Machine Learning

Lecture 1: Introduction

Sibei Yang

SIST, ShanghaiTech

Email: yangsb@shanghaitech.edu.cn

Outline

- Course Information
  - Basic Information
  - Plan of the Course
- Introduction to Machine Learning
  - What is Machine Learning?
  - Examples
  - Types of Machine Learning
  - Applications

1.1 Course Information

Short Bio

- Dr. Sibei Yang

  Email: yangsb@shanghaitech.edu.cn

  - Assistant Professor at SIST
  - Office: 1C-403D
  - Research Interests: Computer Vision, Natural Language Processing, Machine Learning and the intersection of them.

Basic Information

- Time: **Monday & Wednesday**
  - **15:00-16:40**
- Place: 教学中心201
- Teaching Assistants:
  - 石骋 shicheng2022@shanghaitech.edu.cn
  - 李志伟 lizhw@shanghaitech.edu.cn
  - 胡修齐 huxq@shanghaitech.edu.cn
  - 胡浩炀 huhy@shanghaitech.edu.cn
  - 宋欣薇 songxw@shanghaitech.edu.cn
- Office Hours: To be announced on BB
- Course Site: 上科大教学互助平台(Blackboard)
  - **Questions/Discussion** on BB
  - Assignment#x/Quiz#x/Project#[team id]/Lecture#x/Others
- Email the teaching team **(ALL)** in **a manner**
  - Subject: [CS282] Assignment#x/Quiz#x/Project#[team id]/Lecture#x/Others

Basic Information

- Prerequisites: calculus (required), algebra (required), probability and statistics (required), programming languages (required), optimization (strongly recommended).

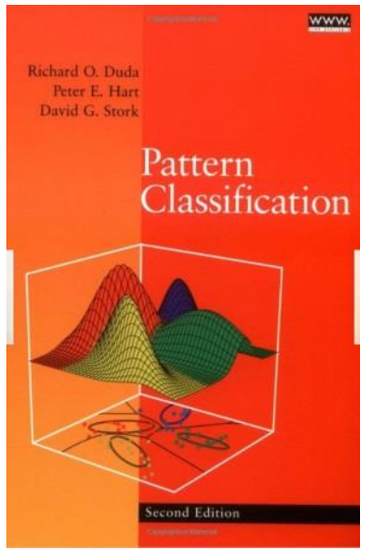
- Will be evaluated in the next quiz (Wednesday)

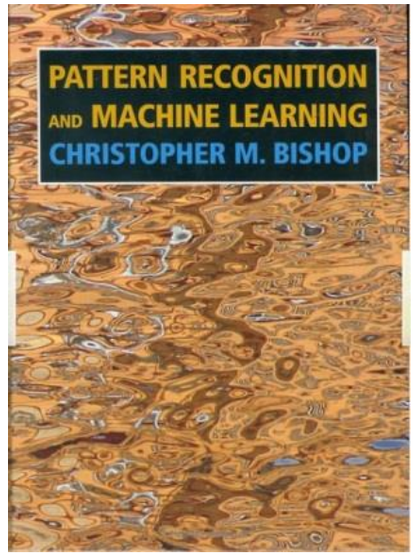
- Course Objectives:
  - Understanding of some of the important machine learning methods, theories, and algorithms.
  - Basic ability to use some machine learning techniques to solve real-world problems.

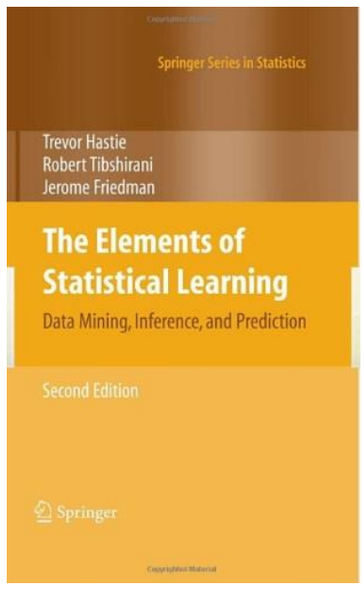
Textbooks and Slides

- 机器学习，周志华/ *“Learning from data.” Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin.*

- **[PC]** R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley, 2000
- **[PRML]** C. M. Bishop, **Pattern Recognition and Machine Learning**, Springer, 2006
- **[Elements]** T. Hastie, R. Tibshirani & J. Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** (2nd ed.), Springer, 2009







Some lectures will be based on these books/papers, but not all of them. Reading the textbooks is not required, but it is recommended. You are not responsible for textbook material that is not covered in lecture.

Acknowledgement: Some lectures are in reference to the course “Machine Learning” given by Dr. Hao Wang and “Learning from data” taught by Prof. Yaser Abu-Mostafa.

Grading Policy

- Evaluation
  - Assignments(40%) + Quizzes(20%) + Project(40%+5%)
- 4 Assignments: 10% x 4 = 40%
- 10 Quizzes (in class): 2% x 10 = 20%
- Final Course Project: 40% (+5%)
  - Proposal
  - Final Report (Conference format)
  - Presentation
  - Bonus points for novel results: 5%
- Late Policy
  - A total of 7 free late days to use, but no more than 4 late days can be used on any single assignment.
  - 如要使用free late days，需在当次due后4天内邮件给助教并说明days，过期无效。
  - After that, 25% off per day late
  - Does not apply to Final course project/Quizzes
- Grade Announcement
  - 7 days to ask questions. After 7 days, the score can not be revised.
- Collaboration Policy
  - Project team: 4~5 students
  - Grading according to each member's contribution (list the contribution percent on the project report)

Academic Integrity

- Academic Dishonesty
  - Plagiarism or unauthorize collaboration, projects, assignments, etc.
  - Getting code/document from the Internet
  - Asking someone else to write the code/document/answers… for you
  - Copying your friend’s code/document/answers
  - …

- Penalties for Violation
  - Zero points on the assignment/quiz in all questions.
  - Repeated violations will result in an F grade for this course as well as further discipline at the school/university level.
  - When one student copies from another student, both students are responsible.

- Plagiarism for assignments: cite your sources to avoid punishments!

- Plagiarism for final project: cite your references!

Course Policies

- Academic Dishonesty
  - No.

- Assignments/Quizzes/Project:
  - Write your own solution

- Submission via Blackboard
  - Email submission/other methods are not accepted, i.e., getting 0.
  - Blackboard上显示收到作业/项目/Quiz的时间作为提交时间。选择最后时刻提交，由于网络等原因造成的分数损失需自己承担。
  - Note: 点击“提交”按钮，而不是“保存”。

Why Take This Course?

- It is Not
  - Easy course with high scores
- You SHOULD:
  - Work hard
  - Be honest

1.2 Plan of this Course

Topics to cover

- **Learning Methods**
  Linear Regression, Logistic Classification, Probabilistic Graph Model, Temporal probability models, kNN, Classification models, Decision Tree, Clustering, Dimension Reduction, DNN (brief introduction)…

- **Learning Theory & Techniques**
  overfitting cross-validation, regularization (ℓ_1, ℓ_2)…

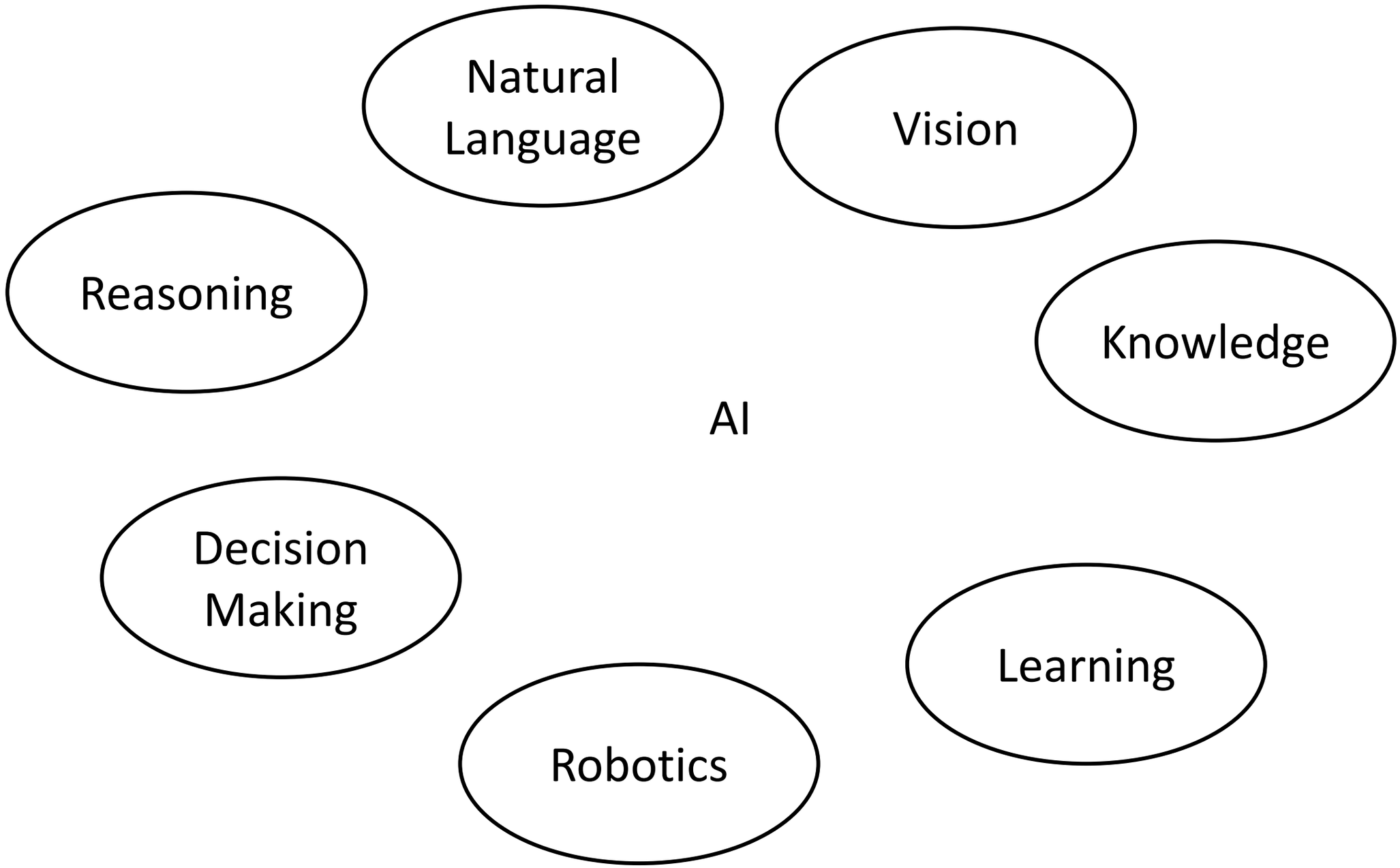
- **Learning Algorithms**
  GD, SGD~~, variance reduction, GP, ADMM, Newton Method, BFGS, IST, Coordinate Descent…~~

Schedule (May be adjusted!)

第一章：引论	课程简介 机器学习简介、定义、类型、案例	第一周
第二章：期望风险极小化	假设函数集合 数据的联合分布 期望风险极小化概念 经验风险极小化概念	第一周
第三章：线性模型：回归	线性回归 法方程 最大似然估计 随机梯度下降法	第二周
第四章：线性模型：分类	感知机 Logistic 回归 Softmax 回归	第三周
第五章：概率图模型	贝叶斯网络 马尔科夫网络 精确推理 近似推理	第四、五周
第六章：Temporal Probability models	马尔科夫模型 隐马尔科夫模型 动态贝叶斯网络 粒子过滤	第六周
第七章：学习理论	偏差-方差分解 过拟合 交叉验证	第七、八周
第八章：分类算法	K近邻 朴素贝叶斯 支持向量机	第八、九周
课程项目	课程项目提案	第九周
第九章：树	划分选择 剪枝处理 多变量决策树 boosting bagging与随机森林	**第十周**
第十章：聚类	层次聚类 K-means 高斯混合模型 谱聚类	第十一周
第十一章：降维	矩阵分解 主成分分析 线性判别分析 局部保留投影	第十二周
课程项目	课程项目分组讨论	第十三、十四周
课程项目	课程项目分组讨论 项目报告与展示	第十五、十六周

2.1 What is Machine Learning?

What is ML and Why ML?



What is machine learning?

- Vast amounts of data are being generated in many fields, and the statisticians' job is to make sense of it all: to extract important patterns and trends, and to understand "what the data says". We call this learning from data.—*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*
  - *"Data Mining" in the subtitle of the book.*

- This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. —*Machine Learning, A Probabilistic Perspective*

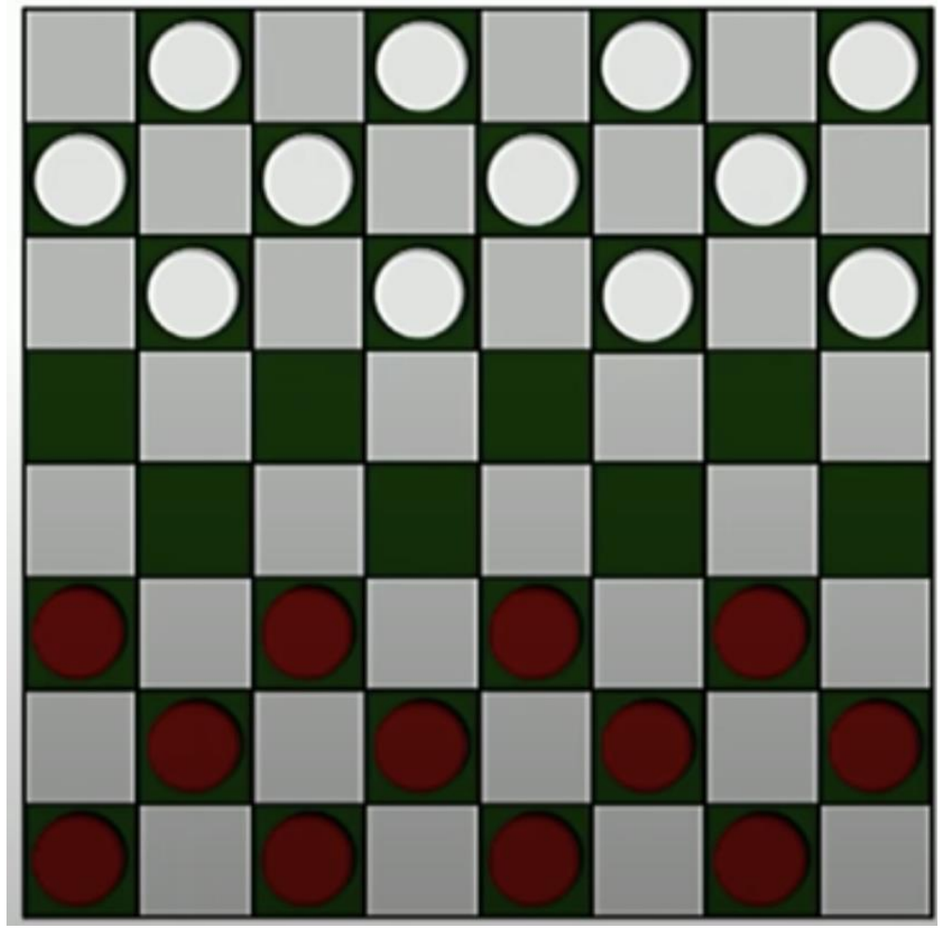
- Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field. —Pattern Recognition and Machine Learning

---

Read: https://machinelearningmastery.com/what-is-machine-learning/

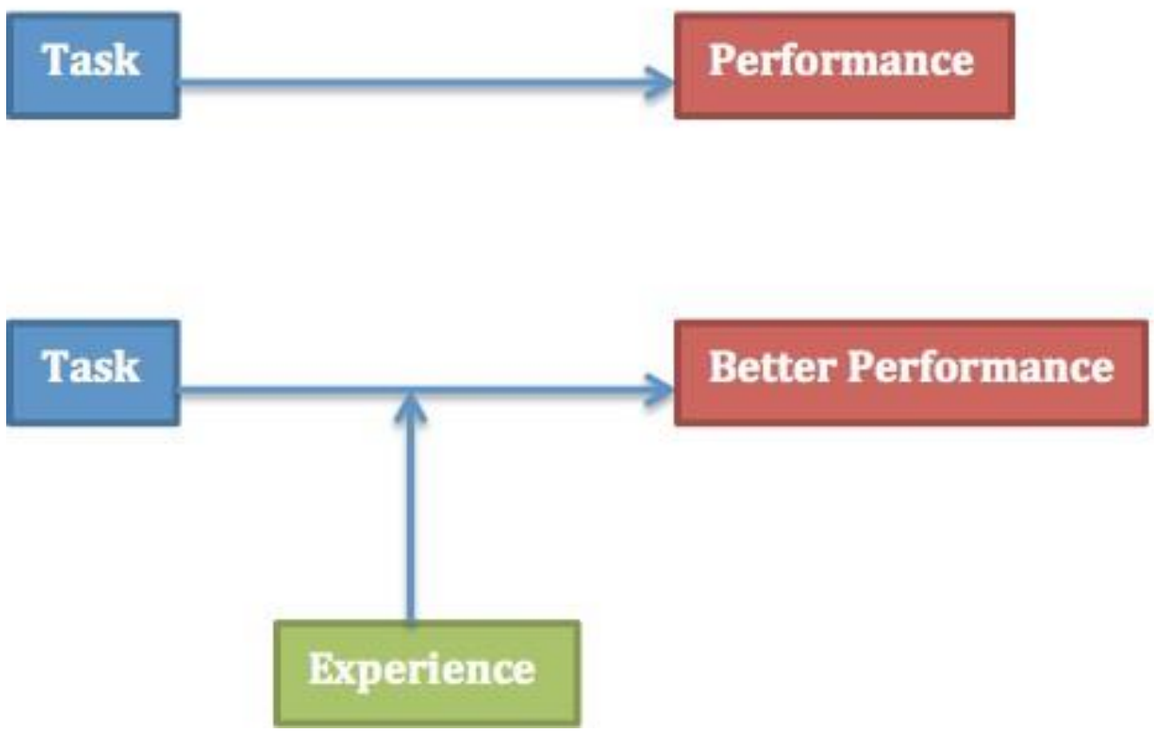
What is machine learning?

- Arthur Samuel (1959) “Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed”.



What is machine learning?

- Tom M. Mitchell (1998): "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".



Experience

2.2 Examples of Machine Learning

Example: Predicting how a viewer will rate a movie

10% improvement > 1 million dollar

The essence of machine learning:

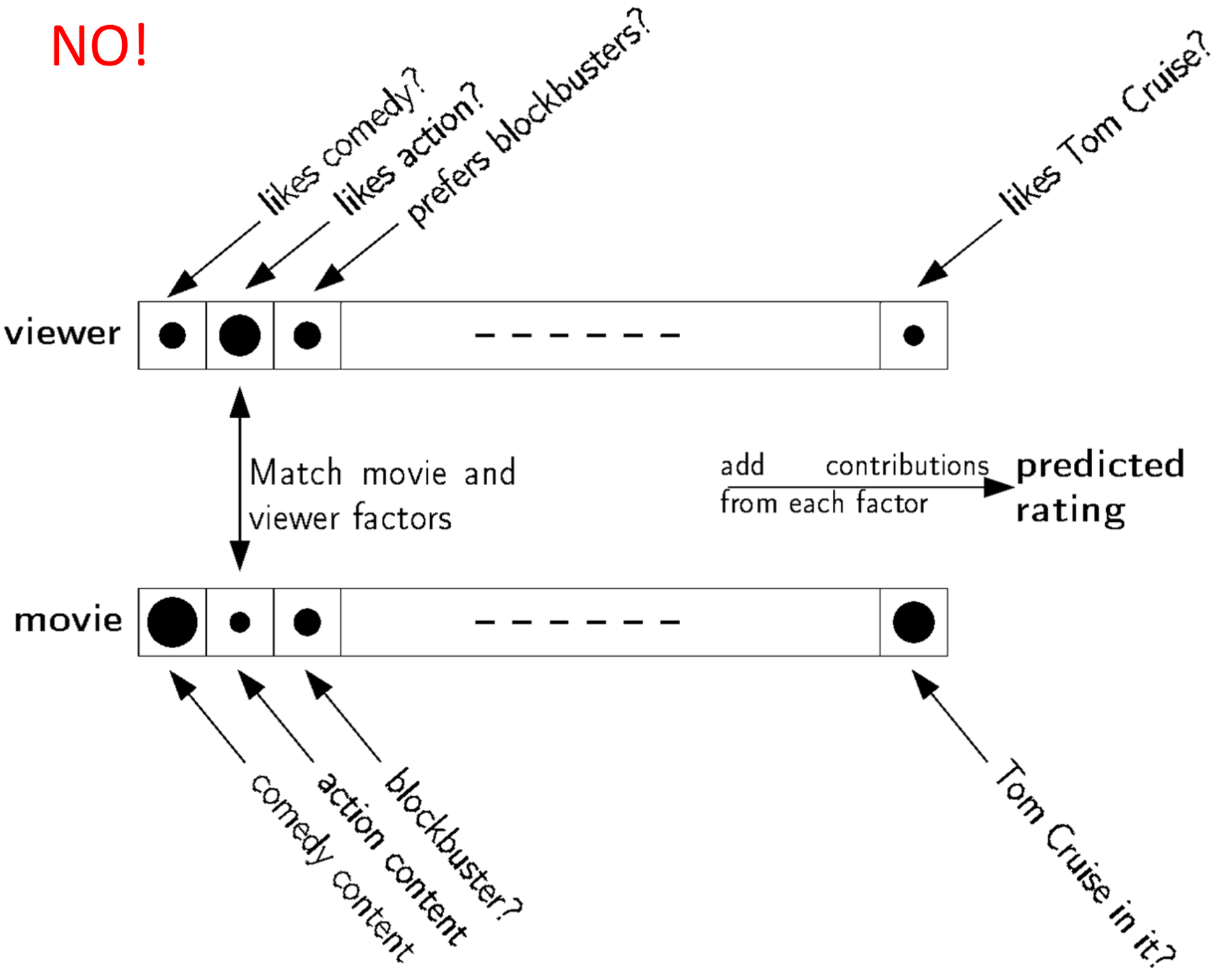
- A pattern exists

- We cannot pin it down mathematically

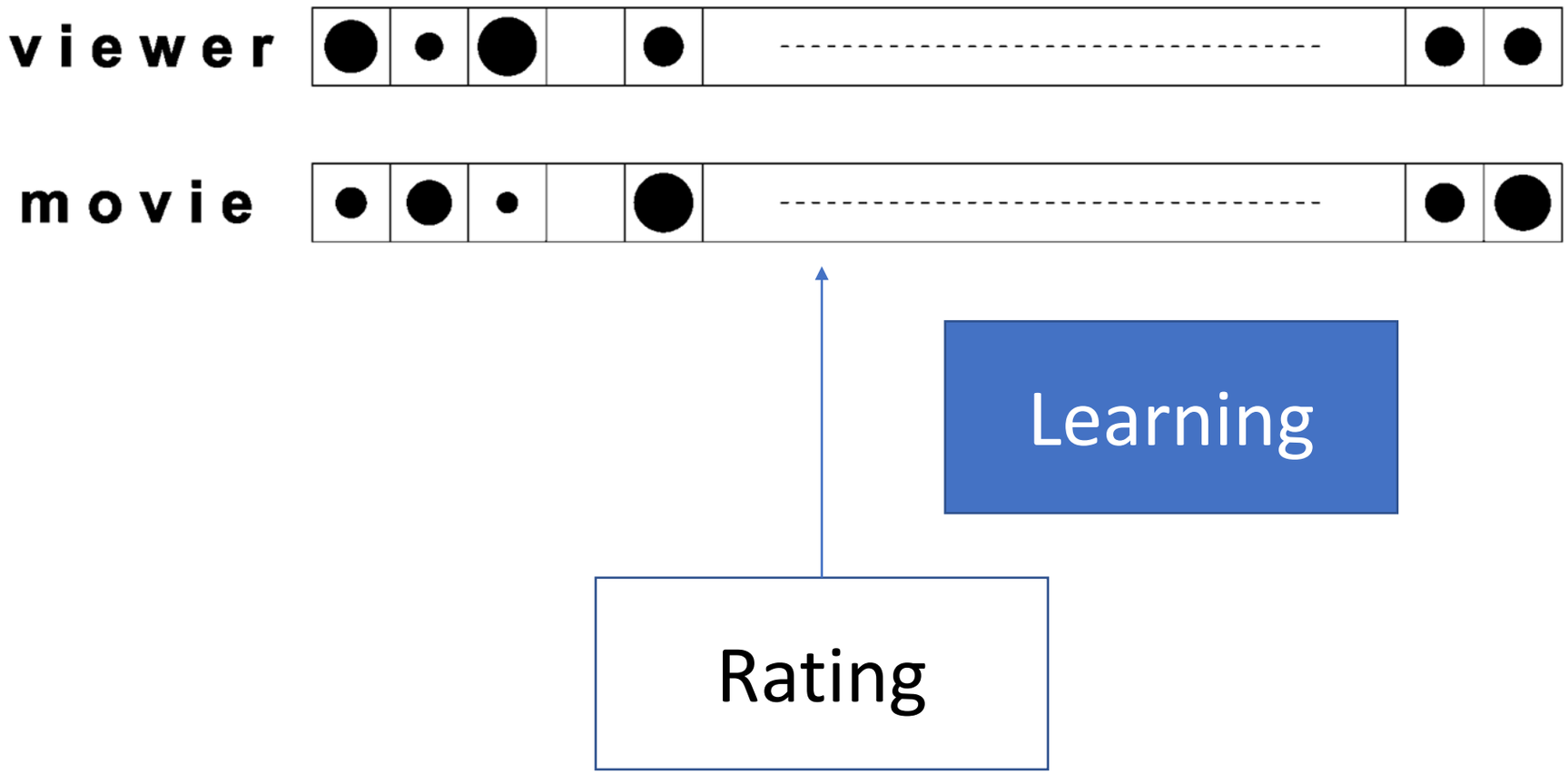
- We have data on it

Example: Predicting how a viewer will rate a movie

NO!



Example: Predicting how a viewer will rate a movie



Components of Learning

Application information:

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

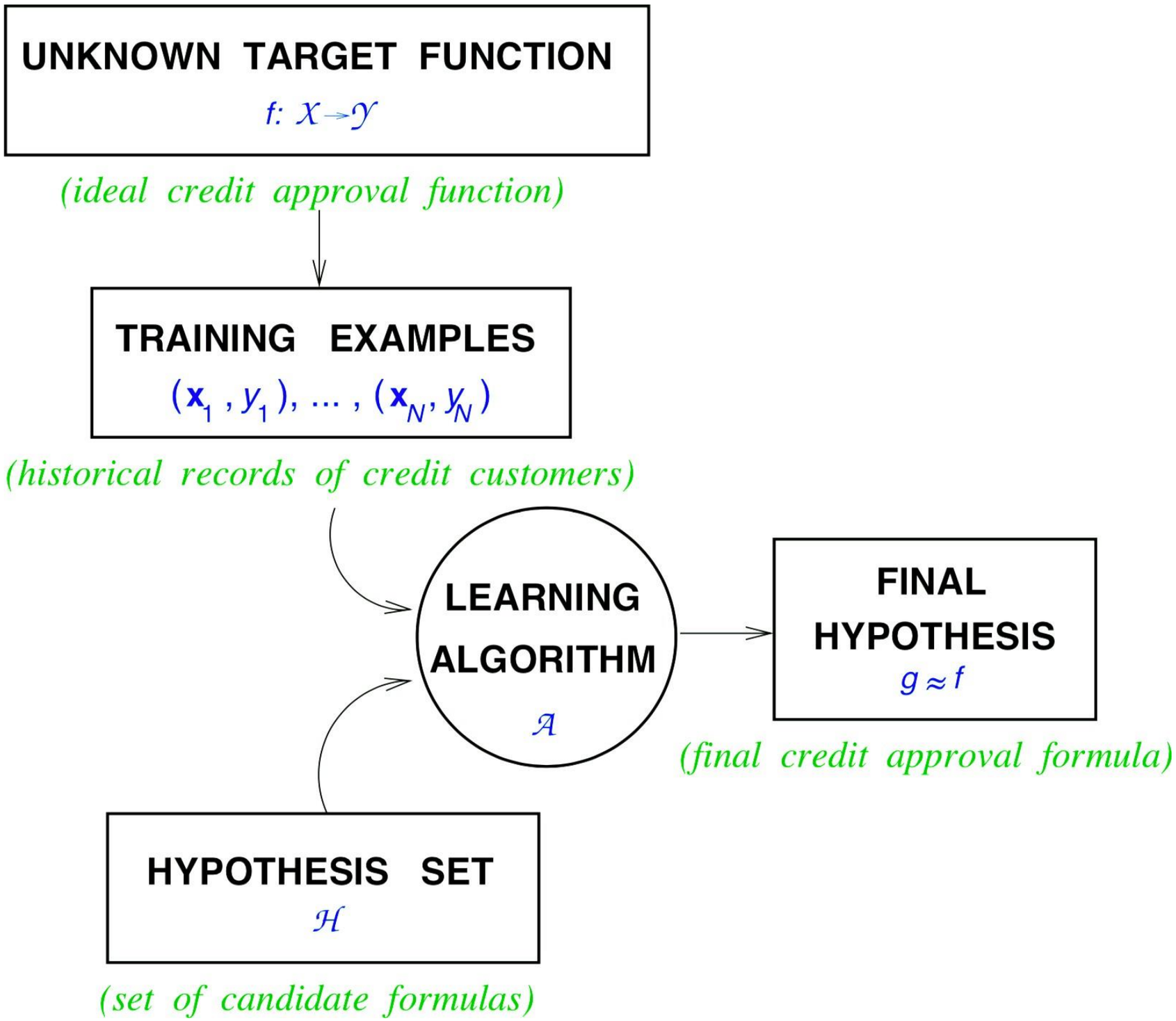
Approve credit?

Components of learning

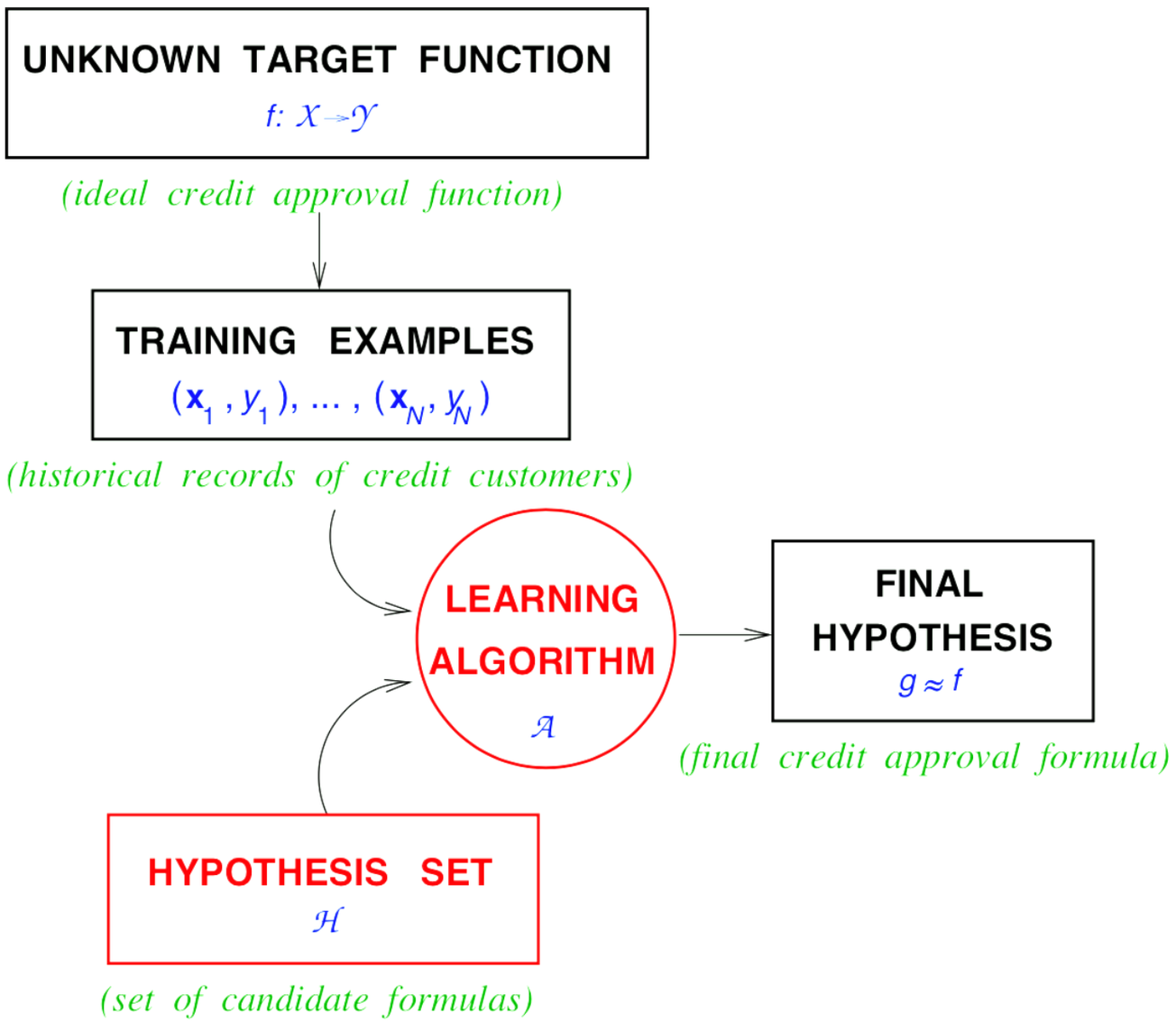
Formalization:

- Input: $\boldsymbol{x}$ (customer application)
- Output: y (good/bad customer?)
- Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ (ideal credit approval formula)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ (historical records)
- Hypothesis: $g : \mathcal{X} \rightarrow \mathcal{Y}$ (formula to be used)

Components of learning



Components of learning



Two solution components of the learning problem:

- The Hypothesis Set:
  $\mathcal{H} = \{h\}, \quad g \in \mathcal{H}$

- The learning algorithm

Together, they are referred to as the learning model.

A simple hypothesis set — the "perceptron"

- For input: $\boldsymbol{x} = (x_1, \ldots, x_d)$ (attributes of a customer)

$$\text{Approval credit if } \sum_{i=1}^{d} w_i x_i > \text{threshold}$$

$$\text{Deny credit if } \sum_{i=1}^{d} w_i x_i < \text{threshold}$$

- This linear formula $h \in \mathcal{H}$ can be written as

$$h(\boldsymbol{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \text{threshold}\right)$$

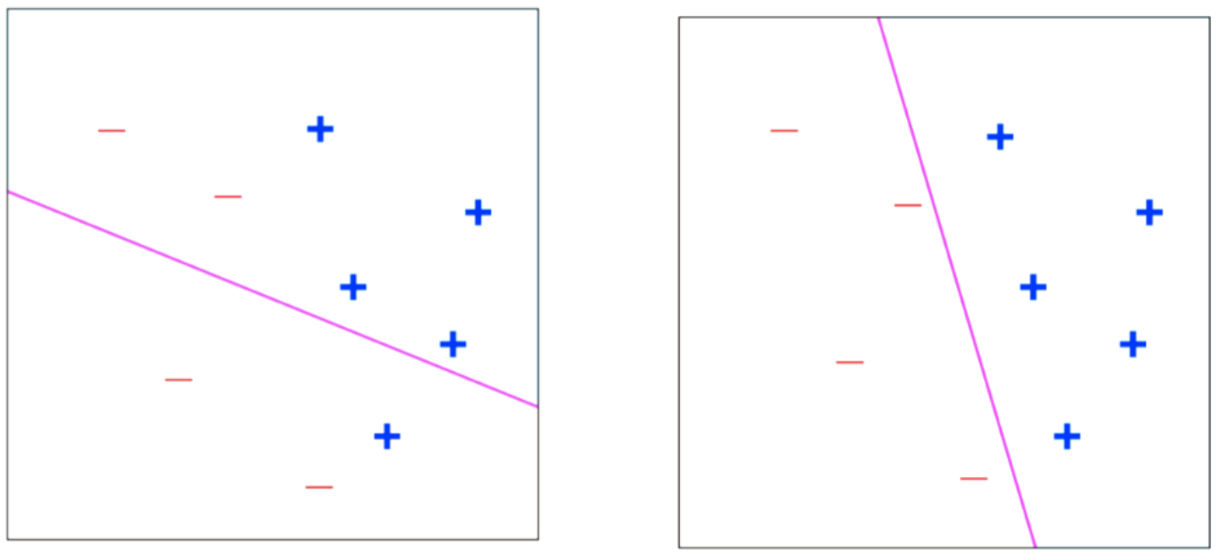
$$h(\boldsymbol{x}) = \text{sign}\Big((\sum_{i=1}^{d} w_i x_i) + w_0\Big)$$

Introduce an artificial coordinate $x_0 = 1$

$$h(\boldsymbol{x}) = \text{sign}\Big(\sum_{i=0}^{d} w_i x_i\Big)$$

In vector form, the perceptron implements

$$h(\boldsymbol{x}) = \text{sign}\big(\boldsymbol{w}^T \boldsymbol{x}\big)$$



'linearly separable' data

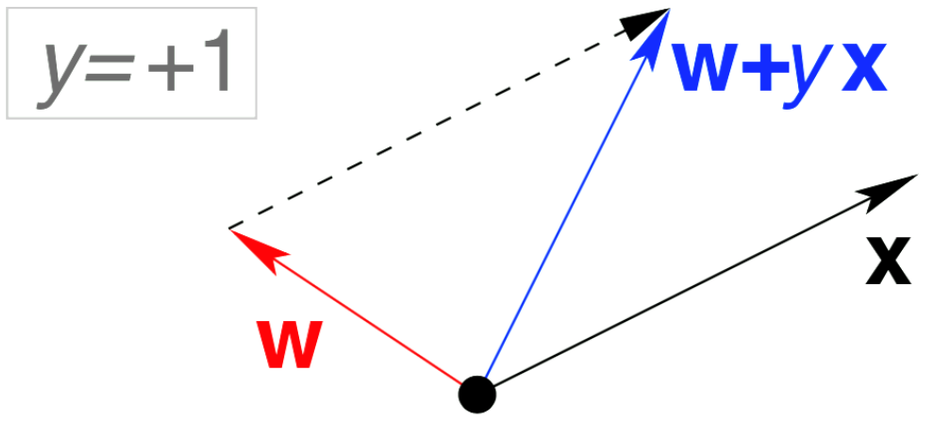
A simple learning algorithm — PLA

The perceptron implements

$h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x})$

Given the training set:

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

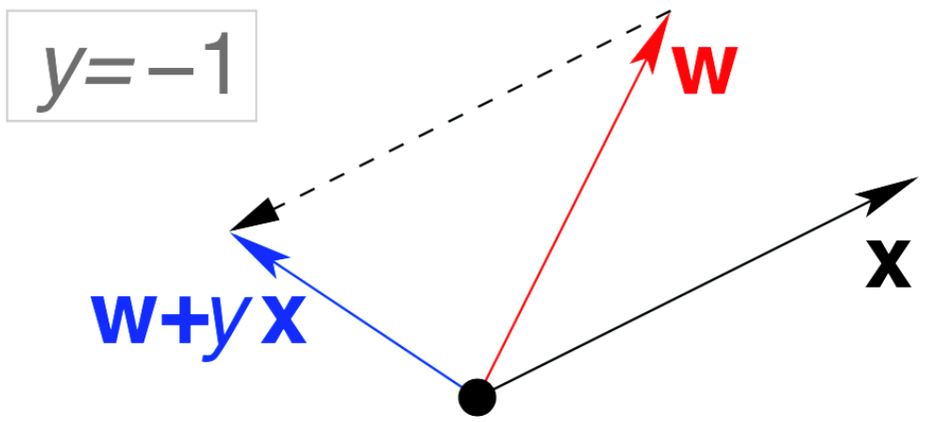


Pick a misclassifled point:

$\text{sign}(\boldsymbol{w}^T \boldsymbol{x}_n) \neq y_n$

and update the weight vector

$$\mathbf{w} \leftarrow \mathbf{w} + y_n \mathbf{x}_n$$



Iterations of PLA

- One iteration of the PLA,

  $$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

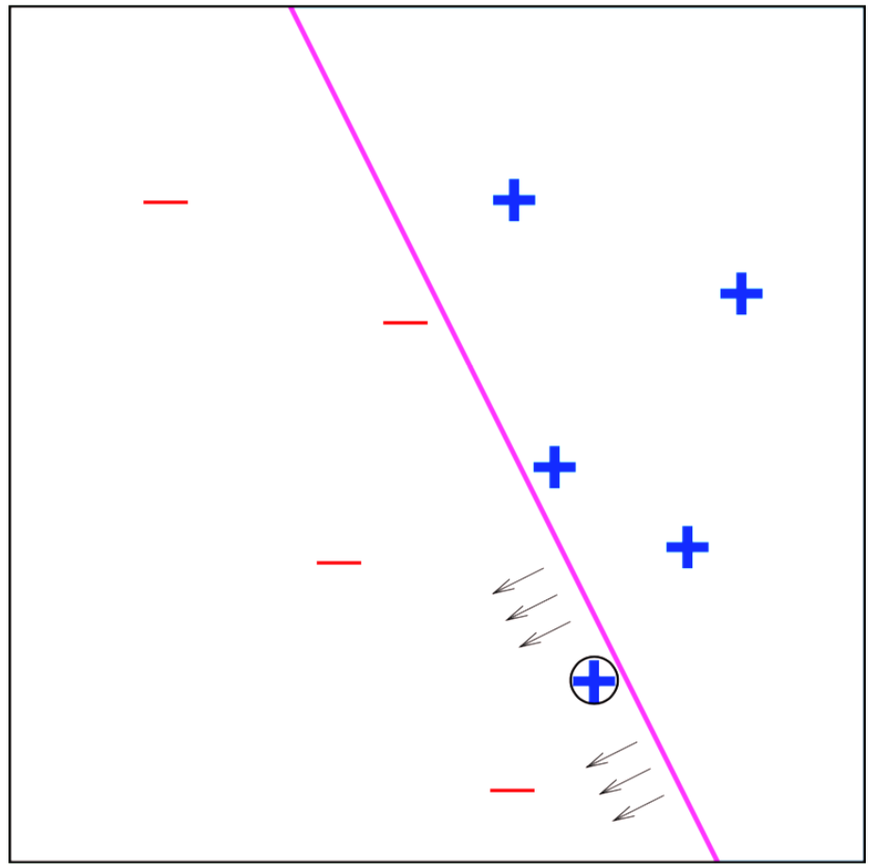
  where $(\boldsymbol{x}, y)$ is a misclassifled training point

- At iteration t = 1,2,3,..., pick a misclassifled point from

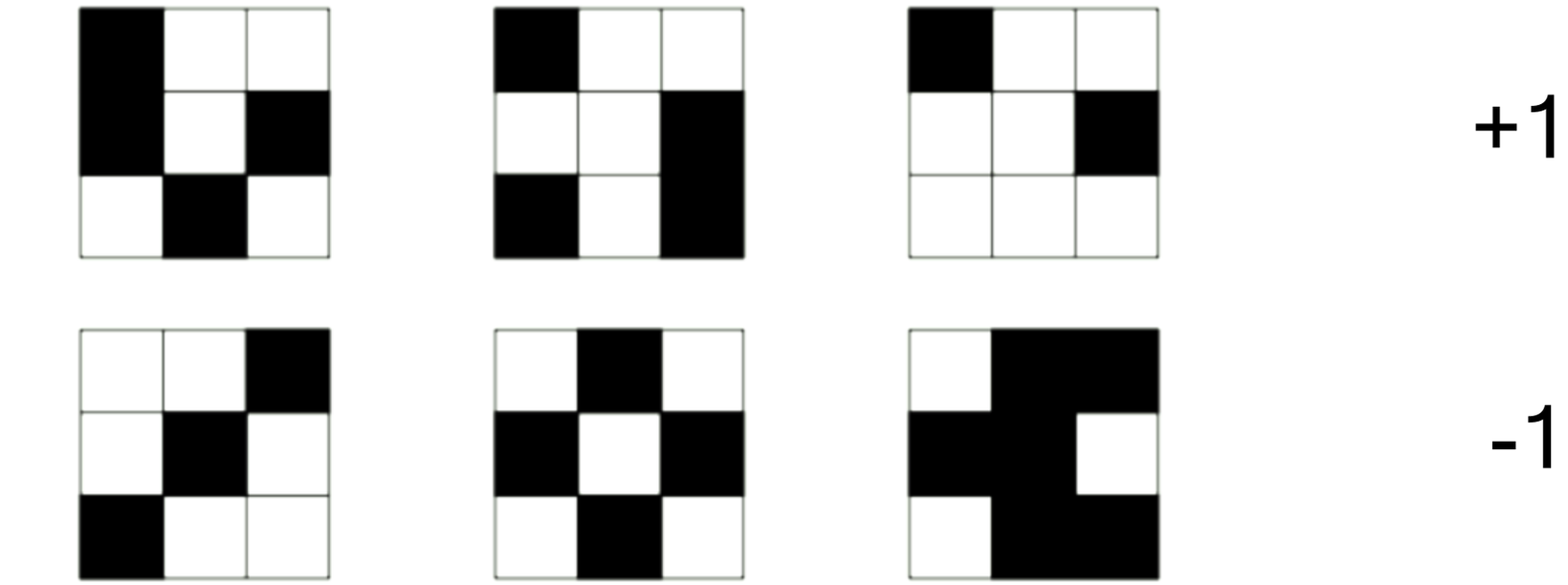
  $$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

  and run a PLA iteration on it

- That's it!

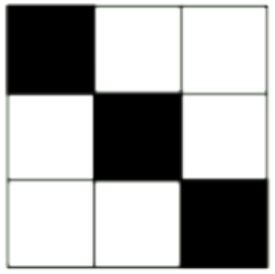


A Learning puzzle



+1

-1



$f(\boldsymbol{x}) = ?$

2.3 Types of Machine Learning

"Using a set of observations to uncover an underlying process"

broad premise ⟶ many variations

- Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system.

- Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

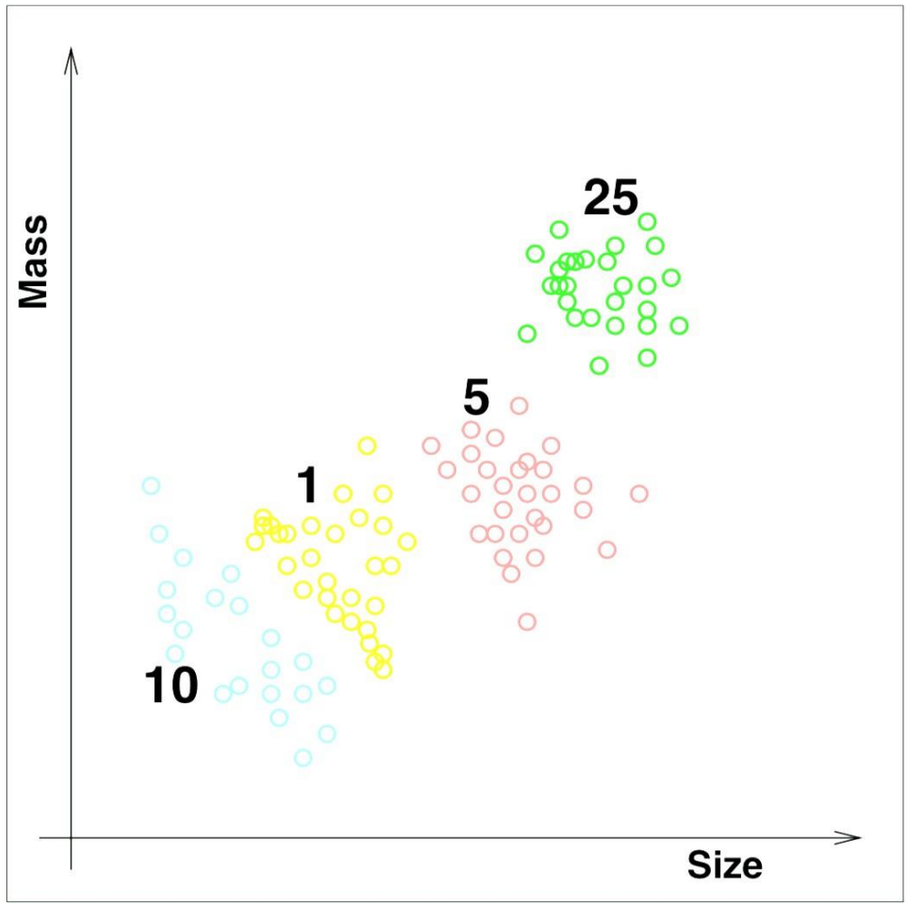
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

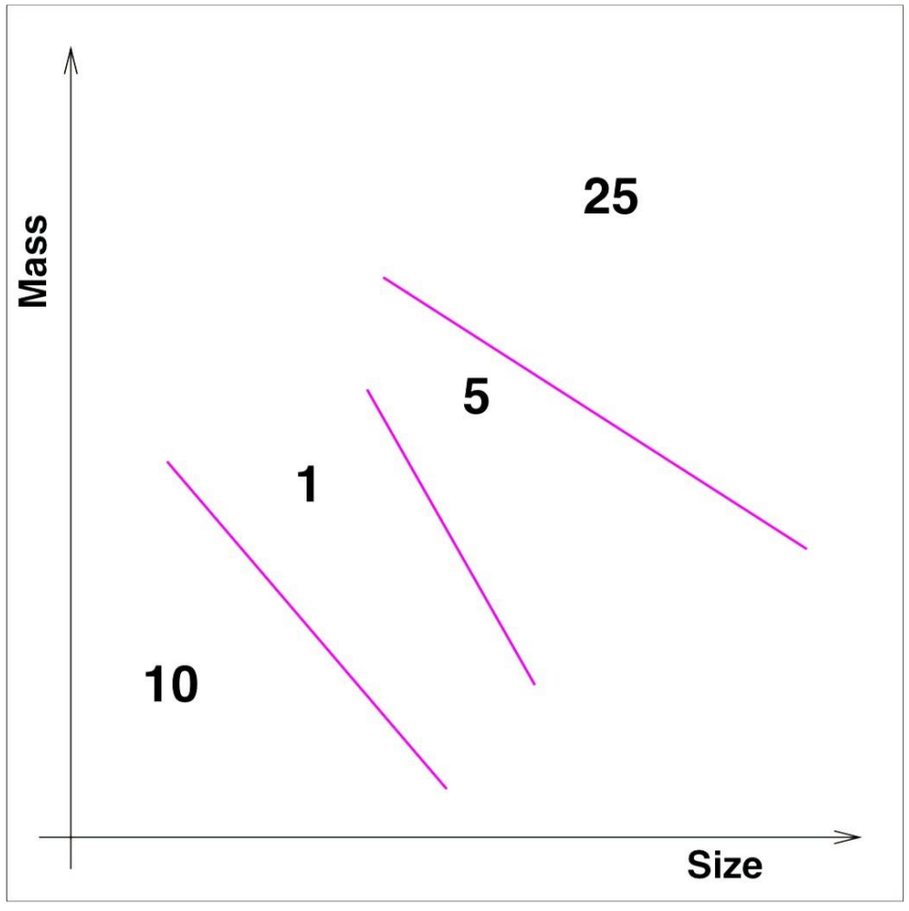
- Between supervised and unsupervised learning is semi-supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing.

- Reinforcement Learning: how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

Example: supervised learning

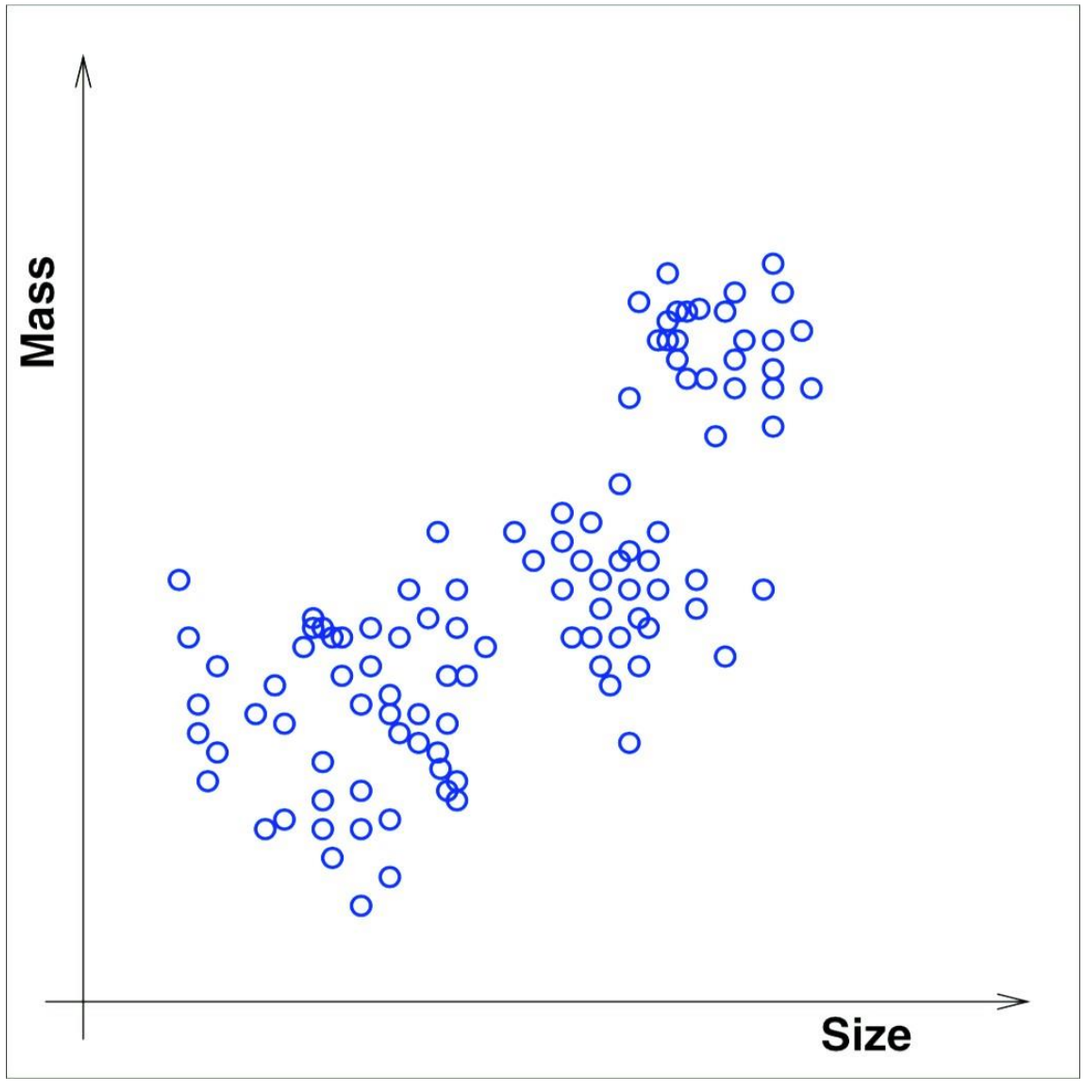
- Example from vending machines – coin recognition





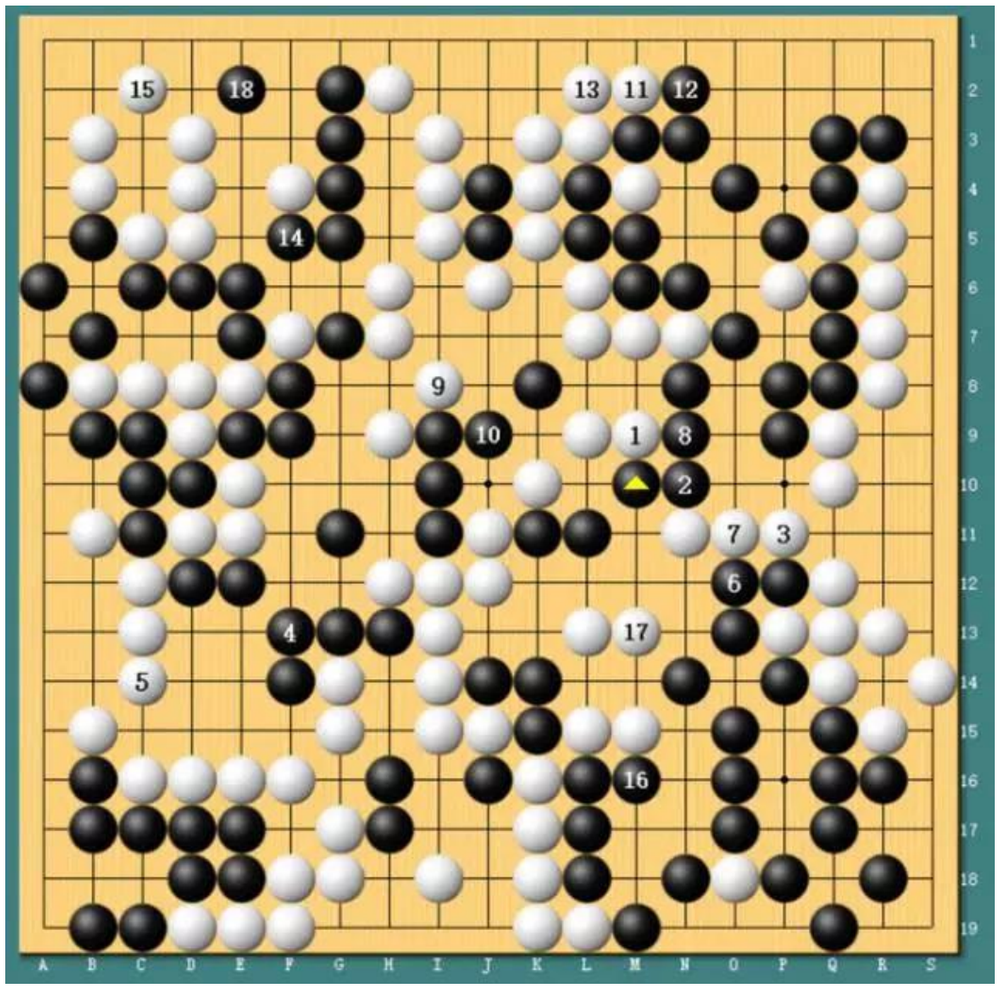
Example: unsupervised learning

- Instead of (input, correct output), we get (input, ?)



Example: reinforcement learning

- Instead of (input, correct output),
  we get (input, *some* output, grade for this output)



Learning Tasks

- In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.

- In regression, also a supervised problem, the outputs are continuous rather than discrete.

- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

- Density estimation finds the distribution of inputs in some space.

- Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics.

Learning Methods

- Regression
- Decision trees
- k−means
- Support vector machine
- Apriori algorithm
- EM algorithm
- PageRank
- kNN
- Naive Bayes
- (Deep) Neural networks

Read: Wu, X., Kumar, V., Ross Quinlan, J. et al. “Top 10 algorithms in data mining.” Knowl Inf Syst (2008) 14: 1.

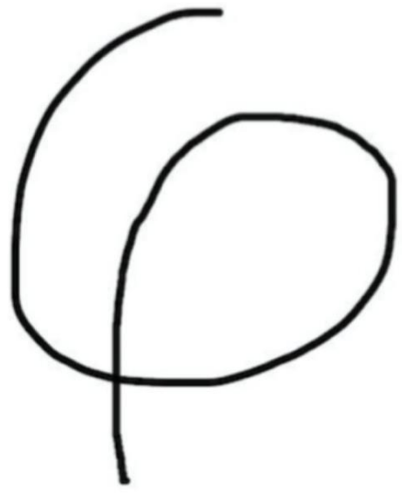
Learning Algorithms

- Gradient Descent Methods
- Online Gradient Methods
- Stochastic Gradient Methods
- Newton method
- Quasi-newton method (BFGS)
- Limited memory BFGS
- Coordinate Descent
- Alternating Direction methods of multipliers
- Penalty method, Augmented Lagrangian
- Gradient Projection method
- Iterative-thresholding method (IST)
- Conditional Gradient method

Read: Wu, X., Kumar, V., Ross Quinlan, J. et al. “Top 10 algorithms in data mining.” Knowl Inf Syst (2008) 14: 1.

2.4. Applications

- Character recognition
  Given an image of a character, correctly identify the character



- Spam recognition
  Given an email, correctly identify the email as spam or not

- Speech recognition
  Given an audio of speech, identify the words being said

- Machine translation
  Given a sample of text in one language, produce text in another language with the same meaning

- Input software

- Computer vision
  Starting with some seminal work on face recognition and continuing to the present with almost every other application in vision, vision has been turned into a largely learning-base field. Instead of trying to figure out geometrically what geometry makes the face, we just give the computer a bunch of faces and let it figure out “In these images, this is what makes up a face”

- Ranking web search results
  Given a search query return a ranking of web pages by relevance/“goodness”

- Recommender systems
  For example: “Netflix movie recommender system”

Netflix movie recommender system

Netflix

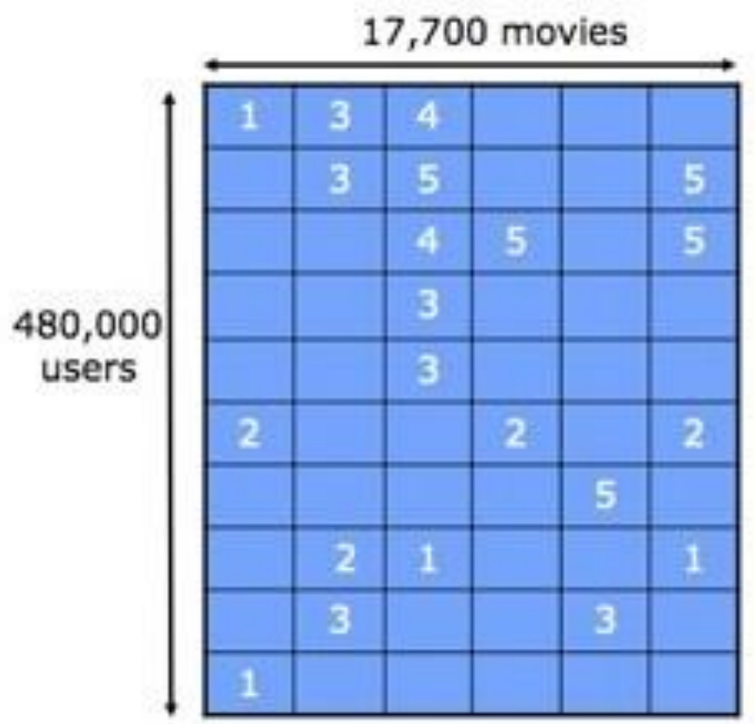
- Movie rentals by DVD (mail) and online (streaming)
- 100k movies, 10 million customers
- Ships 1.9 million disks to customers each day
  - 50 warehouses in the US
  - Complex logistics problem
- Employees: 2000
  - But relatively few in engineering/software
  - And only a few people working on recommender systems
- Moving towards online delivery of content
- Significant interaction of customers with Web site

The $1 Million Question

Netflix Competition

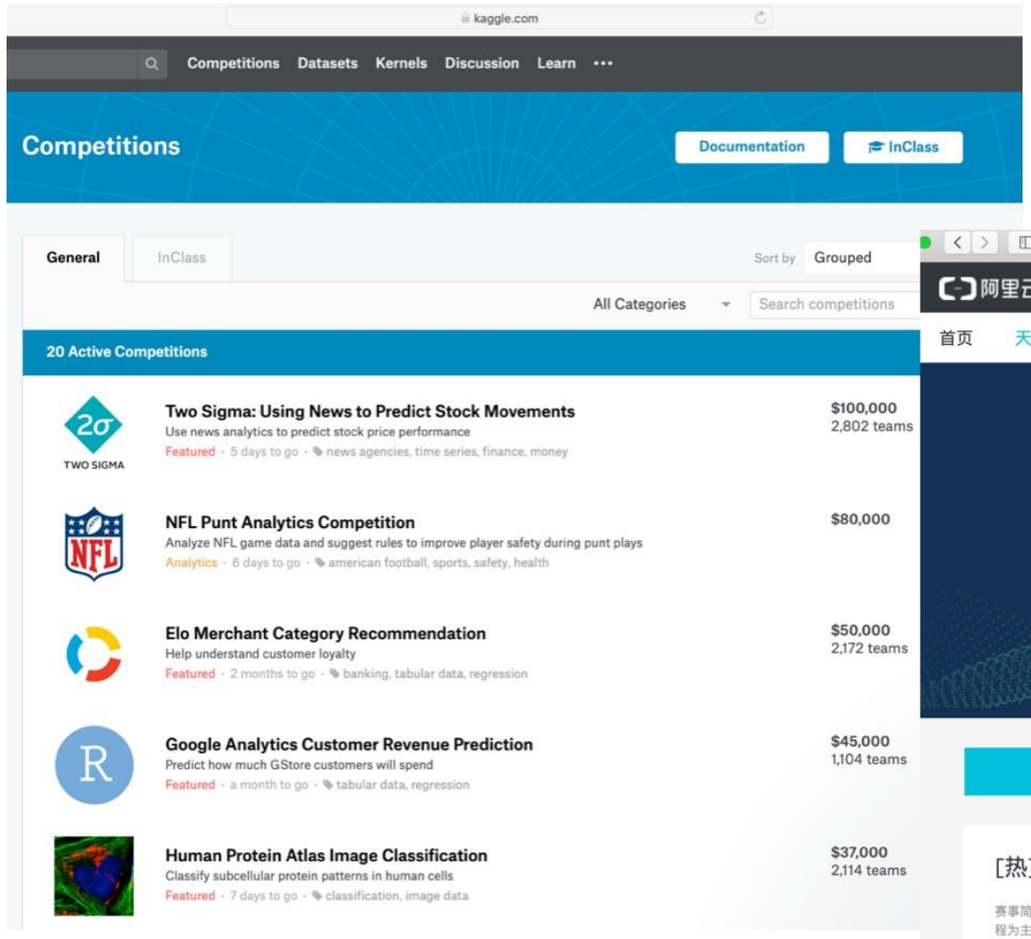
- Training data
  - 100 million ratings
  - 480,000 users
  - 17,770 movies
  - 6 years of data: 2000-2005
- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation criterion: root mean squared error (RMSE)
  - Netflix Cinematch RMSE: 0.9514
- Competition
  - 2700+ teams
  - $1 million grand prize for 10% improvement on Cinematch res
  - $50,000 2007 progress prize for 8.43% improvement

Ratings Data



Million Dollars Awarded Sept 21st 2009

Competitions and prizes are still going on...



If you win a prize on Kaggle/天池, you will get rich, and A+!!