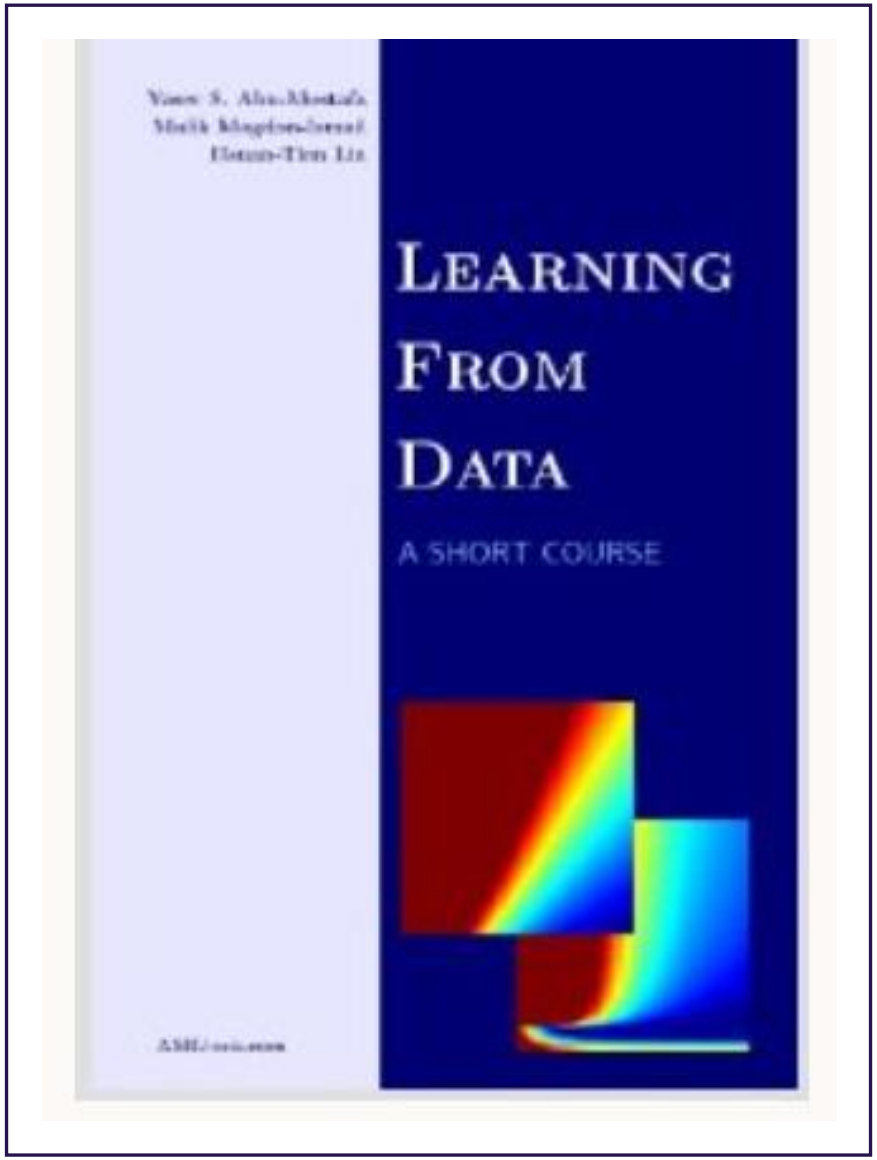
Machine Learning from Data

Lecture 1: Spring 2021

Resources

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



Topics Covered in the Course

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

Today’s lecture

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

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.

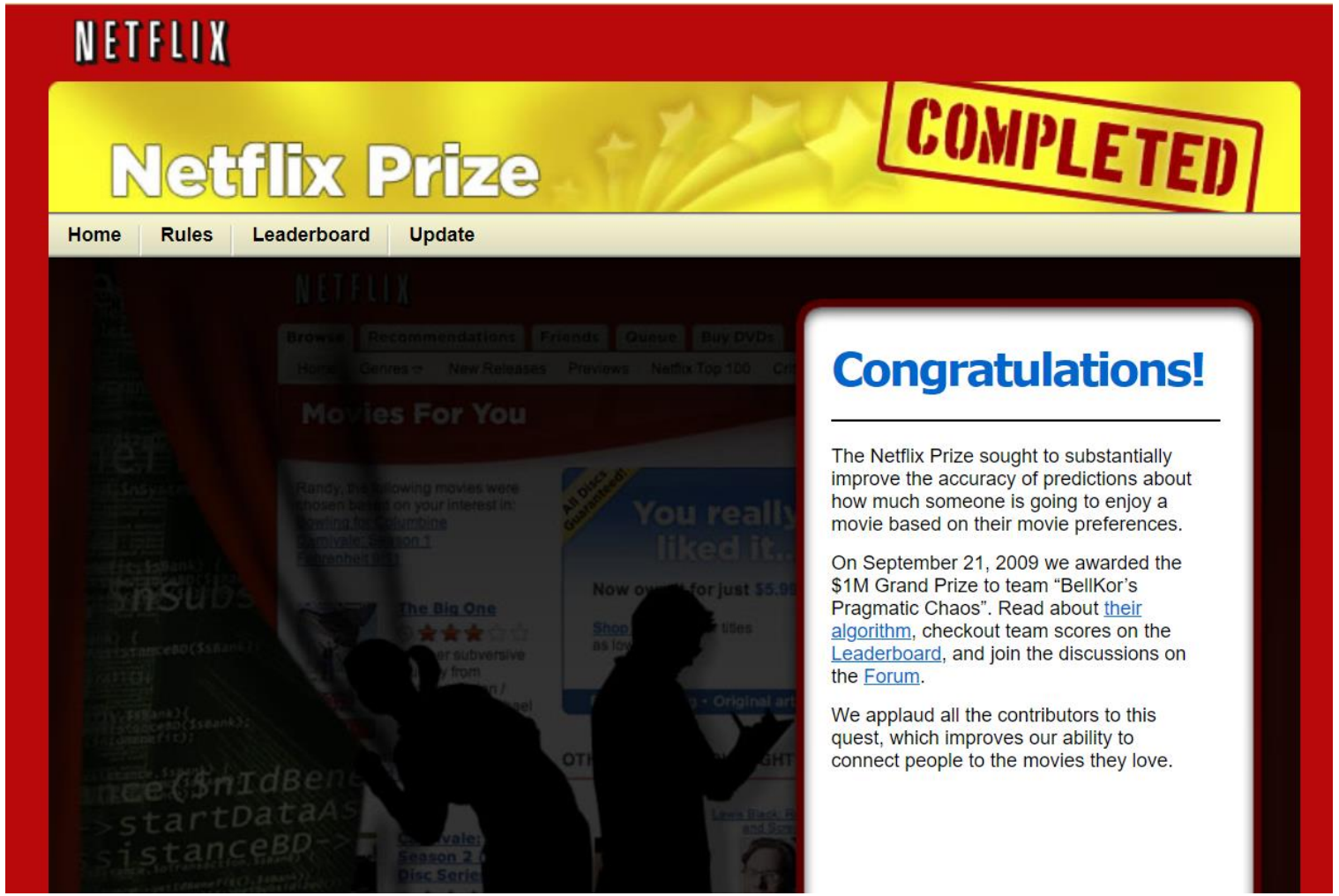
Can we define a dog?

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

Learning: “Which ones are dogs?”

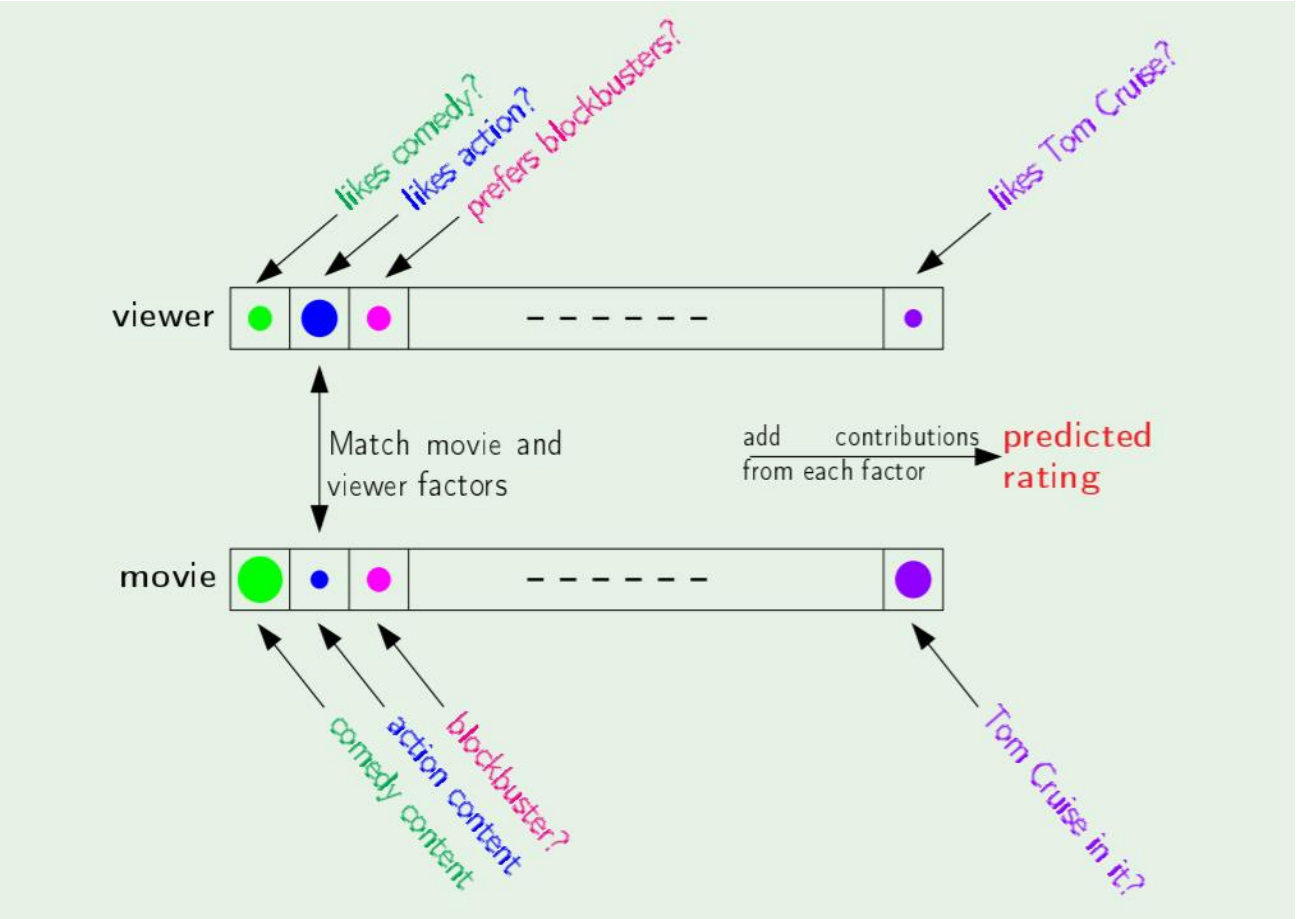
- Defining is hard.
- Recognizing is Easy.
- It is hard to give a mathematical definition of a Dog.
- A 5-year-old can tell the difference (they learned from Data).
- Learning from Data is used when we do not have an analytic solution.
  - We have data to construct an analytic solution.

The Netflix Problem



Problem Setup

- 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

- **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.

age	23 years
gender	male
annual salary	$30,000
years 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:

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

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

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

(The target f is *unknown*.)

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

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

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

The Learning Process

- Start with a set of possible Hypothesis that are most likely to represent the target f.
- $H = \{h_1, h_2, \ldots\}$ is the hypothesis set or the ***model***.
- 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.
- NOTE: We as ML practitioners will choose H and the learning Algorithm.

Summary of the Learning Set-Up