COSC 4368 Fundamentals of Artificial Intelligence

Intro to Machine Learning

September 25th, 2023

What is Machine Learning

> *“Learning is any process by which a system improves performance from experience.”*
>
> \- Herbert Simon

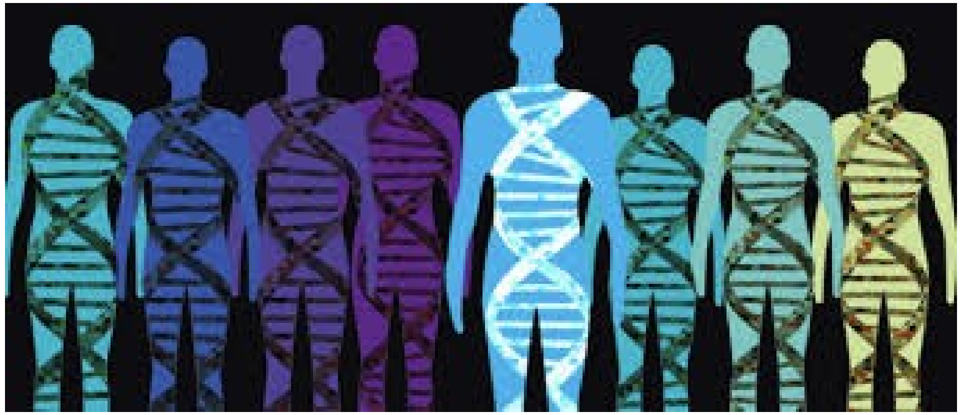
- Definition by Tom Mitchell (1998):
  - Machine learning is the study of algorithms that
    - Improve their performance P
    - At some task T
    - With experience E
  - A well-defined learning task is given by <P, T, E>

When Do We Use Machine Learning

- ML is used when:
  - Human expertise does not exist (navigating on Mars)
  - Humans can’t explain their expertise (speech recognition)
  - Models must be customized (personalized medicine)
  - Models are based on huge amount of data (genomics)
  - Systems need to adapt to changing environments

(1) navigating on Mars

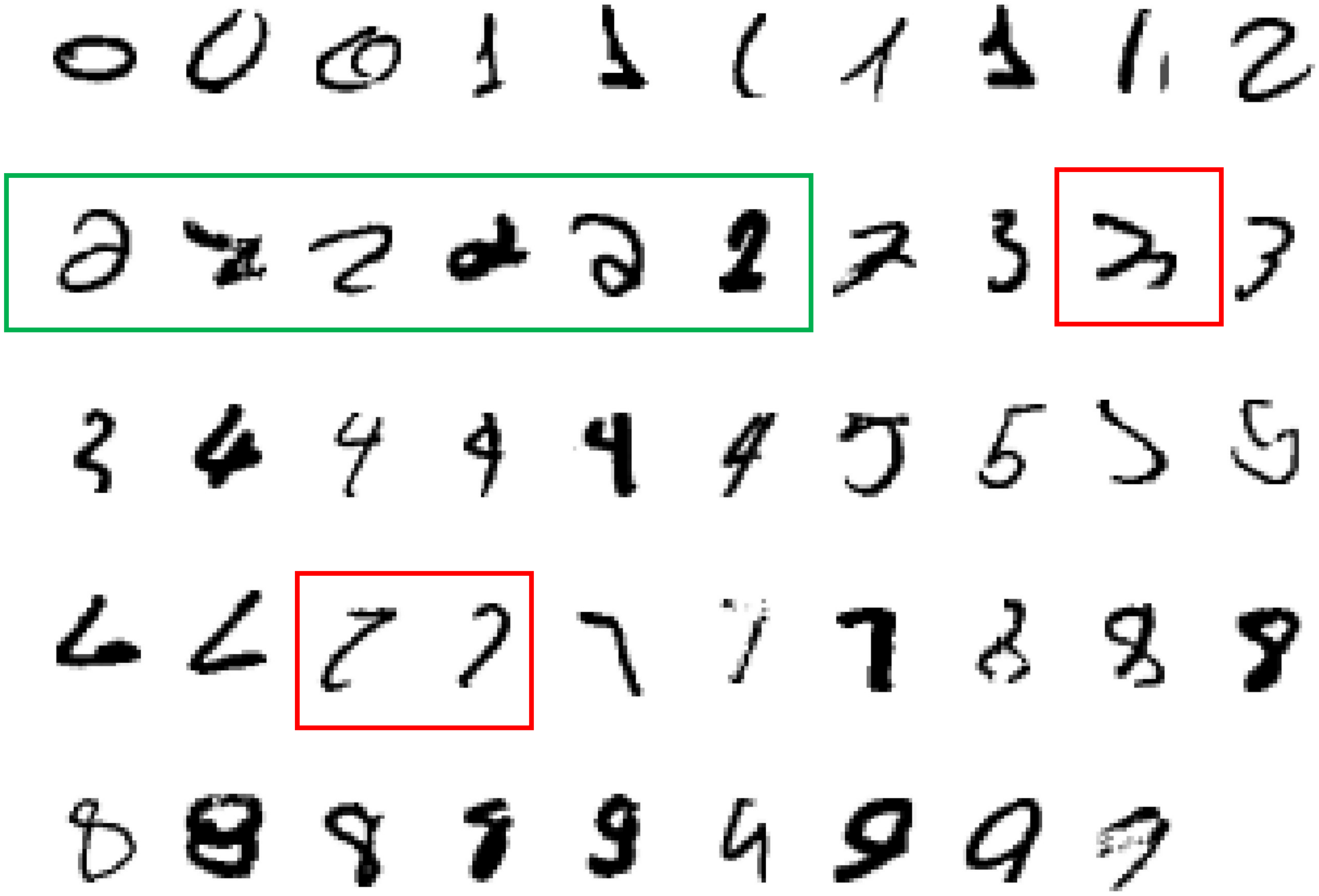
(2) Speech recognition



(3) genomics

- Learning isn’t always useful
  - No need to “learn” to calculate payroll

A Classic Example of Task that Needs ML



Hard to say what makes a 2

Image: 0-9 classification

Types of Machine Learning

- Supervised learning:
  - Given: training data and desired outputs (labels)
  - Learns a mapping from data to label
- Unsupervised learning:
  - Given: training data with no labels
  - Look for ‘interesting’ patterns in the data, e.g., clustering
- Semi-supervised learning:
  - Given: training data and a few labels
- Reinforcement learning:
  - Learning agent interacts with the environment and seeks to maximize the total return

Machine Learning Workflow

- Workflow sketch:
  1. Should I use ML on this problem? Understand domain, prior knowledge and goals
     - Is there a pattern to detect?
     - Can I solve it analytically?
     - Do I have data?
  2. Gather and organize data
     - Preprocessing, cleaning, visualizing, etc.
  3. Establishing a baseline
  4. Choosing a model, loss, regularization, etc.
  5. Optimization: Learn models
  6. Hyperparameter search
  7. Analyze performance (testing)
  8. Deploy discovered knowledge

Various Function Representations (Models)

- Numerical functions
  - Linear models
  - Neural networks
  - Support vector machines
- Symbolic functions
  - Decision trees, etc.
- Instance-based functions
  - Nearest-neighbor, etc.
- Probabilistic graphical model
  - Bayesian networks, Markov networks, etc.

Various Search/Optimization Algorithms

- Gradient descent
- Dynamic programming
- Divide and conquer
- Evolutionary computing
- Etc.

Various Evaluation Metrics

- Accuracy
- Square error
- Likelihood
- Precision and recall
- Cost/Utility
- Entropy
- KL divergence
- Etc.

ML in a Nutshell

- Tens of thousands of machine learning algorithms
  - Hundreds new every year

- Every ML algorithm has three components:
  - Representation
  - Optimization
  - Evaluation

ML in a Nutshell

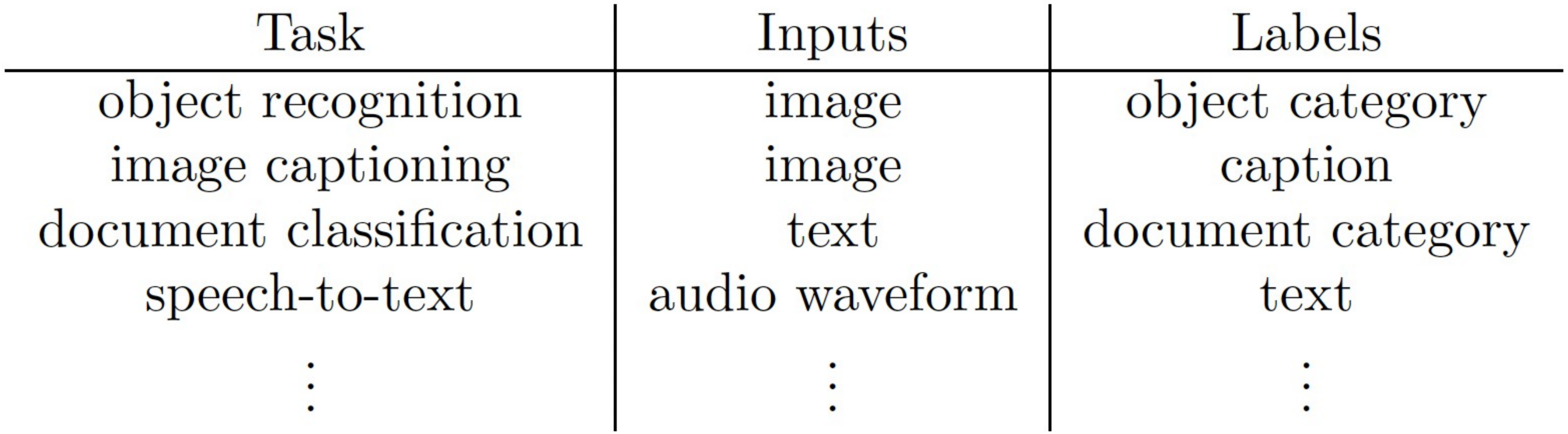
- Learning can be viewed as using direct or indirect experience to approximate a chosen target function

- Function approximation can be viewed as a search through a space of hypotheses (representations of functions) for one that best fits the training data

- Different learning methods assume different hypothesis spaces and/or employ different search techniques

Supervised Learning

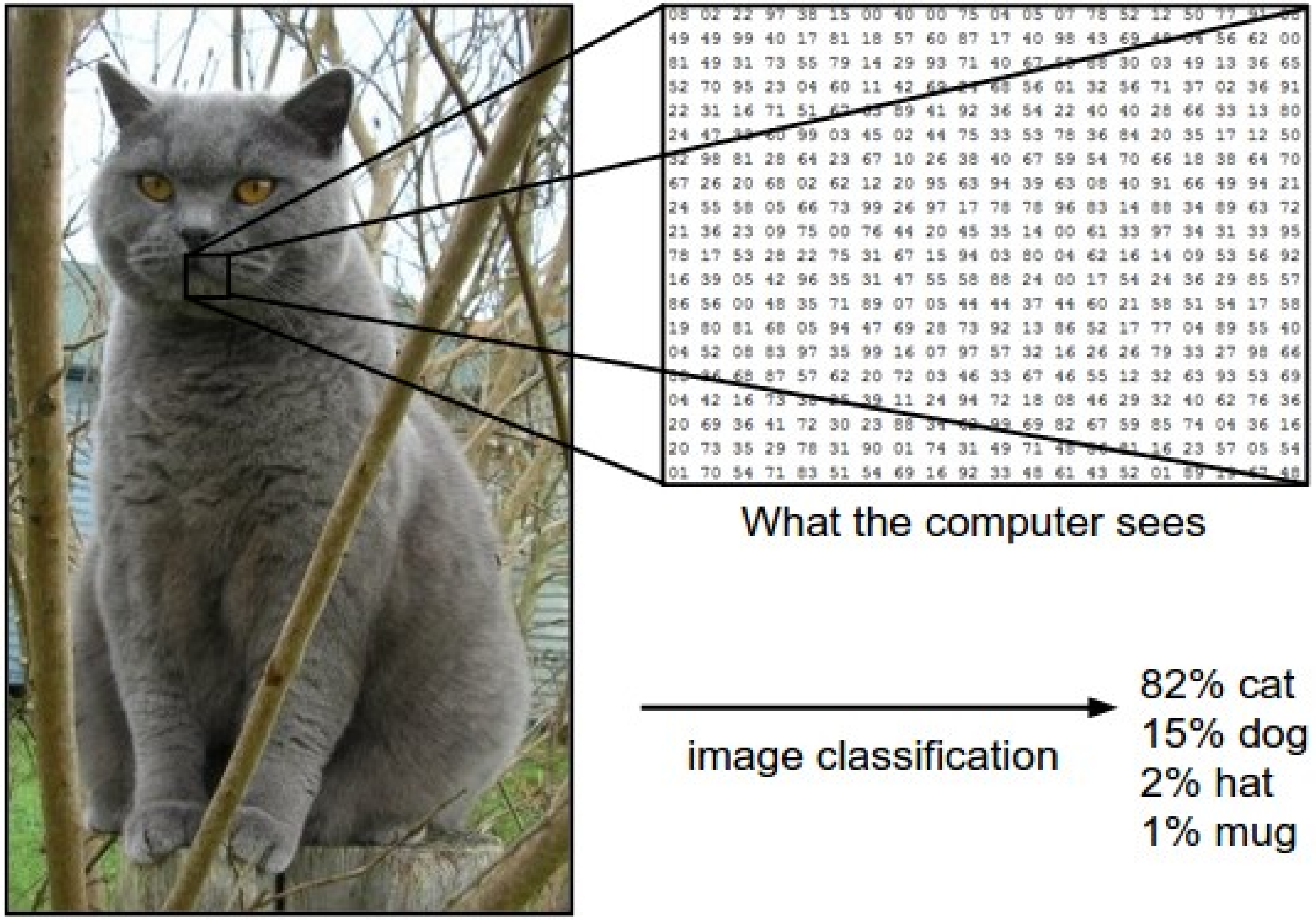
- Focus on supervised learning
- Given a training dataset consisting of inputs and corresponding labels



Task	Inputs	Labels
object recognition	image	object category
image captioning	image	caption
document classification	text	document category
speech-to-text	audio waveform	text
⋮	⋮	⋮

Input Vectors

- What an image looks like to the computer:



```
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
52 70 95 23 04 60 11 42 69 24 68 56 01 32 56 71 37 02 36 91
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 67 48
```

Image: example of representing an image by matrix

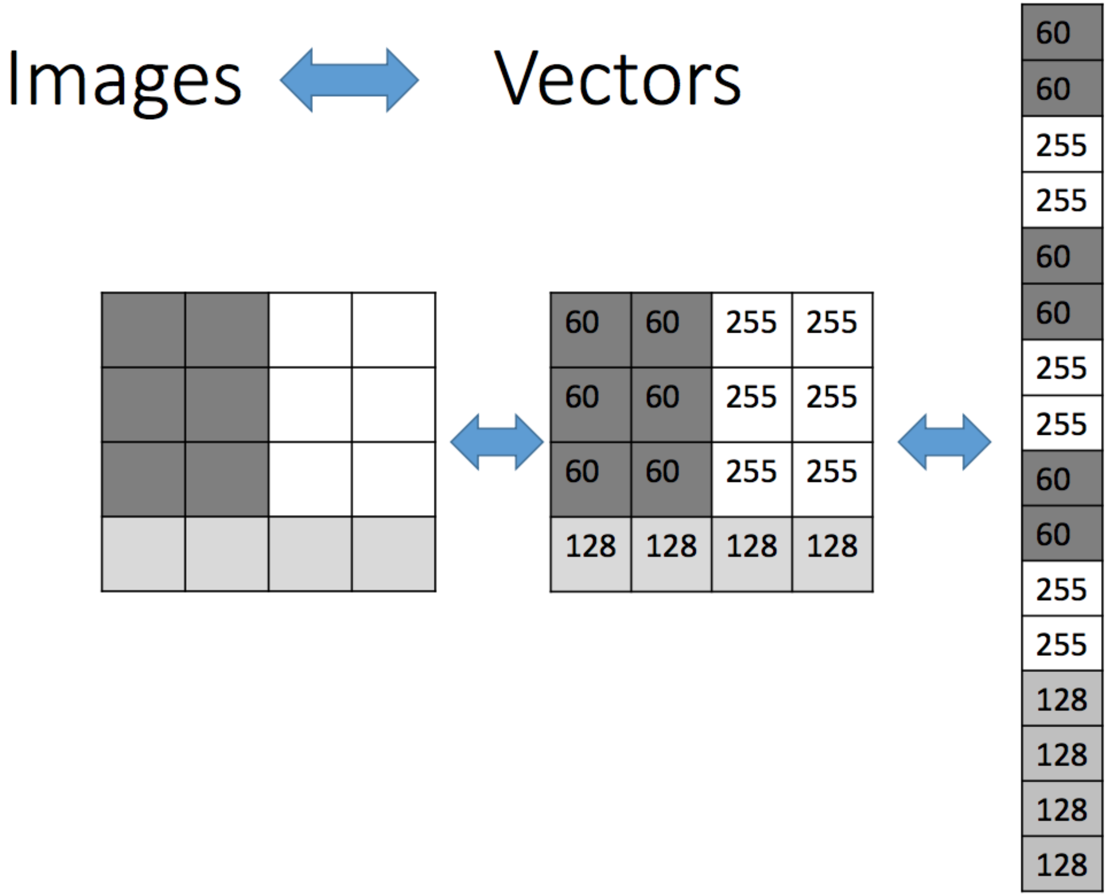
Input Vectors

- Machine learning algorithms need to handle lots of types of data:
  - Images, text, audio waveforms, etc.

- Common strategy: represent the input as an input vector in
  - Map to another space that is easy to manipulate
  - Vectors are a great representation since we do linear algebra

Input Vectors

- Example: image to vector



Images ⟷ Vectors

60	60	255	255
60	60	255	255
60	60	255	255
128	128	128	128

60
60
255
255
60
60
255
255
60
60
255
255
128
128
128
128

Input Vectors

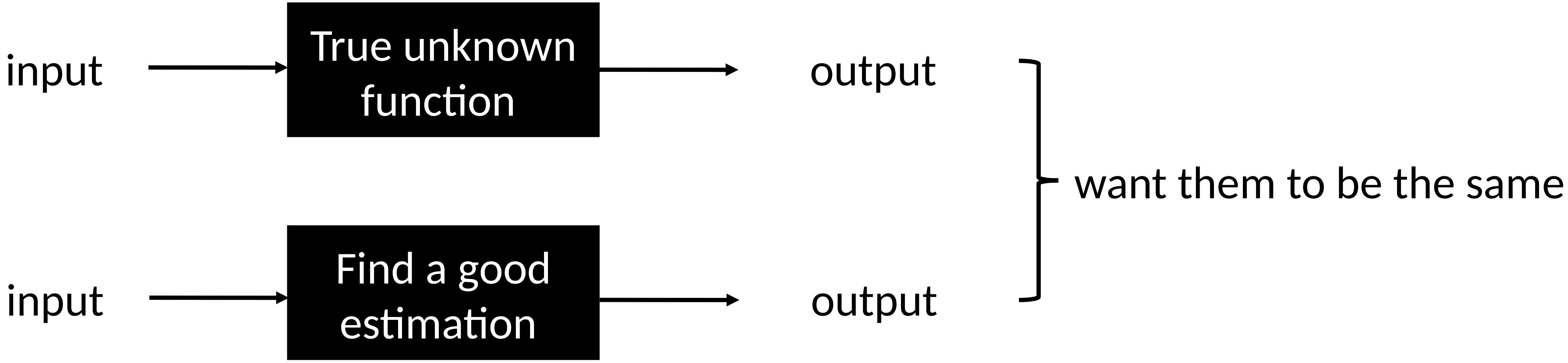
- Mathematically, our training set consists of a collection of pairs of an input vector and its corresponding target or label
  - Regression: is a real number, e.g., stock price
  - Classification: is an element of a discrete set , e.g., which class
  - These days, is often a highly structured output, e.g., image

Hypothesis

- Each was generated by some unknown function :

- Objective: discover a function (hypothesis) that can approximate the true function

- Learning is a search through the space of possible hypotheses



input

True unknown function

output

input

Find a good estimation

output

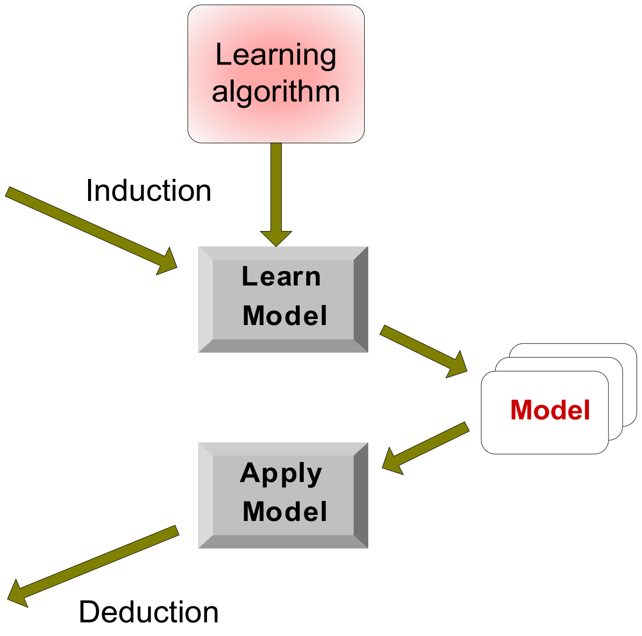
want them to be the same

Hypothesis Evaluation --- Error Rate

- Error rate: the proportion of mistakes the hypothesis makes
  - How many time its prediction for an example
  - Low error rate on the training set does not mean good generalization on unseen data
  - Use a test set of samples to evaluate the accuracy of a hypothesis

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

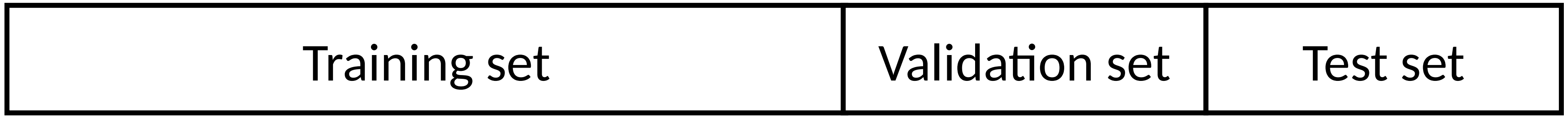
Image: ML procedure

Know the true label for test set

Compare with the predicted label to evaluate the accuracy

Hypothesis Evaluation

- Cross validation
  - Measure performance on unseen data to select a good hypothesis
  - An independent evaluation of the final hypothesis (reported final performance)
  - Split the available dataset into three subsets:
    - Training set: used to learn the model (find the hypothesis)
    - Validation set: used to pick a good hypothesis and tune hyperparameters, evaluate the current the hypothesis and see if we need to improve it
    - Test set: test the final hypothesis, never touch it until you are completely done with learning



Training set

Validation set

Test set

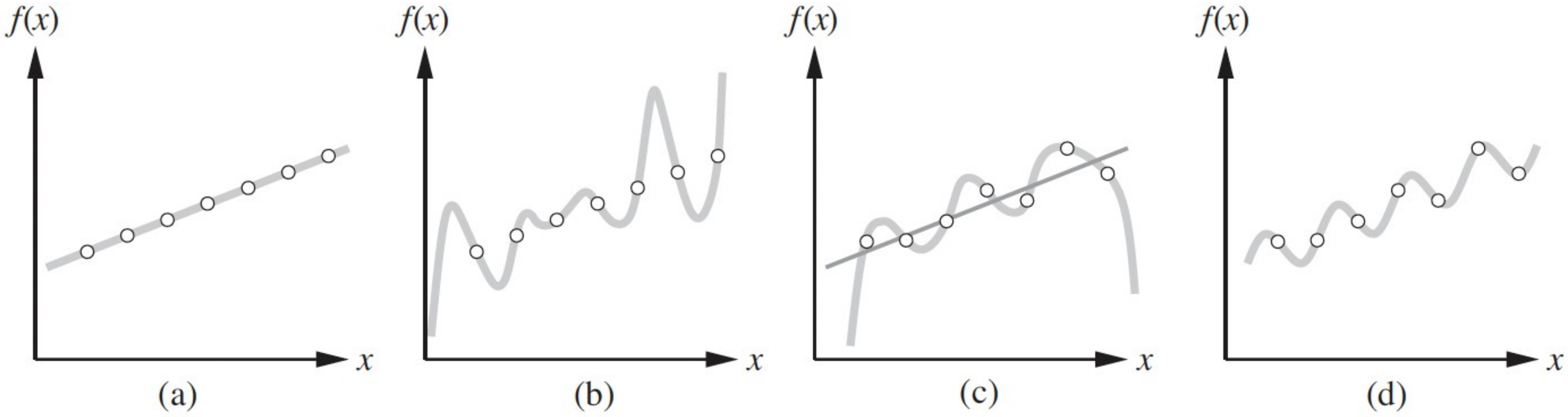
Hypothesis Evaluation and Test Set

- We generally assume that the training and test examples are independently drawn from the same overall distribution of data
    - We call this ‘i.i.d.’: independent and identically distributed

- A hypothesis *generalizes* well if it correctly predicts the labels for novel examples (testing)
    - Generalization error is a measure of how accurately a model is able to predict the labels for previously unseen data

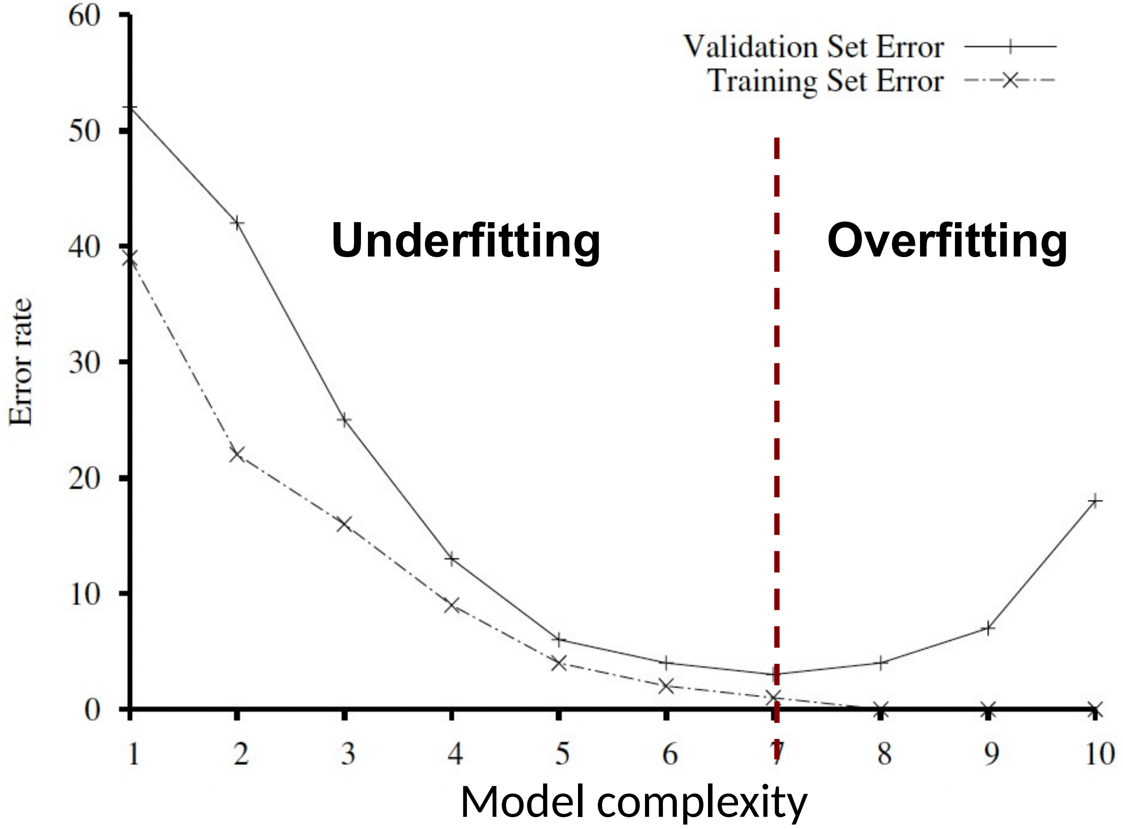
- Special techniques needed to handle non-independent samples and training-testing distribution shift

Model Selections



- (a) linear (b) polynomial (c) polynomial fit or approximate linear fit (d) sinusoidal fit
- Consistent hypothesis: it agrees with all the data
- Two steps: select the hypothesis space and optimize (find the best in the space)
- How to choose among multiple consistent hypothesis?
  - Tradeoff between complex hypotheses that fit the training data well and simpler hypotheses that may generalize better

Underfitting and Overfitting



- **Underfitting:** when the model is too simple, both training and testing errors are large
- **Overfitting:** when the model is too complex, the training error is low, whereas the testing error can increase

Ockham’s Razor

- Given two models of similar generalization errors, one should prefer the simpler model over the more complex model

- For complex models, there is a greater chance that it would also fit the noise in data

- Usually, simple models are more robust with respect to noise

Hypothesis Evaluation --- Loss Function

- The loss function is defined as the amount of utility lost by predicting when the correct answer is for input
- Simplified version
- Examples:
  - Absolute value loss:
  - Squared error loss:
  - 0/1 loss: if , else 1
- Expected generalization loss
  - Assume the underlying data distribution
  - Expected generalization loss:
  - Best hypothesis (function) in hypothesis space:

Hypothesis Evaluation --- Loss Function

- Empirical loss on a set of examples:
  - The underlying data distribution is unknown
  - Use empirical loss to estimate generalization loss
  - Empirical loss:
  - Estimated best hypothesis in hypothesis space :
    - The best we try to find during training
    - May differ from the true function because:
      - may not be realizable, e.g., not in the hypothesis space
      - is learnt based on finite number of samples; sample variance – different datasets lead to different
      - may not be deterministic, e.g., contains noise
      - The hypothesis space is very complex, and it is computationally intractable to find the global optima

Small-Scale Learning vs Large-Scale Learning

- Early stage of machine learning --- small datasets and simple models
    - Generalization error mostly comes from
        - The true function is not in the hypothesis space
        - Large sample variance

- Current stage of machine learning --- large datasets and complex models
    - Generalization error mostly comes from
        - Computationally intractable to find the globally optimal function

Regularization

- The optimization problem in training:

- Regularization: prefer to learn a function that has certain properties
  - Explicitly penalize the functions that do not have that property
  - E.g., can penalize complex functions: push more weights in the function close to zero