Machine Learning & Pattern Recognition

SONG Xuemeng

sxmustc@gmail.com

http://xuemeng.bitcron.com/

Outline

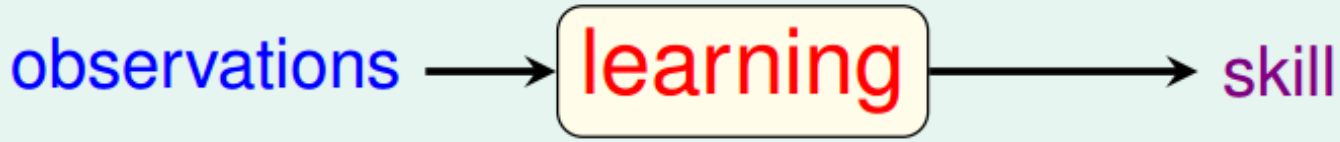
- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

Outline

- **What is Machine Learning?**
- **Applications of Machine Learning.**
- **Components of Machine Learning.**
- **Types of Machine Learning.**

Definition of Machine Learning

learning: acquiring skill
with experience accumulated from observations

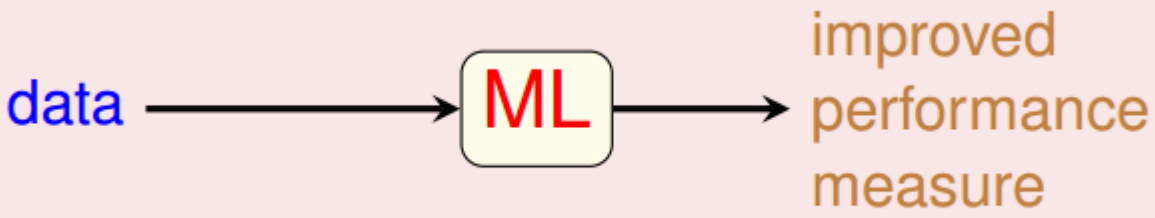


machine learning: acquiring skill
with experience accumulated/**computed** from data

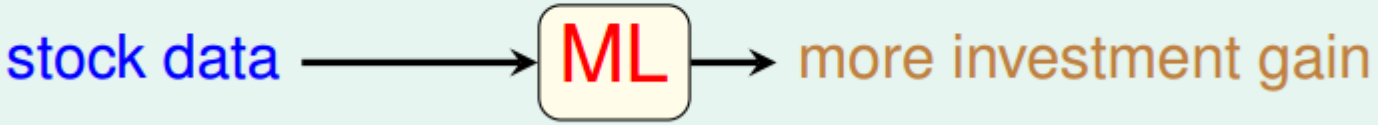
What is skill?

Definition of Machine Learning

machine learning: improving some performance measure
with experience **computed** from data



An Application in Computational Finance



Why use machine learning?

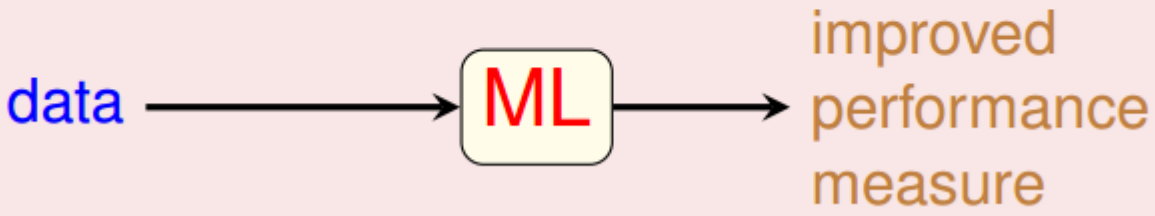
Why Do We Use Machine Learning?

- "Define" trees and hand-program: difficult;
- Learn from data (observations) and recognize;
- "ML-based tree recognition system" can be easier to build than hand-programmed system.

- Human cannot define the solution easily.

Key Essence of Machine Learning

machine learning: improving some performance measure
with experience **computed** from data



1. exists some 'underlying pattern' to be learned
   —so 'performance measure' can be improved
2. but no programmable (easy) definition
   —so 'ML' is needed
3. somehow there is data about the pattern
   —so ML has some 'inputs' to learn from

key essence: help decide whether to use ML

Exercise

Which of the following is best suited for machine learning?

1. predicting whether the next cry of the baby girl happens at an even-numbered minute or not
2. determining whether a given graph contains a cycle
3. deciding whether to approve credit card to some customer
4. guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years

Reference Answer: (3)

1. no pattern
2. programmable definition
3. pattern: customer behavior;
   definition: not easily programmable;
   data: history of bank operation
4. arguably no (or not enough) data yet

Outline

- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

Outline

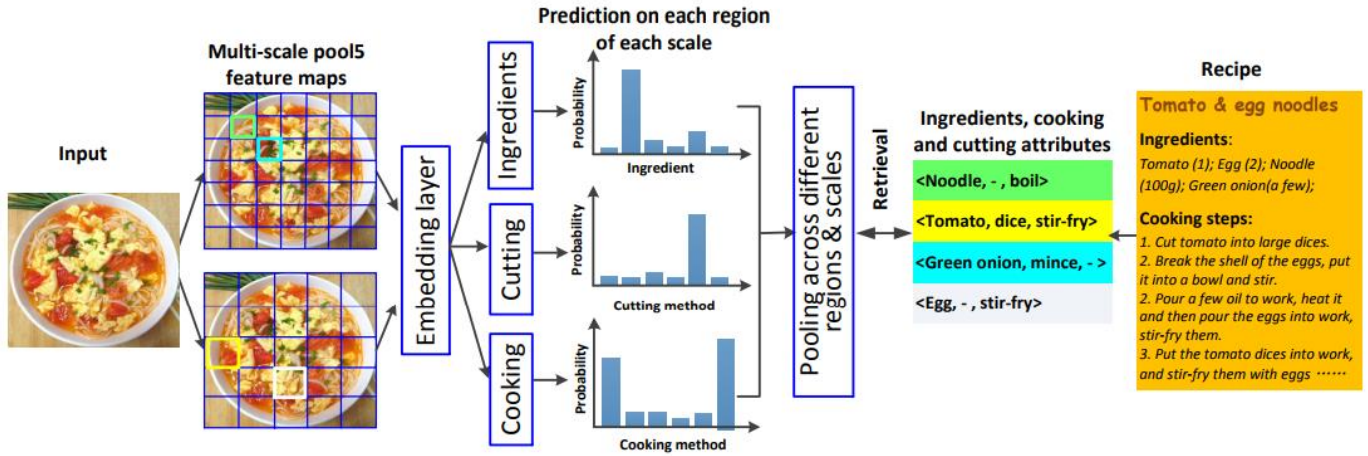
- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

Applications of Machine Learning

- **Daily Needs**
  - ✓ Food
  - ✓ Clothing
  - ✓ Housing
  - ✓ Transportation
  - ✓ Entertainment
  - ✓ Law
  - ✓ Medical
  - ✓ Education

Food

- Ingredient Recognition (Chen et al. 2017)
- Data:
  - The food categories: xiachufang and meishijie.
  - All the images : Baidu and Google image search.



Food

- Preventing Foodborne Illness (Sadilek et al. 2013, 2016)
- Data: Collected from social media (Twitter)

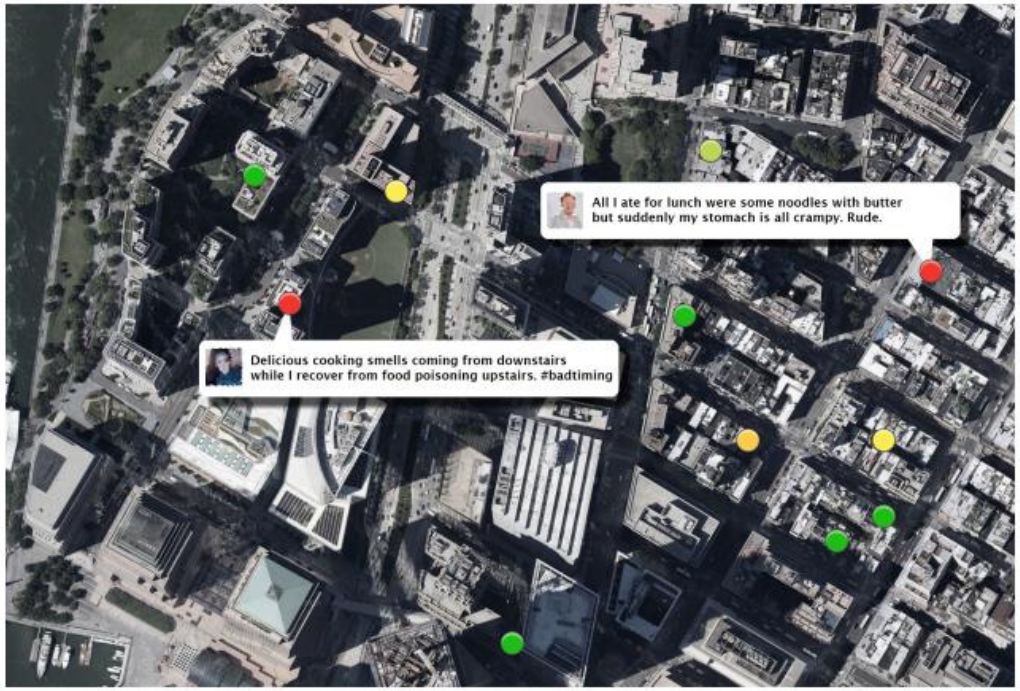


Figure 1: nEmesis analyses people's online messages and reveals individuals who may be suffering from a foodborne disease. Precise geo coordinates embedded in the messages enable us to detect specific restaurants a user had visited prior to falling ill. This fig-

Clothing

- Clothing matching (Song et al. 2018)
- Data: Outfit compositions collected from the Polyvore (website).

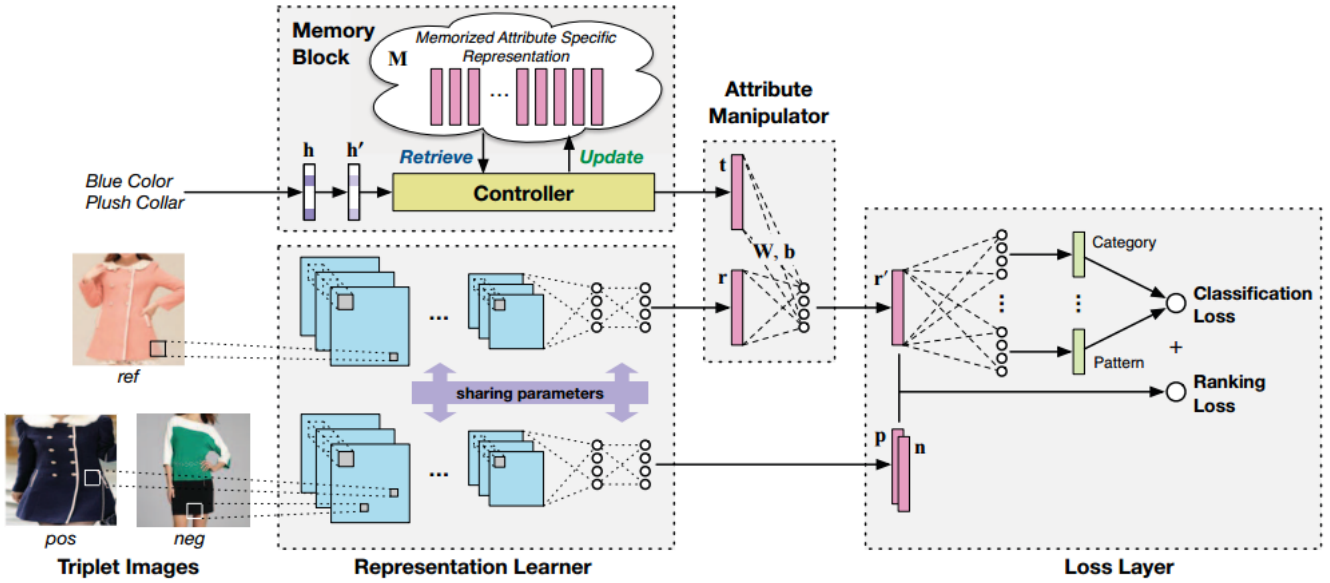
(a) Composition 1. (b) Composition 2. (c) Composition 3.

Figure 1: Example outfit compositions on Polyvore.

Clothing

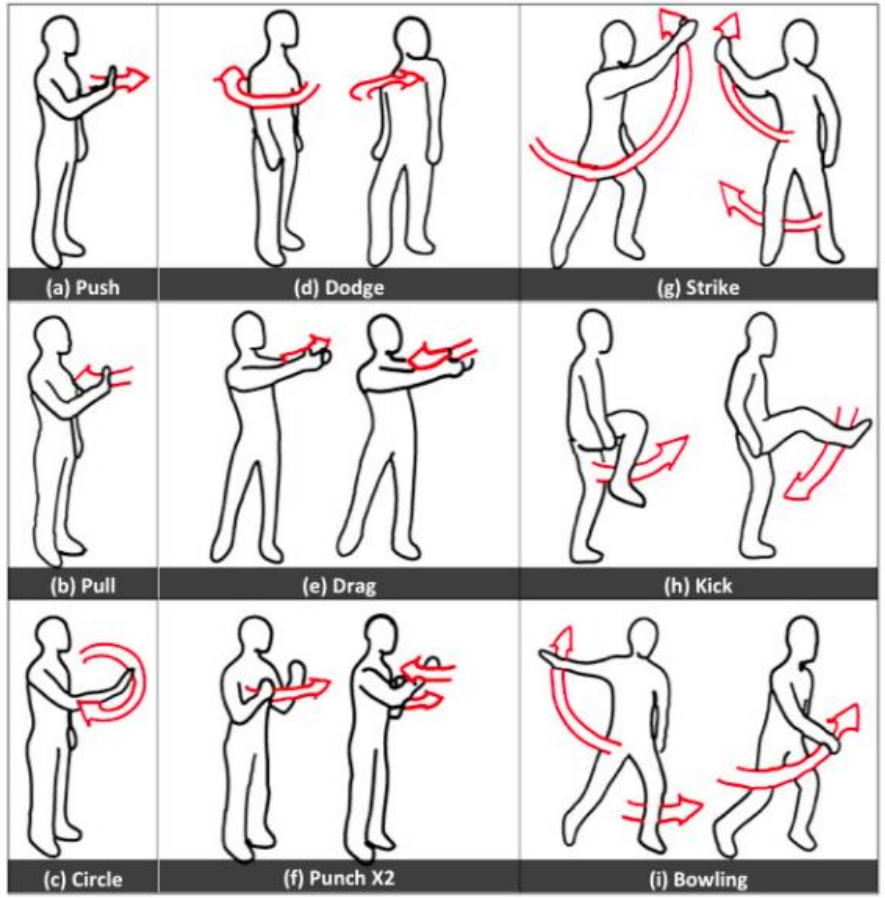
- Interactive Fashion Search(Wu et al. 2017)
  - Data: Public dataset (DARN and DeepFashion). Images with attribute annotations.

Figure 1. Fashion search with attribute manipulation. The user

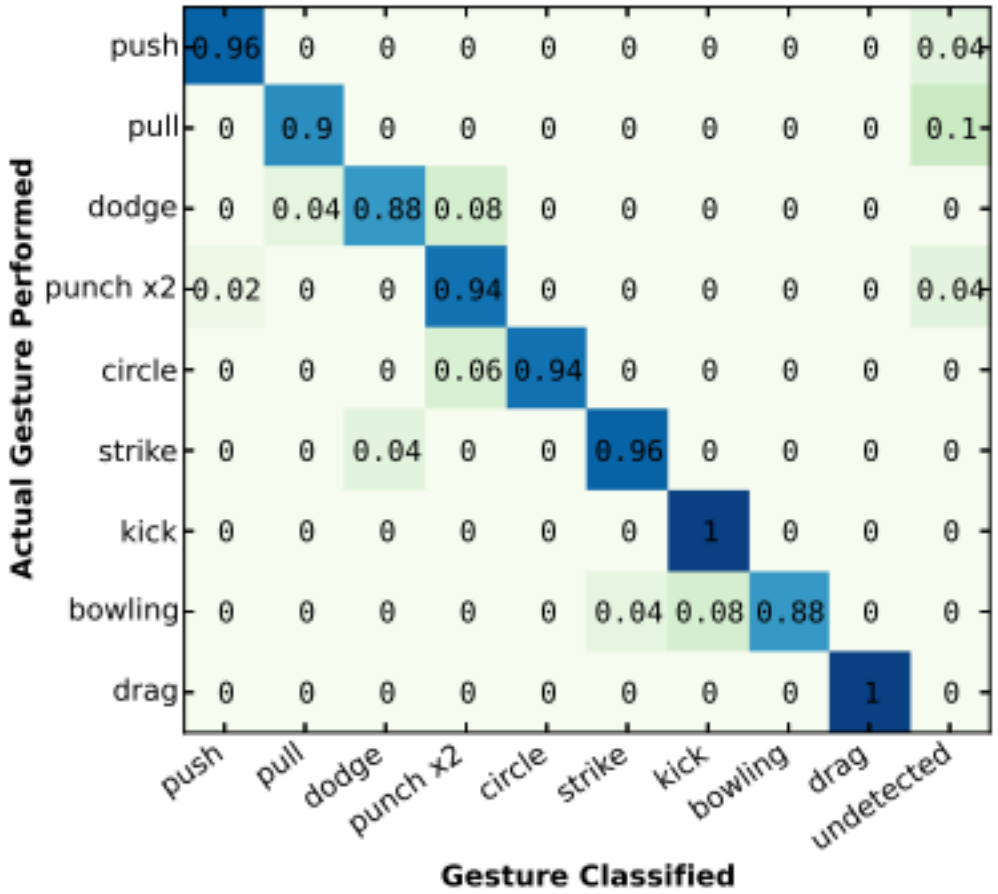


Housing

- Gesture Recognition Using Wireless Signals (Pu et al. 2013, 2015)
  - Data: 5 users perform gestures in an office and a two-bedroom.



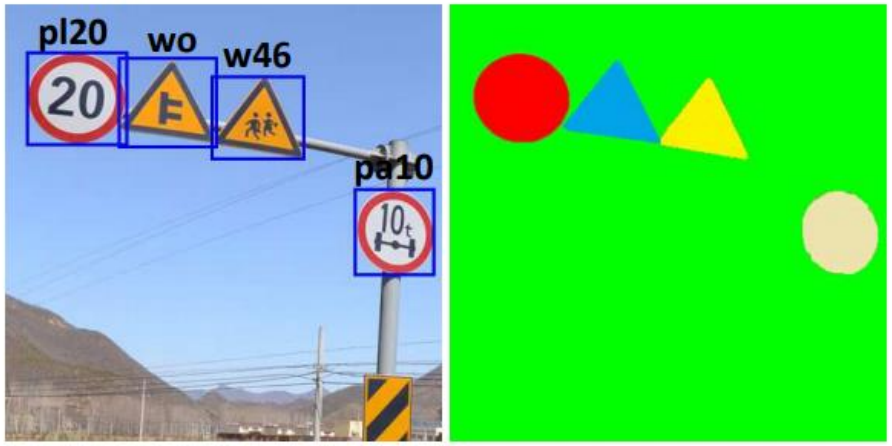
Gesture sketches



Transportation

- **Traffic-Sign Detection (Zhu et al. 2016)**
  - **Data: Tencent street views with manual annotation.**

(a) 8192×2048 panorama from Tencent Street View before slicing vertically into 4 images. Sky and ground at top and bottom have been cropped.



(b) Bounding box and class label (c) Pixel mask annotation

Benchmark

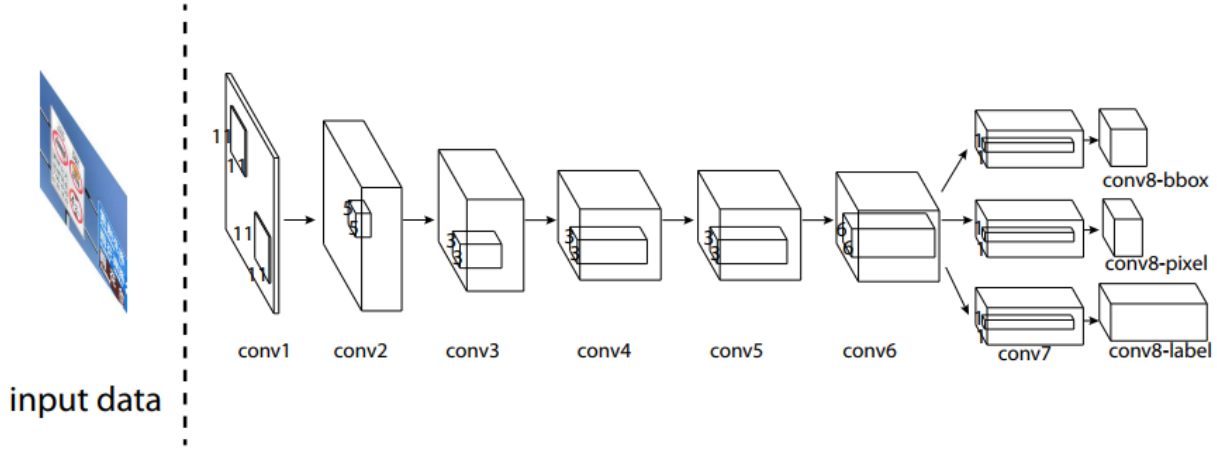


Figure 8. Architecture of our multi-class network. Our network is fully convolutional, and branches after the 6th layer.

Transportation

- **Personalized Tour Recommendation (Zhao et al. 2017)**
  - **Data: Public dataset (Flicker).**

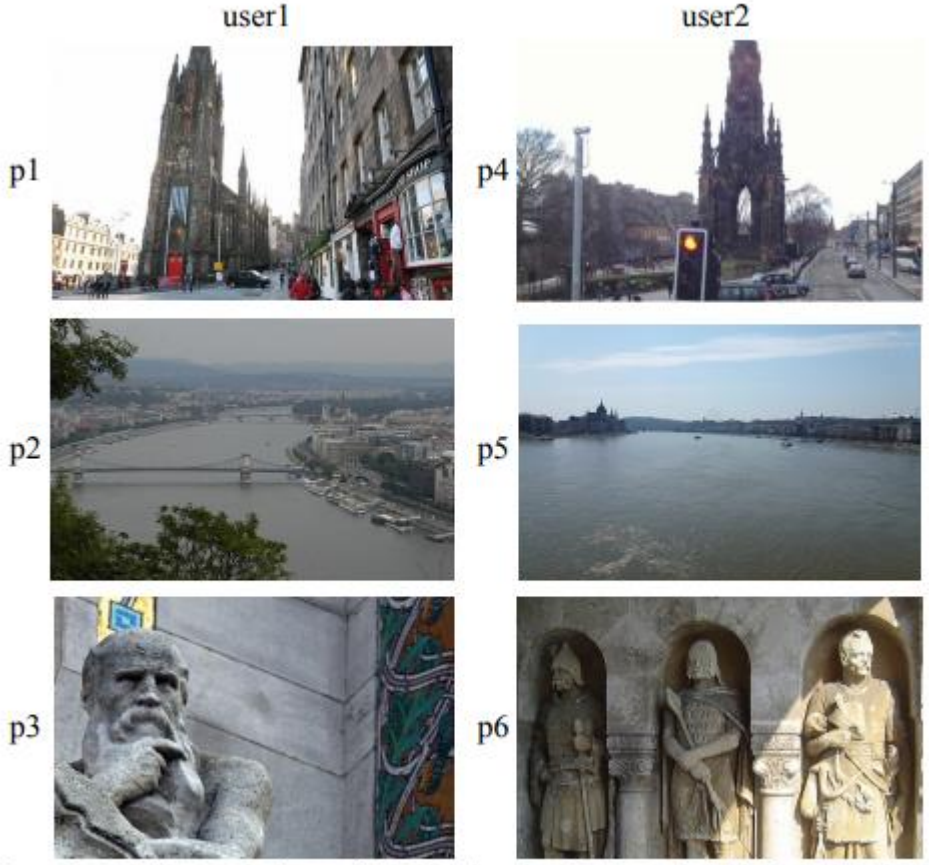


Figure 1: Three pairs of POI photos from six different POIs and visited by two users having similar visual appearances.

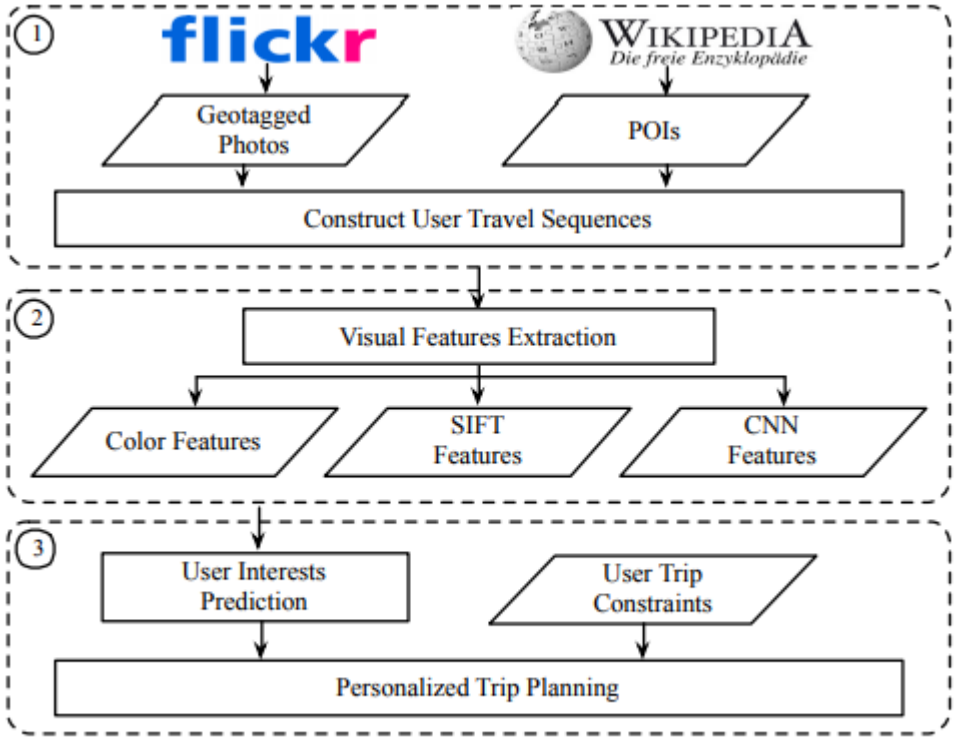
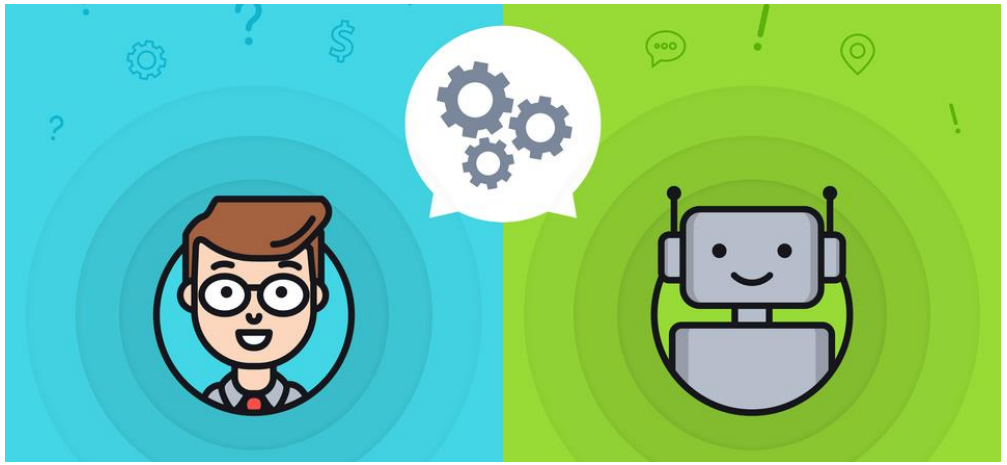


Figure 3: Framework of Photo2Trip Recommender System

Entertainment

- **Chatting Bot (Microsoft 2014, Zhang et al. 2017, Want et al. 2018)**



Harbin Institute of Technology Benben **Microsoft Xiaobing**

Law

- Crime Classification (Want et al. 2018)
  - Data: Collected from the China Judgments Online.
  - Task: Determine the specific articles (as labels) that the evidence (document) violated (multi-label classification).

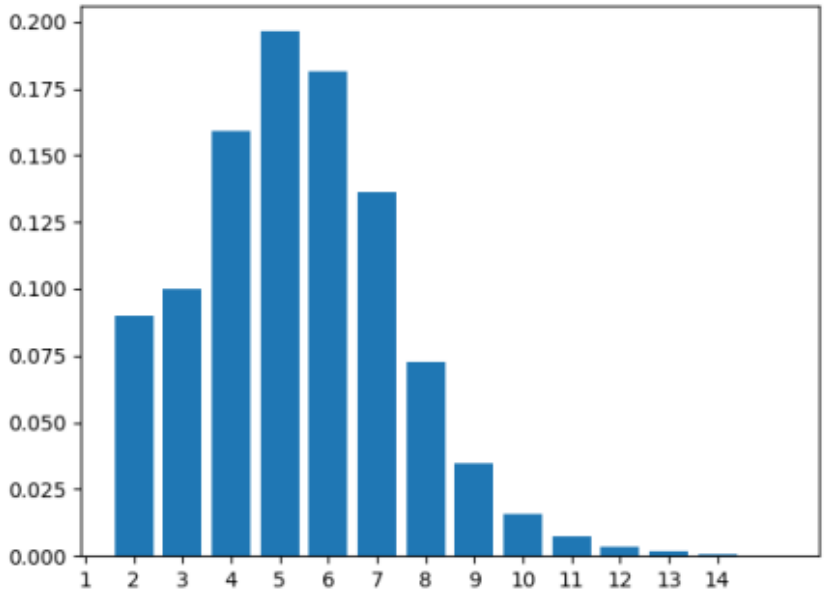
Evidence:	*In late February 1, 2010 10 pm, Li intended to a cinema with his friend Jiang. After a contretemps with the defendant Guo, they gave Guo a beating and Guo ran away. After watched the movie, Li and Jiang were assaulted by Guo and his friends near the cinema. Jiang was stabbed by Guo.....*
Article:	***Article22***: *Preparation for a crime refers to the preparation of the instruments or the creation of the conditions for a crime;* ***Article25***: *A joint crime refers to an intentional crime committed by two or more persons jointly.*

An example of the judgement case, including an evidence and two articles violated.

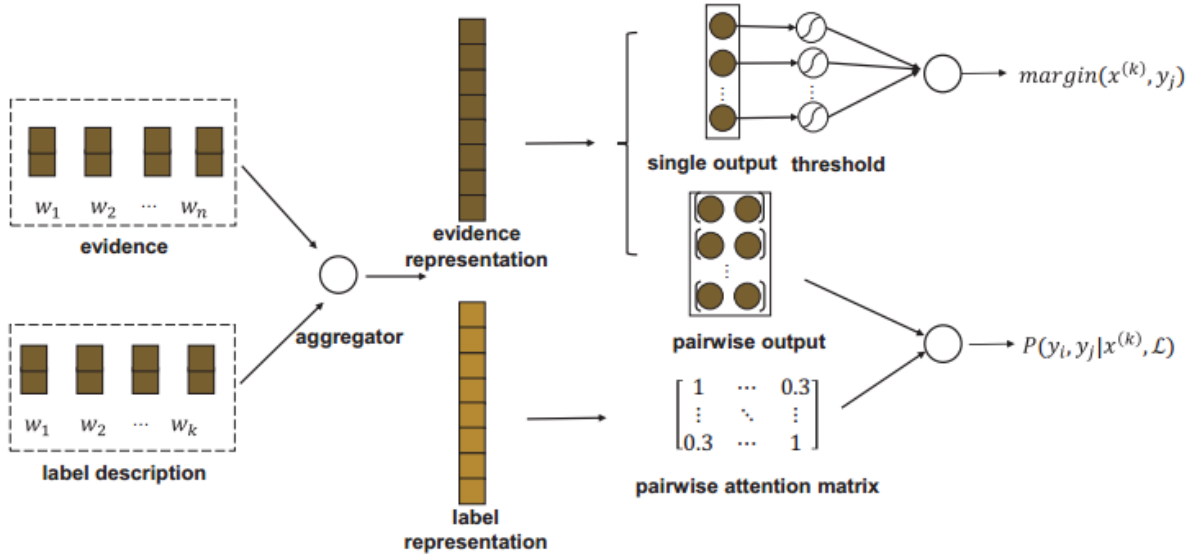
http://wenshu.court.gov.cn/Index

Law

- Crime Classification (Want et al. 2018)
  - Data: Collected from the China Judgments Online.
  - Task: Determine the specific articles (as labels) that the evidence (document) violated (multi-label classification).



Distribution of article set size over evidences.



The overall architecture of the proposed model.

Medical

- **Predicting depression via social media. (Choudhury et al. 2013)**
  - **Data: Social Media Data (Tweets).**
  - **Task: Predict a binary response variable (depressed/not depressed).**

Having a job again makes me happy. Less time to be depressed and eat all day while watching sad movies.
"Are you okay?" Yes…. I understand that I am upset and hopeless and nothing can help me… I'm okay… but I am not alright
"empty" feelings I WAS JUST TALKING ABOUT HOW I I HAVE EMOTION OH MY GOODNESS I FEEL AWFUL
I want someone to hold me and be there for me when I'm sad.
Reloading twitter till I pass out. *lonely* *anxious* *butthurt* *frustrated* *dead*

Example posts from users in the depression class.

	precision	recall	acc. (+ve)	acc. (mean)
engagement	0.542	0.439	53.212%	55.328%
ego-network	0.627	0.495	58.375%	61.246%
emotion	0.642	0.523	61.249%	64.325%
linguist. style	0.683	0.576	65.124%	68.415%
dep. language	0.655	0.592	66.256%	69.244%
demographics	0.452	0.406	47.914%	51.323%
all features	0.705	0.614	68.247%	71.209%
dim. reduced	**0.742**	**0.629**	**70.351%**	**72.384%**

Performance metrics in depression prediction in posts using various models.

Medical

- Classification of skin cancer. (Esteva et al. 2017)
  - Data: The ISIC Dermoscopic Archive, the Edinburgh Dermofit Library and data from the Stanford Hospital (129,450 clinical images).

DERMOFIT IMAGE LIBRARY

High quality skin lesion images for use as a research tool in computer science and medical imaging

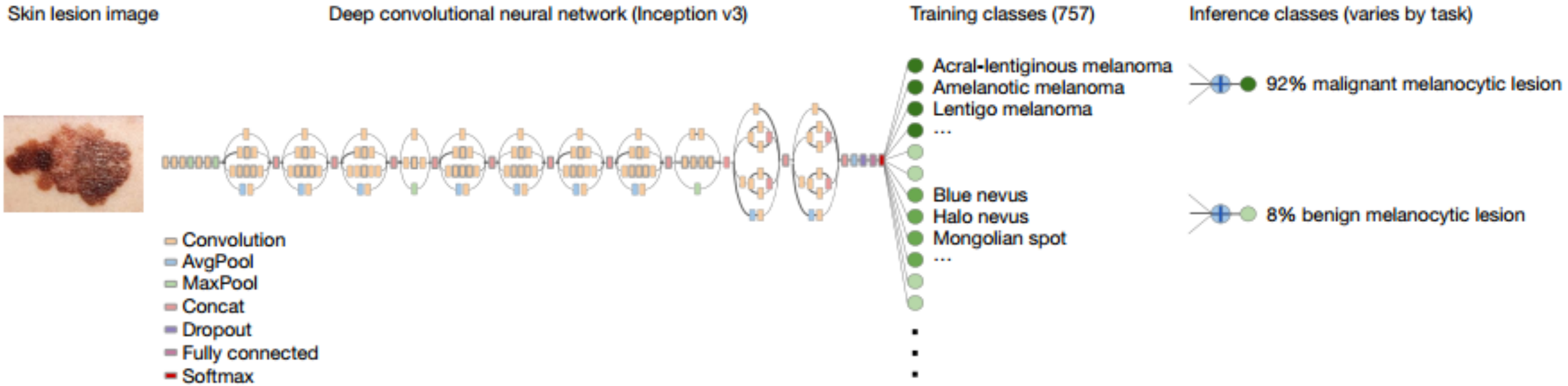
Stanford HEALTH CARE

Open-access dermatology repositories

Data from the hospital.

Medical

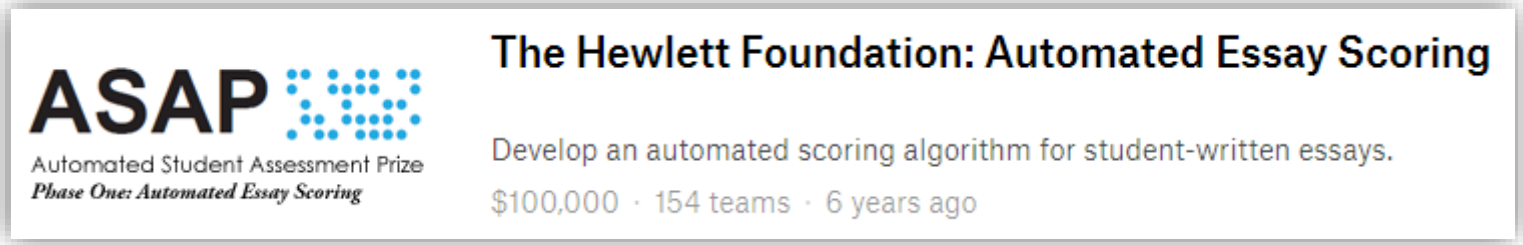
- Classification of skin cancer. (Esteva et al. 2017)
  - Task: Two binary classification use cases:
    - Keratinocyte carcinomas vs benign seborrheic keratoses;
    - Malignant melanomas vs benign nevi.



Proposed framework.

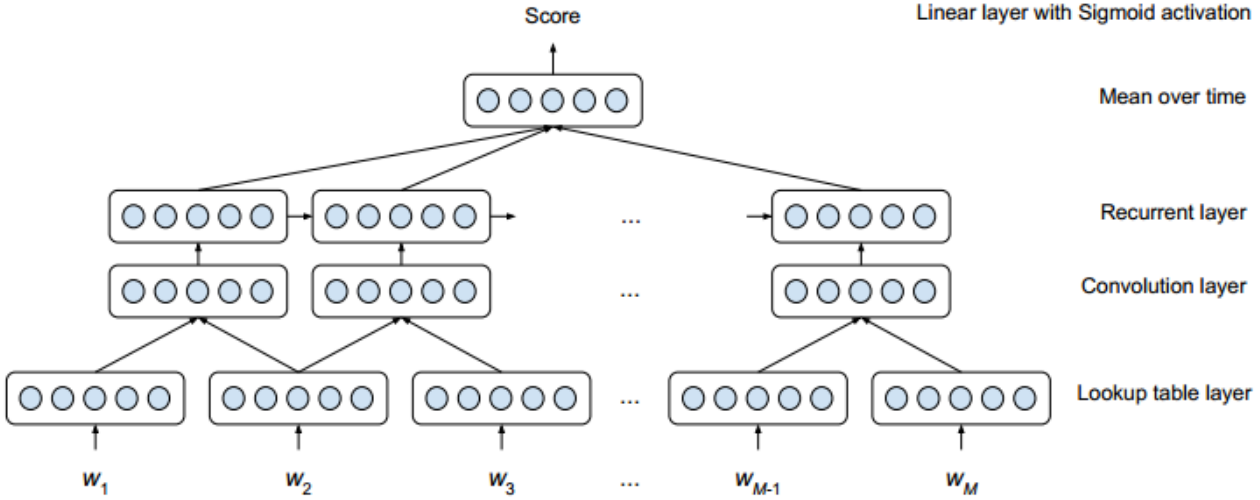
Education

- Automated Essay Scoring. (Taghipour et al. 2016)
  - Data: Public dataset provide by the following competition (2012).



Prompt	#Essays	Avg length	Scores
1	1,783	350	2–12
2	1,800	350	1–6
3	1,726	150	0–3
4	1,772	150	0–3
5	1,805	150	0–4
6	1,800	150	0–4
7	1,569	250	0–30
8	723	650	0–60

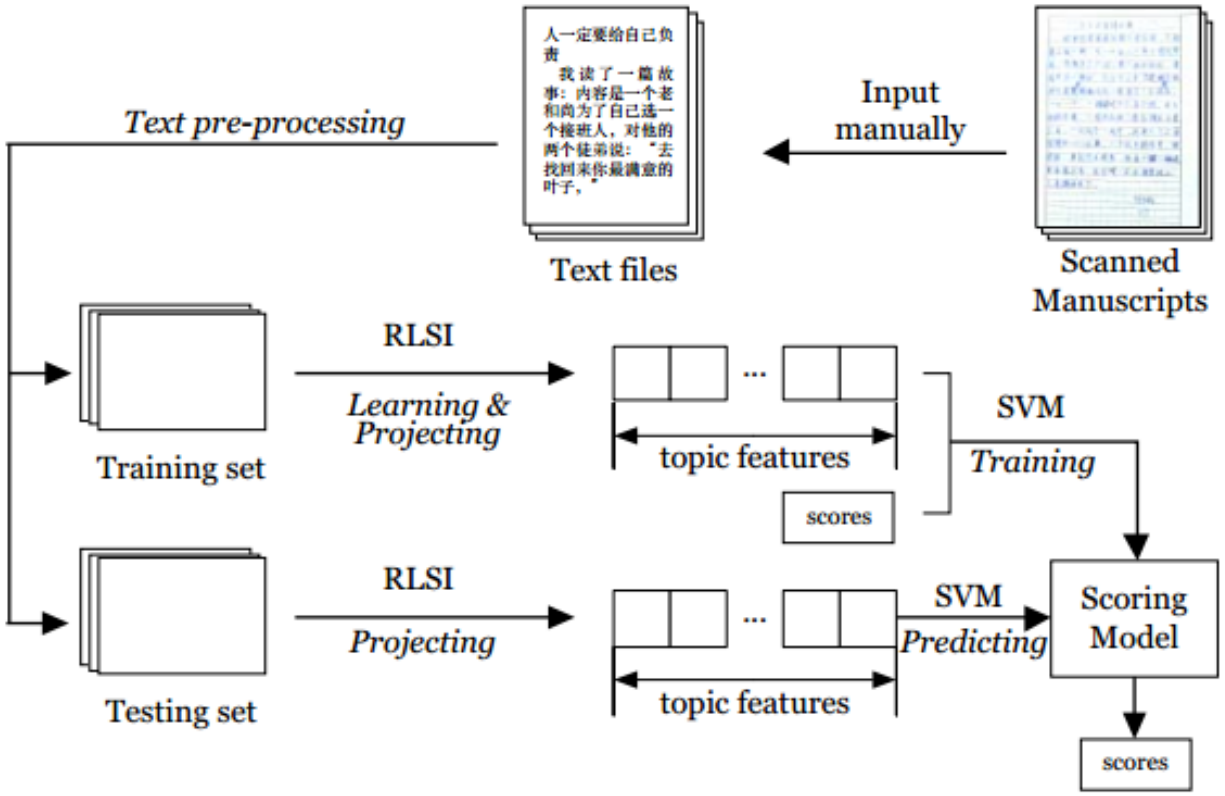
Table 1: Statistics of the ASAP dataset.



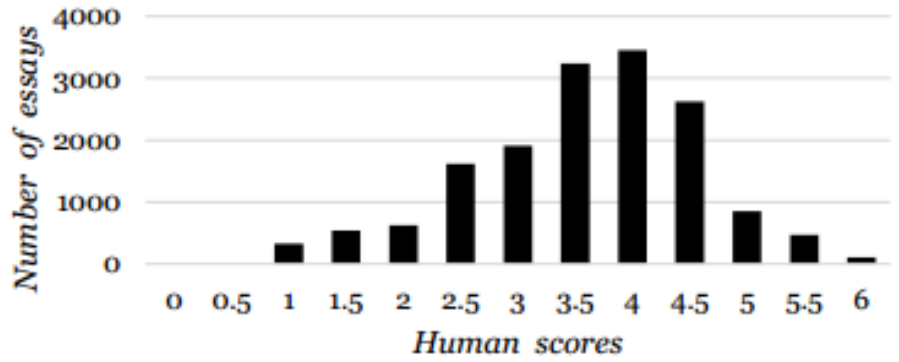
Education

- Automated Essay Scoring. (Hao et al. 2014)
  - Data: Essays from the MHK (the minorities-oriented Chinese level test).

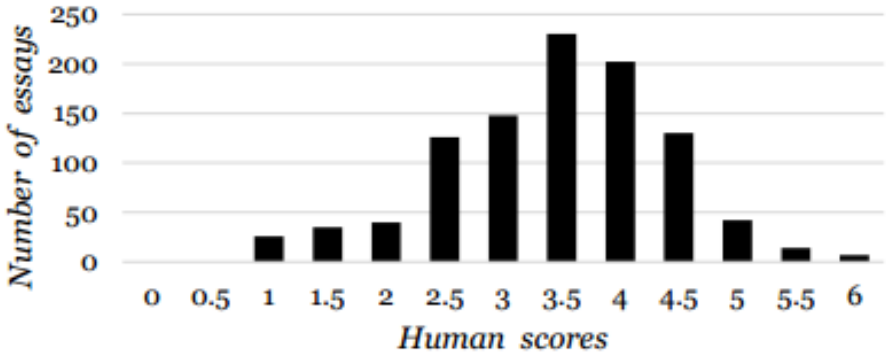
中国少数民族汉语水平等级考试



The flowchart of the proposed method.



Training set (total 15,776 essays)



Testing set (total 1,000 essays)

Outline

- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

Outline

- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

Basic Notations

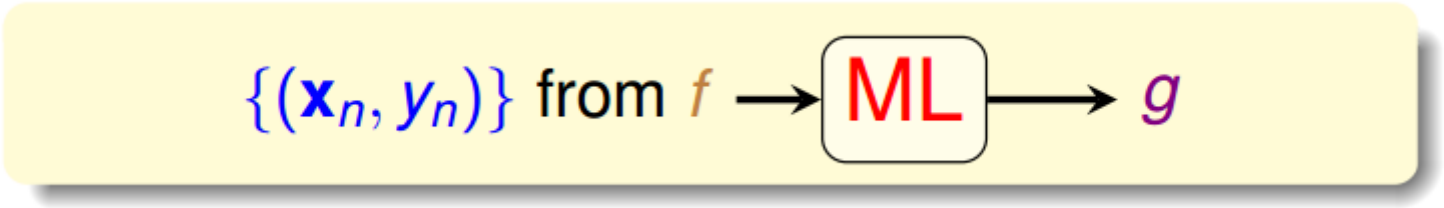
- Bold capital letters (e.g., $\boldsymbol{X}$) $\rightarrow$ Matrices;
- Bold lowercase letters (e.g., $\boldsymbol{x}$) $\rightarrow$ Vectors;
- Non-bold letters (e.g., x) $\rightarrow$ Scalars;
- Greek letters (e.g., β) $\rightarrow$ The parameters.

Basic Notations

- **Input:** $\mathbf{x} \in \mathcal{X}$ **(**$\mathcal{X}$: input/sample space**)**
- **Output:** $y \in \mathcal{Y}$ **(**$\mathcal{Y}$**:** output/label space**)**
- **Unknown pattern** to be learned, i.e., target function:

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

- **Data** ⇔ Training examples (sample/instance).
  - $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_m}, y_m)\}$, where $\boldsymbol{x_i} = (\mathrm{x_{i1}}, \mathrm{x_{i2}}, \ldots, \mathrm{x_{id}}) \in \mathbb{R}^d$, $\boldsymbol{x_i} \in \mathcal{X}$; x_{ij} is the value of the j-th attribute/feature of $\boldsymbol{x_i}$.
  - d**: dimensionality** of the feature space.
- **Hypothesis:** skill with hopefully good performance:
  $g: \mathcal{X} \to \mathcal{Y}$ ('learned' formula to be used)



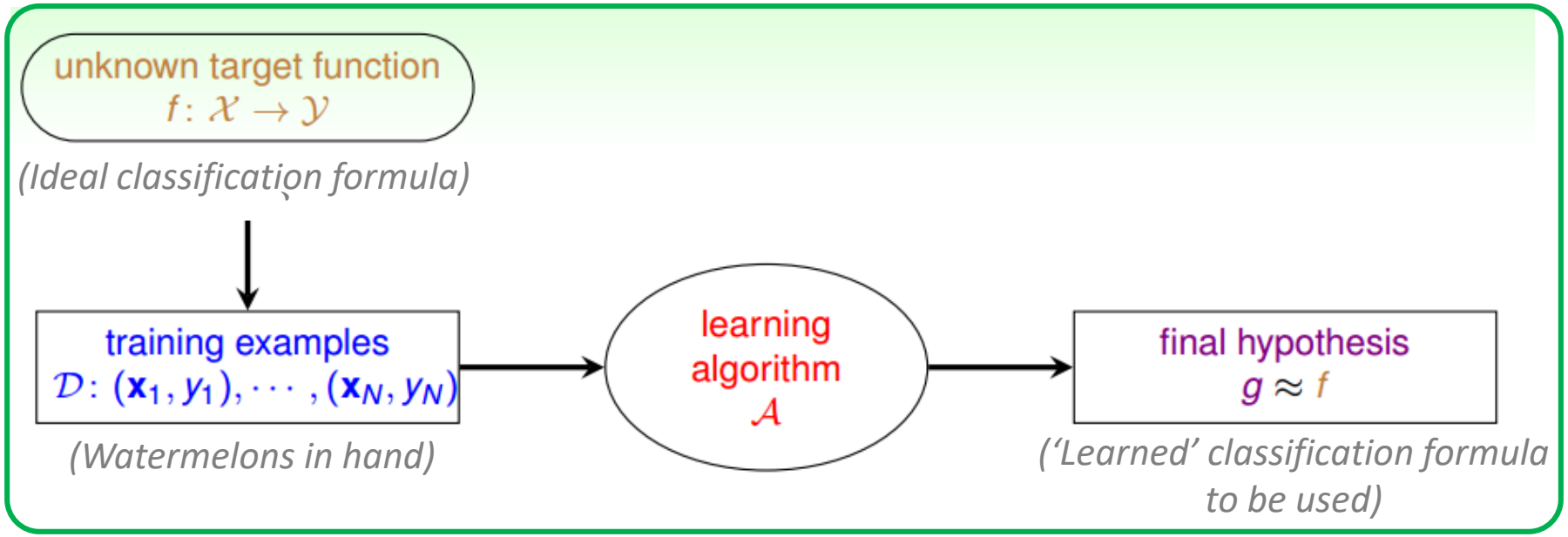
Watermelon Classification

表 1.1　西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

Is the given watermelon good?

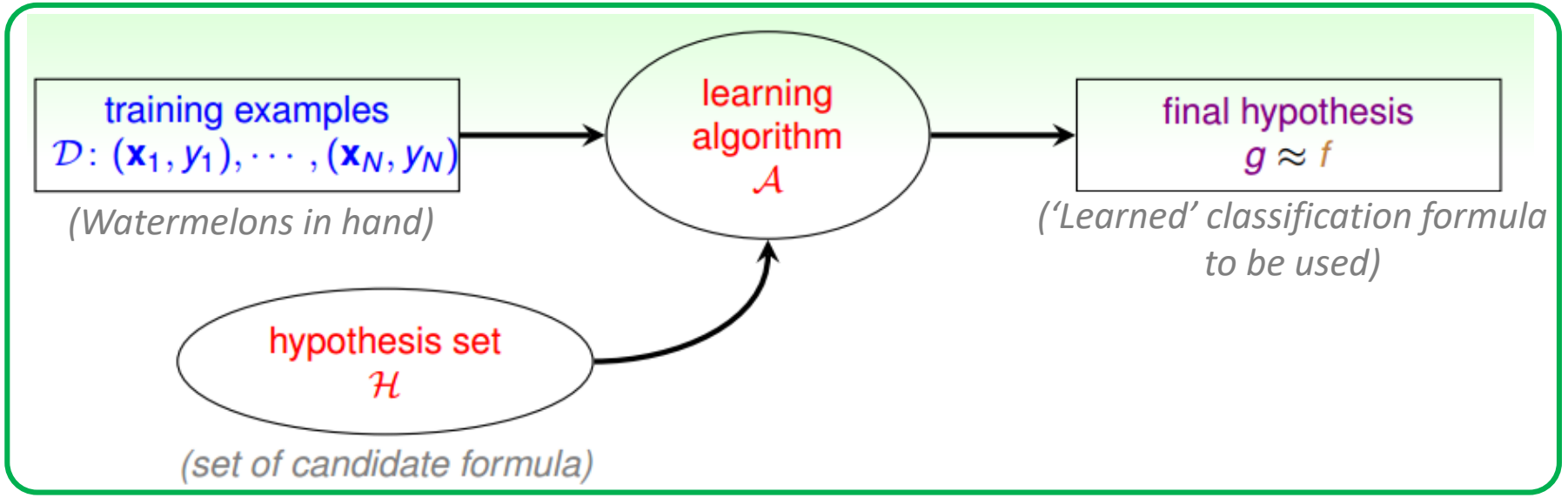
Learning Flow



- target f **unknown**
  (i.e. no programmable definition)
- hypothesis g hopefully $\approx f$
  but possibly **different** from f
  (perfection 'impossible' when f unknown)

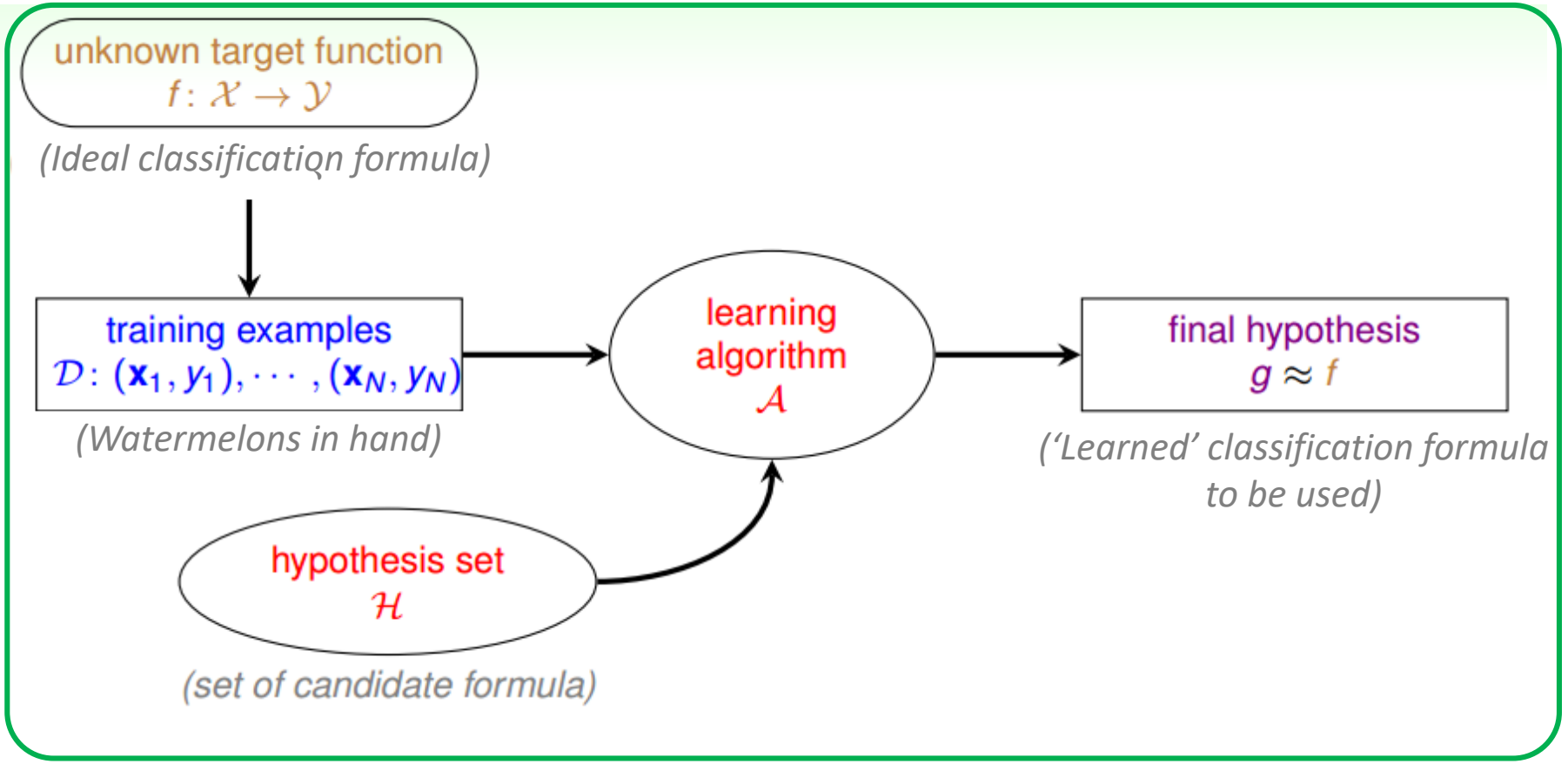
What does *g* look like?

Learning Model



- Assume g $\in \mathcal{H} = \{h_k\}$, i.e., classify a given watermelon is good if
  - h_1: (色泽="青绿")∧(根蒂="蜷缩") ∧(敲声="浊响")
  - h_2: (色泽=*)∧(根蒂="蜷缩") ∧(敲声="浊响")
  - h_3: (色泽=∗)∧(根蒂="蜷缩") ∧(敲声=*)
  - ⋮
- Hypothesis set $\mathcal{H}$:
  - Can contain good or bad hypotheses
  - Up to the learning algorithm *A* to pick the 'best' one as *g*

Practical Definition of Machine Learning



machine learning:
use data to compute hypothesis g
that approximates target f

Outline

- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

Outline

- What is Machine Learning?
- Applications of Machine Learning.
- Components of Machine Learning.
- Types of Machine Learning.

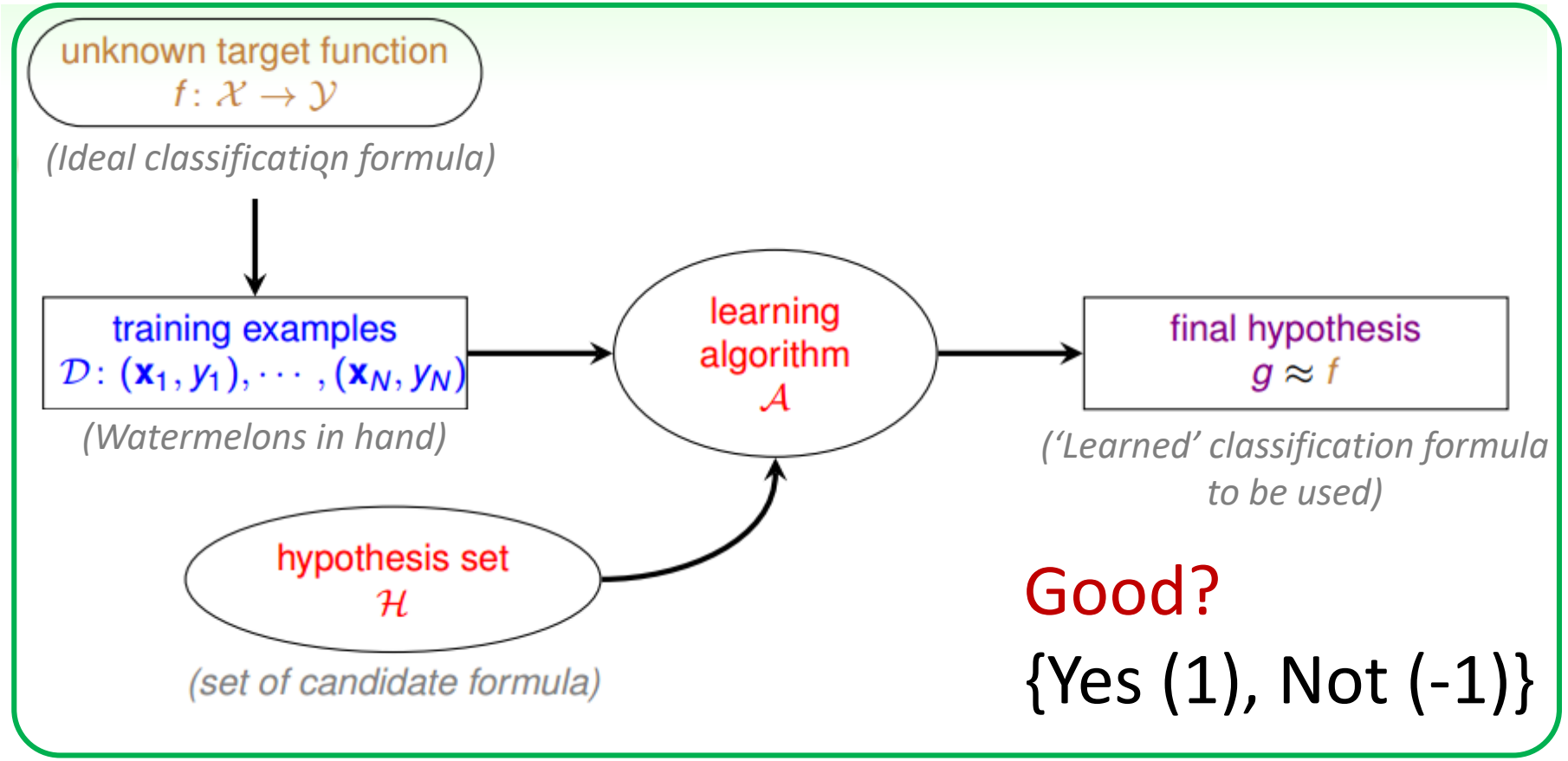
Categories of Machine Learning

- Learning with different output space Y
- Learning with different data label y_n
- Learning with different protocol f (x_n, y_n)
- Learning with different input space X

Categories of Machine Learning

- Learning with different output space Y
- Learning with different data label y_n
- Learning with different protocol f (x_n, y_n)
- Learning with different input space X

Binary Classification

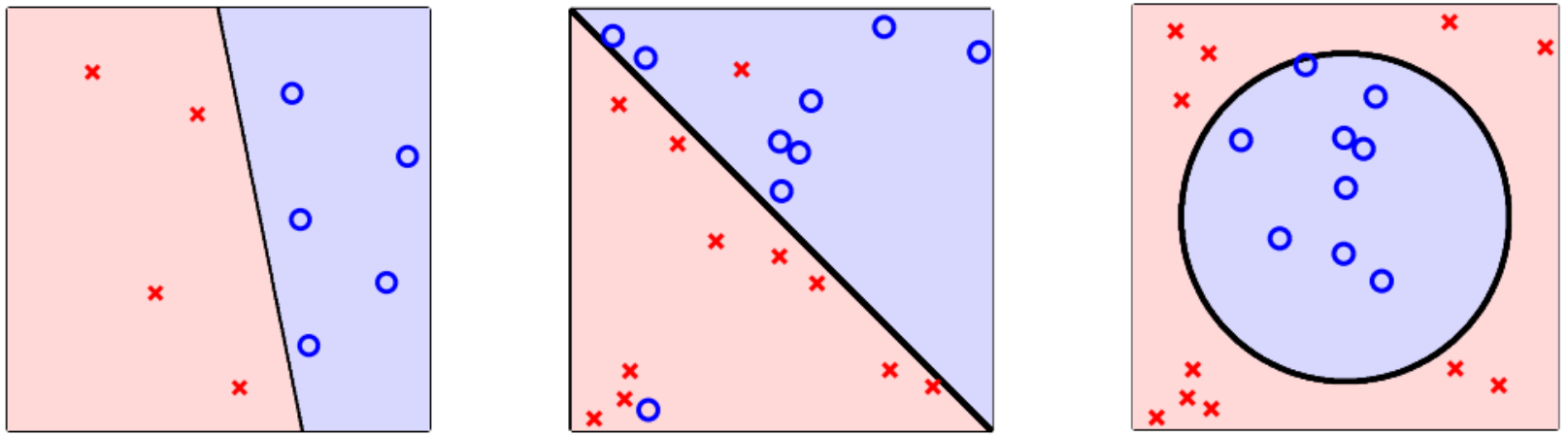


$y = \{-1,1\}$ or $\{0,1\}$: binary classification.

Binary Classification

$y = \{-1,1\}$ or $\{0,1\}$

Finding decision boundaries

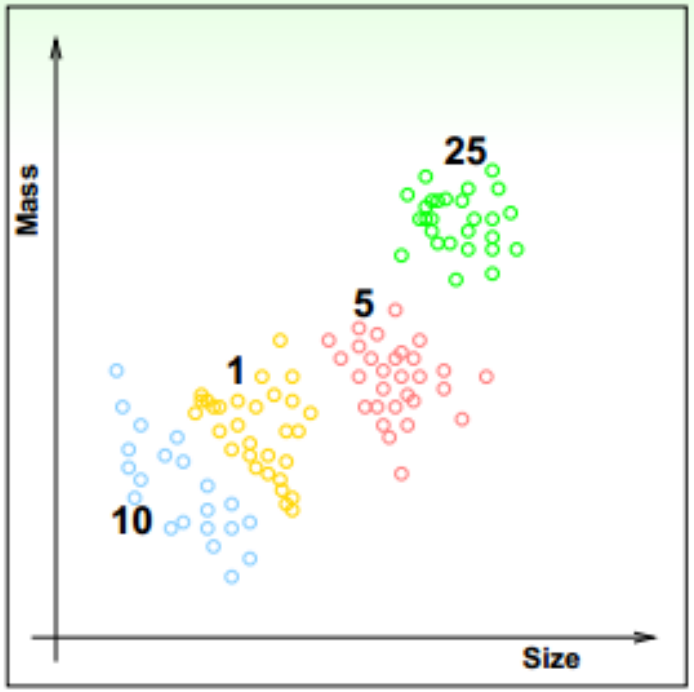


Other Binary Classification Problems

- Credit approve/disapprove
- Email spam/non-spam
- Patient sick/not sick
- Ad profitable/not profitable

Multiclass Classification

$|y| > 2.$



- classify US coins (1c, 5c, 10c, 25c) by (size, mass)
- $\mathcal{Y} = \{1c, 5c, 10c, 25c\}$, or $\mathcal{Y} = \{1, 2, \cdots, K\}$ **(abstractly)**
- binary classification: special case with $K = 2$

Other Multiclass Classification Problems

- written digits $\Rightarrow 0, 1, \cdots, 9$
- pictures $\Rightarrow$ apple, orange, strawberry
- emails $\Rightarrow$ spam, primary, social, promotion, update (Google)

Multiclass vs Multi-label Classification

What color is the cat in this photo?

Cailco

Orange Tabby

Tuxedo

Multiclass vs Multi-label Classification

Multiclass classification refers to the setting when there are > 2 possible class labels (e.g., calico, orange tabby, tuxedo).

x_1	x_2	x_3	x_4	y
1.01	-4.26	7.99	-0.03	Calico
2.50	1.00	4.87	5.95	Orange Tabby
-2.34	-1.24	-0.88	-1.31	Tuxedo
0.55	0.59	-3.08	1.27	Orange Tabby
2.08	-3.46	4.62	-1.13	Gray Tabby
…	…	…	…	…

Multiclass vs Multi-label Classification

What color and sex is the cat in this photo?

Cailco
Female

Orange Tabby
Male

Tuxedo
Male

Multiclass vs Multi-label Classification

Multi-label classification refers to the setting when there > 1 label you want to predict.

x_1	x_2	x_3	x_4	y_1	y_2
1.01	-4.26	7.99	-0.03	Calico	Female
2.50	1.00	4.87	5.95	Orange Tabby	Male
-2.34	-1.24	-0.88	-1.31	Tuxedo	Male
0.55	0.59	-3.08	1.27	Orange Tabby	Male
2.08	-3.46	4.62	-1.13	Gray Tabby	Female
…	…	…	…	…	…

Multiclass vs Multi-label Classification

Multiclass classification

- It’s possible to create multiclass classifiers out of binary classifiers.
  - One vs Rest (One vs All)
    - Each classifier predicts whether the instance belongs to the target class.
  - All pairs
    - Trains a binary classifier for every pair of classes. Whichever class “wins” more pairwise classifications will be the final prediction.

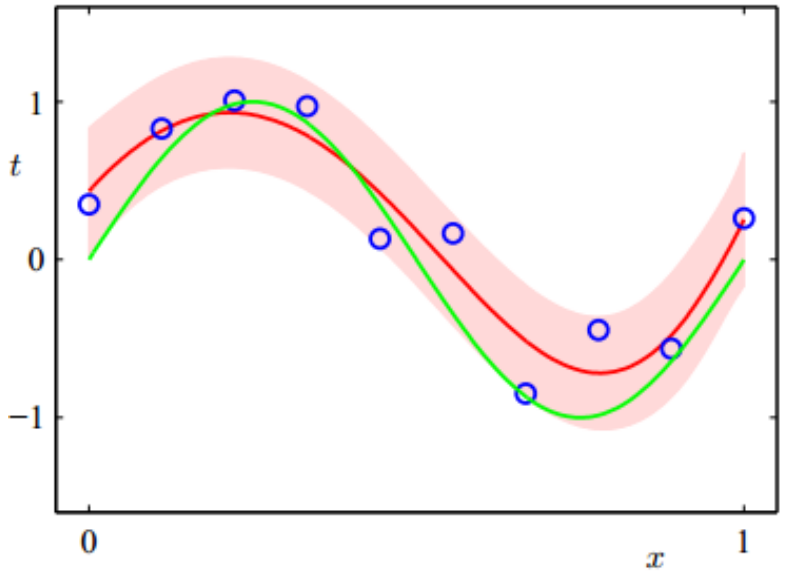
Multi-label classification

- Train separate classifiers for each label.
- There might be correlations between the classes.
  - Calico cats are almost always female
  - Orange cats are more often male.

Regression: Patient Recovery Prediction Problem

- binary classification: patient features $\Rightarrow$ sick or not
- multiclass classification: patient features $\Rightarrow$ which type of cancer
- regression: patient features $\Rightarrow$ **how many days before recovery**
- $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [\text{lower}, \text{upper}] \subset \mathbb{R}$ (bounded regression)
  —**deeply studied in statistics**

Regression: fitting a curve/plane to data



Regression: Patient Recovery Prediction Problem

- binary classification: patient features $\Rightarrow$ sick or not
- multiclass classification: patient features $\Rightarrow$ which type of cancer
- regression: patient features $\Rightarrow$ **how many days before recovery**
- $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [\text{lower}, \text{upper}] \subset \mathbb{R}$ (bounded regression)
  —**deeply studied in statistics**

Other Regression Problems

- Images (Instagram) $\Rightarrow$ Popularity prediction
- E-commerce product $\Rightarrow$ Sale prediction

Mini Summary

Learning with different output space $\mathcal{Y}$

- Binary classification: $\mathcal{Y} = \{-1, +1\}$
- Multiclass classification: $\mathcal{Y} = \{1, 2, \ldots, K\}$
- Regression: $\mathcal{Y} = \mathbb{R}$

Exercise

What is this learning problem?

The entrance system of the school gym, which does automatic face recognition based on machine learning, is built to charge four different groups of users differently: Staff, Student, Professor, Other. What type of learning problem best fits the need of the system?

1. binary classification
2. multiclass classification
3. regression
4. multilabel classification

Categories of Machine Learning

- Learning with different output space Y
- Learning with different data label y_n
- Learning with different protocol f (x_n, y_n)
- Learning with different input space X

Supervised Learning

Every $\boldsymbol{x_n}$ comes with corresponding $\boldsymbol{y_n}$.

$\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_m}, y_m)\}$

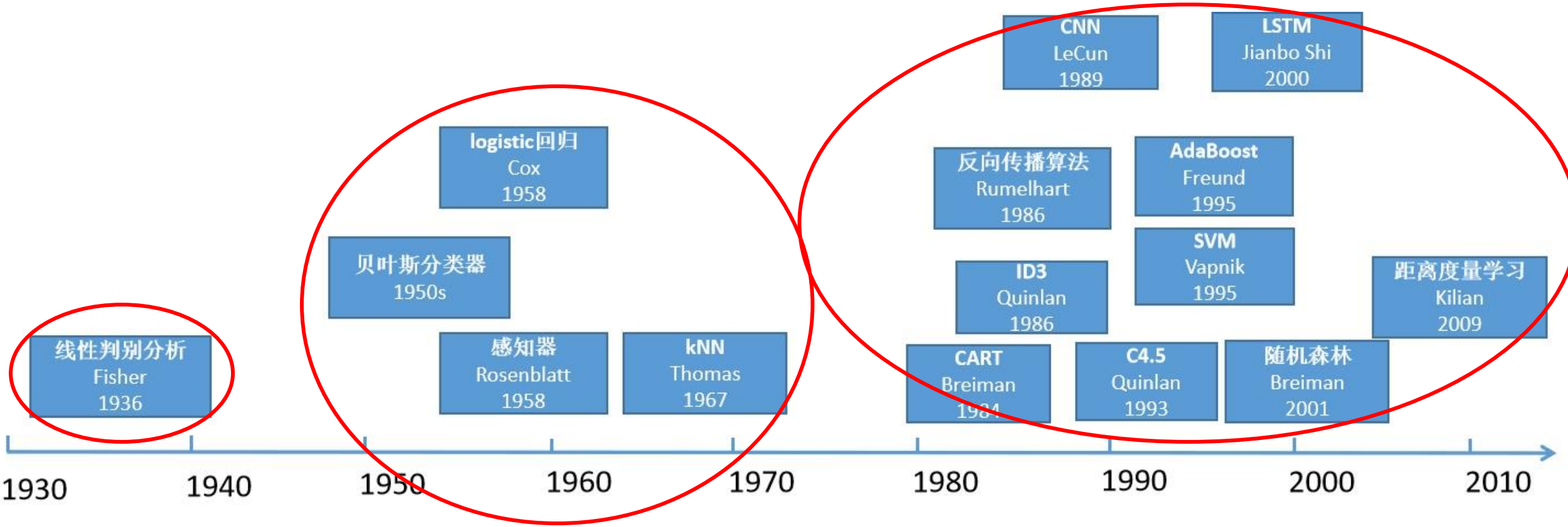
编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

Supervised Learning

Characteristics of the Supervised Learning

- We are primarily interested in prediction.
- The possible values of what we want to predict are specified, and we have some training cases for which its value is known.
- $\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_m}, y_m)\}$

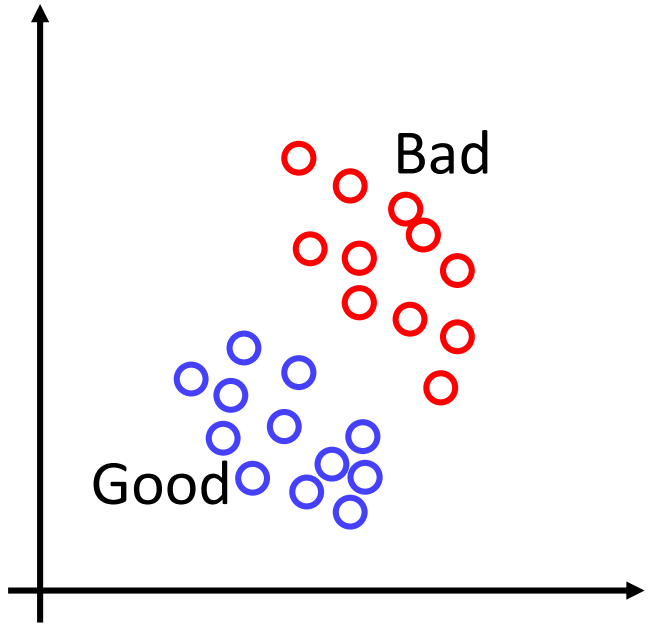
Supervised Learning



Supervised Learning vs Unsupervised Learning

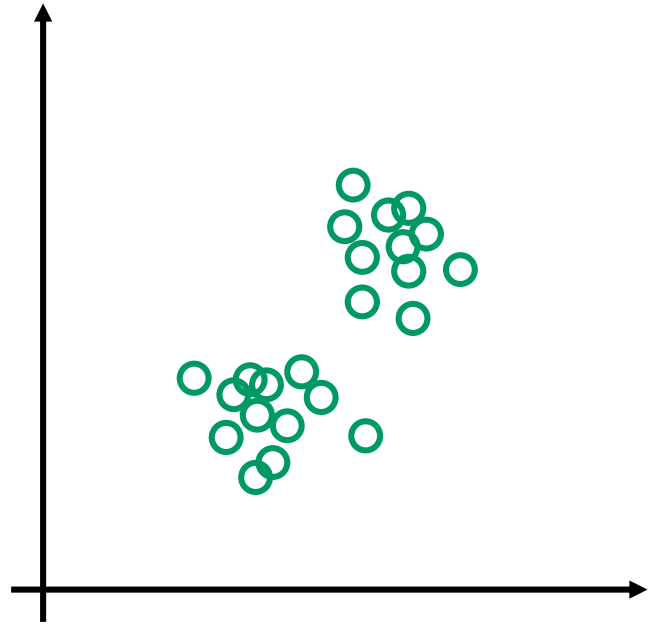
Every $\boldsymbol{x_n}$ comes with corresponding $\boldsymbol{y_n}$.
$\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_m}, y_m)\}$

Assume that we do not have the $\boldsymbol{y_n}$ for every $\boldsymbol{x_n}$. $\boldsymbol{\mathcal{D}} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}\}$



Supervised classification

(Ground Truth) Labels: Good; Bad.



Unsupervised clustering

(Learned) Latent concepts: Dark; Light.

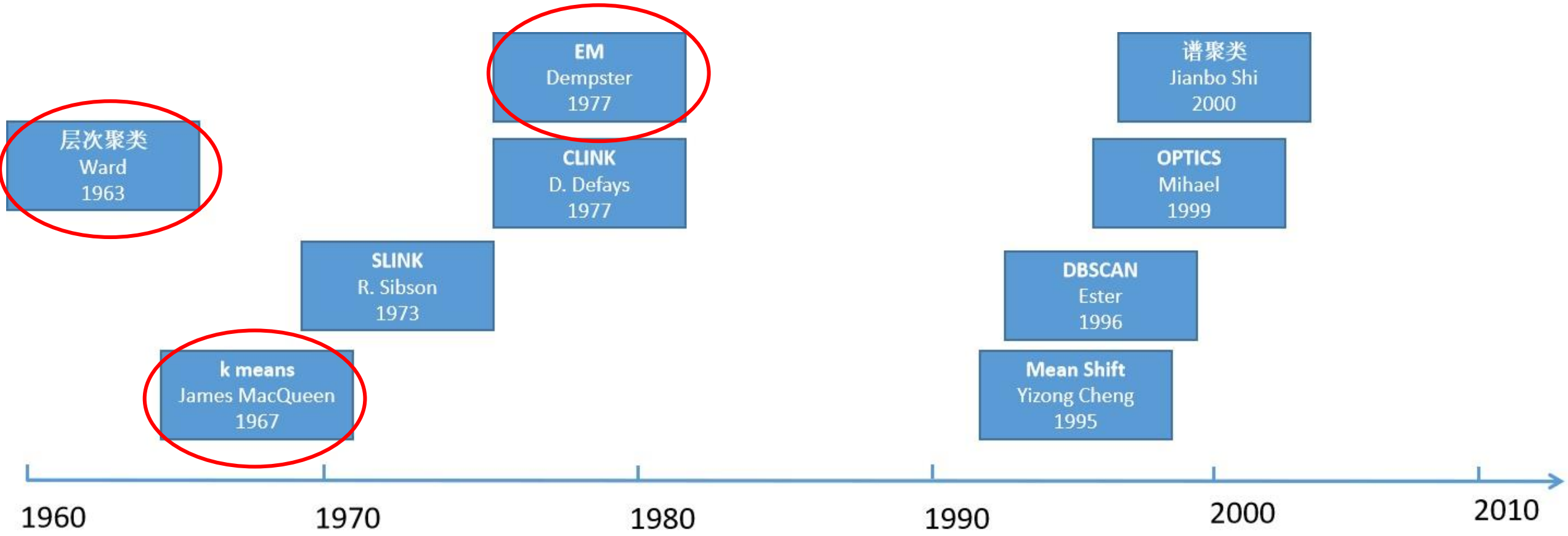
Unsupervised Learning

For an unsupervised learning problem, we do not focus on prediction of any particular thing, but rather try to find interesting aspects of the data $\mathcal{D} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m}\}$.

Examples

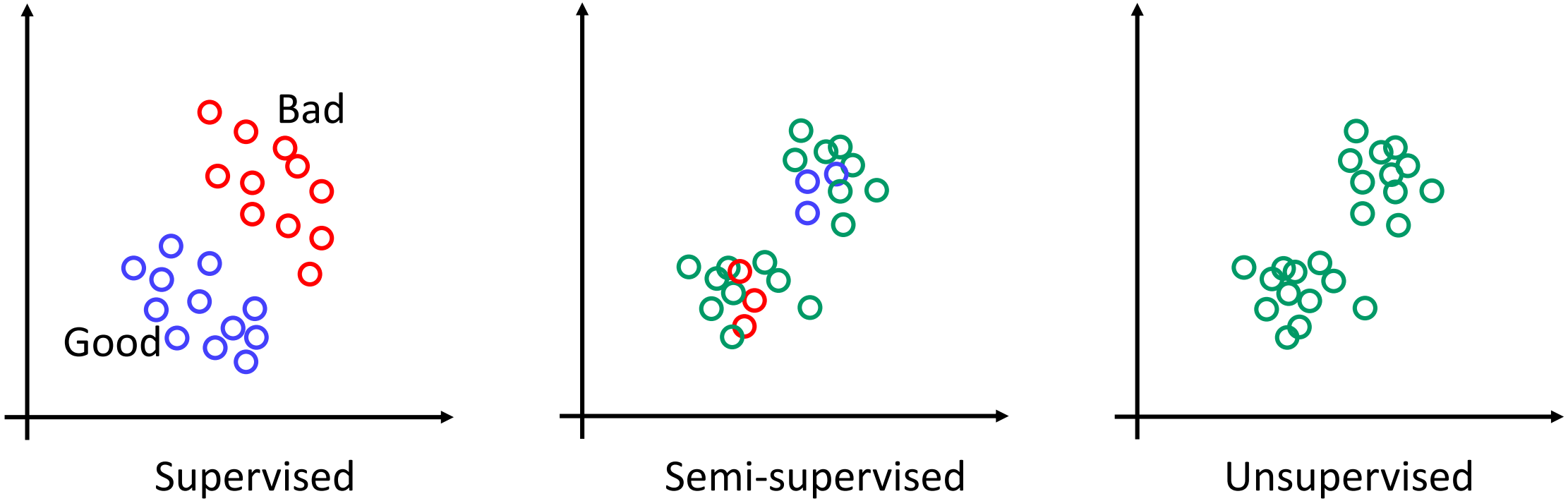
- We may find clusters of *patients* with similar *symptoms* (diseases).
- We may find clusters of *images* with the similar *visual characteristics*.
- We may find clusters of *people* with the similar *interests*.
- We may find clusters of *articles* with the similar *topics*.

Unsupervised Learning



Semi-supervised Learning (with some y_n)

$$\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_k}, y_k), \boldsymbol{x_{k+1}}, \boldsymbol{x_{k+2}}, \ldots, \boldsymbol{x_m}\}.$$



Semi-supervised Learning (with some y_n)

$$\mathcal{D} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_k}, y_k), \boldsymbol{x_{k+1}}, \boldsymbol{x_{k+2}}, \ldots, \boldsymbol{x_m}\}.$$

Semi-supervised learning: leverage unlabeled data to avoid "expensive" labeling.

Examples

- face images with a few labeled $\Rightarrow$ face identifier (Facebook)
- medicine data with a few labeled $\Rightarrow$ medicine effect predictor

Transductive vs Inductive Learning

- Transductive:
  - Unlabeled test data is available during training.
  - Produce label only for the available unlabeled data. The output of the method is not a classifier.
- Inductive (traditional supervised learning):
  - Unlabeled test data is not available during training.
  - Not only produce label for unlabeled data, but also produce a classifier (function f).

Transductive vs Inductive Learning

Inductive

$$\min_{\mathbf{w}_s,\boldsymbol{\alpha}} \frac{1}{2N}\left\|\mathbf{y}-\sum_{s=1}^{S}\alpha_s\mathbf{X}_s\mathbf{w}_s\right\|^2+\frac{\mu}{4N}\sum_{s=1}^{S}\sum_{s'\neq s}\left\|\mathbf{X}_s\mathbf{w}_s-\mathbf{X}_{s'}\mathbf{w}_{s'}\right\|^2$$

$$+\frac{\lambda}{2}\sum_{s=1}^{S}\left\|\mathbf{w}_s\right\|^2+\frac{\beta}{2}\left\|\boldsymbol{\alpha}\right\|^2, \qquad (4)$$

Transductive

$$\min_{\mathbf{f},\mathbf{L}(\mathbf{S_0})}\sum_{i=1}^{N}(y_i-f_i)^2+\mu\mathbf{f}^T\mathbf{L}(\mathbf{S_0})\mathbf{f}$$

$$+\lambda\sum_{k=1}^{K}\left\|\frac{1}{tr(\mathbf{L}(\mathbf{S_0}))}\mathbf{L}(\mathbf{S_0})-\frac{1}{tr(\mathbf{L}(\mathbf{S_k}))}\mathbf{L}(\mathbf{S_k})\right\|_F^2,$$

Predicted results: $\mathbf{f}=\{f_1,f_2,\cdots,f_N,f_{N+1},f_{N+2},\cdots,f_{N+M}\}^T\in\mathbb{R}^{N+M}$

Ground truth labels: $\mathbf{y}=\{y_1,y_2,\cdots,y_N\}^T\in\mathbb{R}^N$

Reinforcement Learning

- Thus far: Learning from examples
- Missing: actions/decisions (Learn from interaction)

Teach Your Dog How to Sit: We say 'Sit Down'

- Action1: The dog pees on the ground.
  - We cannot easily show the dog that y_n =sit when x_n = 'sit down', but we can punish this action.
- Action2: The dog sits down.
  - Cannot easily show the dog that y_n = sit when x_n ='sit down', but we can reward this action.

Reinforcement Learning

Scenario 1
(Punish)

Scenario 2
(Reward)

What Makes Reinforcement Learning Different?

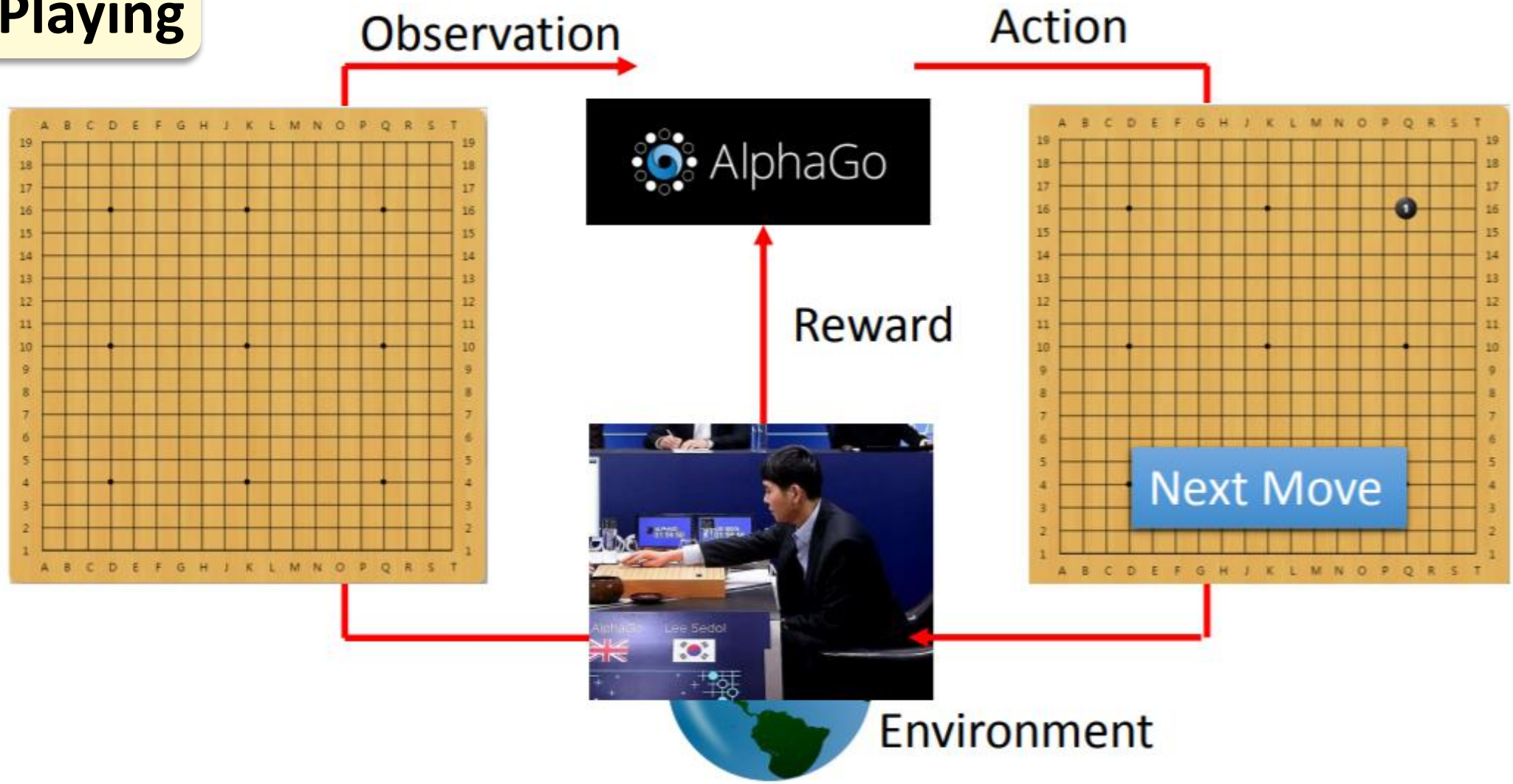
- There is no supervisor, only a reward signal.
- Time really matters (sequential).
- Agent's actions affect the subsequent data it receives.
- Reinforcement learning is based on the reward hypothesis.
- All goals can be described by the maximization of expected cumulative reward.
- Agent goal: maximize cumulative reward.
  - Select actions to maximize the expected cumulative reward.

Reinforcement Learning

- Fly stunt manoeuvre in a helicopter
  - + reward for following desired trajectory
  - - reward for crashing
- Manage a investment portfolio
  - + reward for each dollar in bank
- Make a humanoid robot walk
  - + reward for forward motion
  - - reward for falling over
- Recycling robot
  - + reward for finding cans
  - - reward for running out of battery

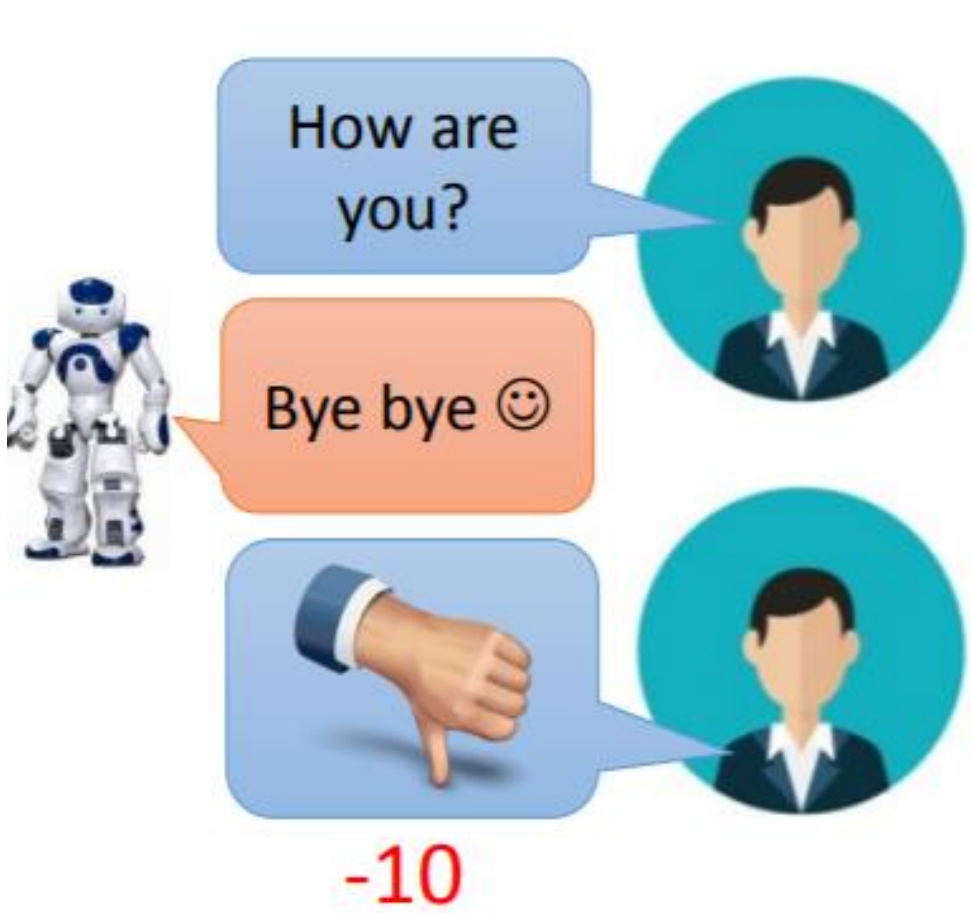
Reinforcement Learning

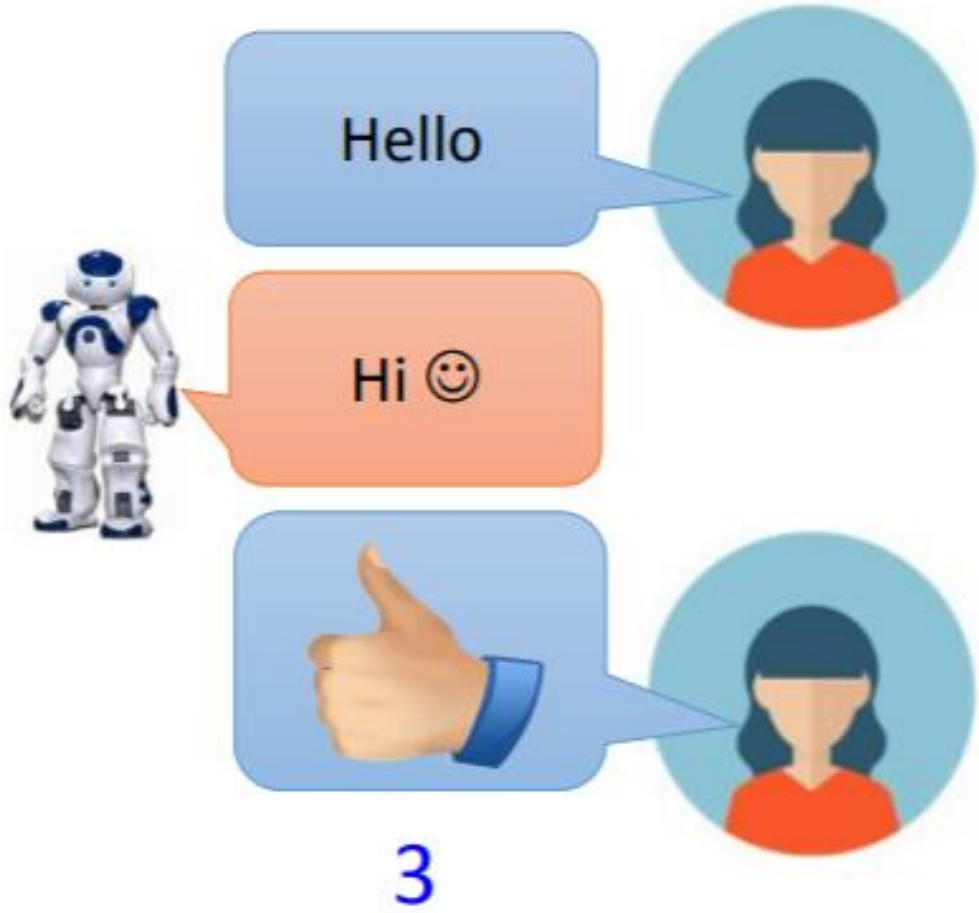
Go Playing



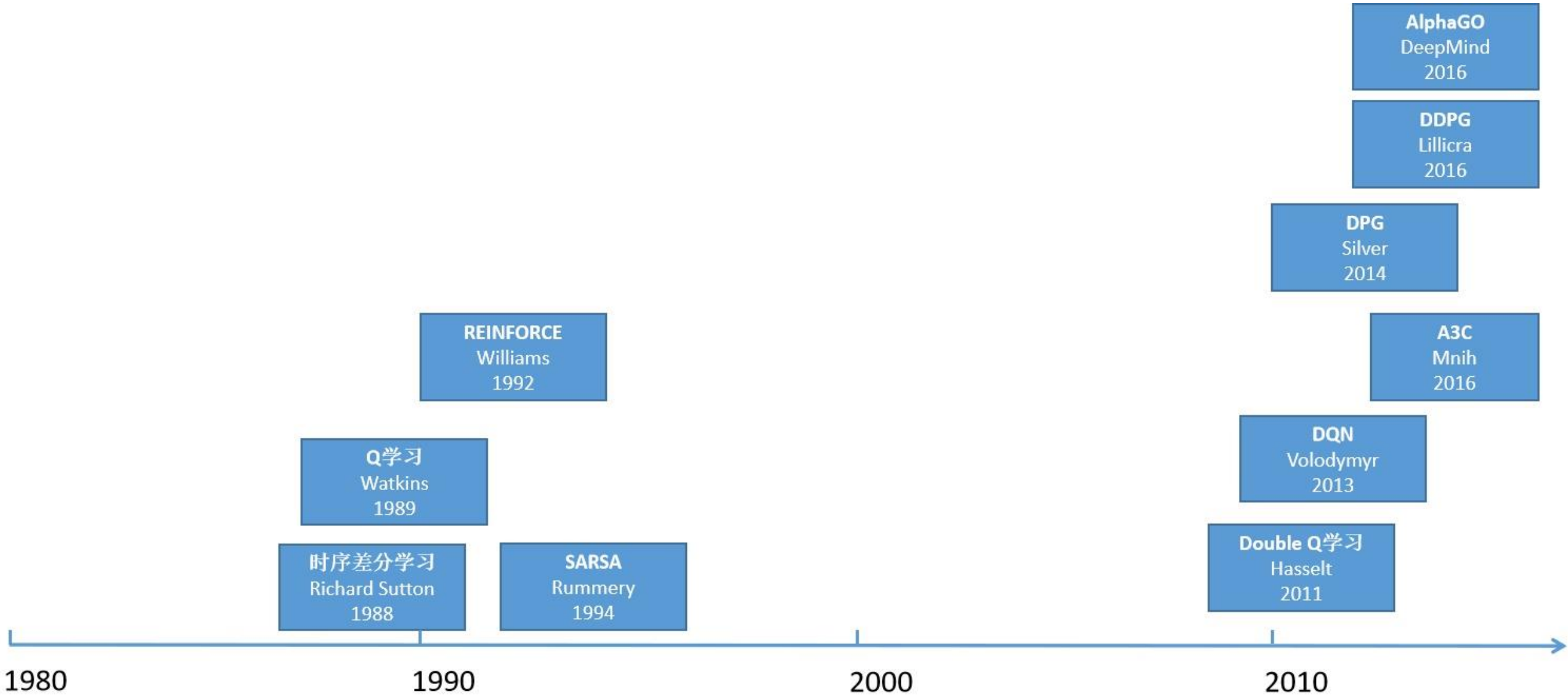
Reinforcement Learning

Chat-bot





Reinforcement Learning



Mini Summary

- **Supervised learning:**
  - using ***X_train*** and ***Y_train***, learn a general classifier to label any point.
- **Semi-supervised learning:**
  - using ***X_train*** and ***Y_train***, and ***X_unlabeled*** learn a general classifier to label any point.
- **Transductive learning:**
  - using ***X_train*** and ***Y_train*** and ***X_unlabeled***, infer labels only for that set of ***X_unlabeled***.
  - If a new point (sample) is given, then ?

Apply the algorithm from the beginning on all data.

Exercise

What is this learning problem?

To build a tree recognition system, a company decides to gather one million of pictures on the Internet. Then, it asks each of the 10 company members to view 100 pictures and record whether each picture contains a tree. The pictures and records are then fed to a learning algorithm to build the system. What type of learning problem does the algorithm need to solve?

1. supervised
2. unsupervised
3. semi-supervised
4. reinforcement

Categories of Machine Learning

- Learning with different output space Y
- Learning with different data label y_n
- Learning with different protocol f (x_n, y_n)
- Learning with different input space X

Batch Learning

Batch (offline) learning: learn from *all known* data (a very common protocol).

Examples

- Batch of (email, spam yes/no) ⇒ spam filter
- Batch of (patient info, cancer yes/no) ⇒ cancer prediction
- Batch of (video, location label) ⇒ video location prediction
- Batch of customer data ⇒ group of customers (interest)
- Batch of documents ⇒ group of documents (topic)

Batch Learning

Batch (offline) learning: learn from *all known* data (a very common protocol).

We typically assume that:

- The learning algorithm $\mathcal{A}$ is deterministic,
- $\mathcal{A}$ does not depend on the ordering of the points in the training set.

Online Learning

Online learning: learn from the *sequential* data.

Examples

- Online Spam filter ($\mathcal{A}$ is the learning algorithm)
  - Observe an email $\boldsymbol{x}_t$;
  - Predict spam label with current $g_t = \mathcal{A}(\boldsymbol{x}_t)$
  - Receive 'ground truth' y_t from the user, and then update g_t with $(\boldsymbol{x}_t, y_t)$, i.e., $g_{t+1} = \mathcal{A}(g_t, (\boldsymbol{x}_t, y_t))$

Online Learning

Online learning: learn from the *sequential* data.

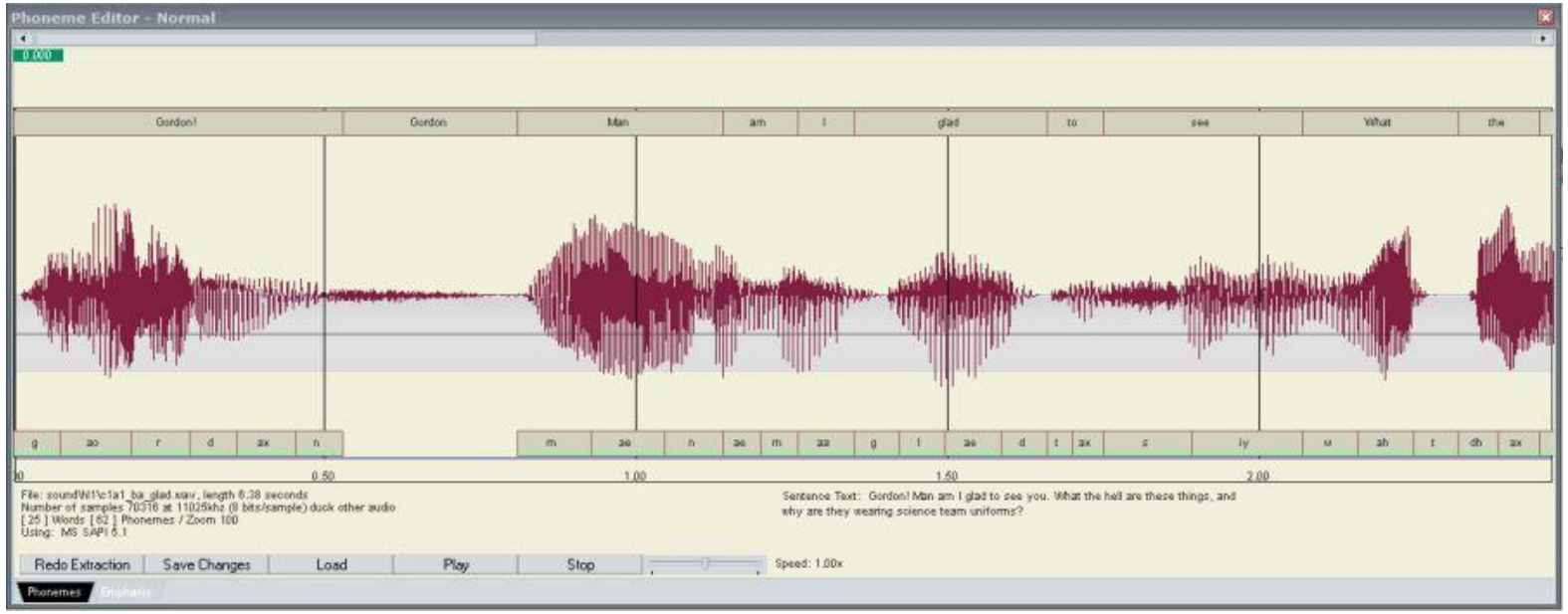
Other Examples

- A recommendation system (e.g., IMDB) that is constantly learning from the ratings given by users and making appropriate recommendations to users.
- The Facebook production ranking systems used in ads ranking and newsfeed ranking use a combination of online learning and offline learning to provide the best results.

Active Learning

- Active learning is well-motivated in modern machine learning problems where data may be abundant but labels are scarce or expensive to obtain.

Speech Recognition.
(Phonemes Annotation).

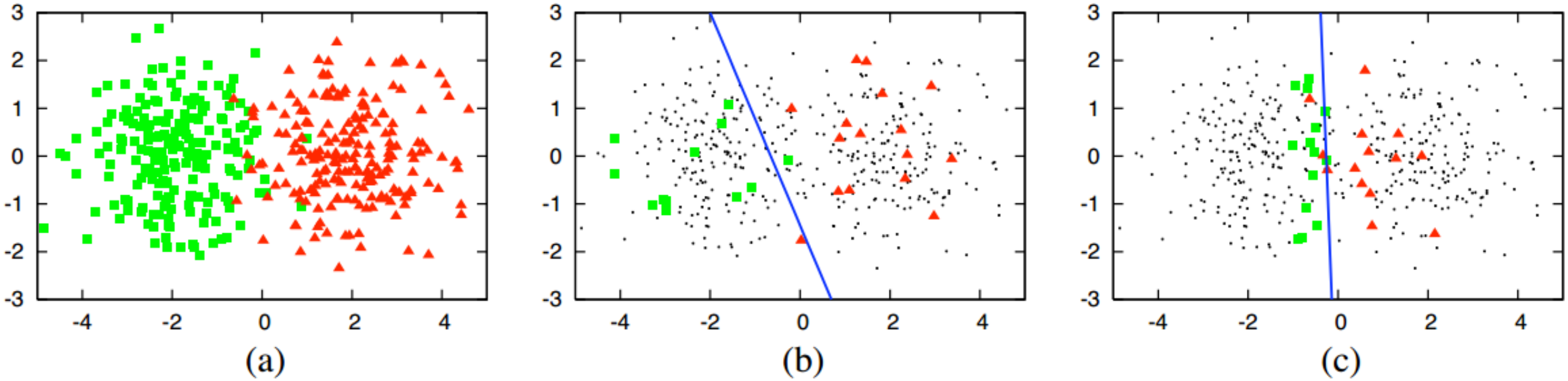


Active Learning

- Active learning is well-motivated in many modern machine learning problems where data may be abundant but labels are scarce or expensive to obtain.

- Active learning (sequentially) queries the y_n of the *strategically chosen* x_n, which is to be labeled by an *oracle* (e.g., a human annotator).
- *Key hypothesis:* if the learning algorithm is allowed to choose the data from which it learns, it will perform better with less training.

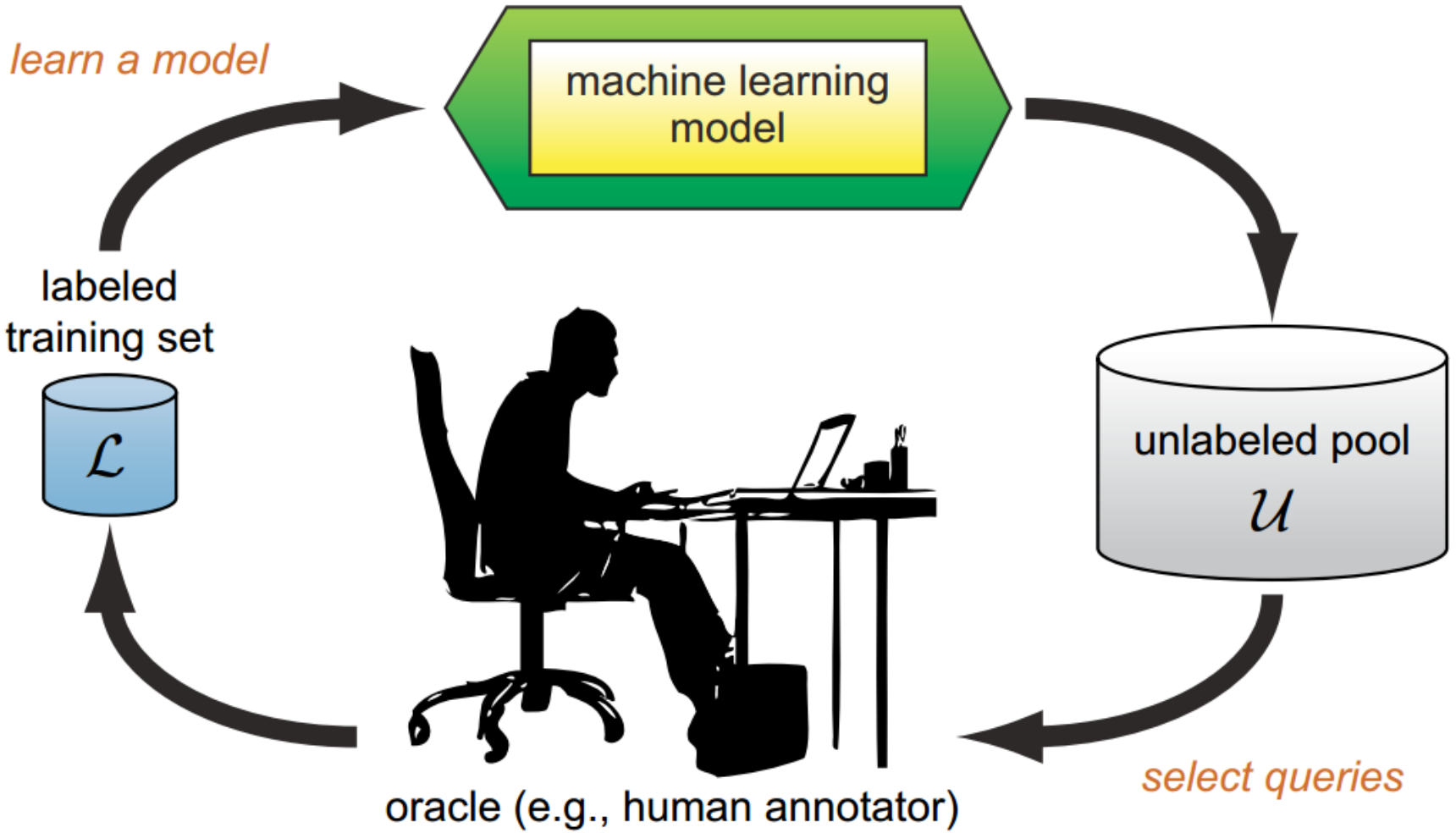
Active Learning



An illustrative example of pool-based active learning. (a) A toy data set of 400 instances. (b) A logistic regression model trained with 30 labeled instances using *random sampling*. (c) A logistic regression model trained with 30 *actively* queried instances using *uncertainty sampling*.

Uncertainty sampling: label those points for which the current model is least certain as to what the correct output should be.

Active Learning



The pool-based active learning cycle .

Categories of Machine Learning

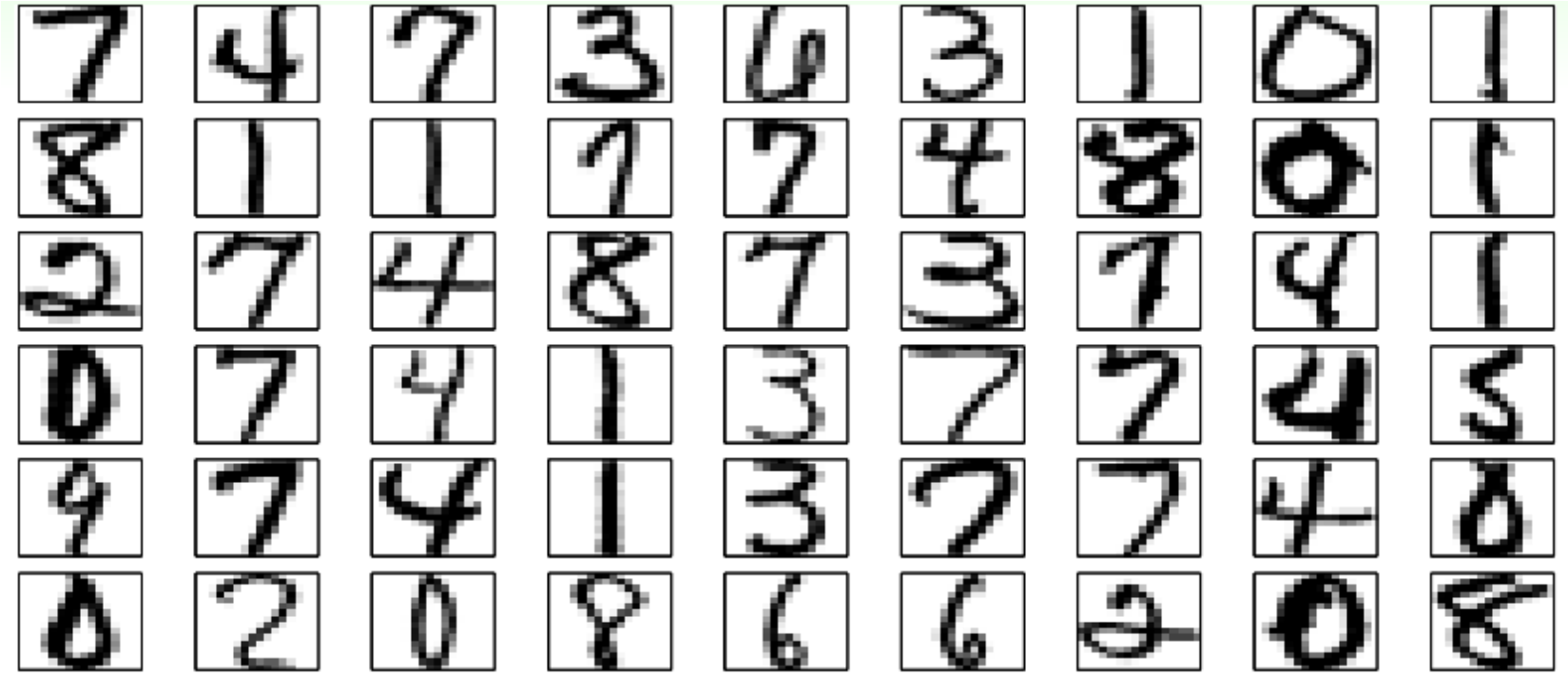
- Learning with different output space Y
- Learning with different data label y_n
- Learning with different protocol f (x_n, y_n)
- Learning with different input space X

Concrete Features

- (size, mass) **for coin classification**
- (MFCC, Zero Crossing Rate, etc.) **for speech recognition**
- Customer info (gender, occupation, etc.) **for credit approval**
- Patient info (height, weight, etc.) **for cancer diagnosis**
- **Often including "human intelligence" on the learning task**

Concrete features: the "easy" ones for ML.

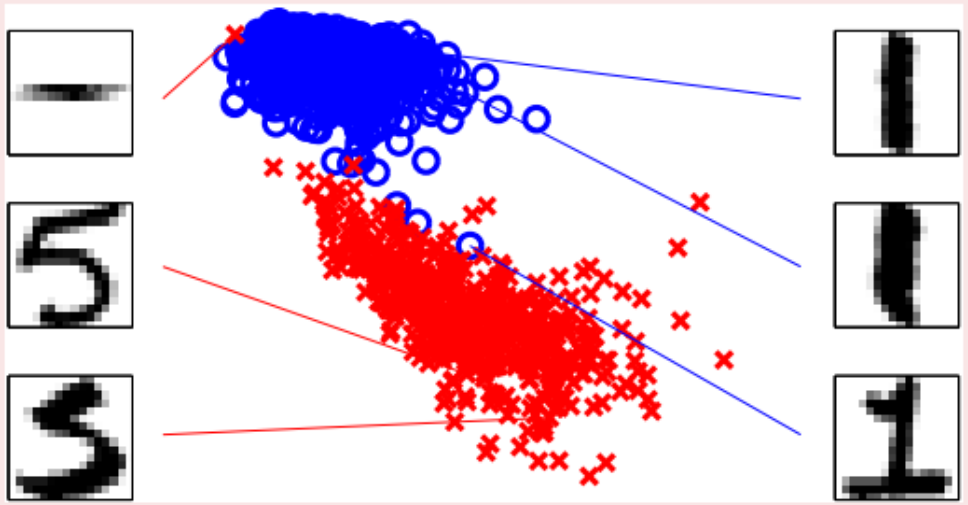
Raw Features



- Digit recognition problem: features ⇒ {0,1,2,…,9}.
- (Supervised multiclass classification problem).

Raw Features

by Concrete Features



$\mathbf{x}$ =(symmetry, density)

by Raw Features

- 16 by 16 gray image $\mathbf{x} \equiv (0, 0, 0.9, 0.6, \cdots) \in \mathbb{R}^{256}$
- '**simple** physical meaning'; thus more difficult for ML than concrete features

- Raw features (image pixels, speech signal, etc.): often need human or machines (e.g., neural networks) to convert to concrete ones.

Abstract Features

Rating Prediction Problem (KDDCup 2011)

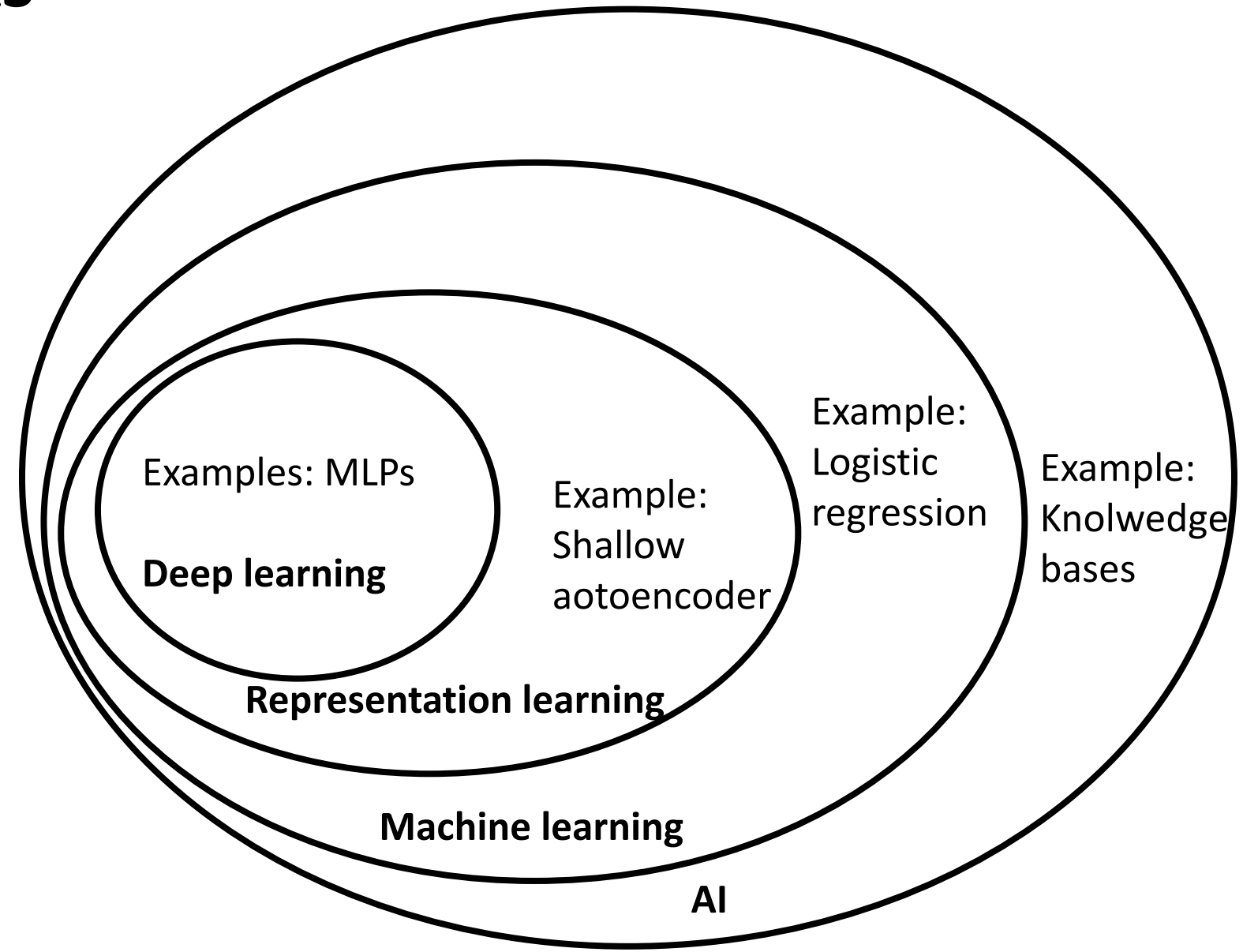
- given previous (userid, itemid, rating) tuples, predict the rating that some userid would give to itemid?
- a regression problem with $\mathcal{Y} \subseteq \mathbb{R}$ as rating and $\mathcal{X} \subseteq \mathbb{N} \times \mathbb{N}$ **as (userid, itemid)**
- '**no** physical meaning'; thus even more difficult for ML

- Abstract Features also need 'feature conversion/construction' by human or machines.

Types of Machine Learning

- Learning with different output space Y
  - [classification], [regression]
- Learning with different data label y_n
  - [supervised], un/semi-supervised, transductive, reinforcement
- Learning with different protocol f (x_n, y_n)
  - [batch], online, active
- Learning with different input space X
  - [concrete, raw], abstract

Remarks


Examples: MLPs
Deep learning
Example:
Shallow
aotoencoder
Representation learning
Example:
Logistic
regression
Machine learning
Example:
Knolwedge
bases
AI