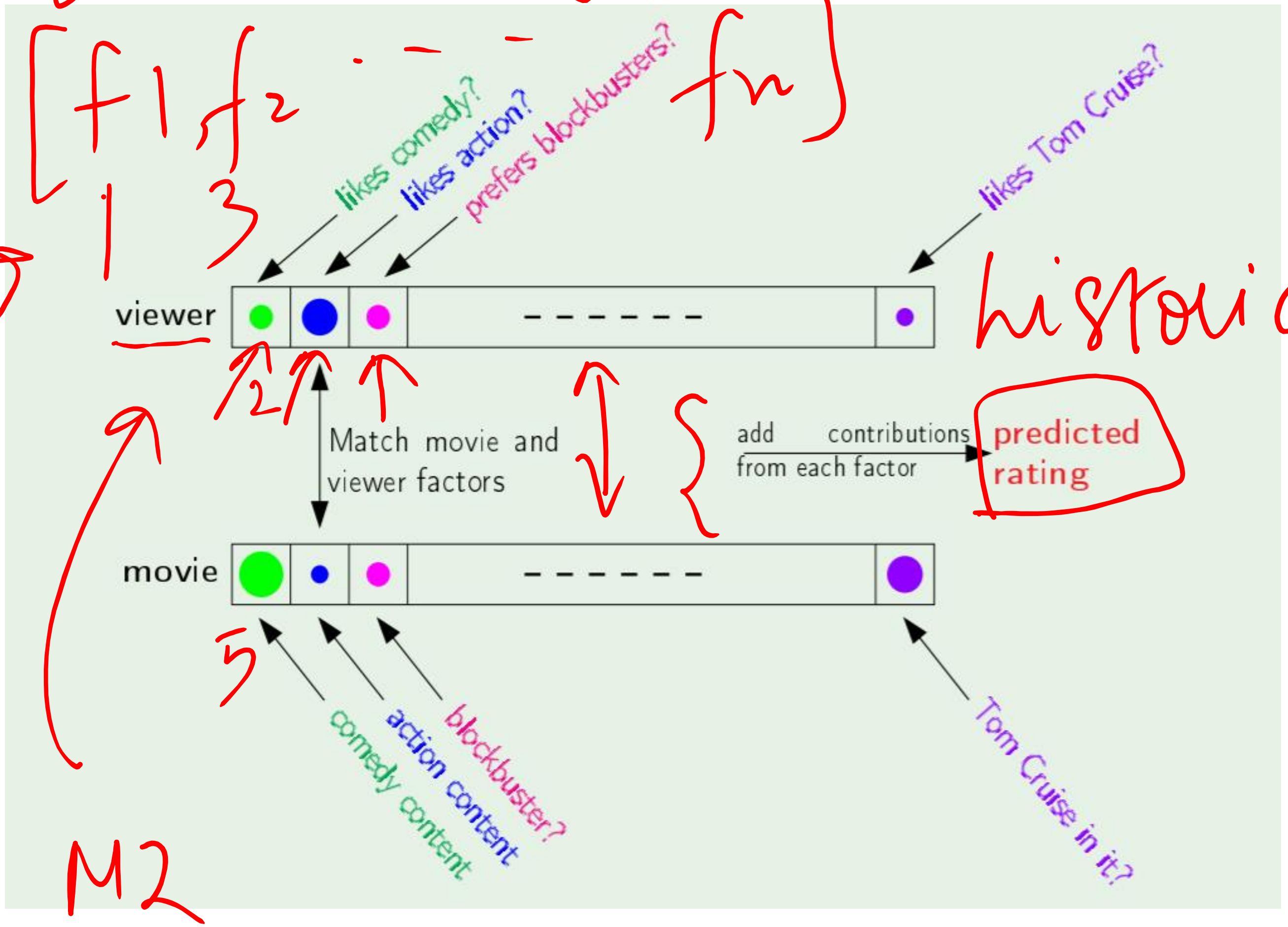
Problem Setup

$$V1 = [\overset{2}{f1}, \overset{5}{f2} \ldots fn]$$

$$M1 = [\underset{1}{f1}, \underset{3}{f2} \ldots fn]$$

Empirical solution.

- Netflix Problem: Predict recommendations, get more subscriptions.
- Criteria used to rate movies – Unknown/Complex
- Create user and movie profiles.
- Calculate predicted rating.
- The learning algorithm 'reverse engineers' the factors based on past ratings (starting with random factors mostly).
- It tunes these factors to make them more aligned with real ratings of viewers.



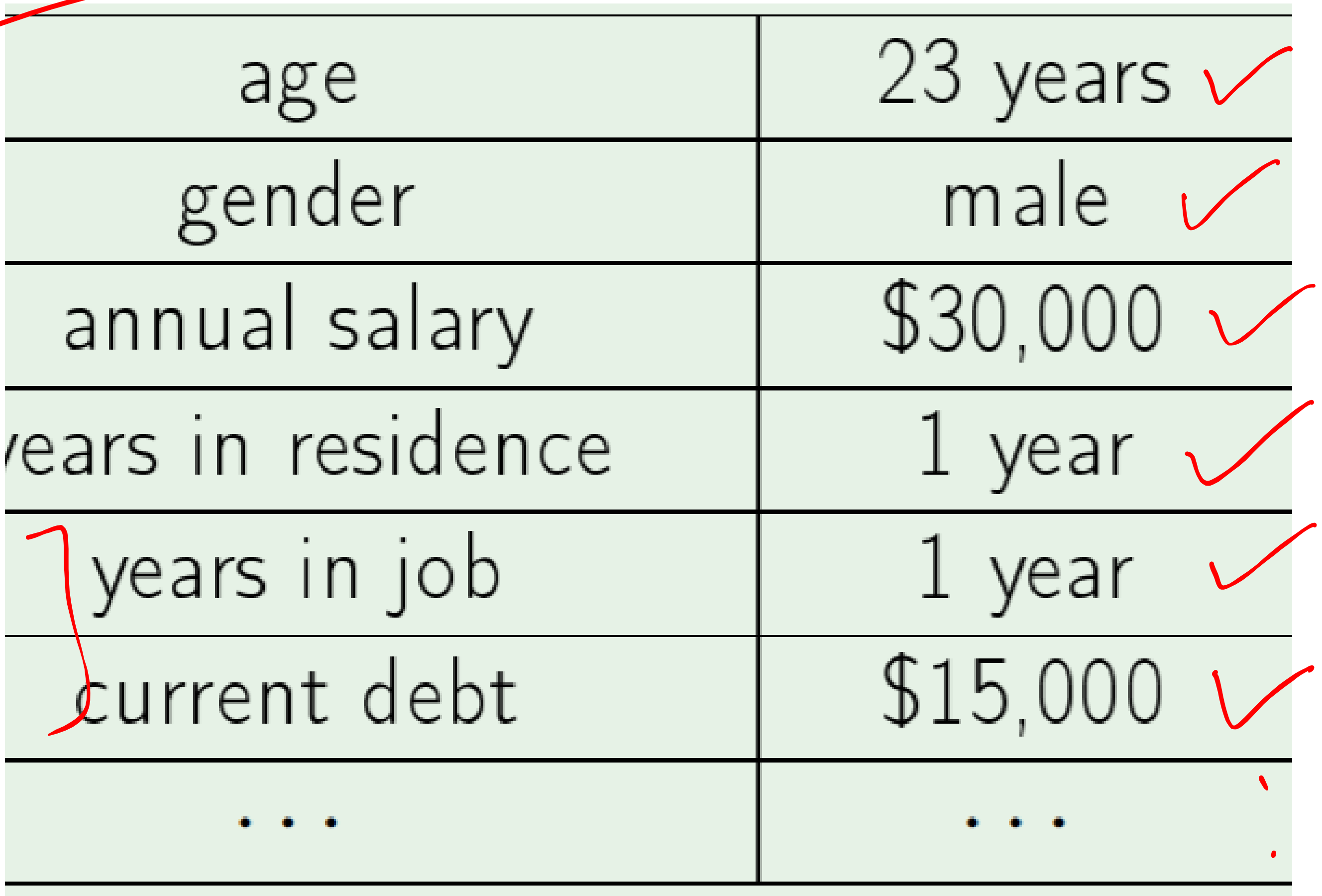
Components of Learning

Banks

Analytic

- **The Credit Approval Problem:**
- Approve or not?
- No magic formula exists.
- Banks have data: customer information like salary and debt; whether they defaulted on their credit or not.

C1 = [- - -]



age	23 years ✓
gender	male ✓
annual salary	$30,000 ✓
ears in residence	1 year ✓
years in job	1 year ✓
current debt	$15,000 ✓
...	...

Key Takeaway

- A pattern exists ✓
- We do not know it ✓
- We have data to learn it ✓

Formalize Components of Learning:

Vector → factors / attributes

N, Y

input $\mathbf{x} \in \mathbb{R}^d = \underline{\mathcal{X}}$.

output $y \in \{-1, +1\} = \underline{\underline{\mathcal{Y}}}$.

Real number

$\underline{\textit{target function}}\ f : \underline{\mathcal{X}} \mapsto \underline{\underline{\mathcal{Y}}}$.

(The target f is *unknown*.)

c_1 c_N

data set $\mathcal{D} = (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$.

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

$$\boxed{y_n = f(x_n)}$$

$y_i = f(x_i)$

$x_1, x_2, x_3 \ldots$

- **Input**: Salary, debt, years
  Output: Approve or not
  Target function: Relationship between X and Y
- $\boxed{\text{Data on customers}}$ *
- X, Y and D will be given by the learning problem.

credit level

The Learning Process

$H = \{\text{Universe of all functions}\}$

one

- Start with a set of possible Hypothesis that are most likely to represent the target f.

function → f $\quad g$ ←

- $\boxed{H = \{h_1, h_2, \ldots\}}$ is the hypothesis set or the ***model***.

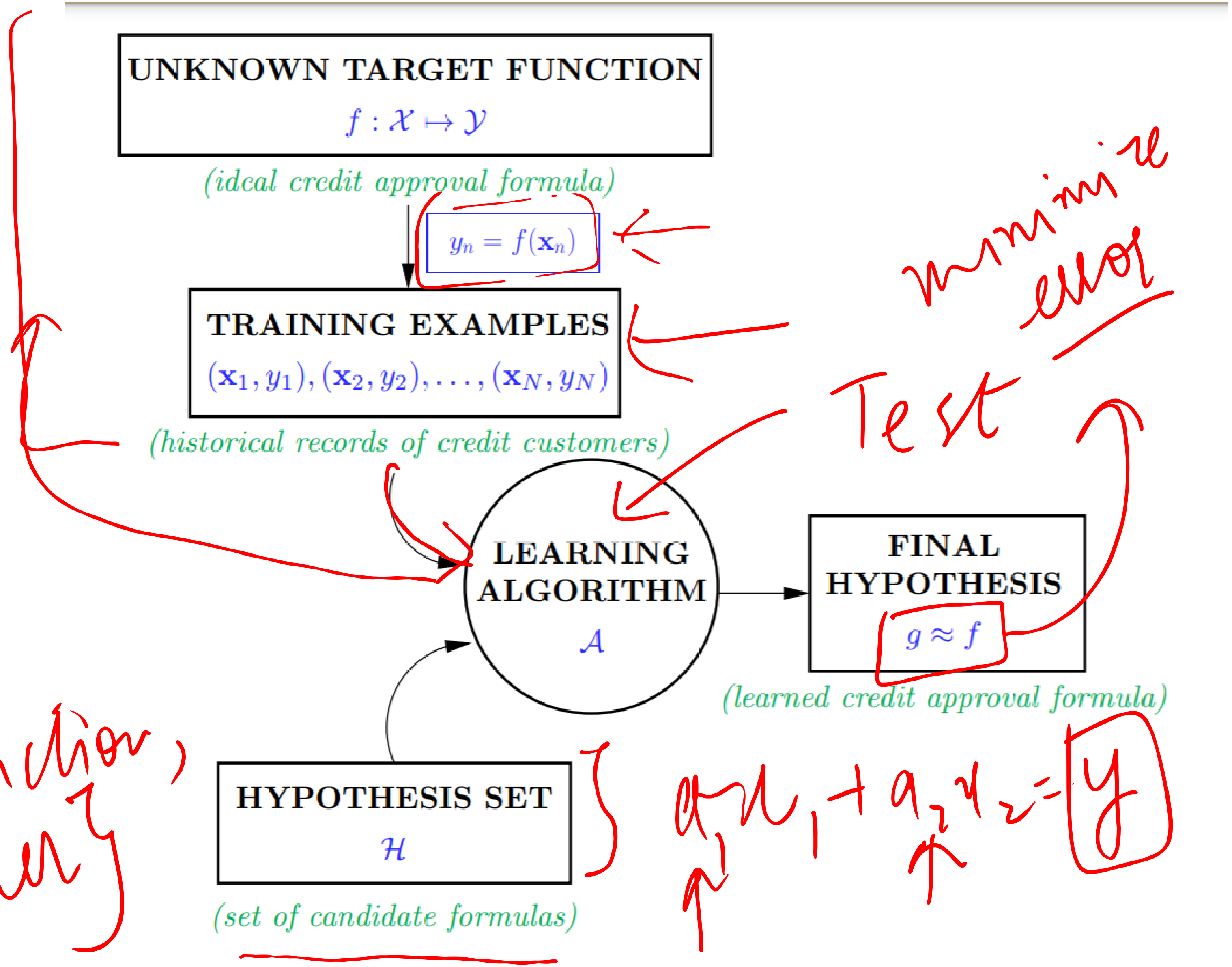
all possible

- Select a hypothesis g from H. The way we did this selection (process) is the ***learning algorithm***.
- Use this selected hypothesis to predict for new data (new customers). Our goal is to bring g as close to f as possible. The target f is fixed but unknown.

$y = g(x) \qquad g \simeq f$

- NOTE: We as ML practitioners will choose H and the learning Algorithm.

Summary of the Learning Set-Up



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)

minimize error

Test

{ linear function, 2nd order }

$a_1x_1 + a_2x_2 = \boxed{y}$