Xiang Deng
Professor
Computer Science Department
Harbin Institute of Technology (Shenzhen)

Research Interests: Embodied AI, Multimodal Large Language Models

Course: Advanced Machine Learning

Instructors: Xianming Liu (csxm@hit.edu.cn)，Xiang Deng (dengxiang@hit.edu.cn)

Grading Policy:

The course will be graded based on participation (10%), project (40%), and final exam (50%).

Lesson1: Introduction to Machine Learning

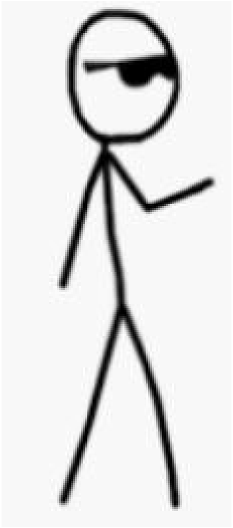
Outline

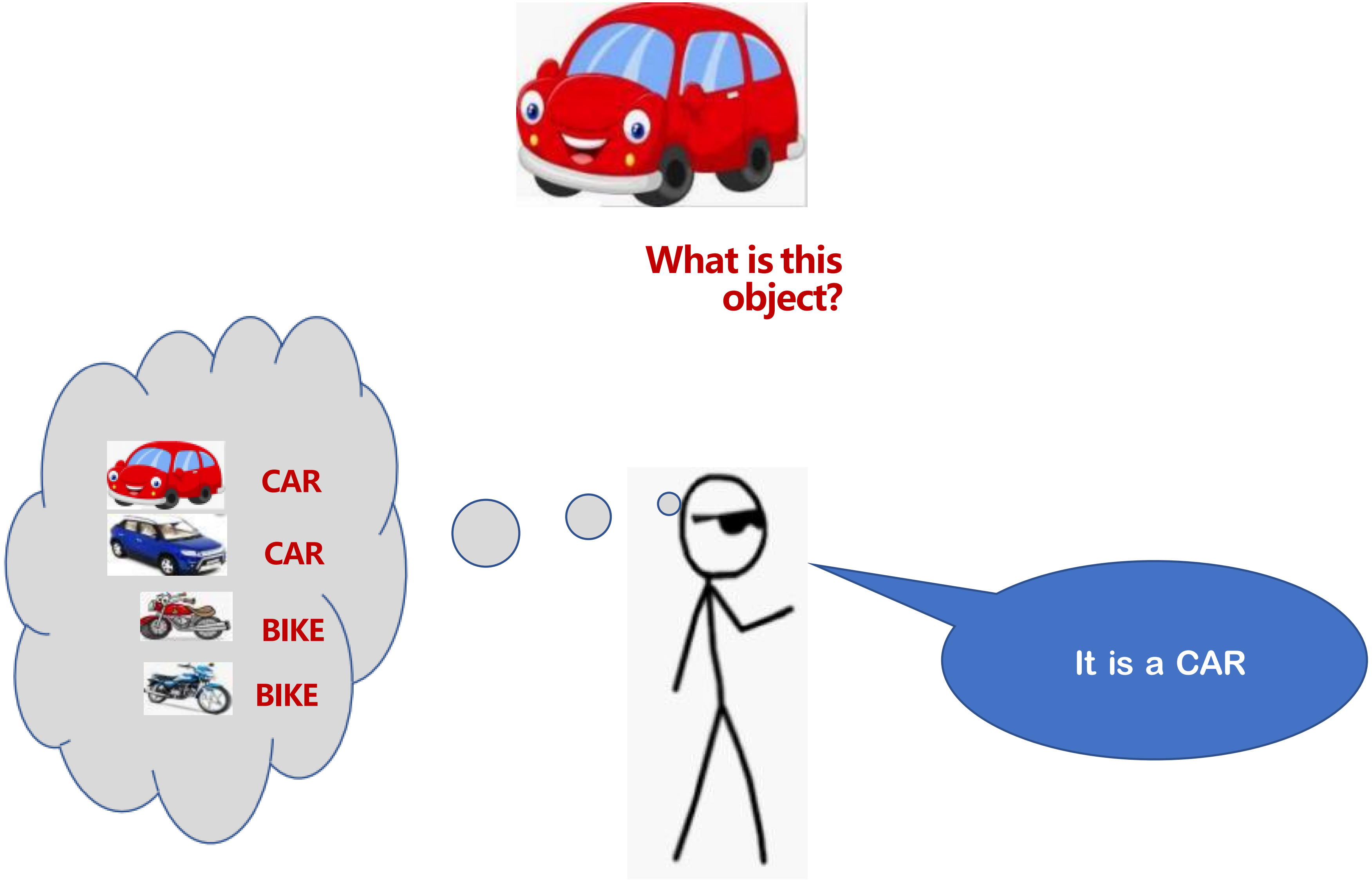
- Introduction
- Terminology
- Hypothesis Space
- Inductive Bias
- Brief History
- Application Status
- Further Reading

What is Machine Learning?
(Layman’ s term)

Human can learn from past experience
and make decision of its own

What is this object?




What is this object?
CAR
CAR
BIKE
BIKE
It is a CAR

Let us ask the same question to him

What is this object?

Let us ask the same question to him

What is this
object?

[But, he is a human being. He can observe and learn]

Let us make him learn

show him

Let us make him learn

show him

CAR

CAR

BIKE

BIKE

Let us ask the same question now

What is this
object?

CAR

CAR

BIKE

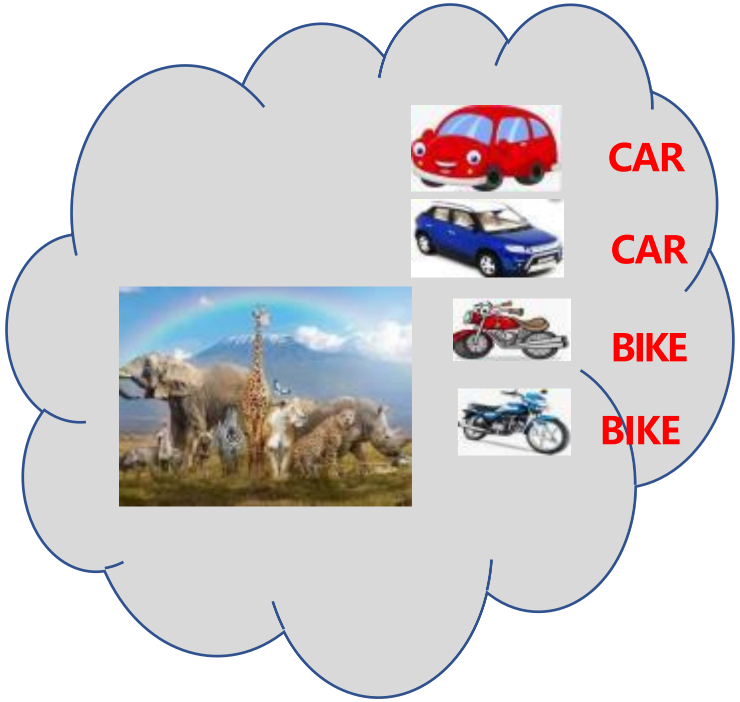
BIKE

Past experience

Let us ask the same question now

CAR

What is this object?



What about a Machine ?

Machines follow instructions

[It can not take decision of its own]

What about a Machine ?

We can ask a machine

- To perform an arithmetic operations such as
  - Addition
  - Multiplication
  - Division

Machines follow instructions

What about a Machine ?

- Comparison
- Print
- Plotting a chart

Machines follow instructions

What is Machine Learning?

[We want a machine to act like a human]

What is Machine Learning?

[to identify this object.]

What is Machine Learning?

Price in 2026?

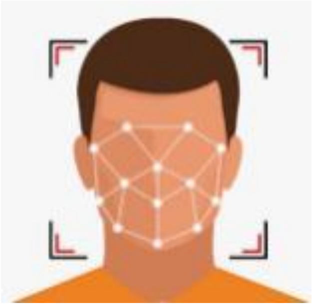
[predict the price in future]

What is Machine Learning?

I ~~made~~ met him yesterday

[Natural Language understand, and correct grammar]

What is Machine Learning?



recognize face

[Recognize Faces]

What is Machine Learning?

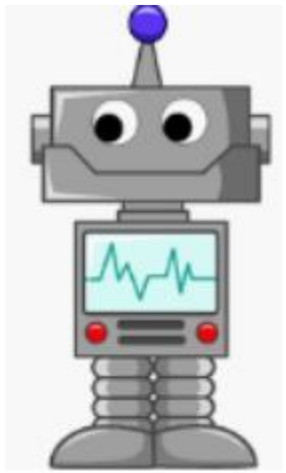
[What do we do?

Just like, what we did to human,

we need to provide experience to the machine.

]

What is Machine Learning?



+

Dataset

[
This what we called as Data or Training dataset

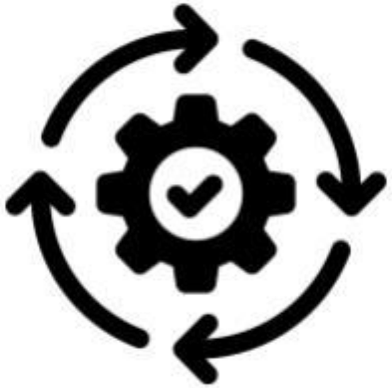
So, we first need to provide training dataset to the machine
]

What is Machine Learning?

+

Datase
t

+



[Then, devise algorithms and execute programs on the data

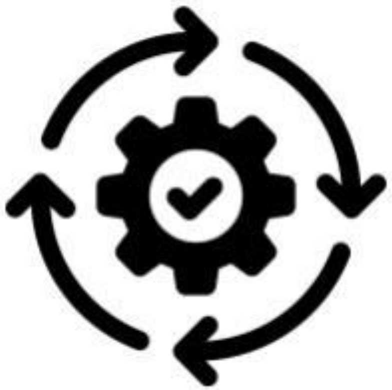
With respect to the underlying target tasks]

What is Machine Learning?

+

Datase
t

+



+

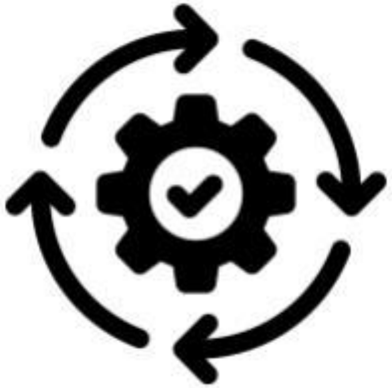
[Then, using the programs, Identify required rules]

What is Machine Learning?

+

Datase
t

+



+

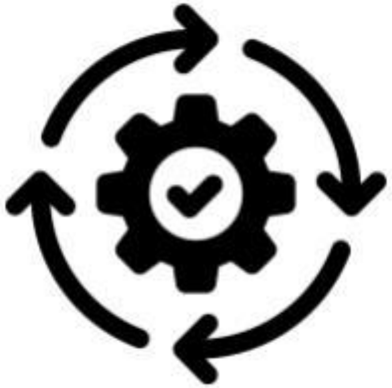
[extract required patterns]

What is Machine Learning?

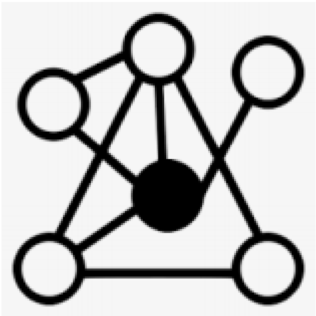
+

Dataset

+



+



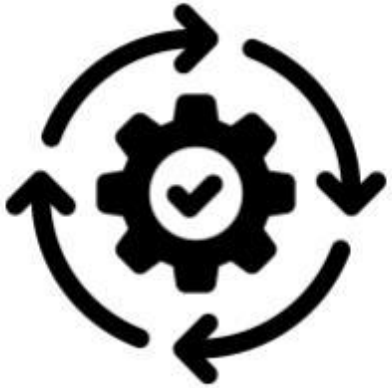
[Identify relations]

What is Machine Learning?

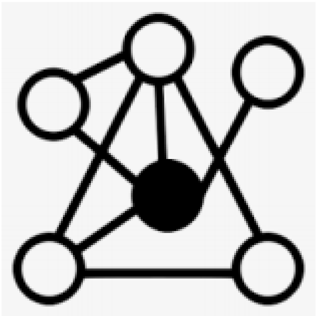
+

Dataset

+



+



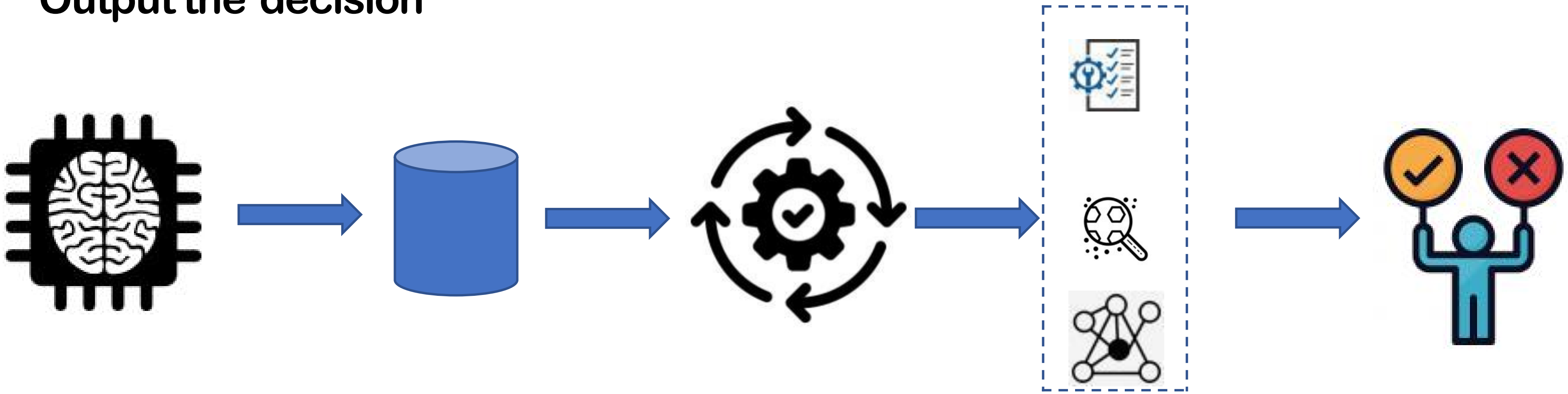
=

[So that machine can derive inferences from the data]

In summary, what is machine learning?

Given a machine learning problem

- Identify and create the appropriate dataset
- Perform computation to learn
    - Required rules, pattern and relations
- Output the decision



What is Supervised Learning?

CAR

CAR

\+

BIKE

BIKE

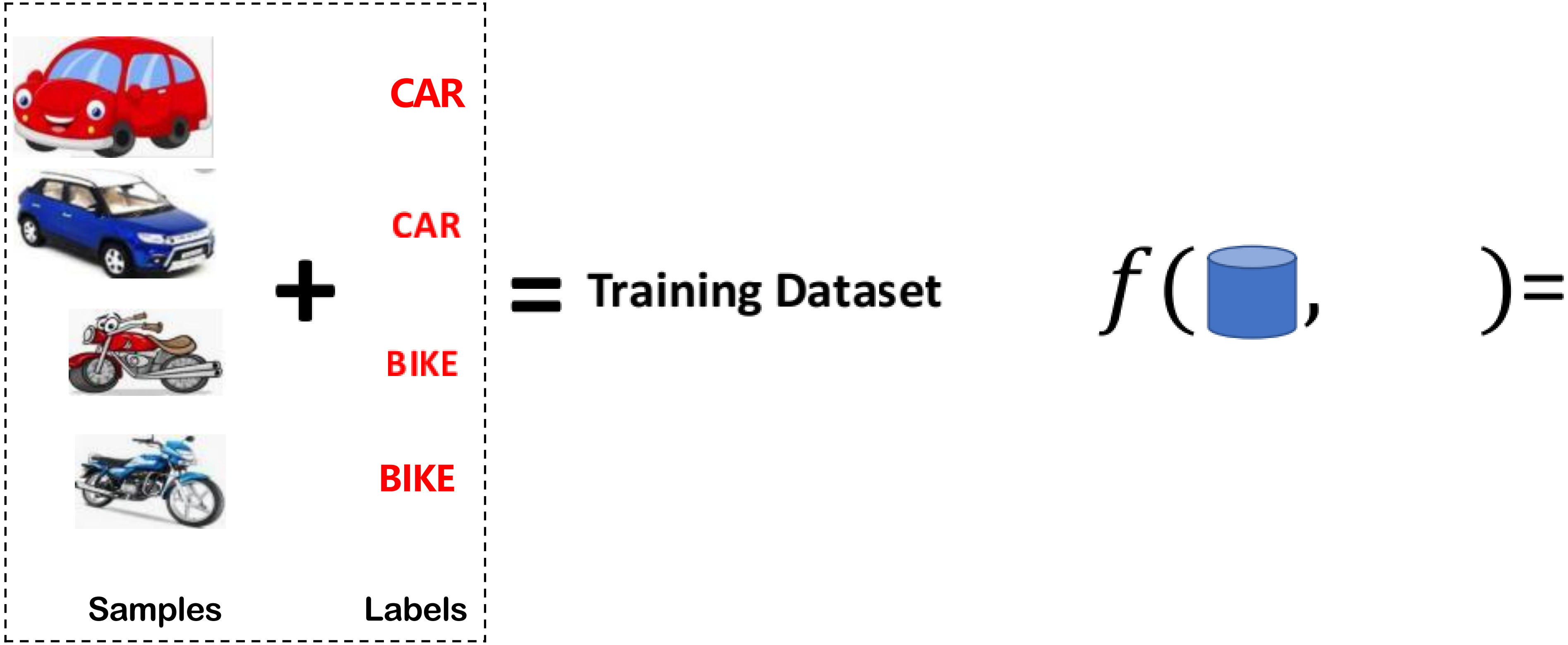
Samples

Labels

= Training Dataset

[In supervised learning, we need some thing called a Labelled Training Dataset]

What is Supervised Learning?



[Given a labelled dataset, the task is to devise a function which takes the dataset, and a new sample, and produces an output value.]

What is Supervised Learning?

CAR

CAR

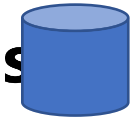
+

BIKE

BIKE

Samples Labels

= Training Datas



Test data

)=

[Given a labelled dataset, the task is to devise a function which takes the dataset, and a new sample, and produces an output value.]

What is Supervised Learning?

CAR

CAR

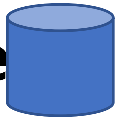
BIKE

BIKE

Samples

Labels

+

= Training Datase   ,)= CAR

Test data

[Given a labelled dataset, the task is to devise a function which takes the dataset, and a new sample, and produces an output value.]

What is Supervised Learning?

CAR

CAR

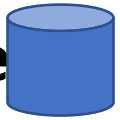
BIKE

BIKE

Samples

Labels

+

= Training Datase   ,)= CAR

Classification

Test data

[If the possible output values of the function are predefined and discrete/categorical, it is called Classification

What is Supervised Learning?

CAR

CAR

BIKE

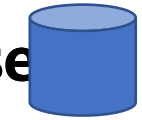
BIKE

Samples

Labels

+

= Training Datase



Classification

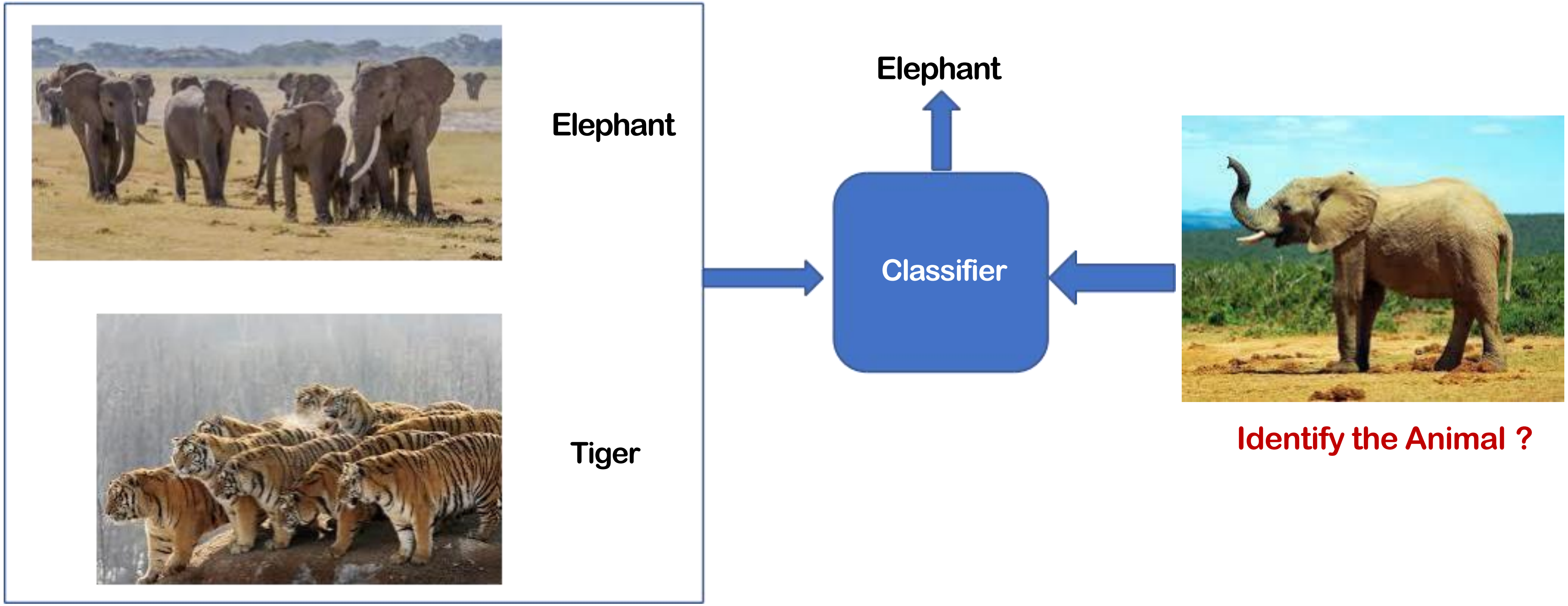
Test data

,)=

CAR

[Predefined classes means, it will produce output only from the labels defined in the dataset. For example, even if we input a bus, it will produce either CAR or BIKE]

Classifier

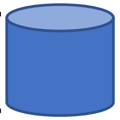


Dataset

Regression

Dataset

Regression

f( , )= 20500.50

[If the possible output values of the function are continuous real values, then it is called Regression]

[

The classification and Regression problems are supervised, because the decision depends on the characteristics of the ground truth labels or values present in the dataset, which we define as experience

]

What is Unsupervised Learning

~~CAR~~

~~CAR~~

~~BIKE~~

~~BIKE~~

Dataset

[In the unsupervised learning, we do not need to know the labels or Ground truth values]

What is Unsupervised Learning

Dataset

Clustering

[The task is to identify the patterns like group the similar objects together]

What is Unsupervised Learning

Association Rules Mining

Dataset

[Association rules like]

More Example Unsupervised Learning

Datase
t

More Example Unsupervised Learning

Datase
t

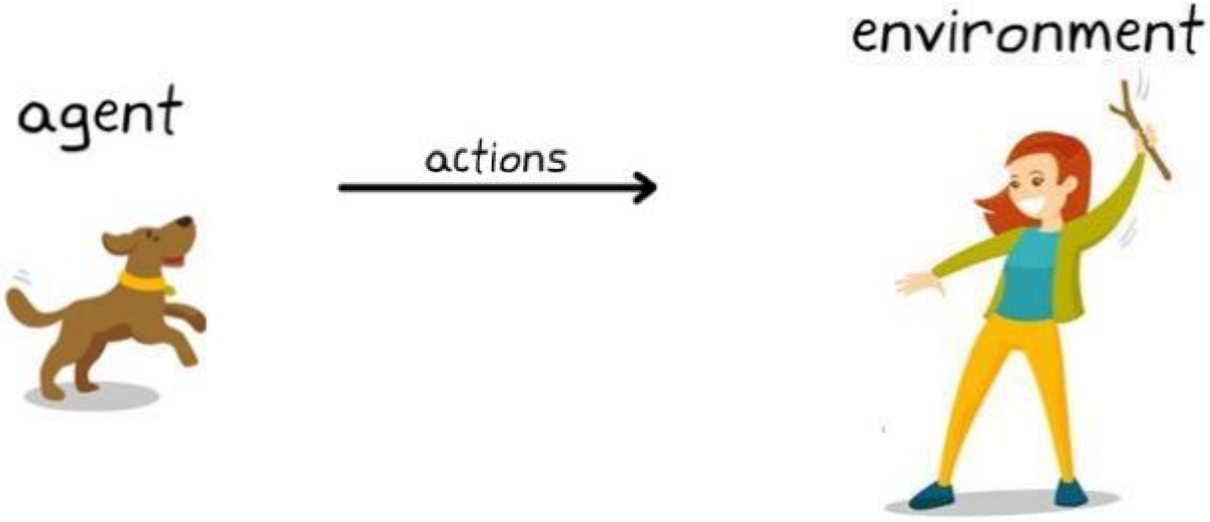
More Example Unsupervised Learning

What is Reinforcement Learning

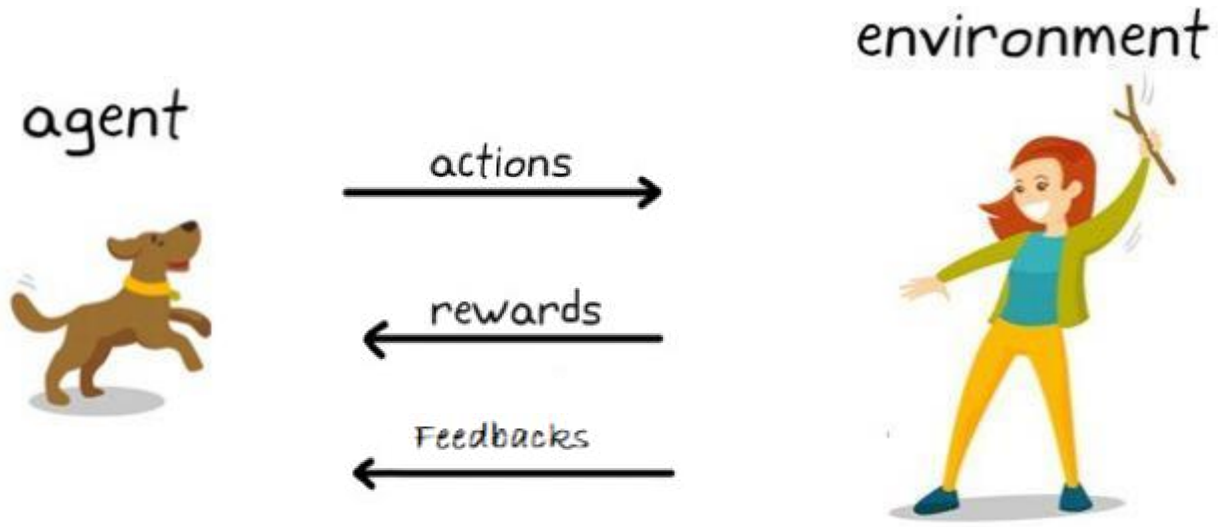
[It is also known as learning from trials and errors]

What is Reinforcement Learning

What is Reinforcement Learning



What is Reinforcement Learning



Another Example

Agent

Task

Environment

Reinforcement Learning

Punishment

Reinforcement Learning

Rewar
d

Reinforcement Learning

Rewar
d

Baby Learn from the Trials and Errors

Reinforcement Learning

Outline

- Introduction
- Terminology
- Hypothesis Space
- Inductive Bias
- Brief History
- Application Status
- Further Reading

Terminology-Data

Watermelon Classification

feature | label

	ID	color	root	sound	ripe
Training set	1	green	curly	muffled	true
	2	dark	curly	muffled	true
	3	green	straight	crisp	false
	4	dark	slightly curly	dull	false
Testing set	1	green	curly	dull	?

Terminology-Task

- Labeled or unlabeled information
  - Supervised learning: classification, regression
  - Unsupervised learning: clustering
  - Semi-supervised learning: a combination of the above two

Terminology-generalization ability

The objective of machine learning is to learn models that can work well on the "new samples", rather than the training examples. The ability to work on the new samples is called the *generalization* ability.

We generally assume that all samples in a sample space follow a distribution $\mathcal{D}$, and all samples are independently sampled from this distribution, that is, *independent and identically distributed* (*i.i.d.*). Generally speaking, the more samples we have, the better-generalized model we can learn.

Outline

- Introduction
- Terminology
- Hypothesis Space
- Inductive Bias
- Brief History
- Application Status
- Further Reading

Hypothesis Space

ID	color	root	sound	ripe
1	green	curly	muffled	true
2	dark	curly	muffled	true
3	green	straight	crisp	false
4	dark	slightly curly	dull	false

$(\text{color}=?) \wedge (\text{root}=?) \wedge (\text{sound}=?) \leftrightarrow \text{ripe}$

Filtering out all hypotheses that are inconsistent with the training examples.

Hypothesis Space Size：3*4*4+1=49

Outline

- Introduction
- Terminology
- Hypothesis Space
- Inductive Bias
- Brief History
- Application Status
- Further Reading

Inductive Bias

ID	color	root	sound	ripe
1	green	curly	muffled	true
2	dark	curly	muffled	true
3	green	straight	crisp	false
4	dark	slightly curly	dull	false
1	green	curly	dull	?

All hypotheses are consistent with the training examples, whereas these hypotheses may make different prediction on the unseen watermelon (color = green) ∧ (root = curly) ∧ (sound = dull):

(color=*; root=curly; sound=*) — ripe

(color=*; root=*; sound=muffled) — unripe

(color=*; root=curly; sound=muffled) — unripe

(color = green, root= curly, sound = dull)

In this case, which model (or hypothesis) should we use?

Inductive Bias

The bias of a learning algorithm towards a particular class of hypotheses is called the *inductive bias*.

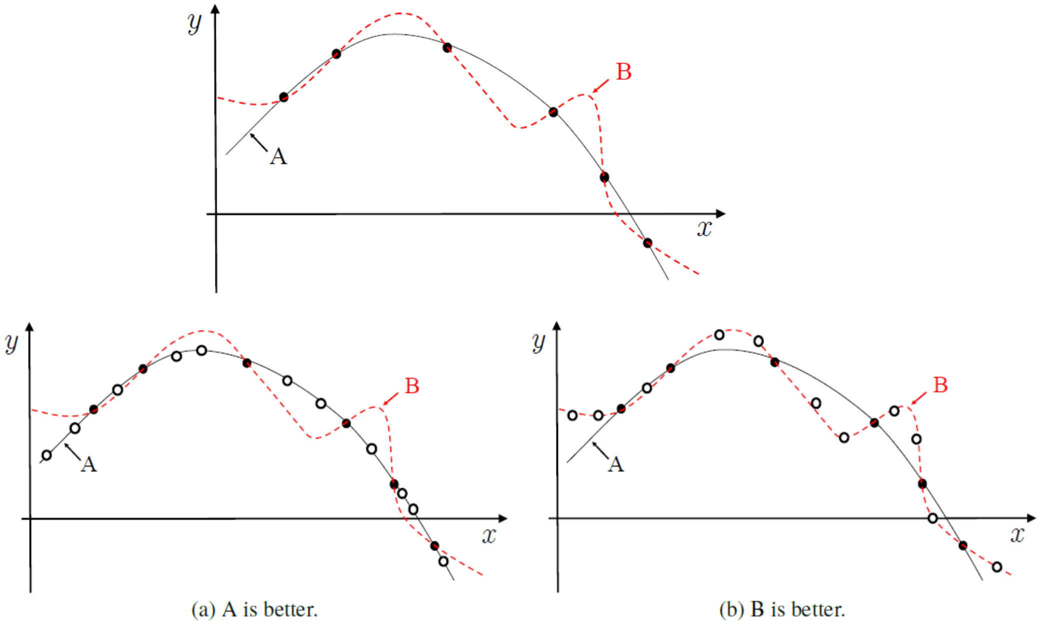


Fig. 1.4: There is no free lunch. (● are training samples; ○ are testing samples)

Inductive Bias

We can regard inductive bias as the heuristic or value philosophy of learning algorithms for search in potentially huge hypothesis spaces.

A fundamental and widely used principle for this question in natural science is the *Occam's razor* principle, which says that we should choose the simplest hypothesis when there is more than one hypothesis consistent with the observations.

In practice, whether this hypothesis matches the specific problem or not usually determines the performance of the model.

No Free Lunch

To simplify the discussion, let both the sample space $\mathcal{X}$ and the hypothesis space $\mathcal{H}$ be discrete. Let $P(h|X, \mathfrak{L}_a)$ denote the probability of getting the hypothesis *h* from the algorithm $\mathfrak{L}_a$ based on the training set *X*, and let *f* be the ground-truth target function that we wish to learn. Then the error on all samples except those in the training set is

$$E_{ote}(\mathfrak{L}_a|X, f) = \sum_h \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \, \mathbb{I}\left(h\left(\boldsymbol{x}\right) \neq f\left(\boldsymbol{x}\right)\right) P\left(h \mid X, \mathfrak{L}_a\right)$$

$\mathbb{I}(\cdot)$ is the indicator function that • returns 1 for true and 0 otherwise.

No Free Lunch

In binary classification problems, the target function could be any functions $\mathcal{X} \mapsto \{0,1\}$ with a function space of $\{0,1\}^{|\mathcal{X}|}$. Summing the errors of f with respect to uniform distribution gives：

$$
\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\ \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))\ P(h \mid X, \mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \\
&= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1 \ .
\end{aligned}
$$

The sum of errors is independent of the learning algorithm！

All learning algorithms are equally good considering all contexts.

Debating “which learning algorithm is better” is meaningless without considering the specific task

Outline

- Introduction
- Terminology
- Hypothesis Space
- Inductive Bias
- Brief History
- Application Status
- Further Reading

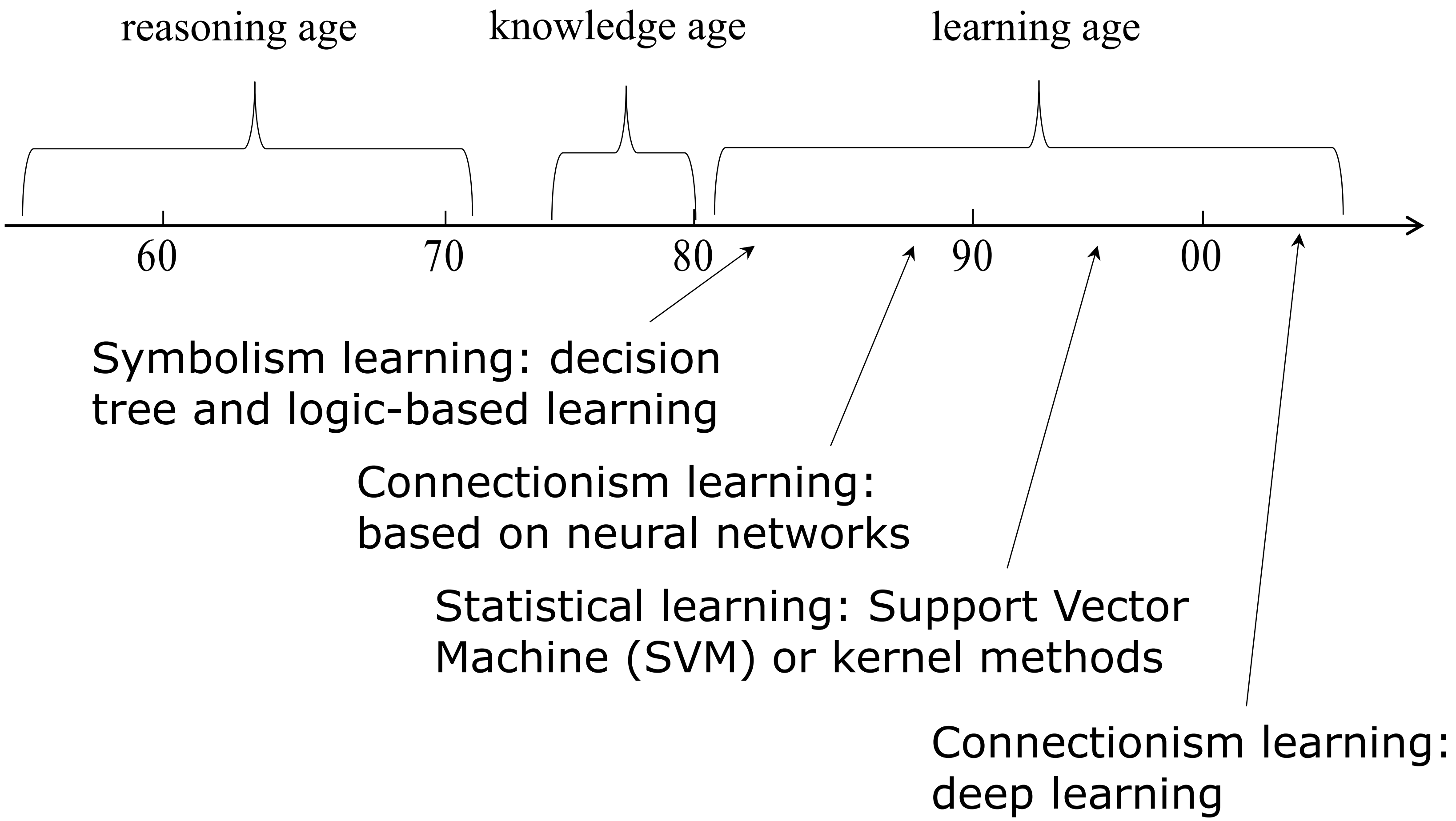
Brief History

- Reasoning age：
  - Seminal works in that period include the Logic Theorist program developed by A. Newell and H. Simon and later on the General Problem Solving program.
  - In 2006, Carnegie Mellon University founded the world’s first school of machine learning, which is directed by Professor T. Mitchell, one of the pioneers in machine learning research.
- Knowledge age：
  - A large number of expert systems with numerous successful applications are developed in a wide range of domains;
  - It is difficult to extract and summarize knowledge into a form that computers can learn.

Brief History

- Learning age：
  - Symbolism learning
    - Decision tree: relies on information theory to simulate the tree-based decision process of humans by minimizing the information entropy.
    - Logic-based learning: employs first-order logic to represent knowledge and induces data by updating and extending the logic expressions.
  - Connectionism learning
    - Neural networks
  - Statistical learning
    - Support Vector Machine (SVM) or kernel methods

Brief History



Outline

- Introduction
- Terminology
- Hypothesis Space
- Inductive Bias
- Brief History
- Application Status
- Further Reading

Application Status

- One of the most current computer science technologies：
  - In 2001, scientists from NASA-JPL published an article in the *Science* magazine pointed out that machine learning is playing an increasingly important role in supporting scientific research.
  - In 2003, DARPA started the PAL project, which puts machine learning to the level of national security.
  - In 2006, Carnegie Mellon University founded the world's first school of machine learning, which is directed by Professor T. Mitchell, one of the pioneers in machine learning research.
- Strongly influences our daily life：
  - Weather forecasting, energy exploration, environmental monitoring, search engines, autonomous vehicles, etc.

Application Status

- Affect the political life of human society: :
  - During the 2012 U.S. election, Obama's machine learning team analyzed various data such as social networks to prompt him for the next campaign action.

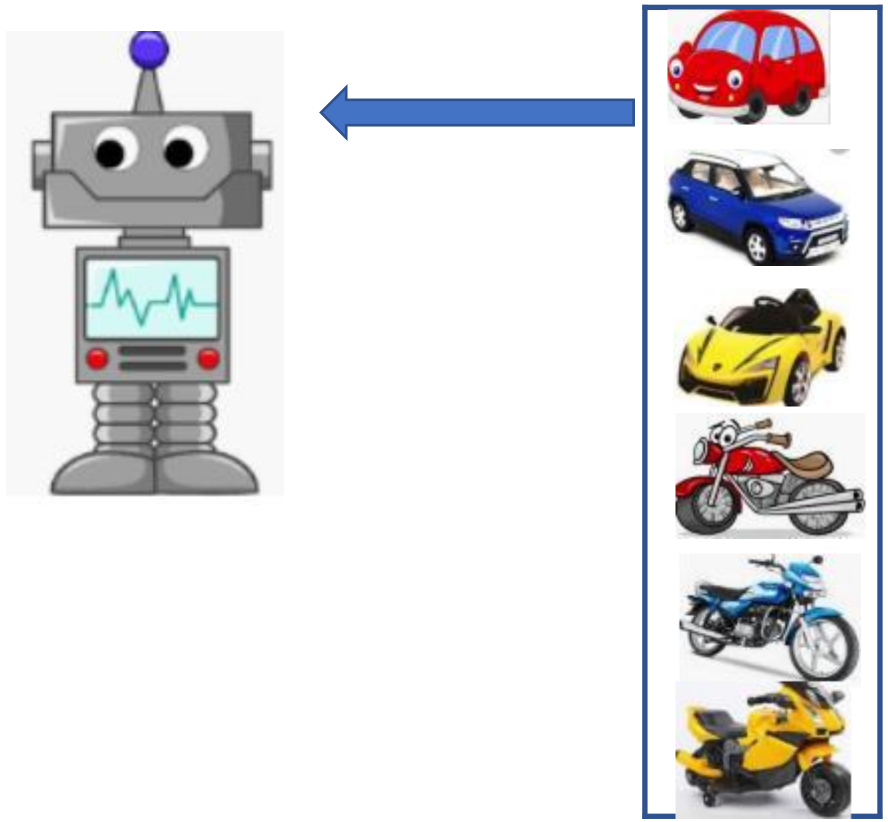
- Sense of exploring the universe like natural science. :
  - The Sparse Distributed Memory (SDM) model was proposed by P. Kanerva in the middle 1980s, there is no intentional imitation to the biological structure of the human brain. However, neuroscience researchers figured out that the sparse encoding mechanism in SDM widely exists in the cortex controlling vision, hearing, and olfactory, thus inspiring more neuroscience research.

Further Reading

- ICML、NeurIPS、ICLR and ect al.

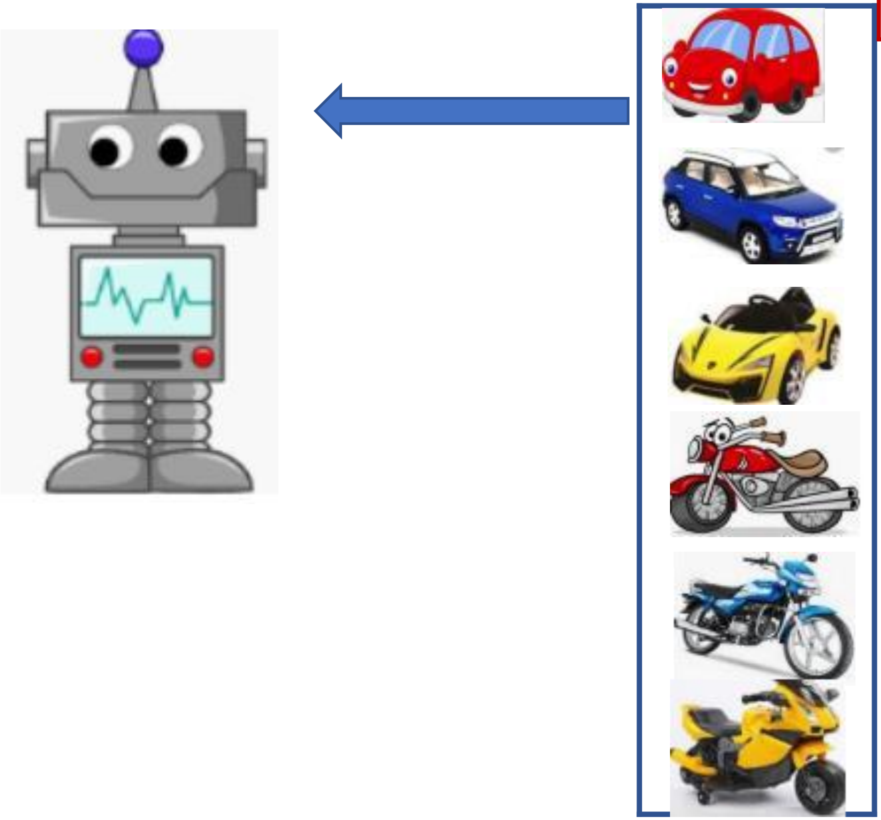
My first machine learning model from Scratch

Teach a machine to identify vehicle types



Represent the sample

#Wheel Height Weight
Color



Represent the sample

#Wheel   Height Weight
Color

Identify the features which can represent the objects

$$F = \{ f_1 \, f_2 \, f_3 \; \ldots \; f_k \}$$

Feature set={ #Wheel   Height   Weight   Color }

Represent the sample

#Wheel Height Weight Color

Identify the features which can represent the objects

$$F = \{f_1 f_2 f_3 \ldots f_k\}$$

For every sample, assign value to corresponding feature

$$v_i = \{w_{i1} w_{i2} w_{i3} \ldots w_{ik}\}$$

where w_{ij} is the value assigned for the feature f_j

Represent the sample

#Wheel	Height	Weight	Color
4	6	500	Red
4	5.5	600	Blue
4	5	550	Yellow
2	3	200	Red
2	3.5	150	blue
2	4	250	Yellow

For every object, assign value to corresponding feature

$$V_i = \{W_{i1} W_{i2} W_{i3} \dots W_{ik}\}$$

where w_{ij} **is the value assigned for the feature** f_j

Vector Space Model

#Wheel	Height	Weight	Color
4	6	500	Red
4	5.5	600	Blue
4	5	550	Yellow
2	3	200	Red
2	3.5	150	blue
2	4	250	Yellow

Features Vectors

This form of representation is called Vector Space Model

Are all features useful?

#Wheel	Height	Weight	Color
4	6	500	Red
4	5.5	600	Blue
4	5	550	Yellow
2	3	200	Red
2	3.5	150	blue
2	4	250	Yellow

Features

Features Vectors

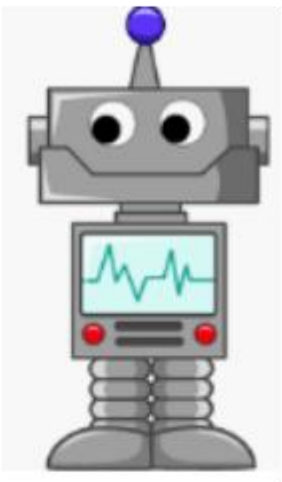
Good Features

- #Wheel
- Height
- Weight

Bad Feature

- Colour

Let us consider single feature



#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE

Training Dataset

Feature vector with Class label

Given the #Wheel, identify the vehicle

#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE

#Wheel	Class Label
2	?

Let us estimate

#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE

Pr(Vehicle type| #Wheel) = ?

Let us estimate the probability (type|#wheel)

#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE

Pr(CAR | 4) = 100%

Pr(BIKE | 4) = 0%

Pr(CAR | 2) = 0%

Pr(BIKE | 2) = 100%

Ask the question now

	#Wheel	Class Label
	4	CAR
	4	CAR
	4	CAR
	2	BIKE
	2	BIKE
	2	BIKE

Pr(CAR| 4) = 100%

Pr(BIKE | 4) = 0%

Pr(CAR| 2) = 0%

Pr(BIKE 2) = 100%

 {2} ?

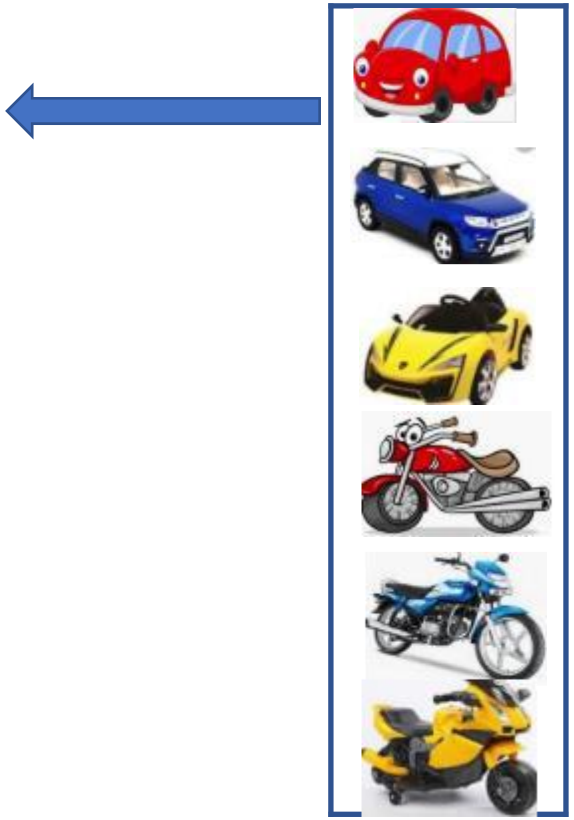
 2

Classifier

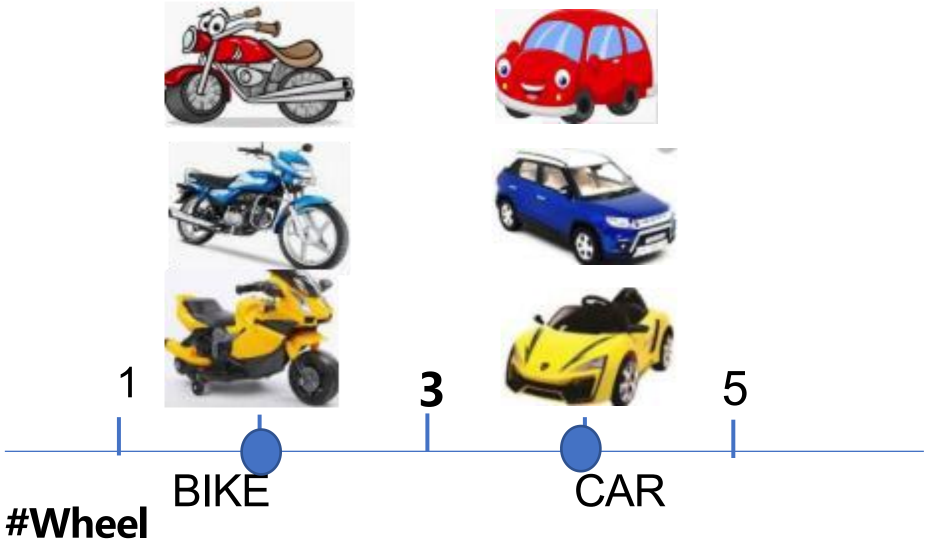
Pr(BIKE|2) > Pr(CAR|2) => BIKE

BIKE

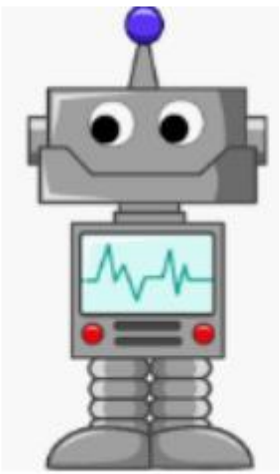
There are multiple ways



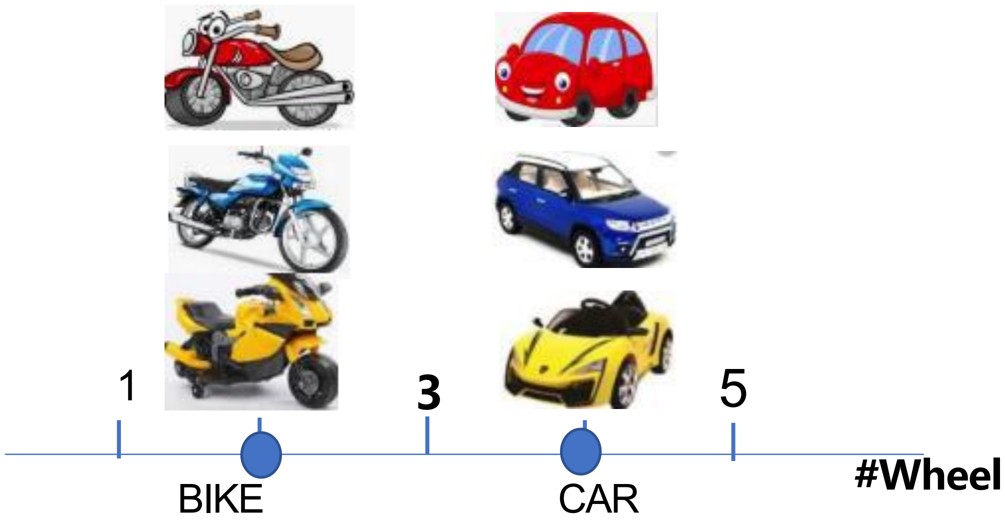
#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE



There are multiple ways



#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE



2 → **Classifier** If #Wheel < 3, then it is BIKE → **BIKE**

If selected feature is not sufficient

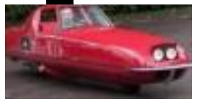
#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE
4	BIKE
2	CAR

$Pr(CAR \mid 4) = 75\%$

$Pr(BIKE \mid 4) = 25\%$

$Pr(CAR \mid 2) = 25\%$

$Pr(BIKE \mid 2) = 75\%$



2 ?

If selected feature is not sufficient

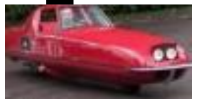
#Wheel	Class Label
4	CAR
4	CAR
4	CAR
2	BIKE
2	BIKE
2	BIKE
4	BIKE
2	CAR

Pr(CAR| 4) = 75%

Pr(BIKE | 4) = 25%

Pr(CAR | 2) = 25%

Pr(BIKE | 2) = 75%



2 BIKE

Pr(BIKE|2) > Pr(CAR|2) => BIKE

More Features

#Wheel	Height	Class Label
4	H	CAR
4	H	CAR
4	H	CAR
2	L	BIKE
2	L	BIKE
2	L	BIKE
4	L	BIKE
2	H	CAR

H: High, height >= 5

L: Low, height < 5

Estimate the probabilities, and ask the same question

#Wheel	Height	Class Label
4	H	CAR
4	H	CAR
4	H	CAR
2	L	BIKE
2	L	BIKE
2	L	BIKE
4	L	BIKE
2	H	CAR

Pr(CAR| 4,H) = 100%

Pr(BIKE | 4,L) = 100%

Pr(CAR | 2,H) = 100%

Pr(BIKE | 2,L) = 100%

Pr(CAR| 4,L) = 0%

Pr(BIKE | 4,H) = 0%

R {2, H}

2,L) = 0%

?

Pr(BIKE 2,H) = 0%

Estimate the probabilities, and ask the same question

#Wheel	Height	Class Label
4	H	CAR
4	H	CAR
4	H	CAR
2	L	BIKE
2	L	BIKE
2	L	BIKE
4	L	BIKE
2	H	CAR

Pr(CAR| 4,H) = 100%

Pr(BIKE | 4,L) = 100%

Pr(CAR | 2,H) = 100%

Pr(BIKE| 2,L) = 100%

Pr(CAR | 4,L) = 0%

Pr(BIKE | 4,H) = 0%

Pr(CAR | 2,L) = 0%

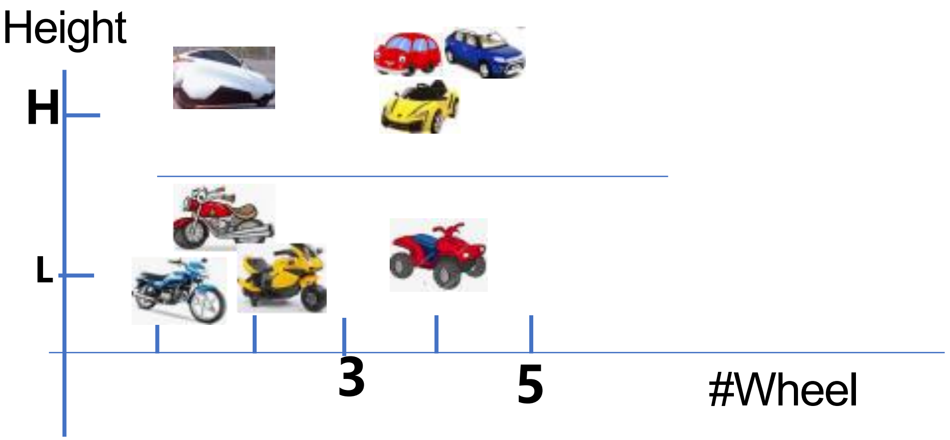
Pr(BIKE | 2,H) = 0%

{2, H}

$\boxed{\text{CAR}}$

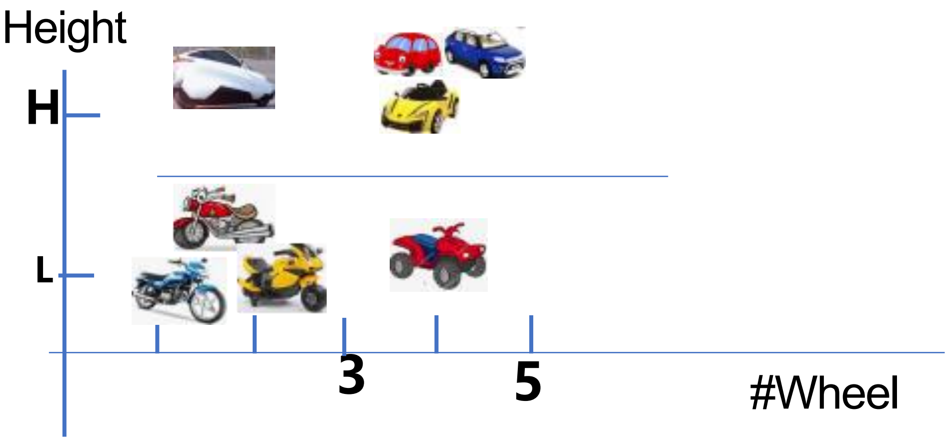
Multiple ways

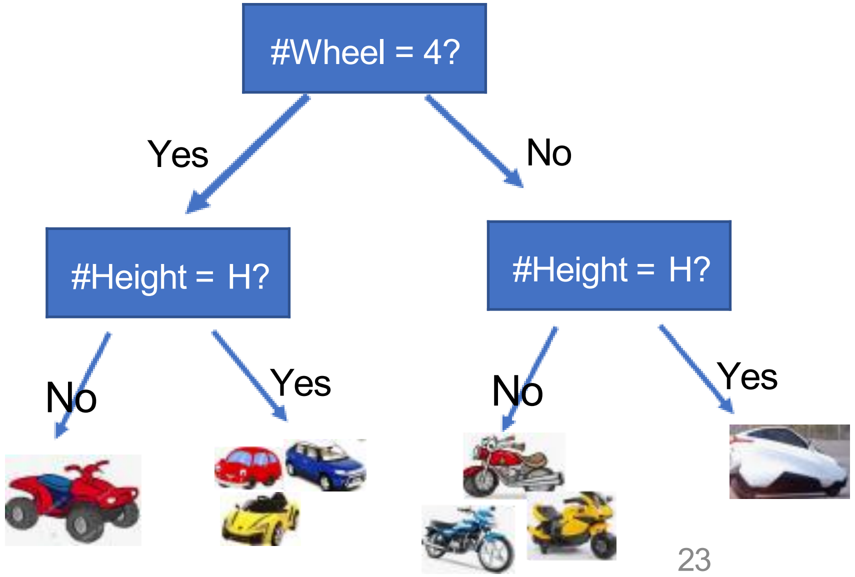
#Wheel	Height	Class Label
4	H	CAR
4	H	CAR
4	H	CAR
2	L	BIKE
2	L	BIKE
2	L	BIKE
4	L	BIKE
2	H	CAR



Multiple ways

#Wheel	Height	Class Label
4	H	CAR
4	H	CAR
4	H	CAR
2	L	BIKE
2	L	BIKE
2	L	BIKE
4	L	BIKE
2	H	CAR





Thanks!