Machine Learning

Lecture 1: Introduction

杨思蓓

SIST

Email: yangsb@shanghaitech.edu.cn

Short Bio

- Dr. Sibei Yang

  Email: yangsb@shanghaitech.edu.cn

  - Assistant Professor at SIST
  - Office: 2-402B
  - Research Interests: Computer Vision, Natural Language Processing, Machine Learning and the intersection of them.

Content

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

1. Course Information

Basic Information

- Time: **Tuesday & Thursday**
  - **8:15am-9:55am**

- Place: 教学中心201

- Teaching Assistants:
  - 田鹏超 tianpch@shanghaitech.edu.cn
  - 苏杭 suhang@shanghaitech.edu.cn
  - 龙俊峰 longjf1@shanghaitech.edu.cn
  - 石骋 shicheng@shanghaitech.edu.cn

- Office Hours: Tuesday 19:00-20:00 1C-309 (苏杭)
  Wednesday 18:00-19:00 1A-406 (田鹏超)

- Course Site: 上科大教学互助平台(Blackboard)
  - Questions/Discussion on BB

Basic Information

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

- Evaluation
  - Assignments/Quizzes(30%) + Project(30%) + Exam(40%)

- Objective:
  - Understandings of some of the important machine learning methods, theory, algorithms.
  - Basic ability to use some machine learning techniques to solve real world problems.

Textbooks and Slides

- Learning theory, supervised learning: *"Learning from data." Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin.* / 机器学习，周志华。

- Learning algorithms:
  Léon Bottou, Frank E. Curtis, Jorge Nocedal, "Optimization Methods for Large-Scale Machine Learning", SIAM Review, Vol. 60, No. 2, pp. 223-331.
  / *Frank E. Curtis, Katya Scheinberg. "Optimization Methods for Supervised Machine Learning: From Linear Models to Deep Learning". In INFORMS Tutorials in Operations Research.*

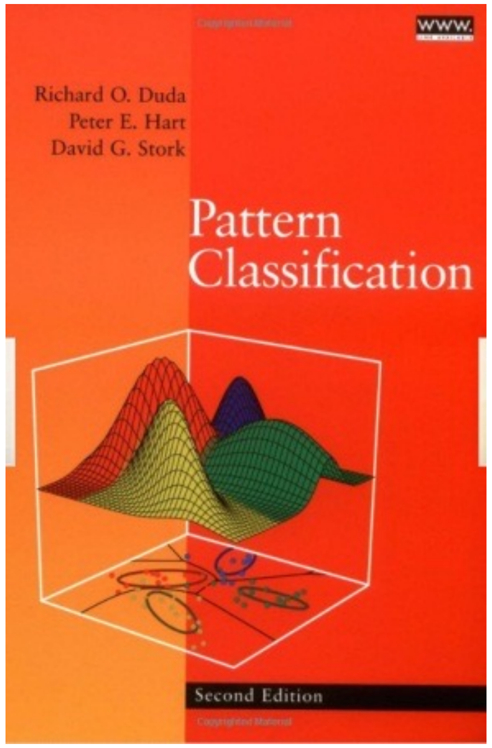
- Others: *For machine learning methods: "Machine Learning, A probabilistic Perspective", Kevin P. Murphy, the MIT Press.*

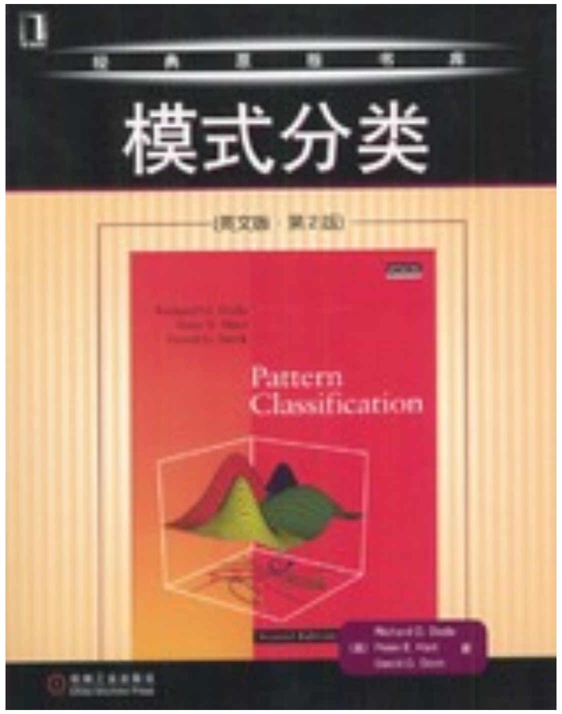
Some lectures will be based on these books/papers, but not all of them. Reading the textbooks is not required, but it is recommended. You are not responsible for textbook material that is not covered in lecture.

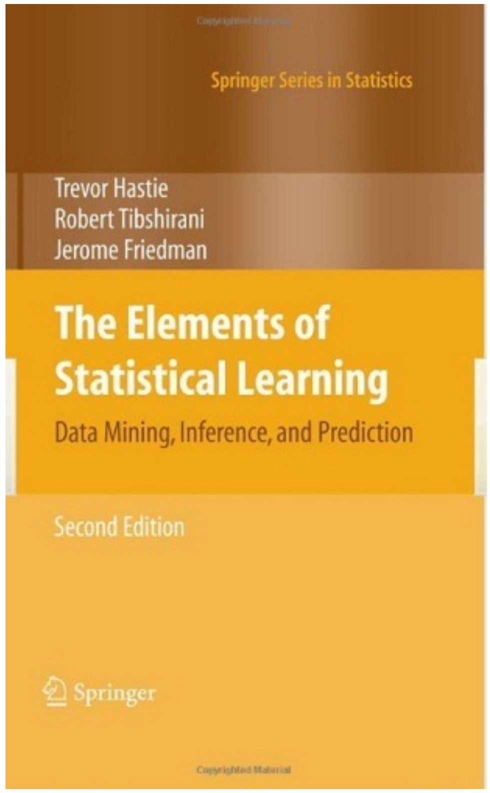
Acknowledgement: Some lectures are in reference to the course "Machine Learning" given by Dr. Hao Wang and "Learning from data" taught by Prof. Yaser Abu-Mostafa.

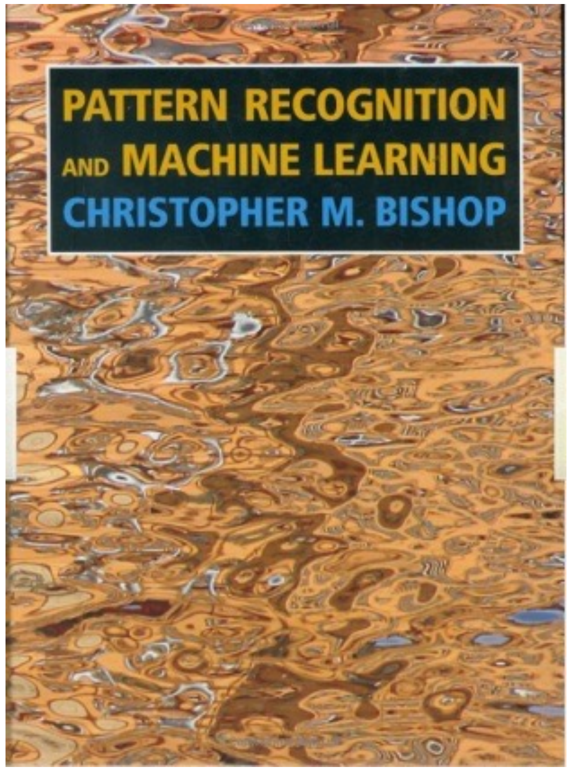
Textbooks and Slides

- [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






Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer



Academic Integrity

- Academic Regulations
- Academic Dishonesty
  - Plagiarism, Cheating on examinations, unauthorize collaboration, project, 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
  - Warning, Lowering grade, failing grade, and more …

- Plagiarism for assignments: cite your sources to avoid punishments

- Plagiarism for final project: cite your references!

Course Policies

- Academic Dishonesty
  - No.

- Assignments/Project:
  - Write your own solution

- Submission via Blackboard
  - Email submission/other methods are not accepted, i.e., getting 0.

- Late Policy:
  - 0~12 hours: 90%
  - 12~36 hours: 50%
  - 36 hours~60 hours: 25%
  - 60 hours~: 0
  - 以上时间为Blackboard上显示收到作业/项目的时间！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。

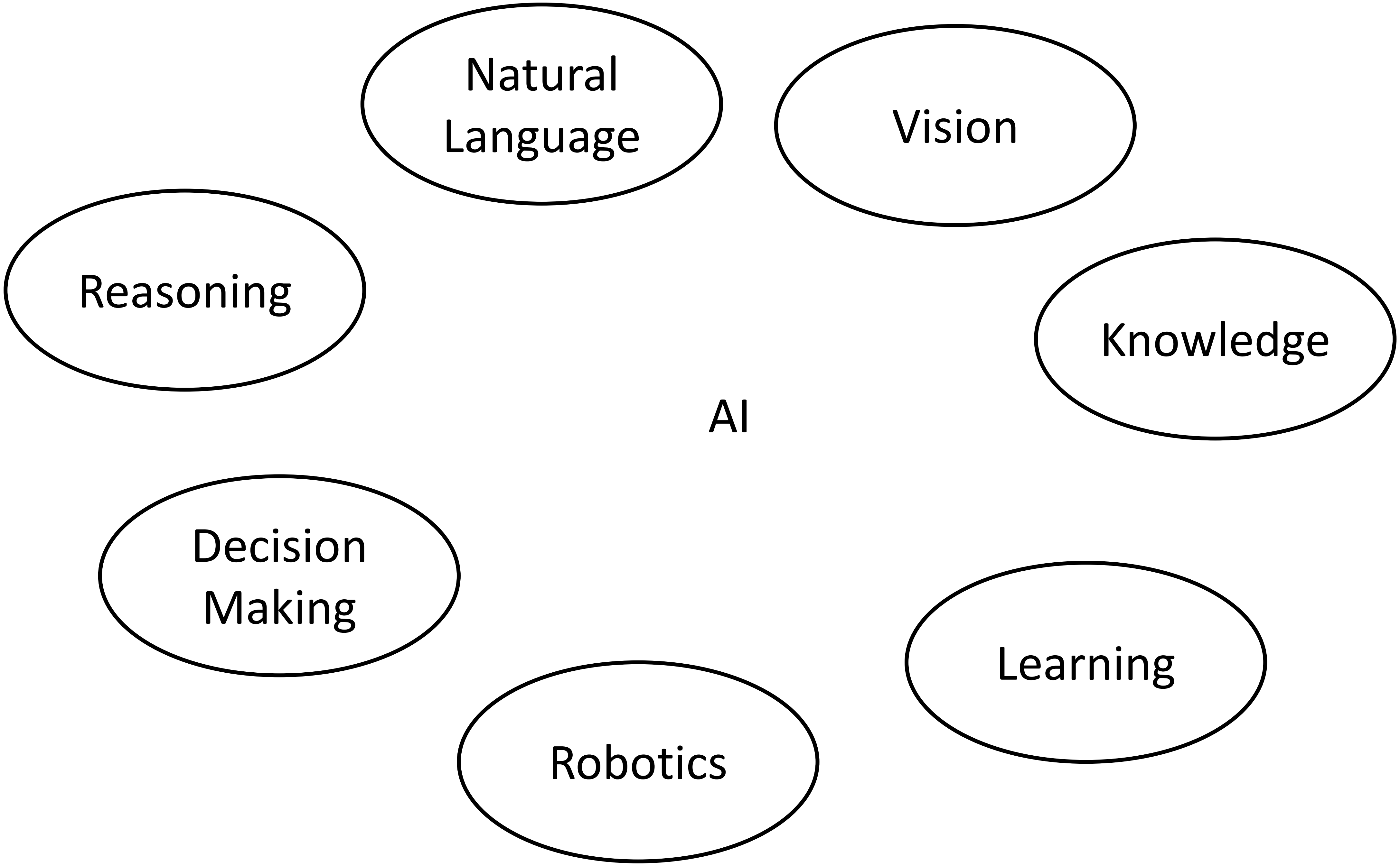
Why Take This Course?

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

2. What is Machine Learning?

What is ML and Why ML?

- Demos



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*

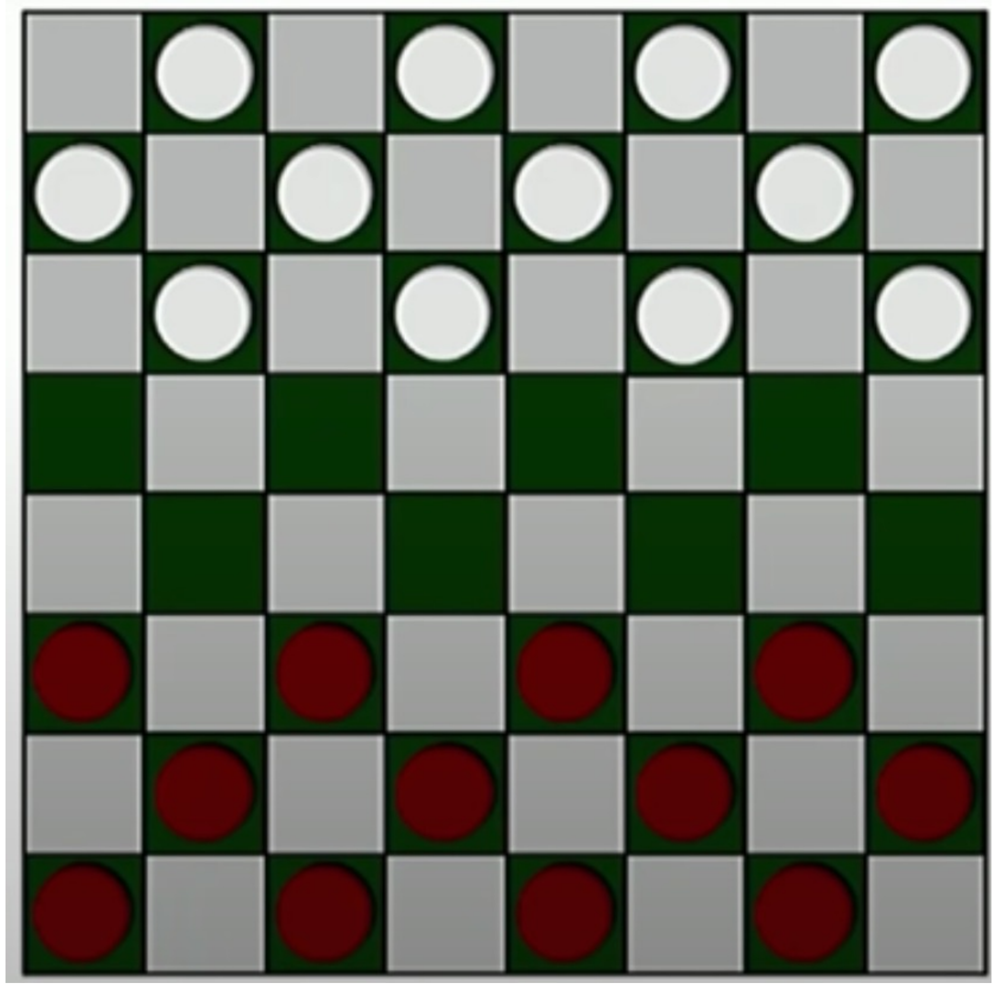
- This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds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

- The job of a statistician is to use the tools of statistics to interpret data in the context of the domain. The authors seem to include all of the field of Machine Learning as aids in that pursuit. Interestingly, they chose to include "Data Mining"...

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

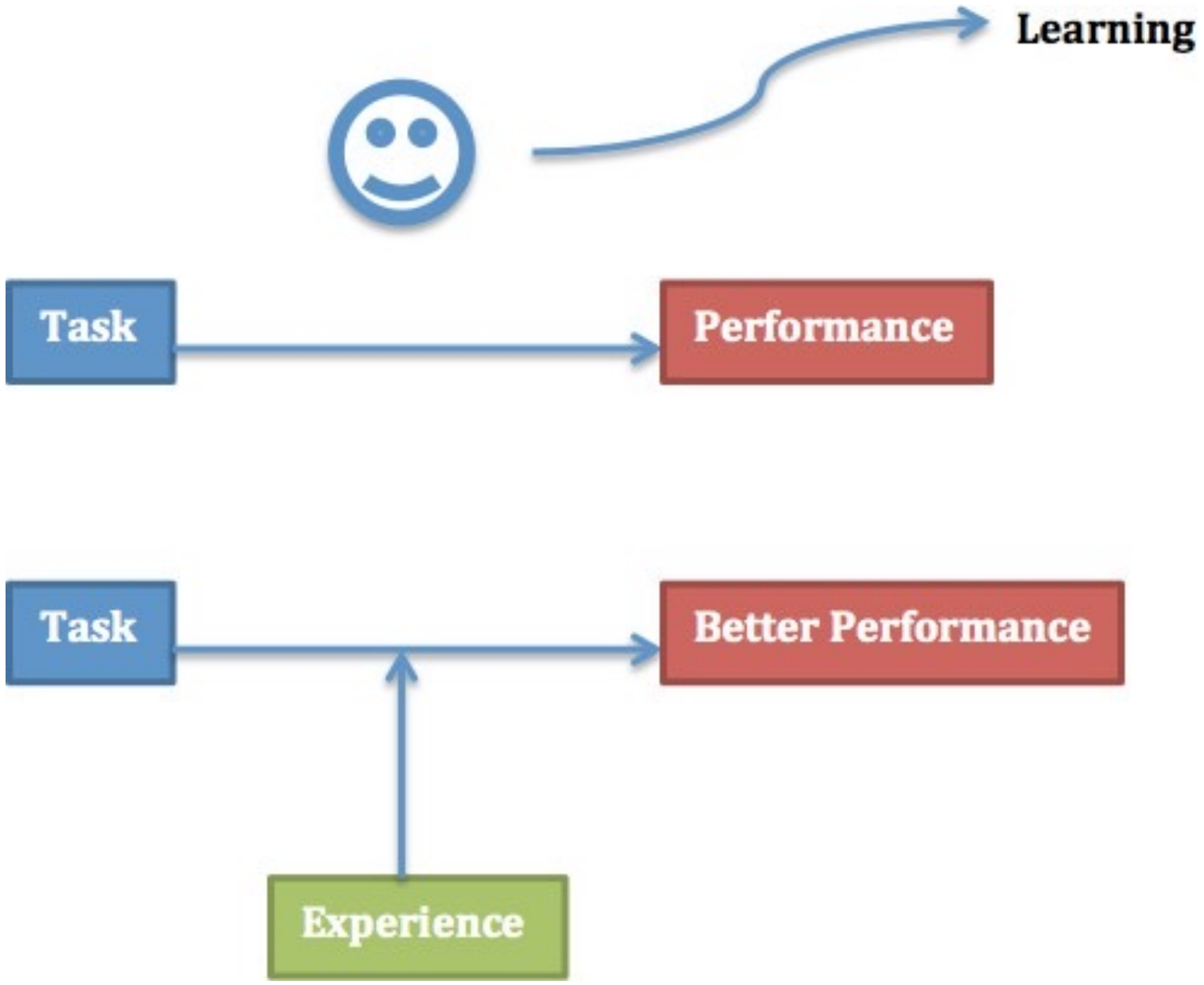
What is machine learning?

- 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

3. Examples of Machine Learning

Example: Predicting how a viewer will rate a movie

10% improvement > 1 million dollar

The essence of machine learning：

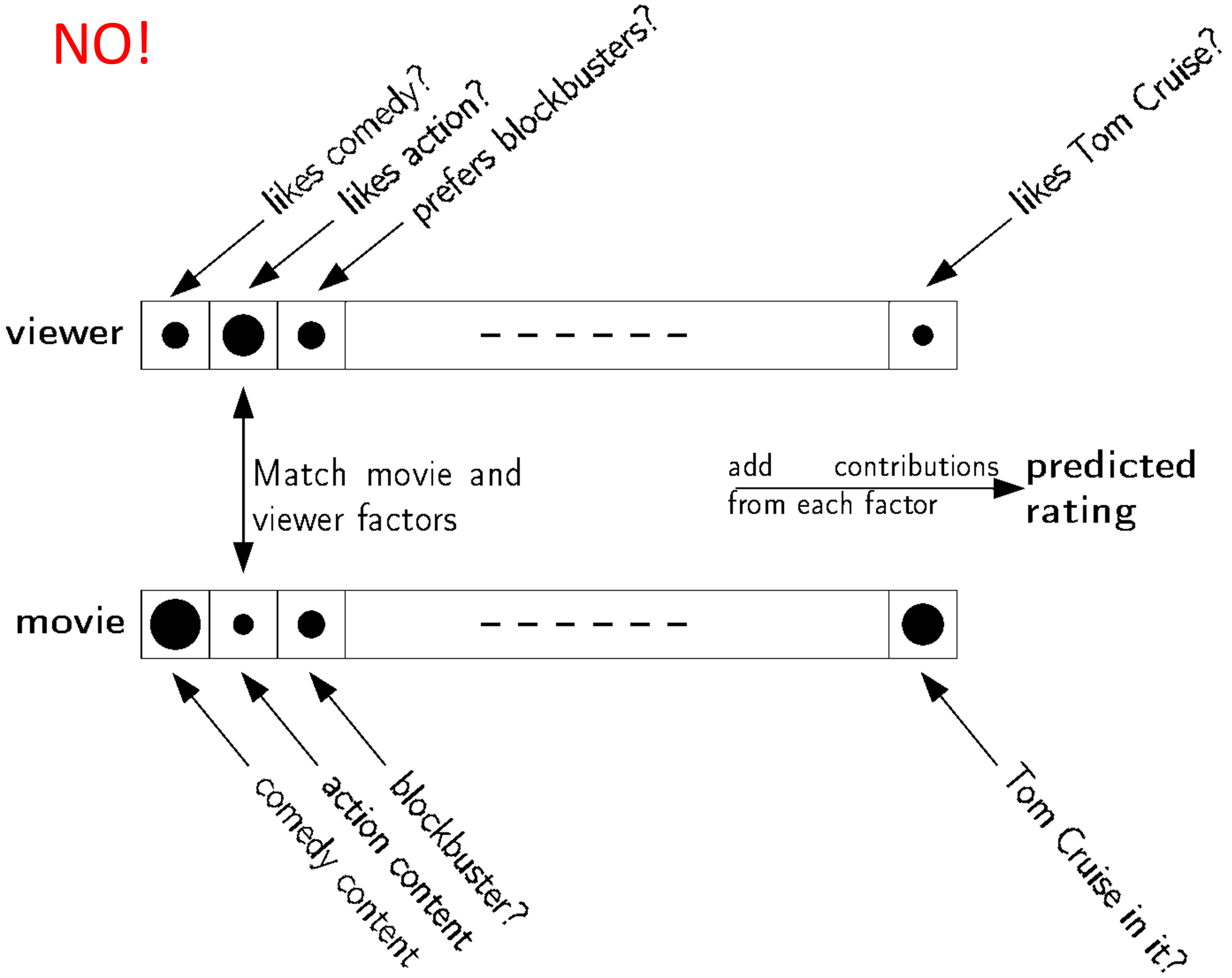
- A pattern exists

- We cannot pin it down mathematically

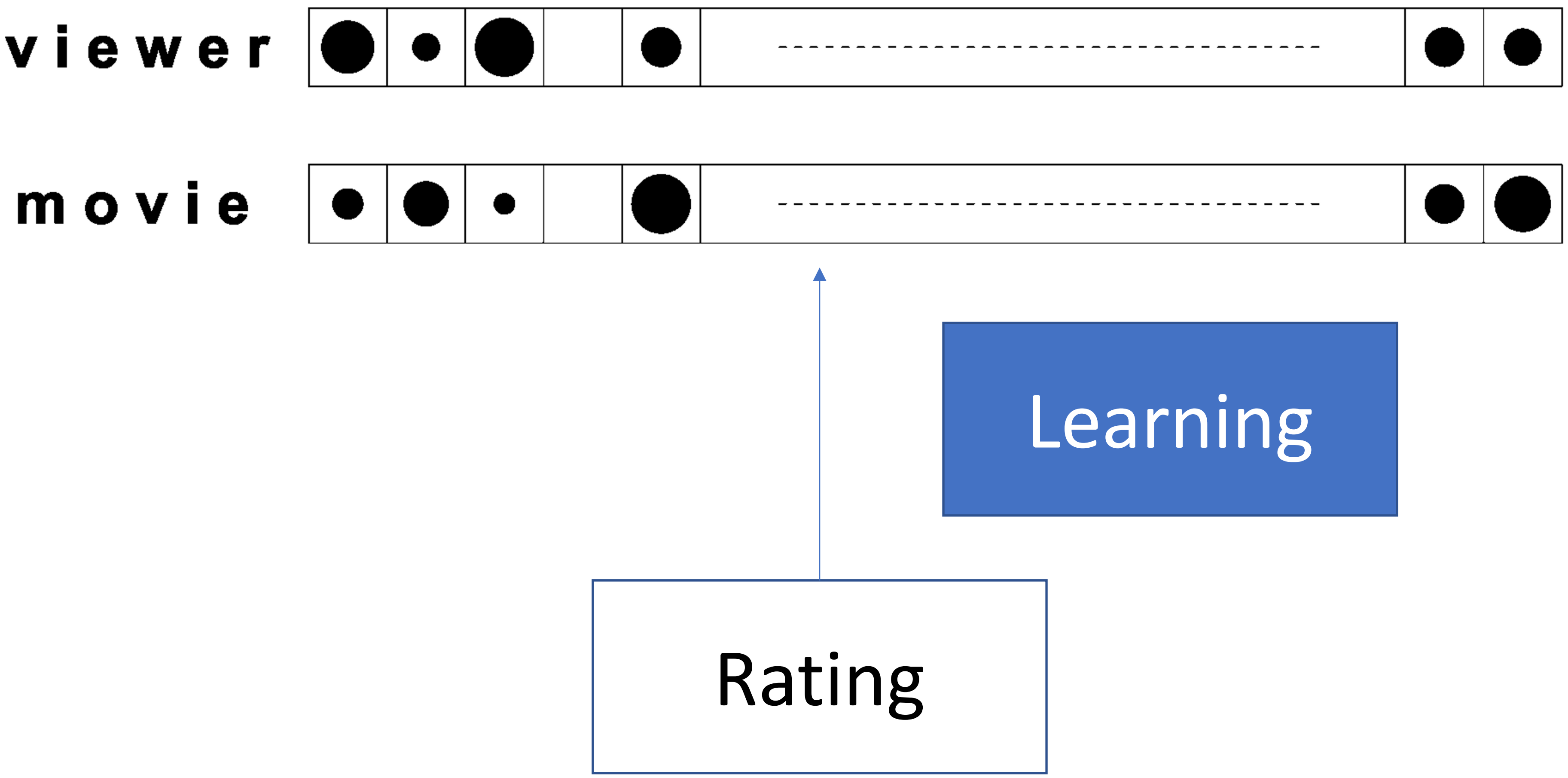
- We have data on it

Example: Predicting how a viewer will rate a movie

NO!

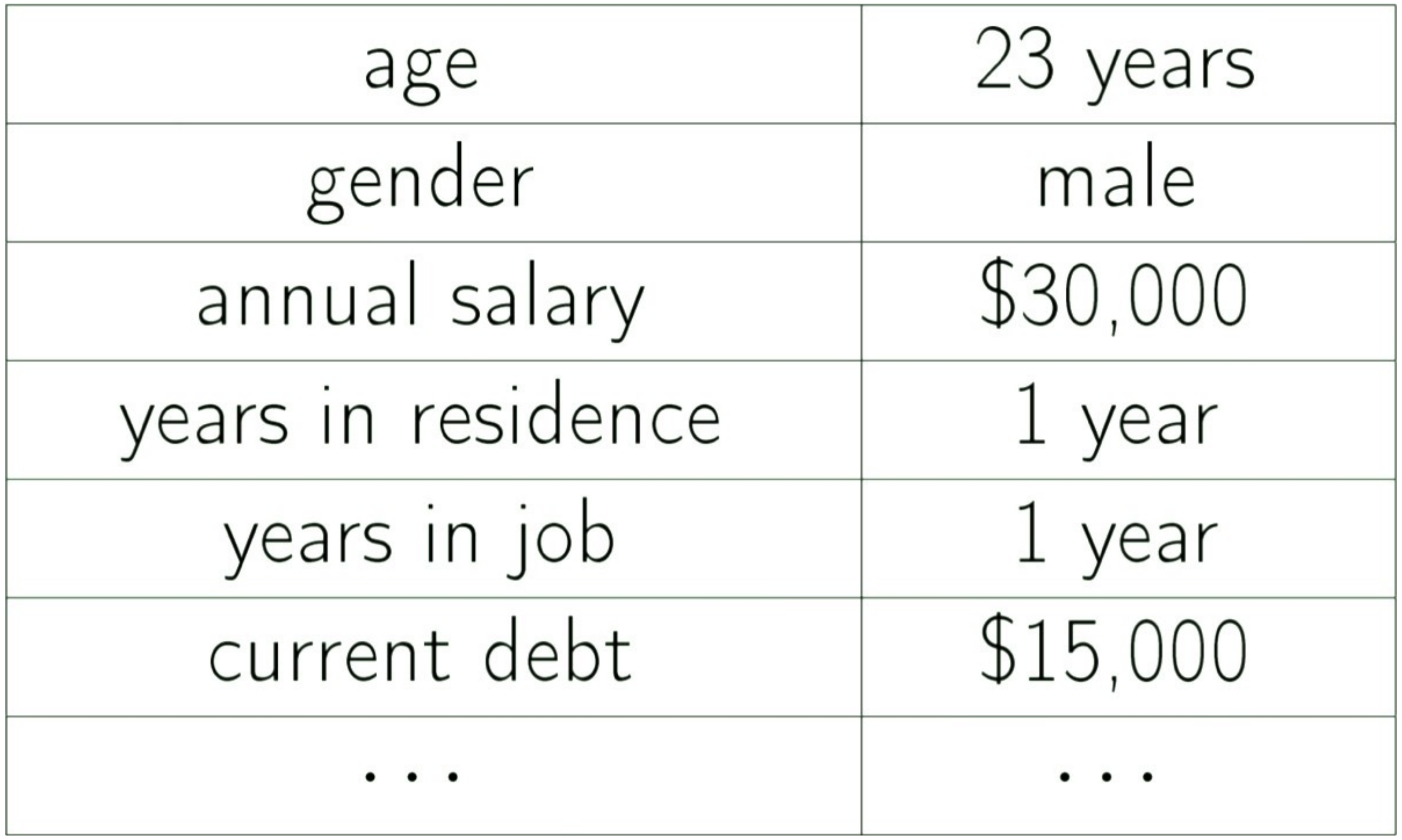


Example: Predicting how a viewer will rate a movie



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?

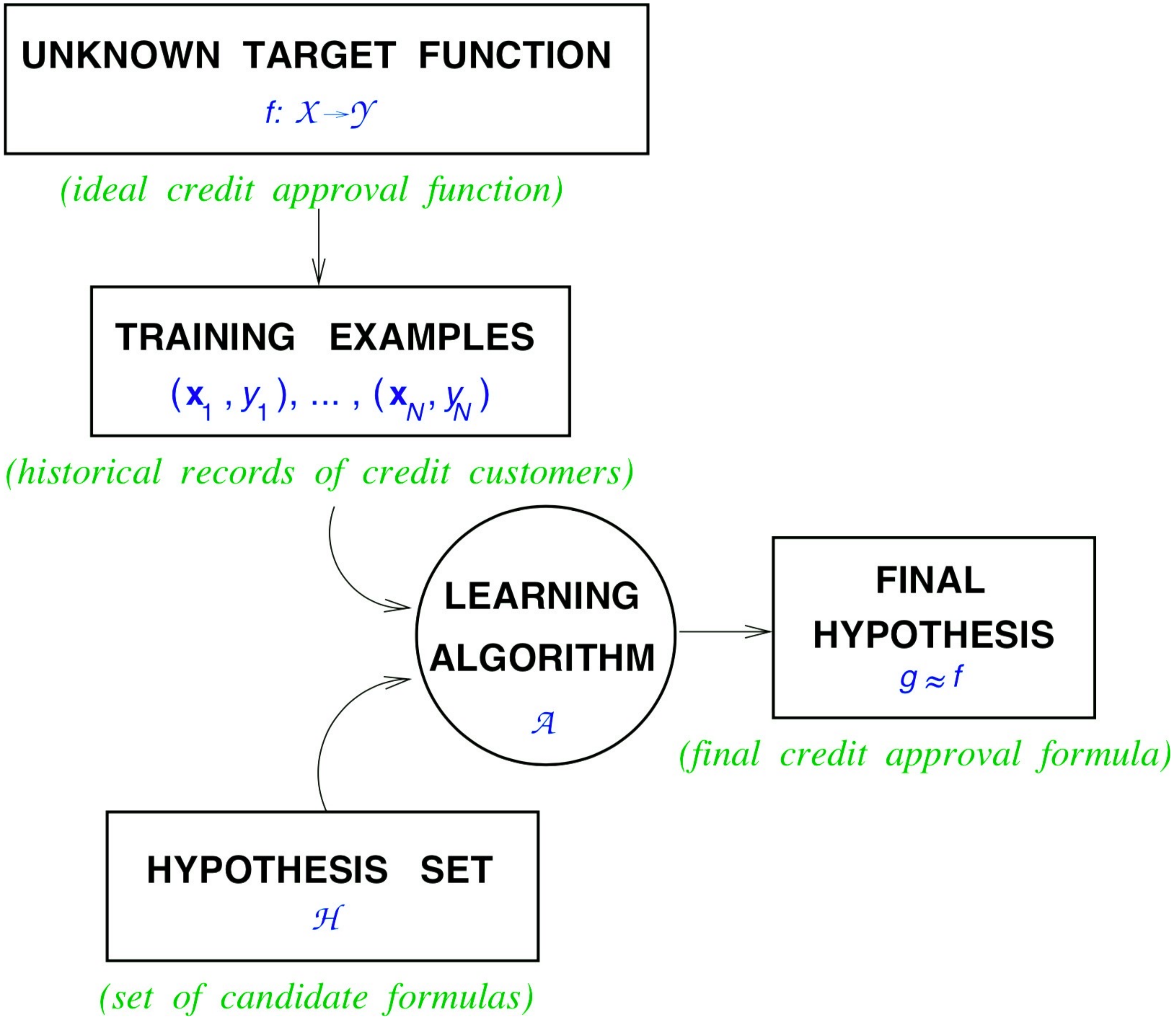
Components of learning

Formalization:

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

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (formula to be used)

Components of learning



UNKNOWN TARGET FUNCTION

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

(ideal credit approval function)

TRAINING EXAMPLES

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

(historical records of credit customers)

LEARNING ALGORITHM

$\mathcal{A}$

FINAL HYPOTHESIS

$g \approx f$

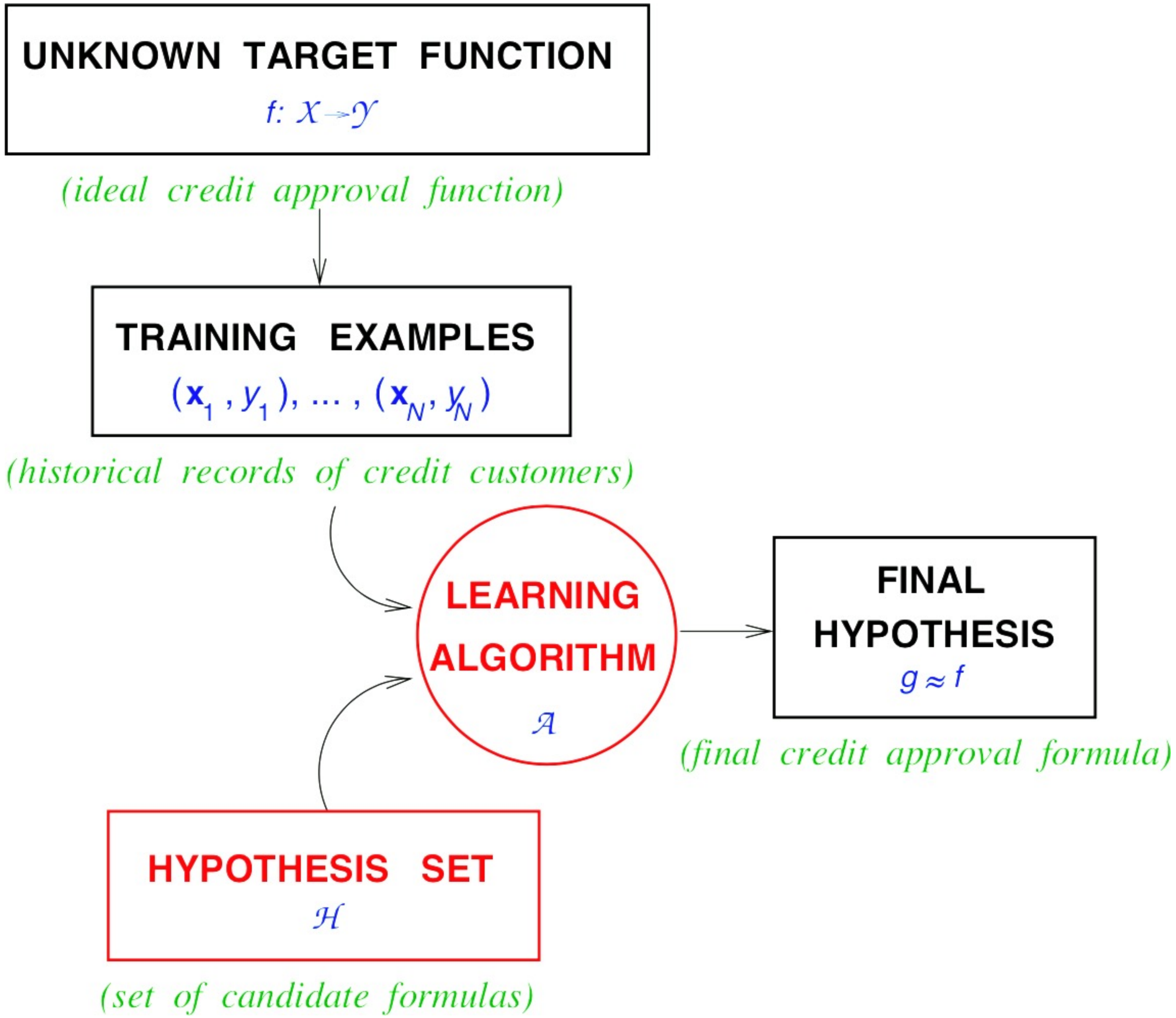
(final credit approval formula)

HYPOTHESIS SET

$\mathcal{H}$

(set of candidate formulas)

Components of learning


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
(ideal credit approval function)
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(final credit approval formula)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

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 $h \in \mathscr{H}$ can be written as

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

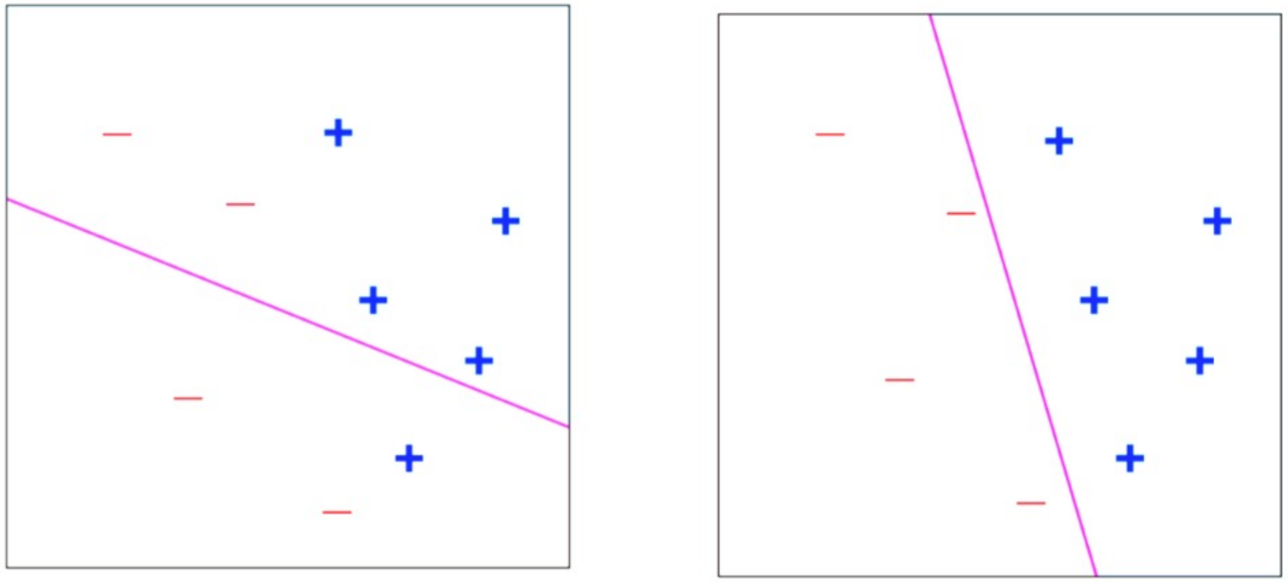
$$h(\boldsymbol{x}) = \text{sign}\Big(\big(\sum_{i=1}^{d} w_i x_i\big) + w_0\Big)$$

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:

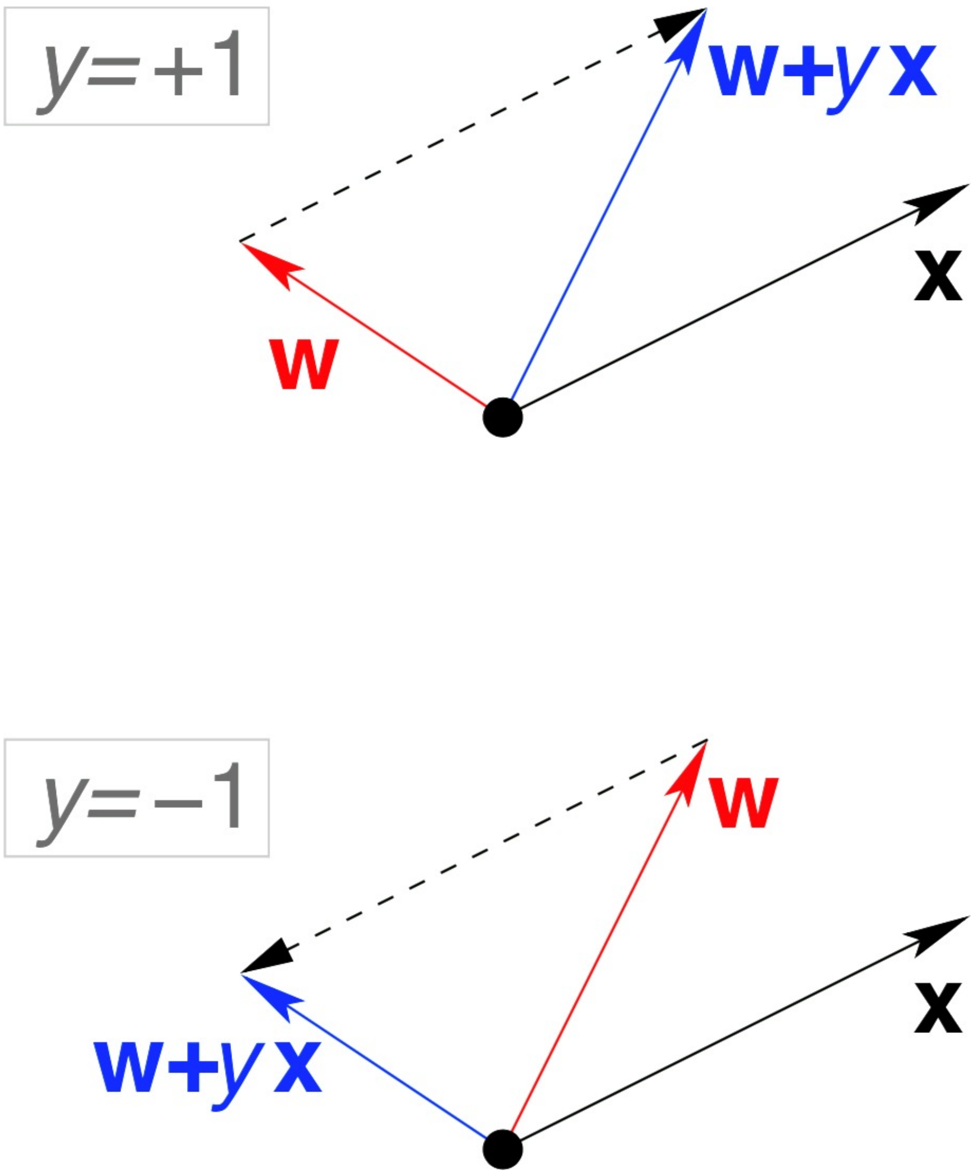
$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)$

Pick a misclassified point:

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

and update the weight vector

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



Iterations of PLA

- One iteration of the PLA,

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

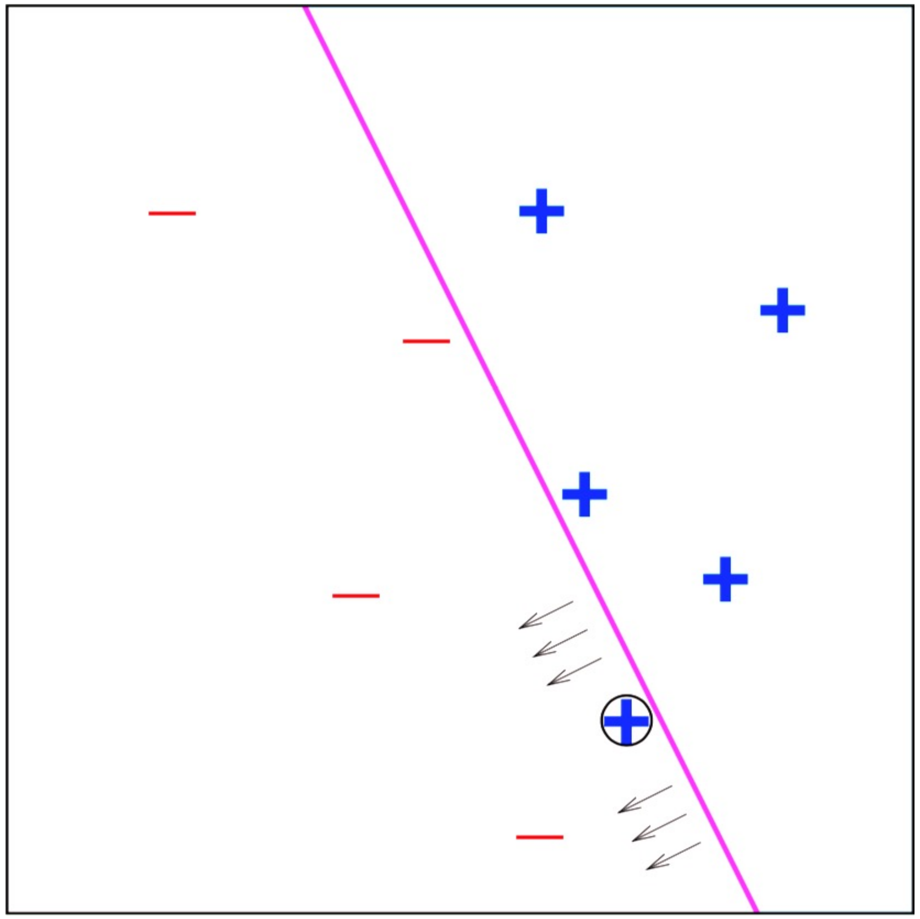
  where (x, y) is a misclassified training point

- At iteration $t = 1,2,3,\ldots$, pick a misclassified point from

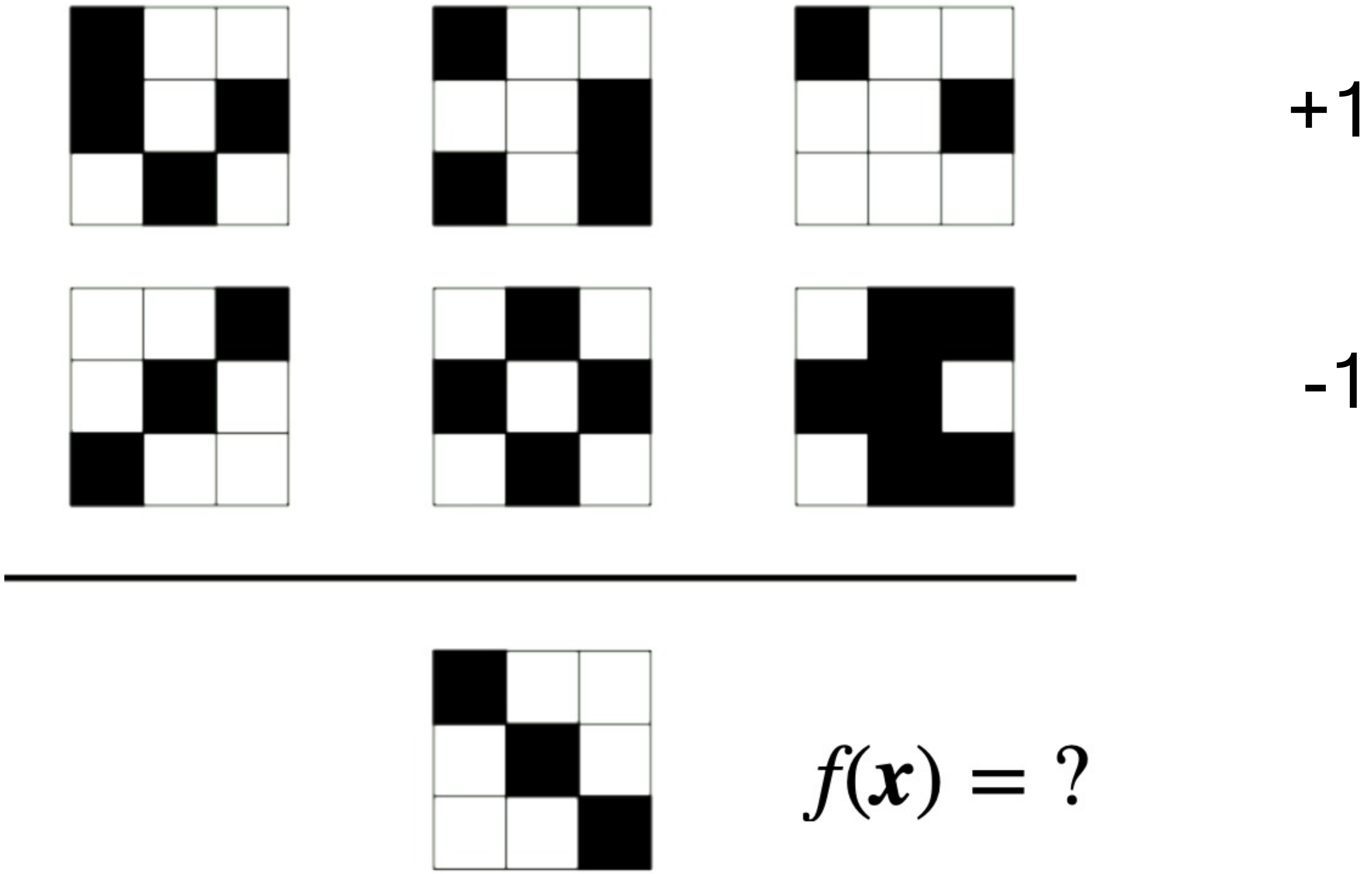
  $$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

  and run a PLA iteration on it

- That's it!



A Learning puzzle



4. 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.

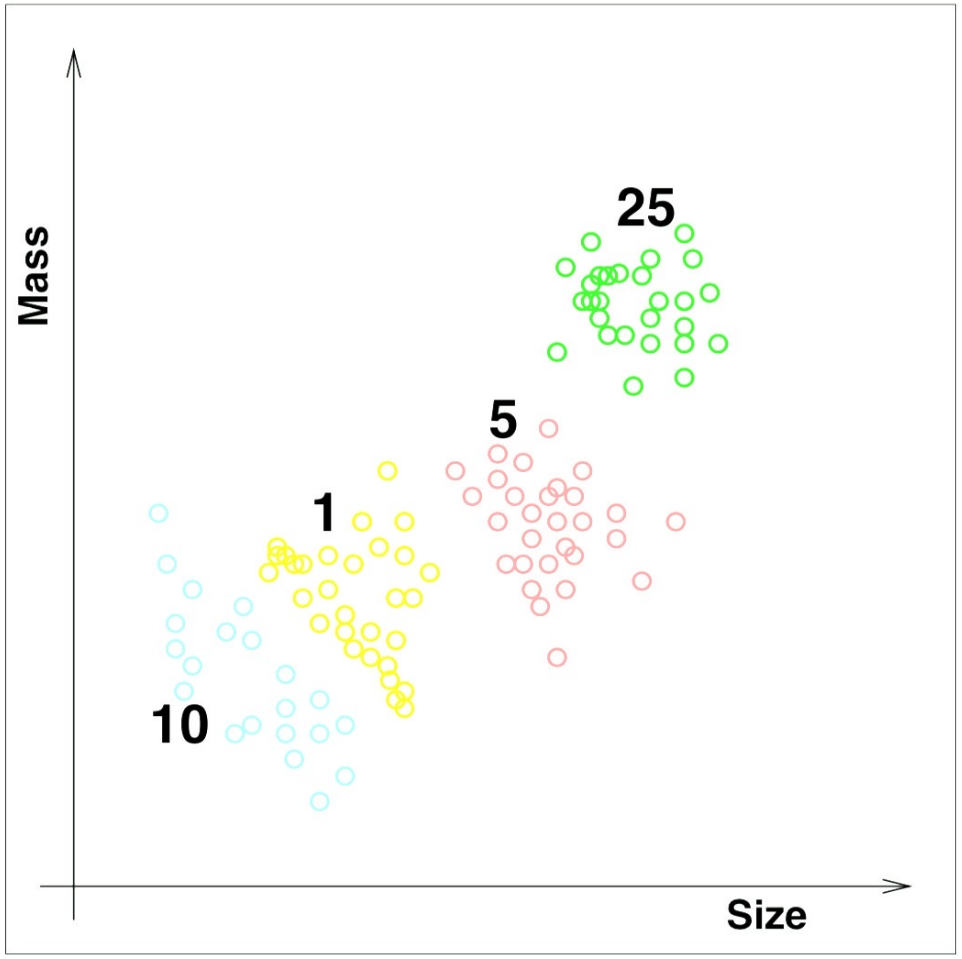
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

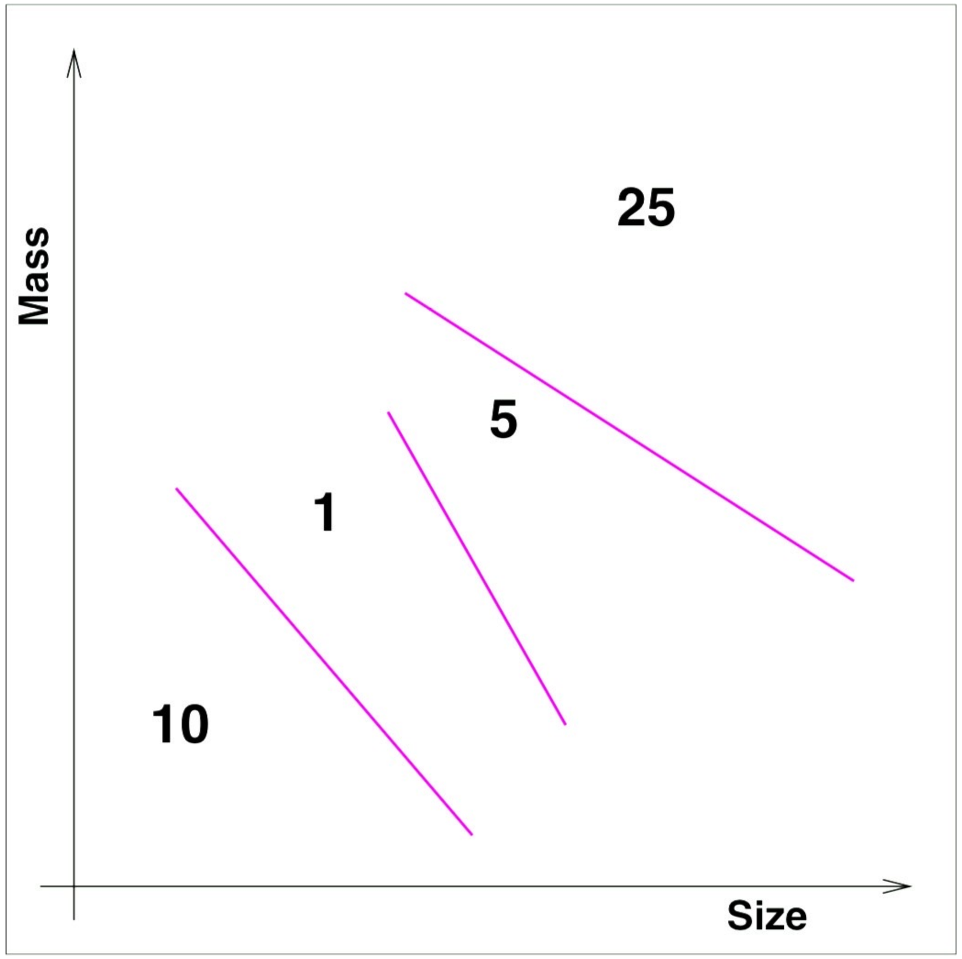
- Between supervised and unsupervised learning is semi-supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing.

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

Example: supervised learning

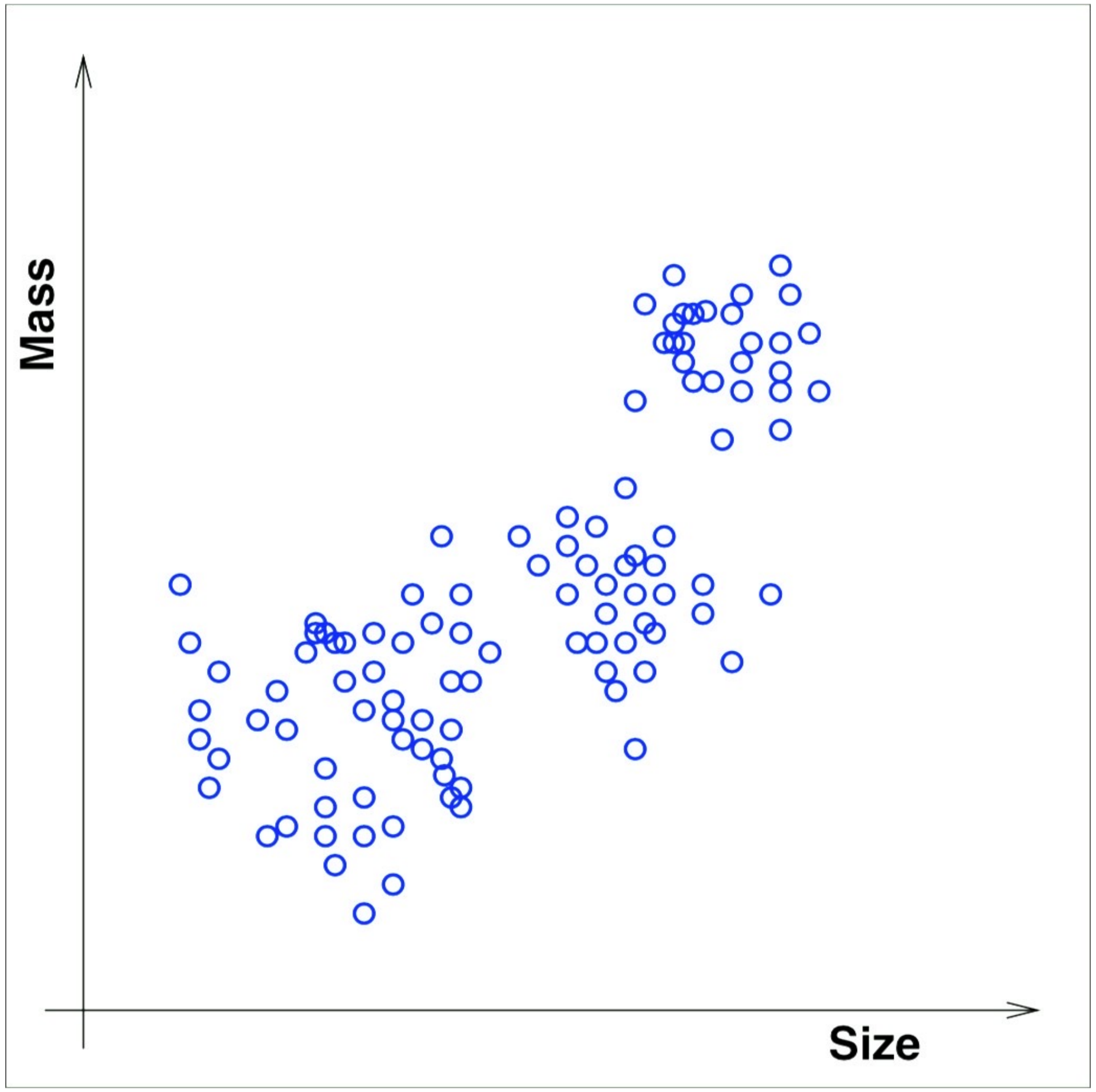
- Example from vending machines – coin recognition





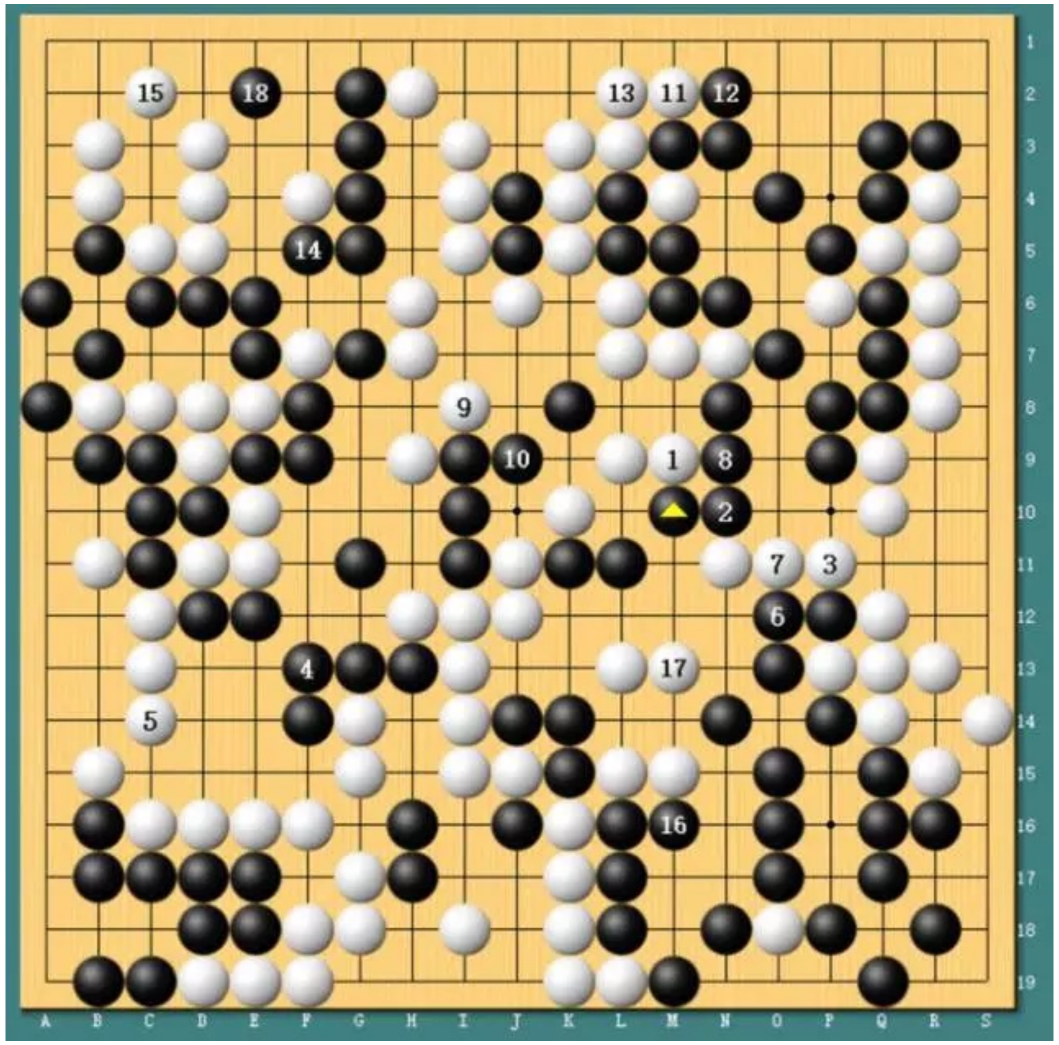
Example: unsupervised learning

- 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

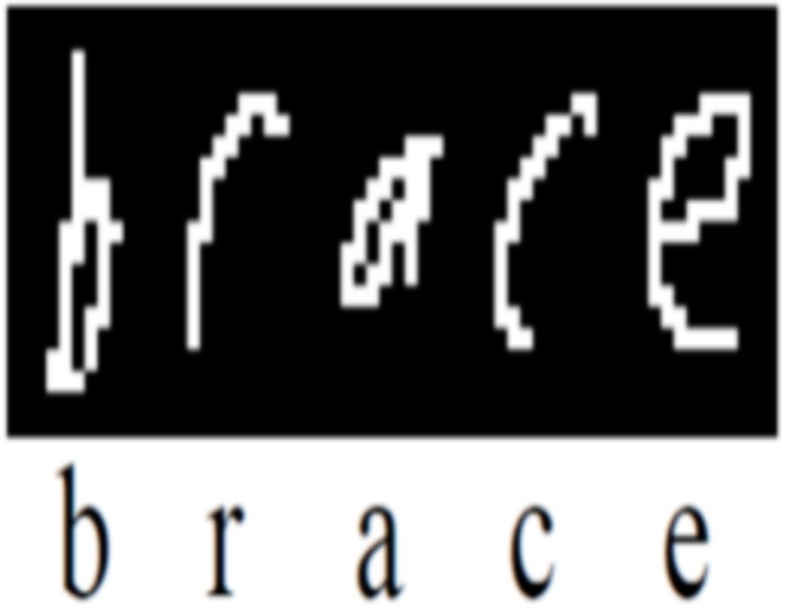
- Gradient Descent Methods
- Online Gradient Methods
- Stochastic Gradient Methods
- Newton method
- Quasi-newton method (BFGS)
- Limited memory BFGS

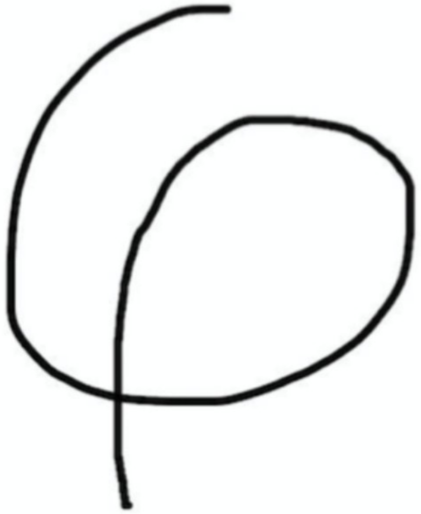
- Coordinate Descent
- Alternating Direction methods of multipliers
- Penalty method, Augmented Lagrangian
- Gradient Projection method
- Iterative-thresholding method (IST)
- Conditional Gradient method

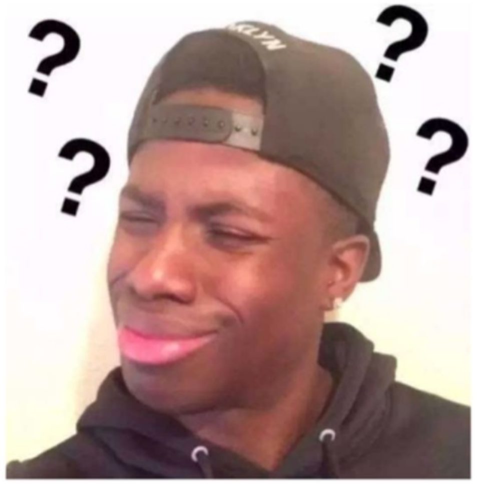
Read: Wu, X., Kumar, V., Ross Quinlan, J. et al. “Top 10 algorithms in data mining.” Knowl Inf Syst (2008) 14: 1.

5. 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

I love SharePoint

أنا أحب SharePoint

Ich liebe SharePoint

Ik hou van SharePoint

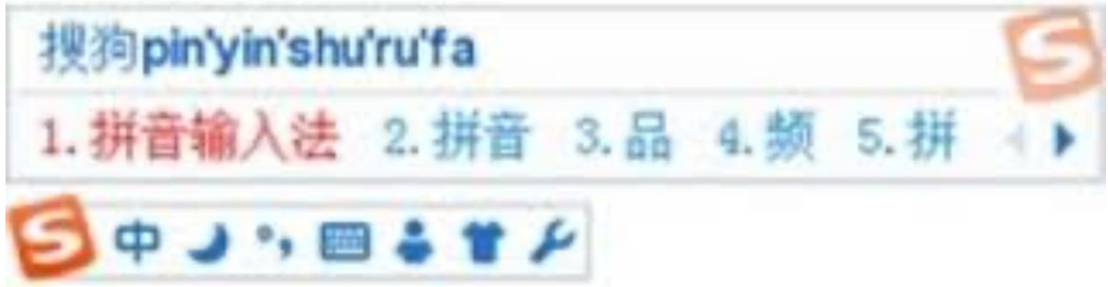
SharePoint を愛する

Me encanta SharePoint

Я люблю SharePoint

machine translation

- 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

Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

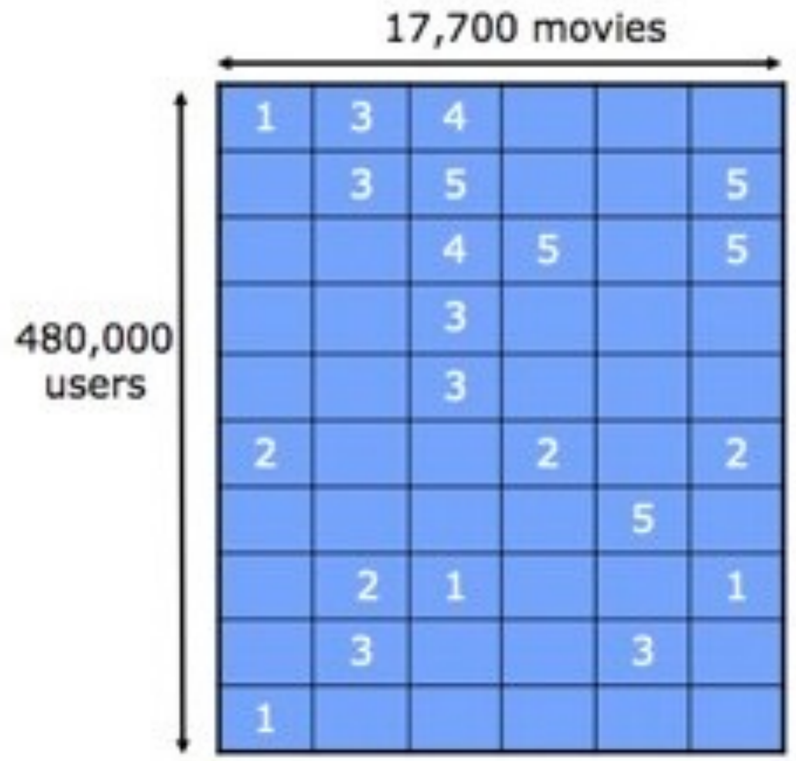
You should also read the frequently asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!

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

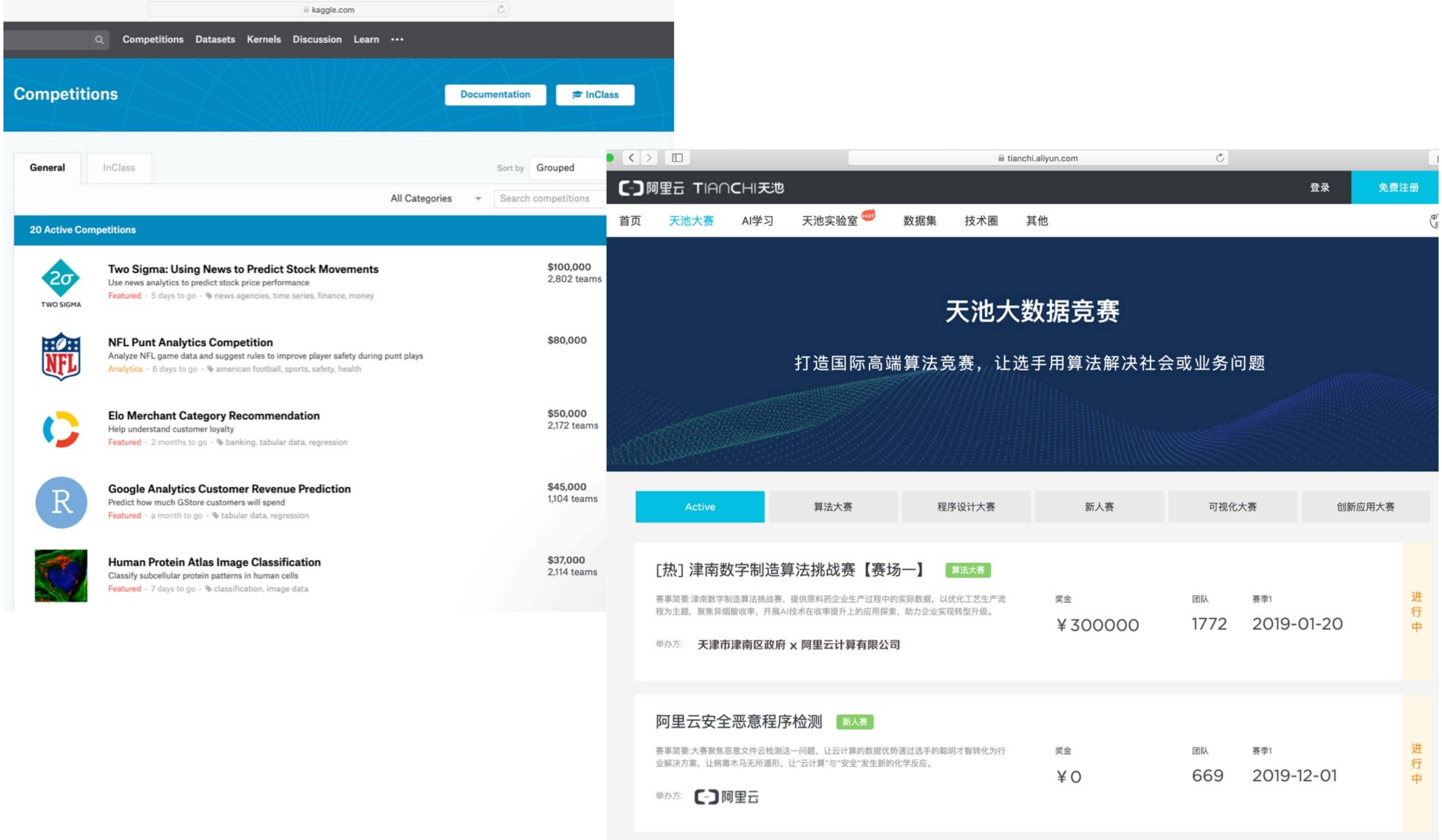
Ratings Data



1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			2		2
				5	
	2	1			1
	3			3	
1					

Million Dollars Awarded Sept 21st 2009

Competitions and prizes are still going on…



20 Active Competitions

Two Sigma: Using News to Predict Stock Movements — $100,000, 2,802 teams
Use news analytics to predict stock price performance
Featured · 5 days to go · news agencies, time series, finance, money

NFL Punt Analytics Competition — $80,000
Analyze NFL game data and suggest rules to improve player safety during punt plays
Analytics · 6 days to go · american football, sports, safety, health

Elo Merchant Category Recommendation — $50,000, 2,172 teams
Help understand customer loyalty
Featured · 2 months to go · banking, tabular data, regression

Google Analytics Customer Revenue Prediction — $45,000, 1,104 teams
Predict how much GStore customers will spend
Featured · a month to go · tabular data, regression

Human Protein Atlas Image Classification — $37,000, 2,114 teams
Classify subcellular protein patterns in human cells
Featured · 7 days to go · classification, image data

天池大数据竞赛

打造国际高端算法竞赛，让选手用算法解决社会或业务问题

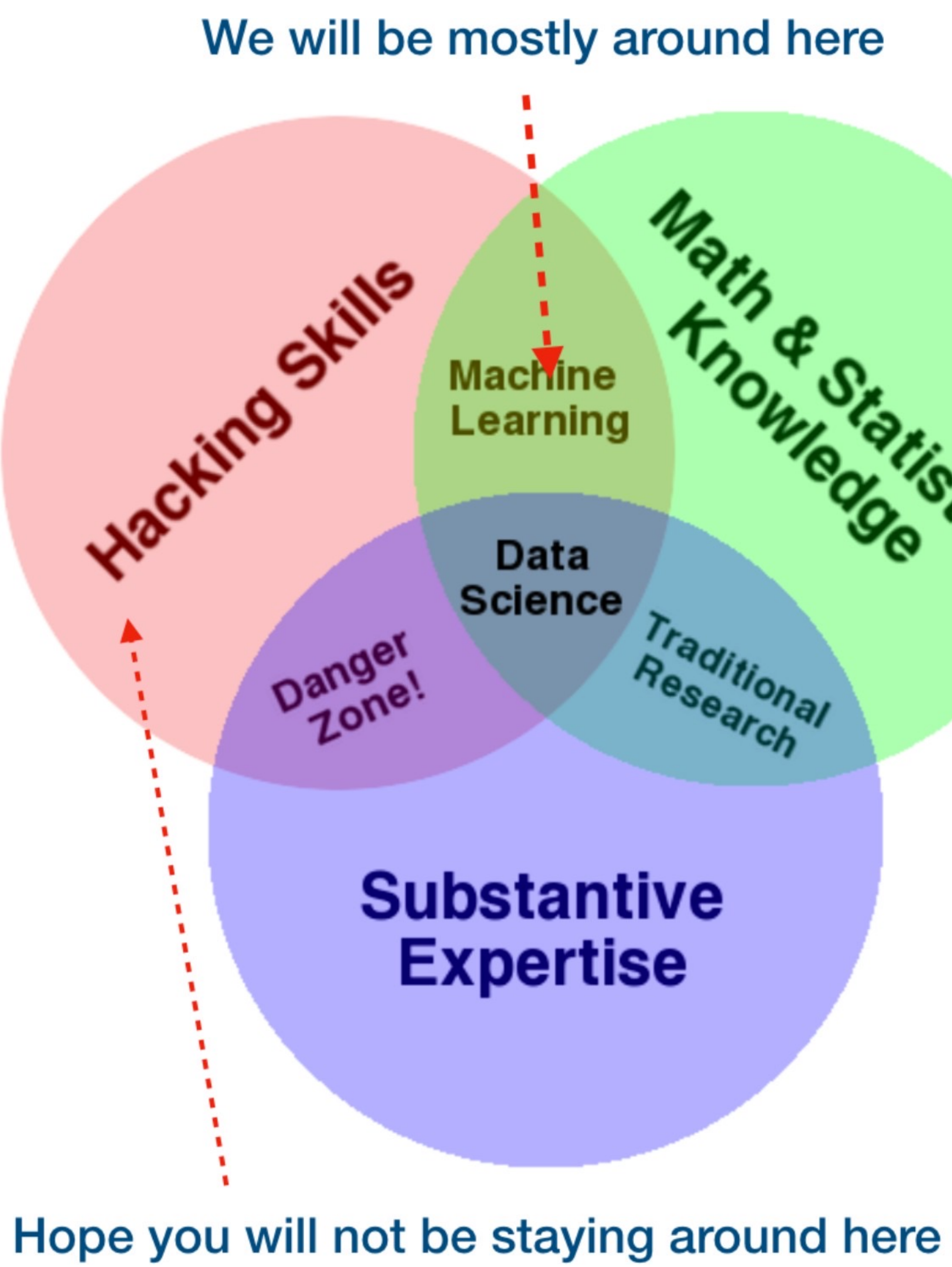
[热] 津南数字制造算法挑战赛【赛场一】 算法大赛
赛事简要:津南数字制造算法挑战赛，提供原料药企业生产过程中的实际数据，以优化工艺生产流程为主题，聚焦异烟酸收率，开展AI技术在收率提升上的应用探索，助力企业实现转型升级。
举办方: 天津市津南区政府 x 阿里云计算有限公司
奖金 ￥300000　团队 1772　赛季1 2019-01-20　进行中

阿里云安全恶意程序检测 新人赛
赛事简要:大赛聚焦恶意文件云检测这一问题，让云计算的数据优势通过选手的聪明才智转化为行业解决方案，让病毒木马无所遁形，让“云计算”与“安全”发生新的化学反应。
举办方: 阿里云
奖金 ￥0　团队 669　赛季1 2019-12-01　进行中

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

6. Plan of this course

What will this course be like?


We will be mostly around here
Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Hope you will not be staying around here

- Data Hacking: SQL, Programming, Hadoop, or whatever
- Substantive Expertise: experience or whatever

Read: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Topics to cover

- **Learning Methods**
  Linear Regression, Logistic Regression, GBDT, SVM, kNN, clustering (k-means), perceptron, DNN (brief introduction), decision tree…

- **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…