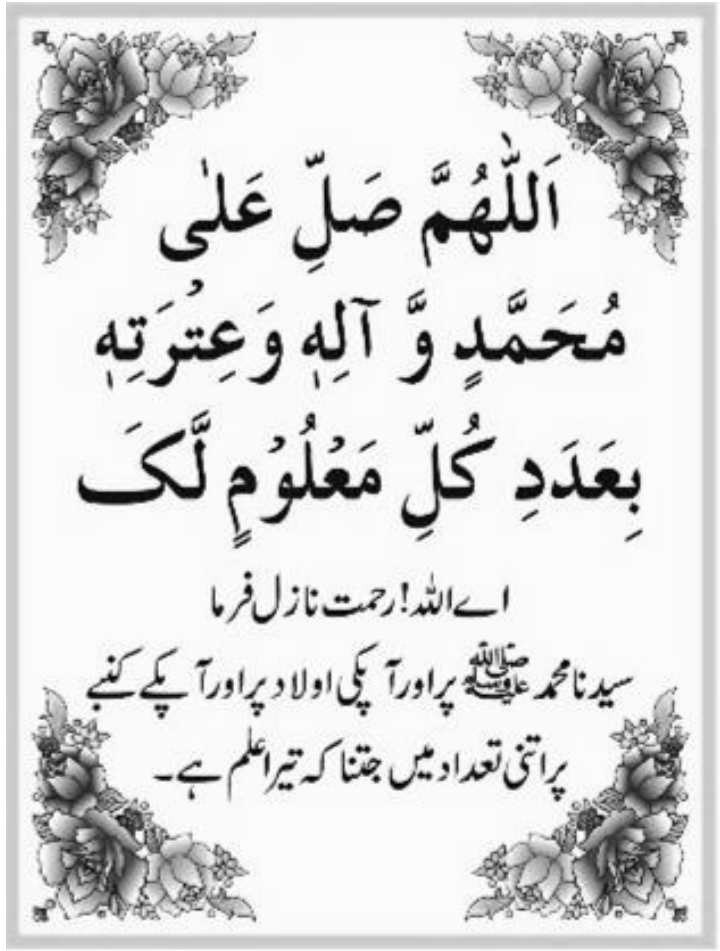
بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

CSC354- Machine Learning

Lecture 01 – Basics of Machine Learning

Dr. Jawad Shafi

Dua – Take Help from Allah Before Starting Any Task



اللَّهُمَّ خِرْ لِيْ وَاخْتَرْ لِي

سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا ؕ

إِنَّكَ أَنْتَ الْعَلِيمُ الْحَكِيمُ

رَبِّ اشْرَحْ لِي صَدْرِي

وَيَسِّرْ لِي أَمْرِي

وَاحْلُلْ عُقْدَةً مِنْ لِسَانِي

يَفْقَهُوا قَوْلِي

Dr. Jawad Shafi – About Me

Assistant Professor

COMSATS University Islamabad, Lahore Campus

Group Member

NLP Group, CUI, Lahore Campus

Course Details

Instructor: ***Jawad Shafi [Ph.D]***

Email: ***jawadshafi@cuilahore.edu.pk***

Office: ***Room no. 124, Faculty Block***

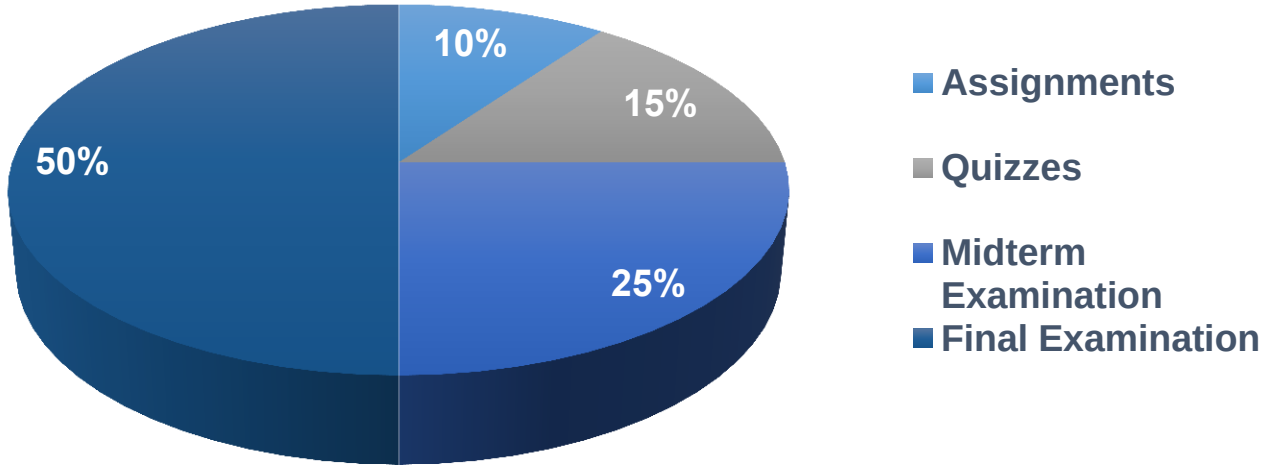
Course Details – For BS Course (Cont.)

Google Classroom Code: zmksjgf4

Note: Join using CUI-Lahore email ID

Office Hours: *Email requests for appointment*

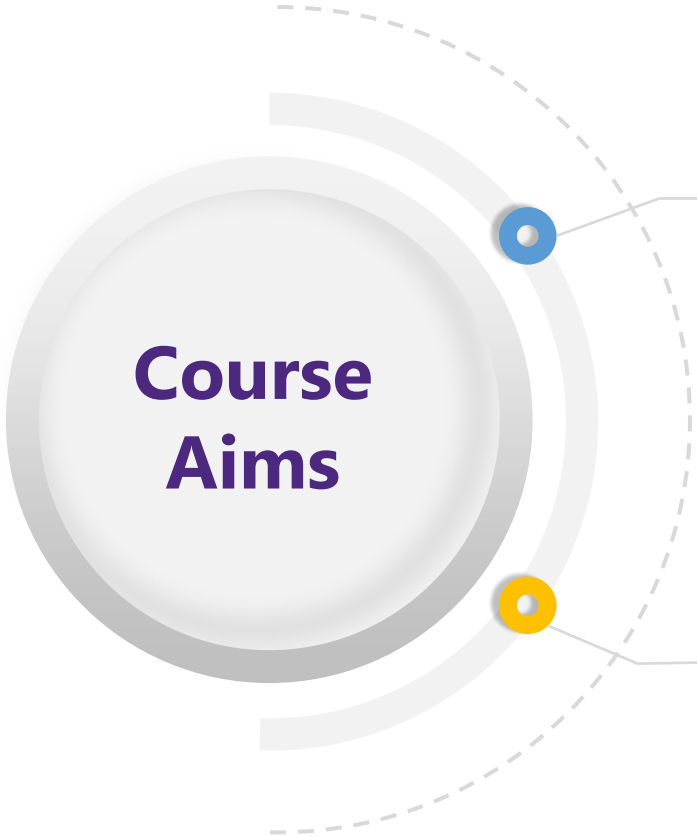
Assessment



Course Aims

1. To introduce with the main concepts that are essential to become a great personality and a great machine learning engineer
2. To develop skills to systematically learn any concept
3. To introduce some of the main topics in machine learning, especially learning from examples (automated concept learning or classification)

Course Aims (Cont...)



4 To develop skills to understand, analyze and pre-process a dataset / corpus to train and test machine learning algorithms

5 To develop skills in designing and building serious artificial intelligence programs.

Course Learning Outcomes (CLOs)

- **By the end of this course, the students should be able to**
  - Understand what daily tasks are important to have a healthy (physically, mentally and socially) and characterful personality
  - Analyze datasets and transform them into representations that will permit machine learning algorithms to learn concepts/discover patterns in them
  - Understand some of the main approaches/algorithms that are used for representing concepts and learning them automatically
  - Evaluate competing machine learning algorithms over the same dataset(s)

Pre-Requisites

Good Programming Skills

python™

scikit learn

jupyter

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

SciPy

NumPy

Course Content – Key Topics

- Basics of Machine Learning
- Concept Learning and General-to-Specific Ordering
- Decision Tree Learning
- Artificial Neural Network
- Deep Learning
- Evaluating Hypothesis
- Bayesian Learning
- Genetic Algorithms
- Instance Based Learning
- Multi-Label Classification
- Course Project Presentations

References

Book 1

T. Mitchell. Machine Learning. WCB/McGraw-Hill, Boston, 1997. http://www.cs.cmu.edu/~tom/mlbook.html

Book 2

I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 3rd Edition, Morgan Kaufmann, San Francisco, 2011.

or

Hands on Machine Learning with Scikit Learn and Tensorflow, Published by O'Reilly Media, 2017, by Aurélien Géron

Book 3

S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach (2nd ed), Pearson Education, 2003.

Book 4

N. J. Nilsson. Introduction to Machine Learning, Draft of a proposed textbook, 1997. Available at: http://robotics.stanford.edu/people/nilsson/mlbook.html.

Course Handbook

Hnadbook

A Course in Machine Learning:

http://ciml.info/

Notes

I will provide you the notes for reading.

Journals and Conferences

- Journal
  Machine Learning
- Journal
  Journal of Machine Learning Research
- Journal
  Neural Computation
- Journal
  Journal of Intelligent Systems
- Conference
  International Conference on Machine Learning
- Conference
  Neural Information Processing Systems (NIPS)

Machine Learning Toolkits

Python

- Scikit-learn (a.k.a. sk-learn)
  http://scikit-learn.org/stable/
- Documentation
  http://scikit-learn.org/stable/documentation.html

Java

- WEKA
  https://www.cs.waikato.ac.nz/ml/weka/
- Documentation
  https://www.cs.waikato.ac.nz/ml/weka/documentation.html

Lecture Outline

- What is Machine Learning
- Learning Input-Output Functions – General Settings
- Types of Machine Learning
- Phases of Machine Learning
- Training Regimes
- Treating a Problem as a Machine Learning Problem – A Step by Step Example

Reading

- Chapter 1 of Mitchell
- Chapter 1 of Witten & Frank

What is Machine Learning

Myth vs Reality

Predicting the future based on the past.

Predicting the unseen based on the already seen.

Ultimate Goal

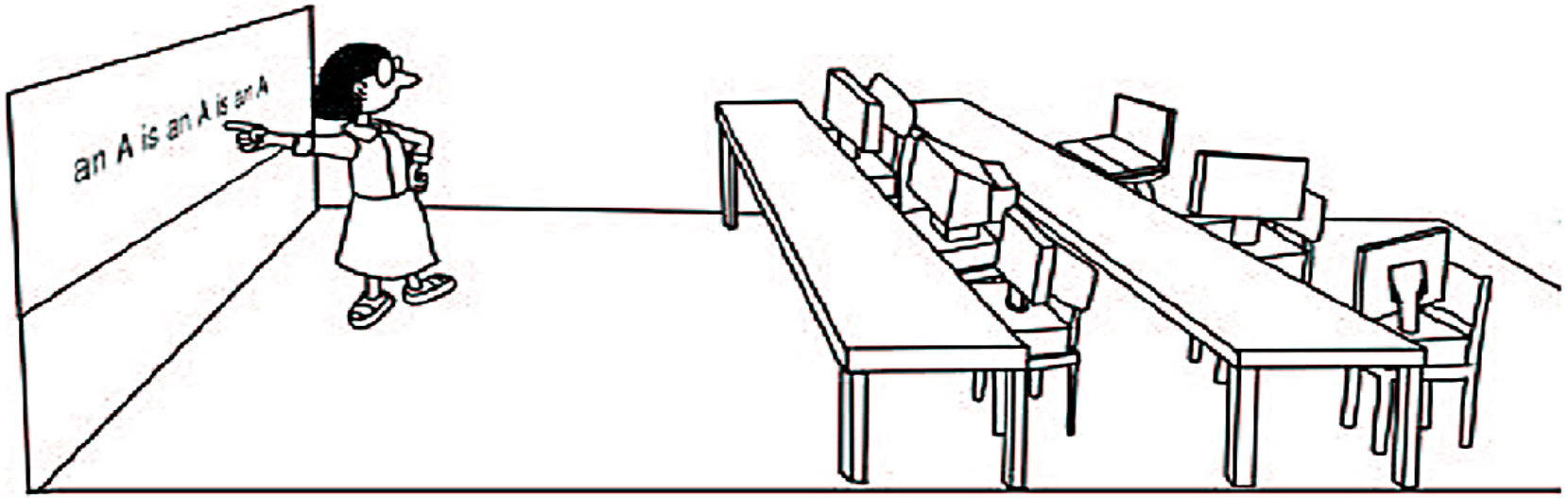
To develop machines which behaves like human(s)

Machine Learning

ML is a branch of artificial intelligence in which computers are trained to learn from data, identify patterns, and make decisions with minimal human intervention

Machine Learning

The study of how to design computer programs whose performance at some task improves through experience



Machine Learning

A computer program is said to be 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

Machine Learning - Example

- Learn to Drive a Car
  - Class of Tasks (T)
    - Starting Car, Changing Gear, Control on Accelerator, Control on Breaks, Parking Car
  - Performance Measure (P)
    - Efficiency
  - Experience (E)
    - No. of hours spent in driving car
- A person is said to be learn to drive a car if:
  - His / her performance P on class of tasks T is improving with experience E

ML History

- Artificial Intelligence (1956) aims to create intelligent machines that can replace or exceed human intelligence

- Machine Learning (1997) is a subset of AI that enables machines to learn from existing data to make decisions or predictions

ML History

- Deep Learning (2017) is a technique within ML which utilizes layers of neural networks to process data and make decisions

- Generative AI (2021) uses DL models (e.g., Generative Adversarial Networks) to generate high-quality text, images, and other content based on the data they were trained on

ML History

- Agentic AI (2024) – (Evolved from 1950's rule-based system) can independently plan, use external tools, and act autonomously to achieve goals.

Machine Learning vs Traditional Programming

Traditional Programming

Machine Learning

Machine Learning vs Traditional Programming

is a field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning

Learning Analogy

To learn something is the one's ability to use previous knowledge to perform future actions

- Suppose you took a new course this semester (e.g., Mathematics)
- You expect to "learn" something from that course
- What is a common way to judge how well you do?
- You did well at learning, if you do well on the exam

Learning Analogy Conti.

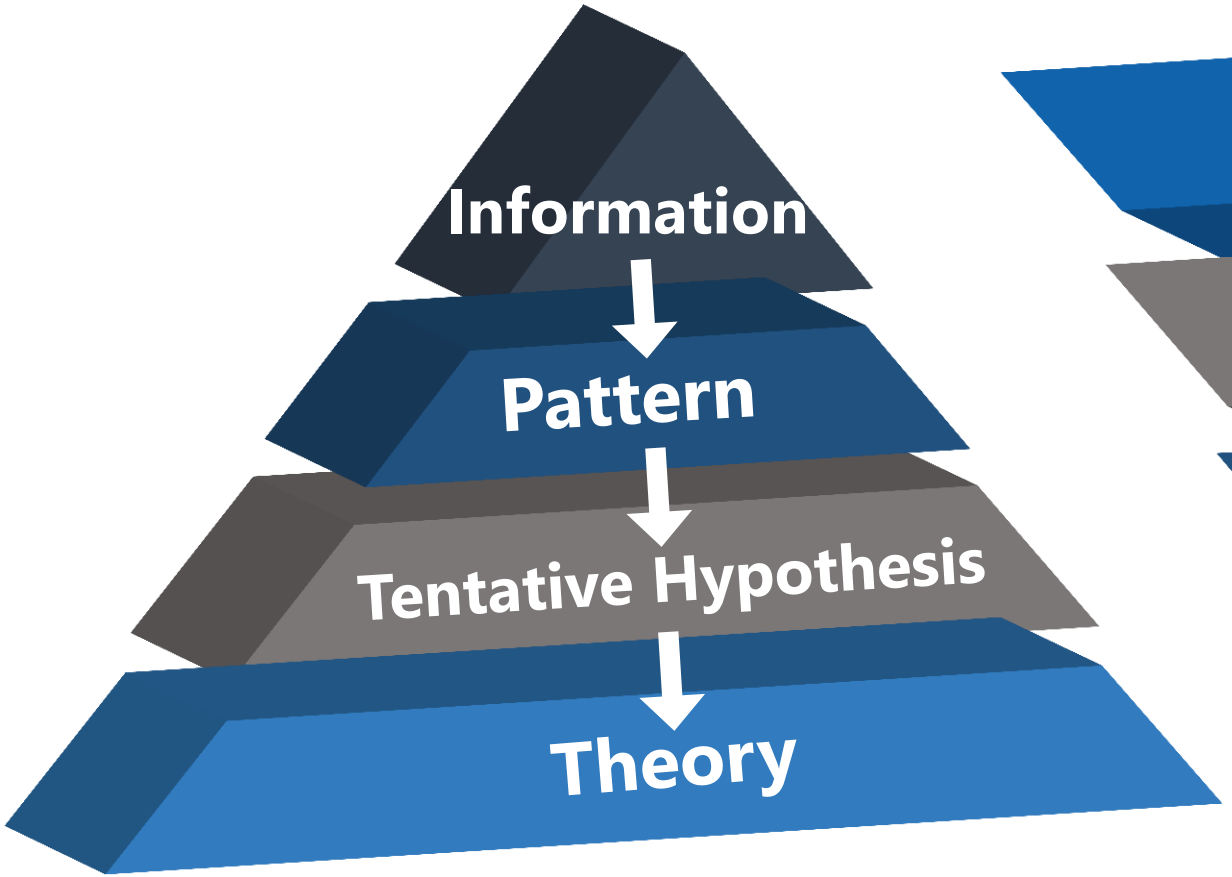
What makes a reasonable exam?

- If it has chemistry questions, it's not representative of your learning
  Remember the course was mathematics
- If it only has questions that were already solved in the lectures, that's a bad test of your learning
- The best practice would be
  You study and understand the concepts with examples during the lectures
- The exam then have "new" but "related" questions

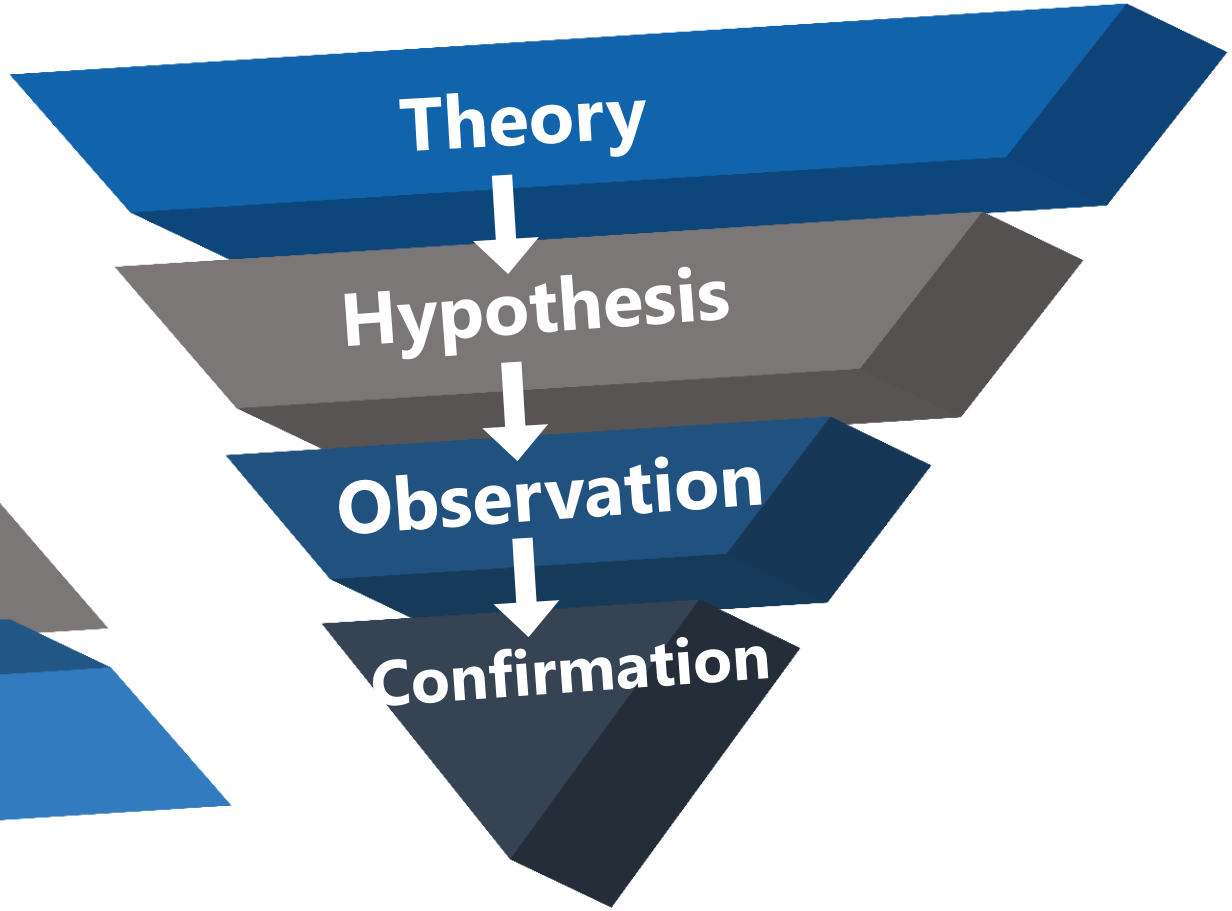
- **The good exam would test your ability to "generalize"**

Types of Learning

Deductive Learning



Inductive Learning



Deductive Learning

Deductive learning

- Works on existing facts and knowledge
- Does not generate "new" knowledge at all
- Makes the reasoning system more efficient

Deductive Learning – Example

Example

- Concept to be Learned - Throwing a ball in air
- How Deductive Learning works?
  - I know Newton's Laws of Gravitation
  - So, I conclude that if I let a ball go, it will certainly fall downwards

Inductive Learning

Inductive learning

- **Takes examples of a concept and generalizes rather than starting with existing knowledge**
- **Generates "new" knowledge**
- **Has "scope of error"**

Several successful systems developed using inductive learning approach

Inductive Learning – Example

- Concept to be Learned - Throwing a ball in air
- How Inductive Learning works?

- Take examples of the concept to be learned
  - 1 example – 1 time I throw ball in the air
  - 50 examples – 50 times I throw ball in the air
  - 100 examples – 100 times I throw ball in the air
- Learn from Examples
  - I throw a ball 100 times in the air and learned that every time (100 times) I throw the ball in the air, it falls downward
- Generalize the Concept Learned from Examples
  - I conclude that if I let a ball go, it will certainly fall downwards

How Learning is represented ?

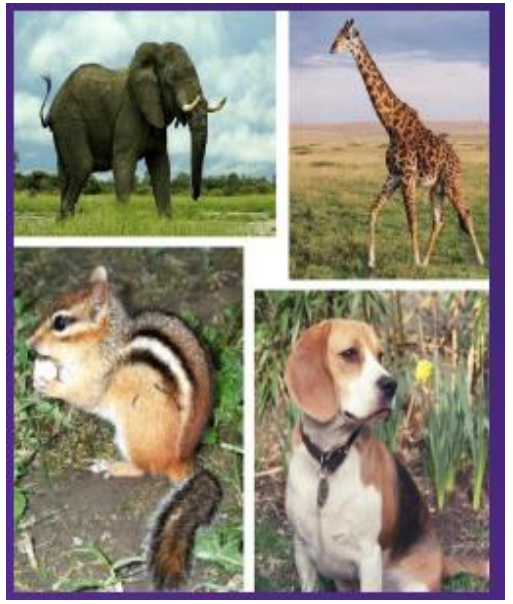
- Learning is captured in a model learned from discovering patterns in data
  - Patterns are discovered through model parameters
  - Data is provided as training examples
- Trained model should be a good and useful approximation of the data
  - The trained model is used later for prediction
- All of the above is achieved using a learning algorithm

How Learning is represented ?

Let's try out an example

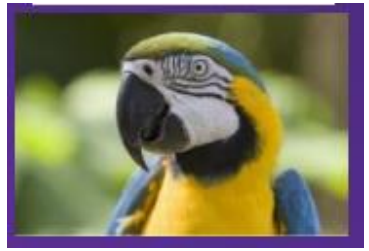
Class A

Class B



How Learning is represented ?

- **Let's Different learning algorithms/models have different hypotheses**



Flying?
Birds?
Mammal?

How Learning is represented ?

- How would you write a program to distinguish a picture of yourself from a picture of someone else?

  ?????

- How would you write a program to distinguish cancerous cells from normal cells?

  ?????

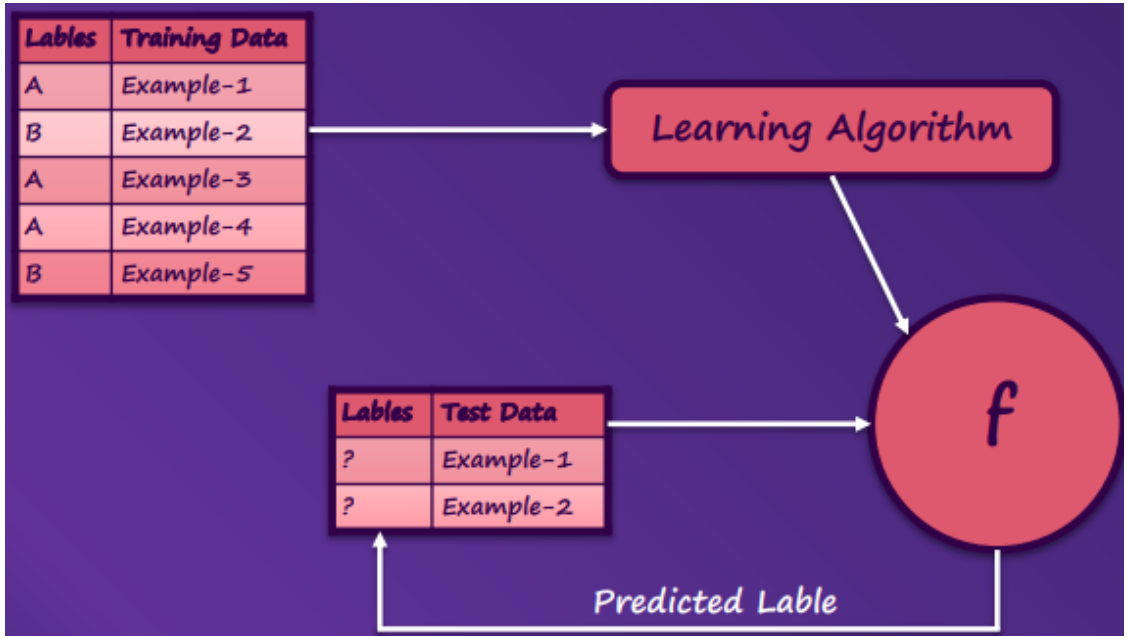
How Learning is represented ?

- How would you write a program to distinguish a picture of yourself from a picture of someone else?
  Provide example pictures of yourself and pictures of other people and let a classifier learn to distinguish the two
- How would you write a program to distinguish cancerous cells from normal cells?
  Provide examples of cancerous and normal cells and let classifier learn to distinguish the two

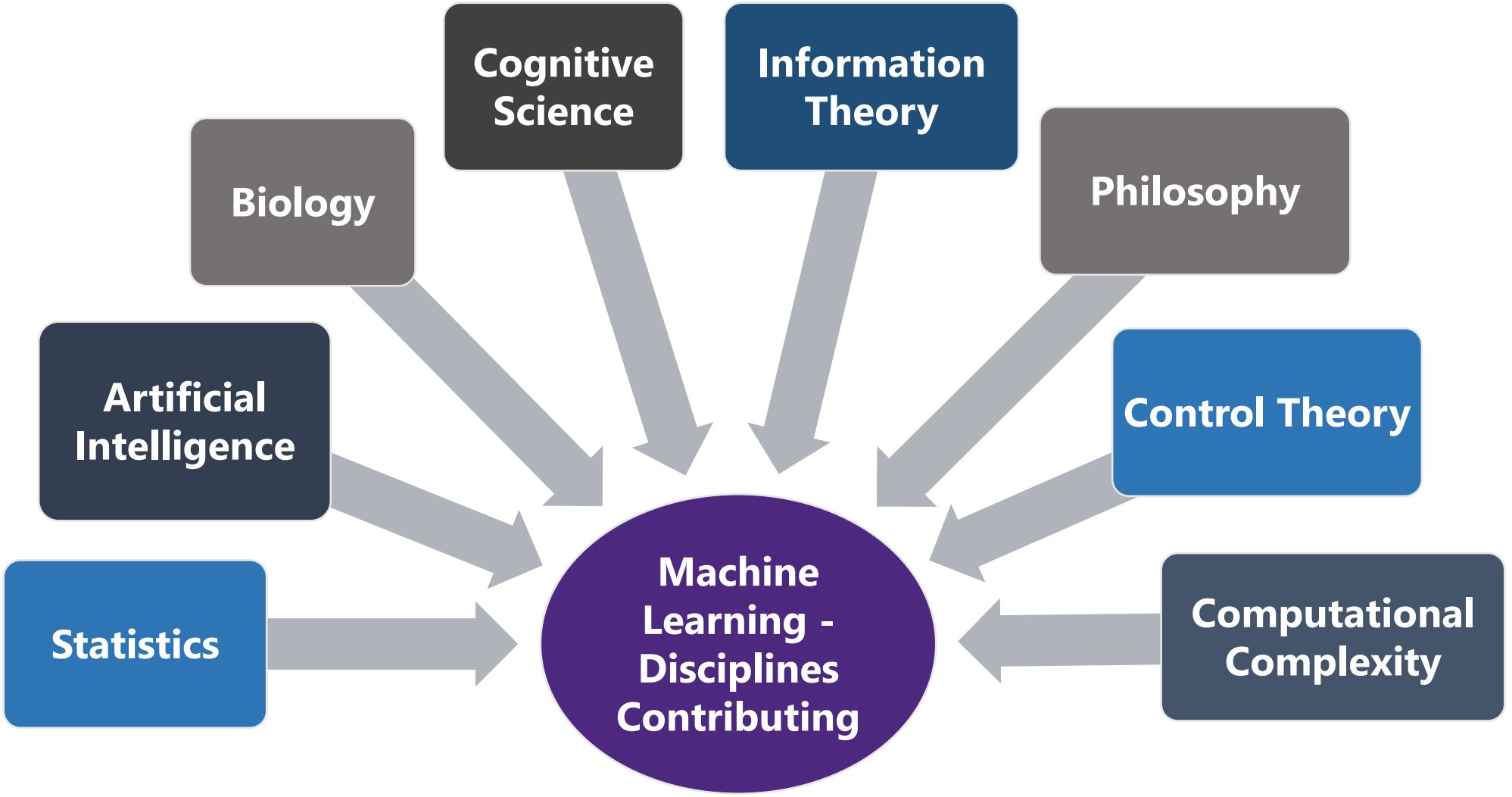
These are known as a classification tasks

How Learning is represented ?

Let's Predict the future based on the past



Machine Learning – Disciplines Contributing



Machine Learning – Why to Study

Why to Study

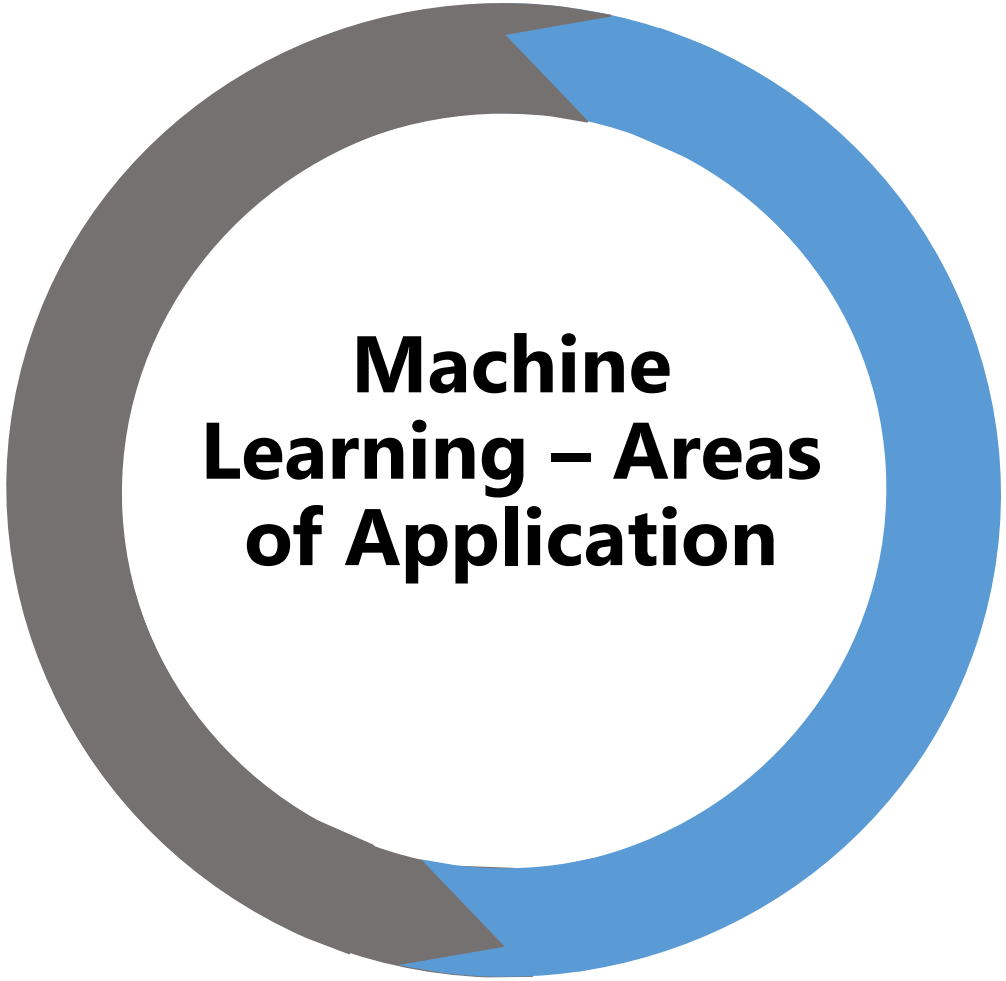
- **Technological or Engineering Motivation**
  - To build computer systems that can improve their performance at tasks with experience (data)
  - Massive growth in on-line data
- **Cognitive Science Motivation**
  - To understand better how humans learn by modeling the learning process

Machine Learning – Why to Study (Cont…)

Why to Study

- To understand better properties of various algorithms for function (or concept to be learned) approximation
  - How the data must be represented?
  - How much data they require?
  - How accurate they can be?
  - How to choose optimal data for training?

Machine Learning - Areas of Application



- Data mining
- Medicine
- Business
- Agriculture
- Computer games
- Software applications
- Personalized / Self
- Customizing program

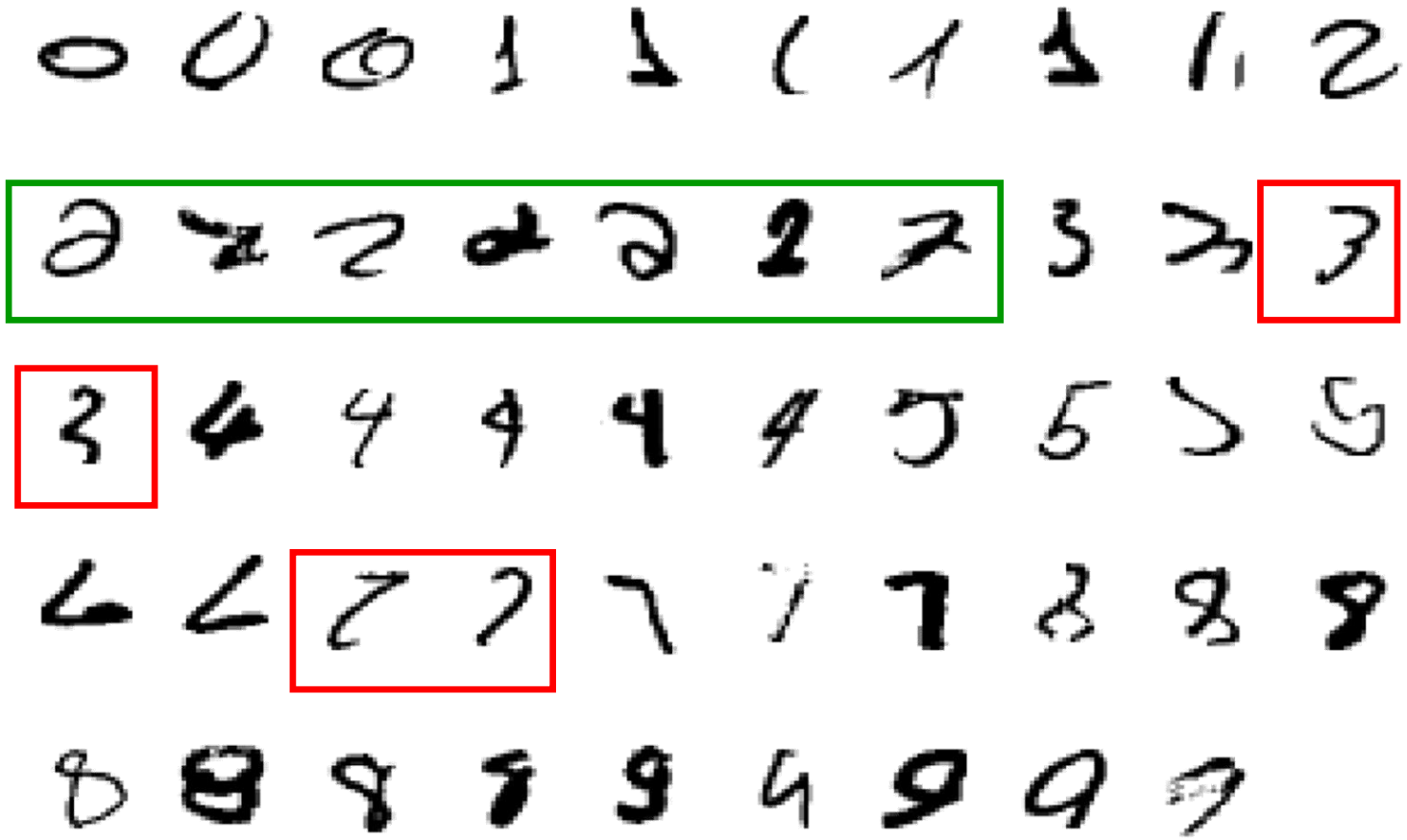
Machine Learning – Why it is Hard

You See

Your ML Algorithm Sees

Machine Learning – Why it is Hard (Cont…)

What is a "2"?



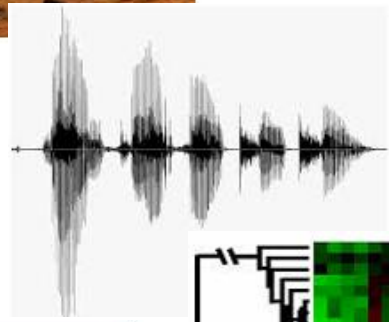
When Do We Use ML?

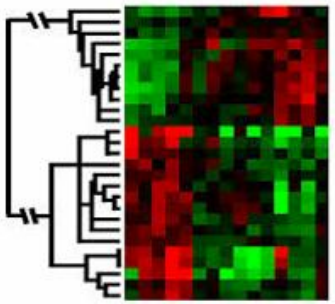
Human expertise does not exist (navigating on Mars).

Humans can't explain their expertise (speech recognition).

Models must be customized (personalized medicine).

Models are based on huge amounts of data (genomics).





Learning isn't always useful:

There is no need to "learn" to calculate payroll.

Machine Learning – When to Use

1. When we have lots of data

2. When patterns exist in our data. even if we don't know what they are?

Machine Learning - Summary

Learn from Data

Data = Model + Error

ML in Practice

- Understand domain, prior knowledge, and goals.
- Data integration, selection, cleaning, pre-processing, etc.
- Learn models.
- Interpret results.
- Consolidate and deploy discovered knowledge.

Learning Input – Output Functions - General Settings

What is to be Learned

- Function
- Program
- Finite state machine
- Grammar
- Problem solving system

Most of ML revolves around
- Learning Input-Output Functions or Function Approximation or Function Learning or Concept Learning

Concept Learning

Concept Learning
- A major subclass of inductive learning

How Concept Learning Works?
- Takes examples of a concept and
- Tries to build a general description of the concept

Very often, the examples are described using attribute-value pairs

Concept Learning

Goal

- Learn a generalized target function f (or target concept c) from a set of examples

Scope of Error in Inductive Learning

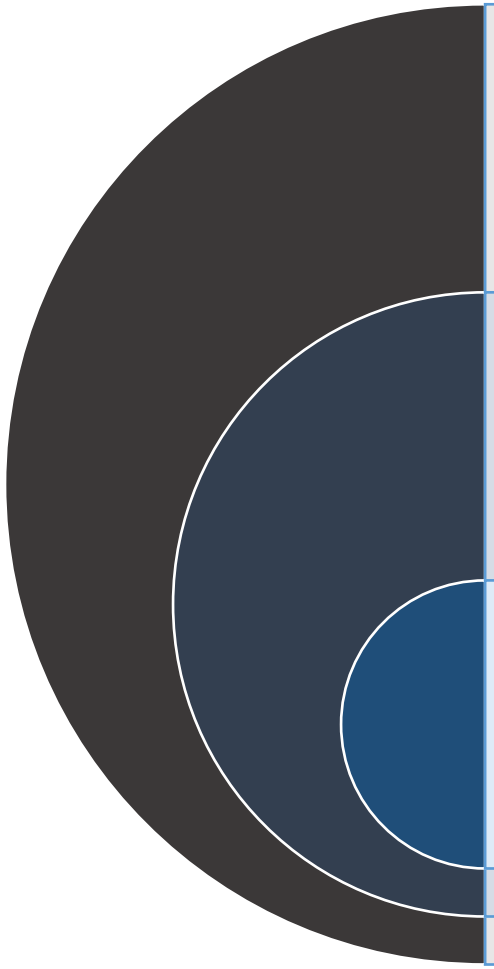
- Target function (f) cannot be completely learned, however it can be approximated
  - Hypothesis (h) – is an approximation of the target function f

Learning Input Output Functions – General Settings

- **Goal - We are trying to learn a function f**
  - f is often referred to as the target function

- **Input ➡ f ➡ Output**
  - f takes a vector-valued input → a n-tuple $x = (x_1, x_2, \ldots x_n)$
  - f itself may be vector-valued → yielding a k-tuple as output
    - Often f produces a single output value
    - E.g. k = 1 (in such cases the output is not thought of as a vector)

Learning Input Output Functions – General Settings



Job of the learner is to output a hypothesis *h* which is its guess or approximation of the target function f

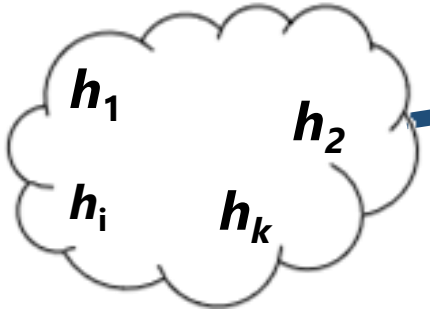
h is assumed a priori to be drawn from a class of functions H

Note that f may or may not be in H and this may / may not be known

Learning Input Output Functions – General Settings

Training Example =
$\{x_1, \ldots. x_m\}$ +$f(x_i)$
for each x_i ϵ TE

Hypothesis Space $\mathcal{H}$



Learner

h*, where *h* ≈*f

Concept Learning - Representation

Machine is dump, need to represent two things

1. **Hypothesis**
   Will be discussed in next lecture
2. **Examples or Data**
   Will be discussed in this lecture

Example

Example (a.k.a. instance, data point, observation)

Example = Input + Output

Example - Representation

Representation of Input

- Attribute-value pair
- Input can be
  - Single valued or vector valued (mostly vector valued)
- "Values of attributes" can be
  - Categorical /Ordinal
    e.g. Male, Female, Yes, No
  - Numeric
    Discrete – e.g.10, 25, 10000
    Continuous – e.g. 3.5, 5.9

Representation of Output

- Attribute-value pair
- Output can be
  - Single valued or vector valued (mostly single valued)
- "Values of attributes" can be
  - Categorical / Ordinal
    e.g. Male, Female, Yes, No
  - Numeric
    Discrete – e.g. 10, 25, 10000
    Continuous – e.g. 3.5, 5.9

Concept Learning - Representation of Examples/Instances

Example

Input	**Human**
Output	**Gender of a Human**
Instance	**Instance = Input + Output**

Concept Learning - Representation of Examples/Instances

Representation of Input

- Set of attributes with possible values

Attribute	Possible Values		
Height	Short	Medium	Tall
Weight	Small	Medium	Heavy
Beard	Yes	No	-

- HINT: Try to identify the most discriminating attributes for a learning problem

Representation of Output

- Single Attribute with possible values
  Gender – Male, Female

Concept Learning - Representation of Examples/Instances

- **Note the difference between**
  - **Attribute**
  - **Attribute Value / Value of Attribute**
  - **e.g. Height is an “attribute” and Tall is “value of attribute”**
- **To Summarize**
  - **Instance is a “vector of attribute values”**

Concept Learning - Representation of Examples/Instances

- **Below are 3 possible instances for the Gender Identification learning problem**

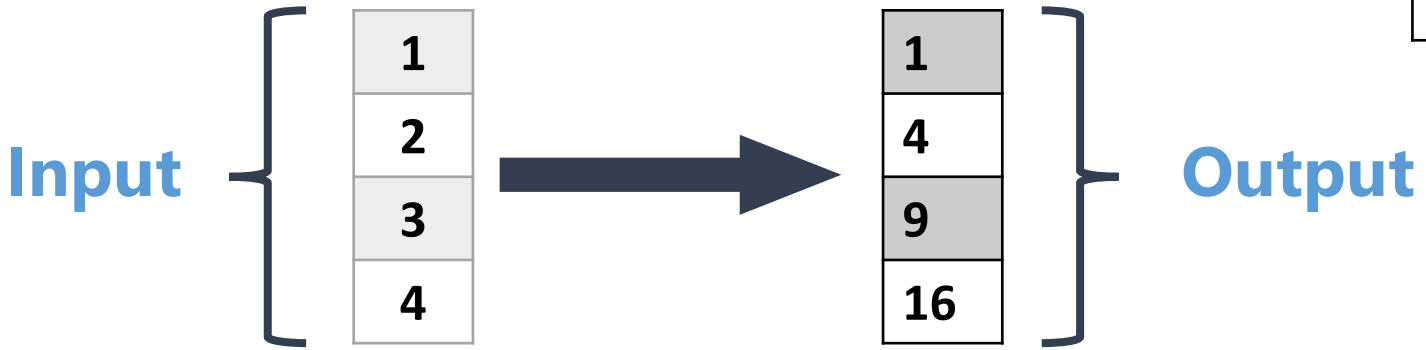
Instance no.	Height	Weight	Beard	Gender
1	Short	Medium	No	Female
2	Tall	Heavy	Yes	Male
3	Medium	Medium	No	Female

Example 01 - Learning Input-Output Functions

Input | **Single**

Output | **Single**

Instance #	Input	Output
1	1	1
2	2	4
3	3	9
4	4	16



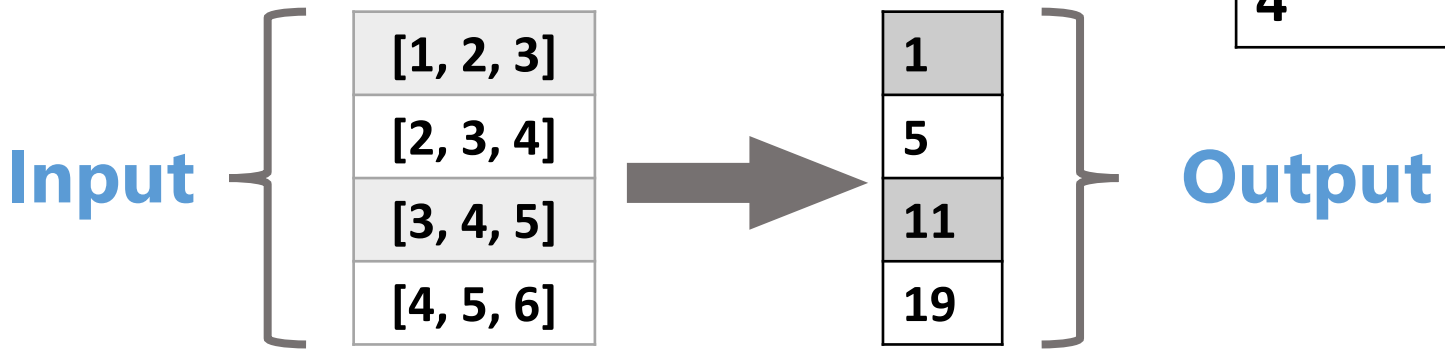
Here the concept to learn is x^2

Example 02 - Learning Input-Output Functions

Input | Vector Valued
- Vector of "input attribute values"

Output | Single

Instance #	Input	Output
1	[1,2,3]	1
2	[2,3,4]	5
3	[3,4,5]	11
4	[4,5,6]	19



Here the concept to learn is: [a,b,c] -> a*c – b

Example 03 - Learning Input-Output Functions

Input | Vector Valued
▯ Vector of "input attribute values"

Output | Single

Set of Input Vectors

Output

Instance #	Height	Weight	Hair Length	Beard	Scarf	Gender
1	180.3	196	Bald	Yes	No	Male
2	170.0	120	Long	No	No	Female
3	178.5	200	Short	No	No	Male
4	163.4	110	Medium	No	Yes	Female
5	175.2	220	Short	Yes	No	Male
6	165.0	150	Medium	No	Yes	Female
7	179.1	185	Long	Yes	No	Male
8	160.5	130	Short	No	No	Female
9	177.8	160	Bald	No	No	Male
10	161.1	100	Medium	No	No	Female

Learning Input-Output Functions - Summary

Machine Learning - Real World Examples

Task 01 – Gender Identification

Input | Human (represented as a fixed set of attributes)

Output | Gender of the person (Male, Female)

Goal | Learn from Input (human - fixed set of attributes) to predict Output (Male, Female)

Machine Learning - Real World Examples

Task 01 – Gender Identification

Examples

Instance no.	Height	Weight	Beard	Gender
1	Short	Medium	No	Female
2	Tall	Heavy	Yes	Male
3	Medium	Medium	No	Female

Summary – Gender Identification Task

- Input – Structured Data (fixed set of attributes)
- Output – Fixed Set

Machine Learning - Real World Examples

Task 02 – Gender Identification

Input | Comment / Review written by a human (Text Data)

Output | Gender of the person who wrote the comment / review (Male, Female)

Goal | Learn from Input (review / comment) to predict Output (Male, Female)

Machine Learning - Real World Examples

Task 02 – Gender Identification

Examples

Instance no.	Comment / Review	Gender
1	iPhone7 is a good mobile	Male
2	Battery of this phone is bad	Female
3	I am using iphone7	Male

Summary – Gender Identification Task

- Input – Unstructured Data (of Variable Length)
- Output – Fixed Set

Machine Learning - Real World Examples

Task 03 – Gender Identification

Input | Image (Image Data)

Output | Gender of the person in the image (Male, Female)

Goal | Learn from Input (image) to predict Output (Male, Female)

Machine Learning - Real World Examples

Task 03 – Gender Identification

Examples

Male

Female

Male

Summary – Gender Identification Task

- Input – Unstructured Data (of Variable Length) in the form of pixels (extracted from input image)
- Output – Fixed Set

Machine Learning - Real World Examples

Task 04 – Sentiment Analysis

Input | Comment / Review written by a human (text data)

Output | Polarity of the comment / review (Positive, Negative, Neutral)

Goal | Learn from Input (comment / review) to predict Output (Positive, Negative, Neutral)

Examples

Instance no.	Comment / Review	Gender
1	iPhone7 is a good mobile	Positive
2	Battery of this phone is bad	Negative
3	I am using iphone7	Neutral

Machine Learning - Real World Examples

Task 04 – Sentiment Analysis

Summary – Sentiment Analysis Task

- Input – Unstructured Data (of Variable Length)
- Output – Fixed Set

Machine Learning - Real World Examples

Task 05 – Machine Translation

Input | Text in English (Source Language)

Output | Text in Urdu (Target Language)

Goal | Learn from Input (source language) to predict Output (Translation in target language)

Machine Learning - Real World Examples

Task 05 – Machine Translation

Examples

Instance no.	Source Language	Target Language
1	Allah is one	اللہ ایک ہے۔
2	Understanding is deeper than Love.	تفہیم محبت سے زیادہ گہری ہے۔
3	The only thing that is constant in this world is change.	اس دنیا میں واحد چیز جو مستقل ہے وہ ہے تبدیلی۔

Machine Learning - Real World Examples

Task 05 – Machine Translation

Summary – Machine Translation Task

- Input – Unstructured Data (of Variable Length)
- Output – Unstructured Data (of Variable Length)

Your Turn

Write input, output and goal for following tasks

- Summarization
- Object Detection in Image
- Vehicles Categorization
- Author Identification
- Plagiarism Detection
- Speaker Identification

Types of Machine Learning

Data vs Information

Data

Raw facts and figures

Data vs Information

Varieties of Data

Structured Data	Unstructured Data	Semi-Structured Data
Data is stored, processed, and manipulated in a traditional Relational Database Management System (RDBMS)	Data that is commonly generated from human activities and doesn't fit into a structured database format	Data doesn't fit into a structured database system, but is none-the-less structured by tags that are useful for creating a form of order and hierarchy in the data

Data vs Information

- **Data**
  - **Main Forms of Data**
    - Text
    - Image
    - Video
    - Audio

- **Information**
  - Processed form of data

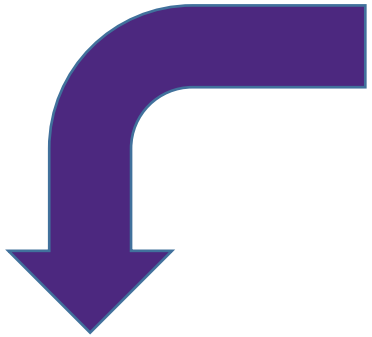
Data Annotation

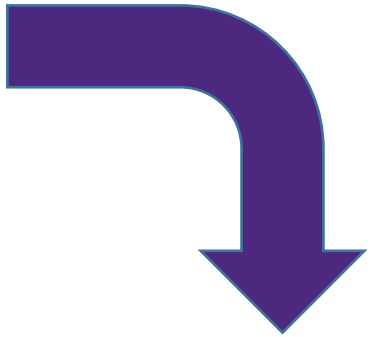
Data annotation

- a.k.a. data labeling / data tagging
- is the process of labeling data to make it usable for machine learning
- is performed by domain experts (humans – a.k.a. annotators / taggers / raters)
- requires a lot of effort, time and cost

Example 01 - Data Annotation

Raw Data
iPhone7 is a good mobile
Battery of this phone is bad
I am using iphone7





Data Annotation – Sentiment Analysis	
Comment / Review	Sentiment
iPhone7 is a good mobile	Positive
Battery of this phone is bad	Negative
I am using iphone7	Neutral

Data Annotation – Gender Identification	
Comment / Review	Gender
iPhone7 is a good mobile	Male
Battery of this phone is bad	Male
I am using iphone7	Female

Example 02 - Data Annotation

Raw Data	Data Annotation – Gender Identification	
		Male
		Female
		Male

Example 02 - Data Annotation

Raw Data	Data Annotation – Emotion Analysis	
		Angry
		Happy
		Surprise

Example 02 - Data Annotation

Raw Data	Data Annotation – Age Group Identification	
		31-60
		19-30
		10-18

Data and Annotation

- For machine learning, data is mainly available as
  - Un-annotated Data
  - Annotated Data
  - Semi-annotated Data

Data and Annotation

Un-annotated Data

- Output is not associated with the inputs

Example	
Input	**Output**
iPhone 7 is a good mobile	-
Battery of this phone is bad	-
I am using iPhone 7	-

Data and Annotation

Annotated Data

- Output is associated with all the inputs

Example	
Input	**Output**
iPhone 7 is a good mobile	Male
Battery of this phone is bad	Male
I am using iPhone 7	Female

Data and Annotation

Semi-annotated Data

- Output is associated with some of the inputs

Example	
Input	**Output**
iPhone 7 is a good mobile	Male
Battery of this phone is bad	-
I am using iPhone 7	Female

Balanced Data vs Unbalanced Data

- For more accurate learning, it is important to have balanced data
- Balanced Data

> For each class, the dataset must contain the same number of instances

Balanced Data vs Unbalanced Data (Cont...)

- **Example – Gender Identification**
  - Two Classes = Male and Female
  - Dataset = 300 examples
  - Unbalanced datasets
    - Male = 100, Female = 200
    - Male = 200, Female = 100
  - Balanced dataset
    - Male = 150, Female = 150

Types of Learning

01 Supervised Learning (or Classification)

02 Unsupervised Learning (or Clustering)

03 Semi Supervised Learning

Supervised Learning (or Classification)

If the training examples have associated output values, then the learning setting is called supervised learning

Supervised Learning (or Classification)

For Supervised Learning

- Train data – is annotated
- Test data – must be annotated

Types of Supervised Learning

Classification

- Output is categorical (or discrete)
- Example – Gender Prediction

Regression

- Output is numeric (or continuous)
- Example – House Price Prediction

Supervised Learning – Example

Set of Input Vectors

Instance #	Height	Weight	Hair Length	Beard	Scarf	Gender
1	180.3	196	Bald	Yes	No	Male
2	170.0	120	Long	No	No	Female
3	178.5	200	Short	No	No	Male
4	163.4	110	Medium	No	Yes	Female
5	175.2	220	Short	Yes	No	Male
6	165.0	150	Medium	No	Yes	Female
7	179.1	185	Long	Yes	No	Male
8	160.5	130	Short	No	No	Female
9	177.8	160	Bald	No	No	Male
10	161.1	100	Medium	No	No	Female

Output

Unsupervised Learning (or Clustering)

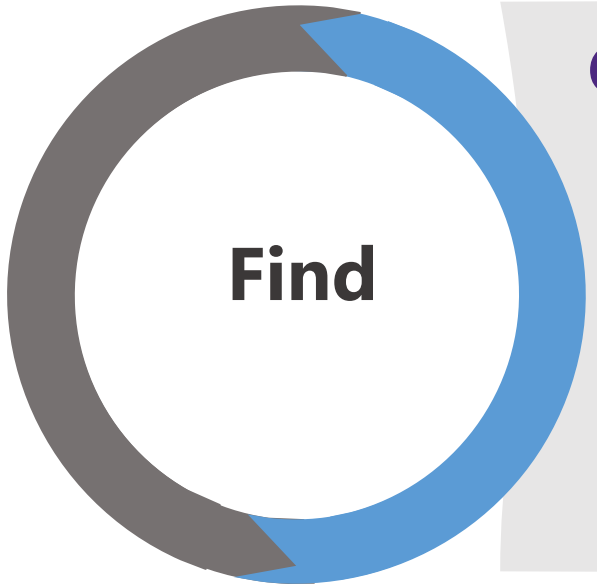
If the training examples do not have associated output values, then the learning setting is called unsupervised learning

Unsupervised Learning (or Clustering)(Cont…)

- Useful for splitting datasets into partitions, which is known as clustering
- For Unsupervised Learning
  - Train data – is not annotated
  - Test data – must be annotated

Example of Unsupervised Learning – Document Clustering

1. A set of text documents (or examples) and
2. A distance measure (Euclidean Distance)



Clusters (sports, showbiz, religion, crime, politics) such that

- Text documents in one cluster are more similar to one another
- Text documents in separate clusters are less similar to one another

Semi Supervised Learning

If some training examples have associated output values, then the learning setting is called semi supervised learning

Semi Supervised Learning (Cont…)

For Semi Supervised Learning

- Train data – is semi-annotated
- Test data – must be annotated

Summary - Points to Consider in any Learning Setting

- For accurate learning, we need
  1. Large amount of data
  2. High quality data
  3. Balanced data
- For any learning setting, test data must be annotated to evaluate the performance of a model

Phases in Machine Learning

Phases in Machine Learning

- **Machine learning has three main phases**
  1. **Training**
  2. **Testing**
  3. **Application**

Data Split

- For both training and testing phases we need data, therefore, we split the available data into
  - 01 Train Data (or Train set)
  - 02 Test Data (or Test set)
- Standard Approach for data split
  - Use 2 / 3 data as train set
  - Use 1 / 3 data as test set
- Note – Train set and test set must be disjoint i.e. example accruing in the train set should not occur in the test set and vice versa

Data Split (Cont…)

- Two main approaches to data split

Data Split (Cont…)

Example 01 – Gender Identification (Balanced Data)

- Dataset = 600 instances (Male = 300, Female = 300)
- Train-Test Split Ratio is 67%-33%

1 Random Split

- Train set = 400 instances (Male = 250, Female = 150)
- Test set = 200 instances (Male = 150, Female = 50)

2 Class Balanced Split

- Train set = 400 instances (Male = 200, Female = 200)
- Test set = 200 instances (Male = 100, Female = 100)

Data Split (Cont…)

Example 02 – Gender Identification (Unbalanced Data)

- Dataset = 900 instances (Male = 600, Female = 300)
- Train-Test Split Ratio is 67%-33%

1 Random Split

- Train set = 600 instances (Male = 500, Female = 100)
- Test set = 300 instances (Male = 100, Female = 200)

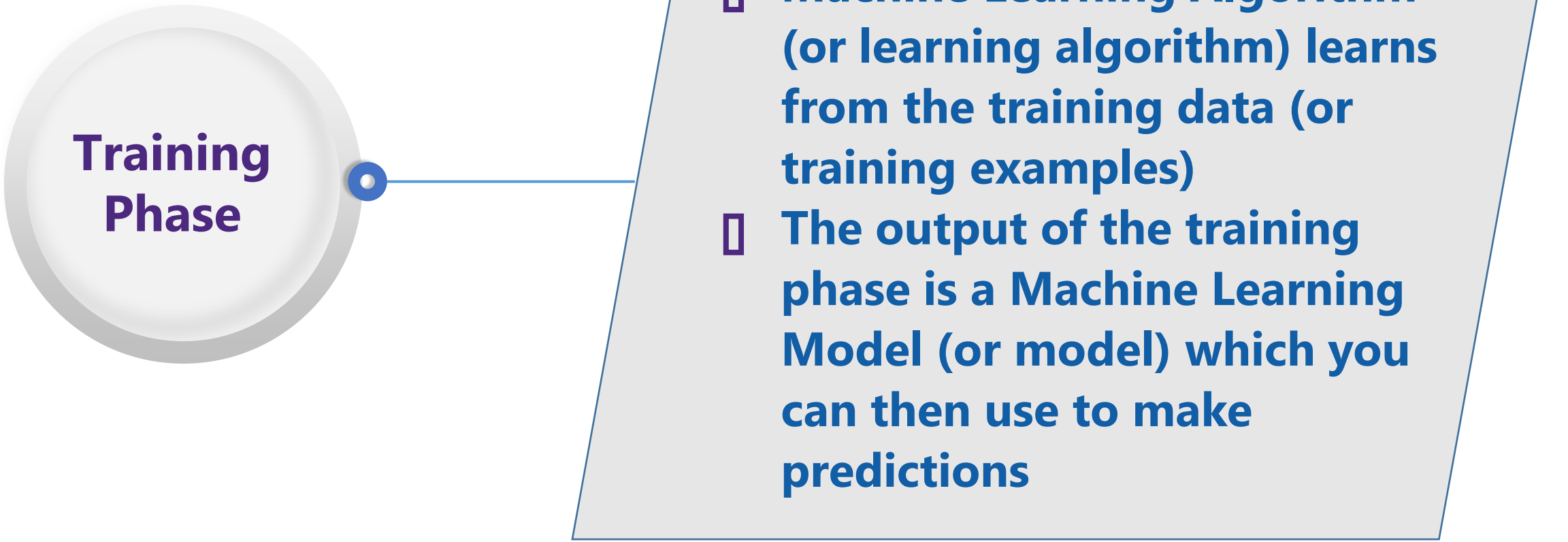
2 Class Balanced Split

- Train set = 600 instances (Male = 400, Female = 200)
- Test set = 300 instances (Male = 200, Female = 100)

Data Split (Cont…)

It is good to split data in a train-test ratio of 67%-33% using the class balanced split approach

Phases in Machine Learning



Phases in Machine Learning (Cont…)

Testing Phase

The performance of the model (created in the training phase) is evaluated on the test data using evaluation measure(s)

- Standard evaluation measures include Accuracy, Precision, Recall, F-measure, Area Under the Curve (AUC), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) etc.

Training and Testing Phases

Recall the Equation

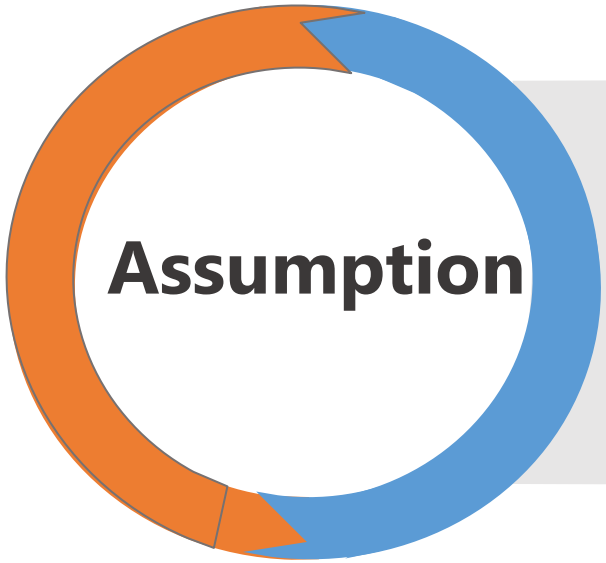
Data = Model + Error

- Training phase creates the “model” using the “data” i.e. training data
- Testing phase checks the “error” in the “model” using the test data

Application Phase

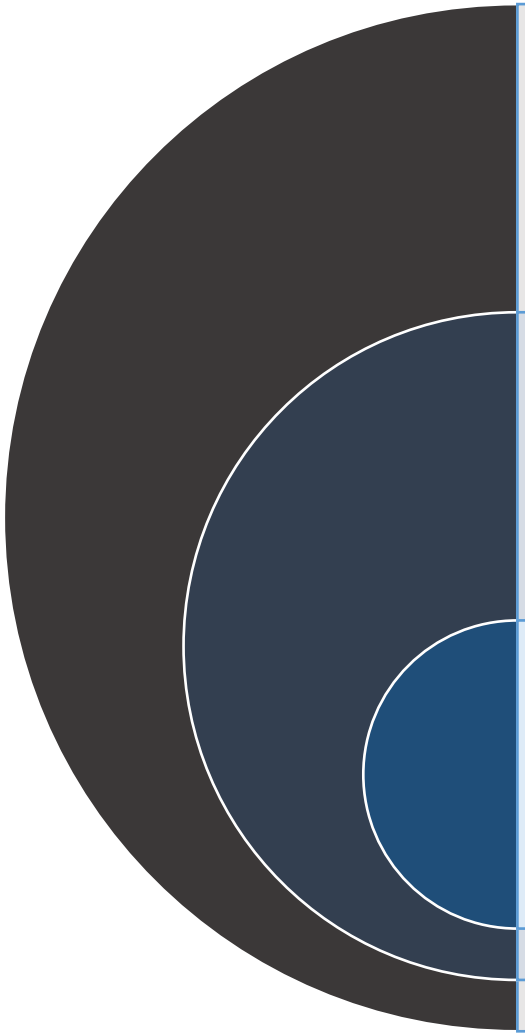
- The trained model (or model) is deployed in the real world to make predictions on new (unseen) data
- Note – If a model produces an Accuracy of 80% on a large test set then we would say that in the real world it will correctly classify 80% of unseen data (or instances)

Machine Learning – Assumption



If a trained model performs well on large test data, it will perform well on real-time data

Summary – Phases in Machine Learning



Machine learning models learn from existing data (or training data) to make predictions on new (unseen) data

Performance of a trained model is evaluated on test data using evaluation measure(s)

If a trained model performs well on large test data, it is likely to perform well in the application phase

Machine Learning - Training Regimes

Training Regimes

- Considerable variation possible in how training examples are presented/used
- Main Types of Training Regimes
  1. Batch Method
  2. Incremental Method
  3. On-line Method

Types of Training Regimes

Batch Method

- All training examples (complete dataset) are available and used all at once to compute h.
- Training a neural network on the entire MNIST dataset (60,000 images) to recognize handwritten digits.

Types of Training Regimes (Cont...)

Incremental Method

- All training examples are used iteratively to refine a current hypothesis until some stopping condition
- In the incremental method one member of the training set is selected at a time and used to modify the current hypothesis
- Members of the training set can be selected at random (with replacement) or the set can be cycled through iteratively

Types of Training Regimes (Cont...)

Incremental Method

- Incremental training is a process where a model is trained on new data in small chunks, rather than retraining the entire model from scratch.
- e.g. we have a model that can recognize 5 types of animals: dog, cat, bird, fish, and horse.
  - Initial Training- Train the model on 1000 images of these 5 animals.
    Incremental Training- Add 500 new images of 2 new animals: tiger and lion.
    Update the model to recognize these 2 new animals, without retraining the entire model. Now can recognize 7 types of animals: dog, cat, bird, fish, horse, tiger, and lion.

Types of Training Regimes (Cont.)

On-line Method

- If training instances become available one at a time and are used as they become available, the method is called an on-line method
- e.g. a robot which is learning a hypothesis from sensory inputs which controls its actions (and hence determines its future sensory inputs)
- The model learns and updates its predictions in real-time, as new data becomes available.

Direct Training vs Indirect Training

Direct Training

- Direct training involves training a model directly on the target task.

Direct Training vs Indirect Training

Direct Training

- Suppose we want to build a model that recognizes pictures of dogs.- We collect a dataset of labeled pictures of dogs and cats.- We train the model directly on this dataset to recognize dogs. The model learns to recognize dogs directly from the labeled data.

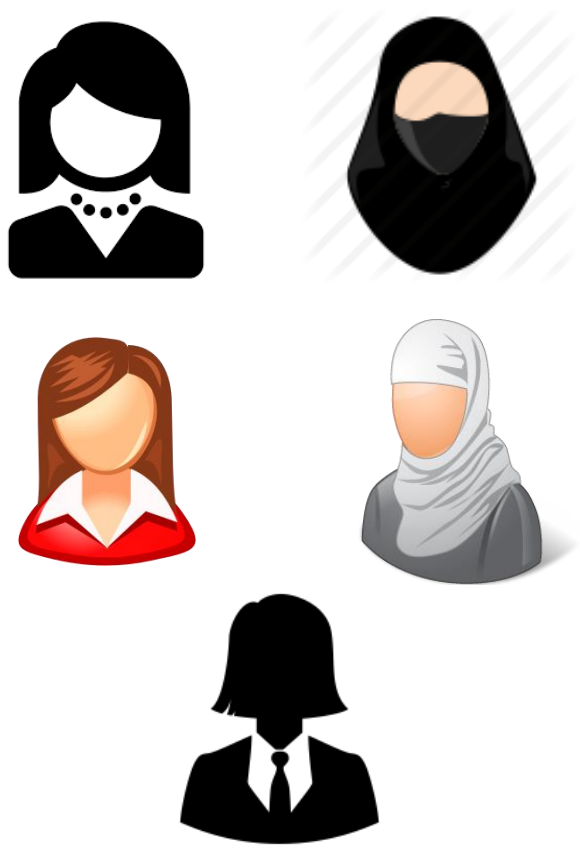
Direct Training vs Indirect Training (Cont.)

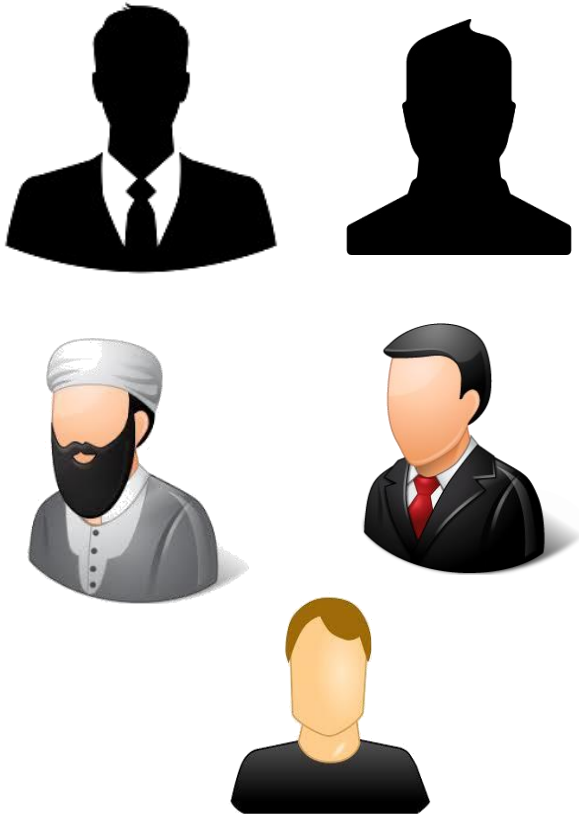
Indirect Training

- **Involves training a model on a related task, which then helps it perform better on the target task.**
- **e.g. Suppose we want to build a model that recognizes pictures of dogs, but we don't have enough labeled data.**
- **- We collect a large dataset of unlabeled pictures of animals.**
- **- We train the model on an auxiliary task, such as recognizing different animal species (e.g., birds, cats, etc.).**
- **- After training on the auxiliary task, we fine-tune the model on our small dataset of labeled dog pictures.**
- **The model learns general features about animals through indirect training, which helps it recognize dogs better when fine-tuned on the target task.**

Treating a Problem as a Machine Learning Problem – A Step by Step Example

Problem – Gender Identification





Treating a Problem as a Machine Learning Problem - Example

Treating a Problem as a Machine Learning Problem - Example

Two Main Approaches to Solve the Problem

01 Manual Approach - Involves Humans

02 Automatic Approach - Involves Machines

Limitations of Manual Approach

It is not practical when we have huge amount of data

Solution – Build automatic intelligent systems to identify the gender of a person

Types of Automatic Approach

Two main types of Automatic Approach

01 Rule Based Approach

- Humans (domain experts) manually extract rules from data

02 Machine Learning Approach (or Statistical / Probabilistic Approach)

- Machines automatically extract rules (or learn) from data

Types of Automatic Approach (Cont…)

Treat the problem of gender identification as a supervised classification task

Supervised Classification Task

- To treat the problem of gender identification as a supervised classification task, we need
  - Large amount of labeled data
  - High quality labeled data
  - Balanced data

Gender Identification - Supervised Classification Task

01 Data Collection and Preparation (Cont…)

- Data Source
- Feature Extraction from Data
- Data Split into Train / Test sets
- Selection of Machine Learning Algorithm(s) and Evaluation Measures

Gender Identification - Supervised Classification Task

A Four Step Process (Cont...)

02 Training Phase

03 Testing Phase

04 Application Phase

Gender Identification - Supervised Classification Task

A Four Step Process

01 Data Collection and Preparation

- Identification of Input and Output
- Representation of Input and Output
  - Selection of Attributes / Features
  - Selection of Attributes Values

Data Collection and Preparation

Identification of Input and Output

Problem – Gender Identification

- Input = Human
- Output = Gender

Data Collection and Preparation (Cont…)

Representation of Input and Output

- Attribute-Value Pair

Representation of Input

- Human – can be represented as a set of attributes / features
- Vector – Valued

Representation of Output

- Gender – can be represented as a single attribute
- Single – Valued

Data Collection and Preparation (Cont…)

Selection of Attributes / Features

- Problem Dependent
- Data Dependent

Two Step Process

01 List down all possible attributes / features

02 Select the most discriminating ones

Data Collection and Preparation (Cont…)

Example – Input Attributes for Gender Identification Problem

01 **A possible set of attributes / features**

- Weight, Height, Hair Length, Scarf, Beard, No. of eyes, No. of Nose, No. of hands, No. of foot

02 **Most discriminating ones may include**

- Weight, Height, Hair Length, Scarf, Beard

Data Collection and Preparation (Cont…)

Selection of Attribute Values

- Depends upon what type of information an attribute / feature carry
- Decide two things
  - Data Type
  - Possible Values

Data Collection and Preparation (Cont…)

Example – Gender Identification

Object / Instance/ Example		
Attribute	**Data Type**	**Possible Values**
Height (cm)	Numeric	Range: 160- 185
Weight (lb)	Numeric	Range: 110 – 220
Hair Length	Categorical	Bald/ Short/ Medium/ Long
Beard	Categorical	True / False
Scarf	Categorical	True / False
Gender	Categorical	Male / Female

Data Collection and Preparation (Cont…)

Data Source

- What will be your main source(s) of data
- Ensure reliability and quality of data source

Example – Gender Identification

- We may take 10 volunteers from Machine Learning class (5 Male and 5 Female) as source of data

Data Collection and Preparation (Cont…)

Feature Extraction from Data

Feature Extraction

- is the process of extracting “values of attributes” from data

Data Collection and Preparation (Cont…)

Feature Extraction from Data

- "Attribute Values" can be extracted through
  - Observation – e.g. No. of hands
  - Measurement – Height
    - Automatic Measurement
    - Manual Measurement
- Extract "values of attributes" for both "input attribute(s)" and "output attribute(s)"

Data Collection and Preparation (Cont…)

Feature Extraction from Data

Instance #	Height	Weight	Hair Length	Beard	Scarf	Gender
1	180.3	196	Bald	Yes	No	Male
2	170.0	120	Long	No	No	Female
3	178.5	200	Short	No	No	Male
4	163.4	110	Medium	No	Yes	Female
5	175.2	220	Short	Yes	No	Male
6	165.0	150	Medium	No	Yes	Female
7	179.1	185	Long	Yes	No	Male
8	160.5	130	Short	No	No	Female
9	177.8	160	Bald	No	No	Male
10	161.1	100	Medium	No	No	Female

Data Collection and Preparation (Cont…)

Data Split into Train / Test Sets

Use a Train-Test split ratio of 67%-33% to split data using the “class balanced split” approach

Data Collection and Preparation (Cont…)

Data Split into Train / Test Sets

Train set

Test set

Instance #	Height	Weight	Hair Length	Beard	Scarf	Gender
1	180.3	196	Bald	Yes	No	Male
2	170.0	120	Long	No	No	Female
3	178.5	200	Short	No	No	Male
4	163.4	110	Medium	No	Yes	Female
5	175.2	220	Short	Yes	No	Male
6	165.0	150	Medium	No	Yes	Female
7	179.1	185	Long	Yes	No	Male
8	160.5	130	Short	No	No	Female
9	177.8	160	Bald	No	No	Male
10	161.1	100	Medium	No	No	Female

Whole Data

Data Collection and Preparation (Cont…)

Selection of Machine Learning Algorithms and Evaluation Measures

Selection of Machine Learning Algorithms

- Depends on type of data
- Example
  - For textual data - Naïve Bayes, Random Forest etc. are reported to be more suitable
  - For image data - Neural Networks are reported to be more suitable

Data Collection and Preparation (Cont…)

Selection of Machine Learning Algorithms and Evaluation Measures

Selection of Evaluation Measures

- Depends on class distribution
  - For balanced data – Accuracy is more suitable
  - For unbalanced data – F_1 and Area Under the Curve are more suitable

Data Collection and Preparation (Cont…)

Selection of Machine Learning Algorithms and Evaluation Measures

Selection of Evaluation Measures (Cont…)

- Depends on Learning Setting
  - For Classification tasks – Accuracy, F_1 etc. are more suitable
  - For Regression tasks – Mean Absolute Error, Root Mean Square Error etc. are more suitable

Data Collection and Preparation (Cont…)

Selection of Machine Learning Algorithms and Evaluation Measures

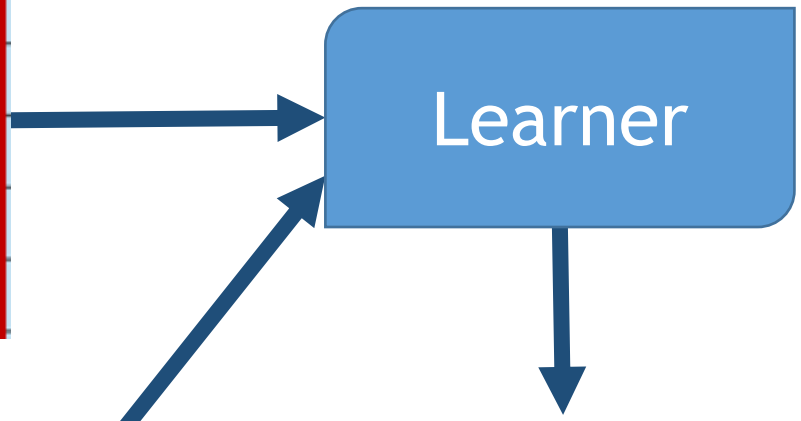
Selection of Evaluation Measures (Cont…)

- Depends on Learning Problem
  - For Gender Identification – Accuracy is more suitable (assuming balanced data)
  - For Plagiarism Detection – Precision, Recall and F_1 are more suitable

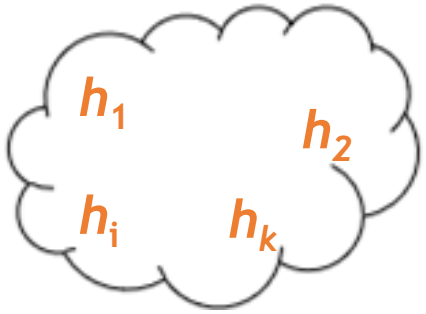
Training Phase

- Training Example = $\{x_1, \ldots. x_m\}$ +$f(x_i)$ for each $x_i \in TE$

180.3	196	Bald	Yes	No	Male
170.0	120	Long	No	No	Female
178.5	200	Short	No	No	Male
163.4	110	Medium	No	Yes	Female
175.2	220	Short	Yes	No	Male
165.0	150	Medium	No	Yes	Female



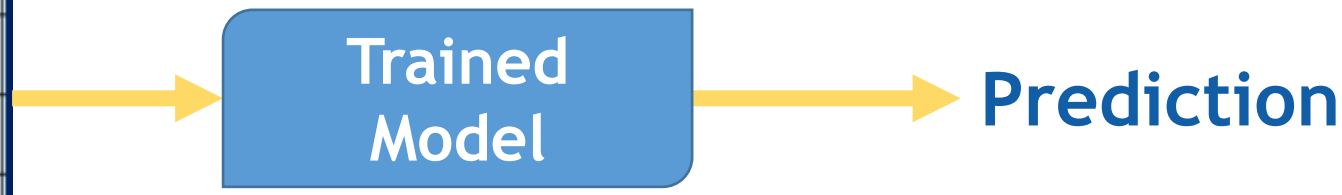
Hypothesis Space



h, where h ≈ f
h best fits the training data

Testing Phase

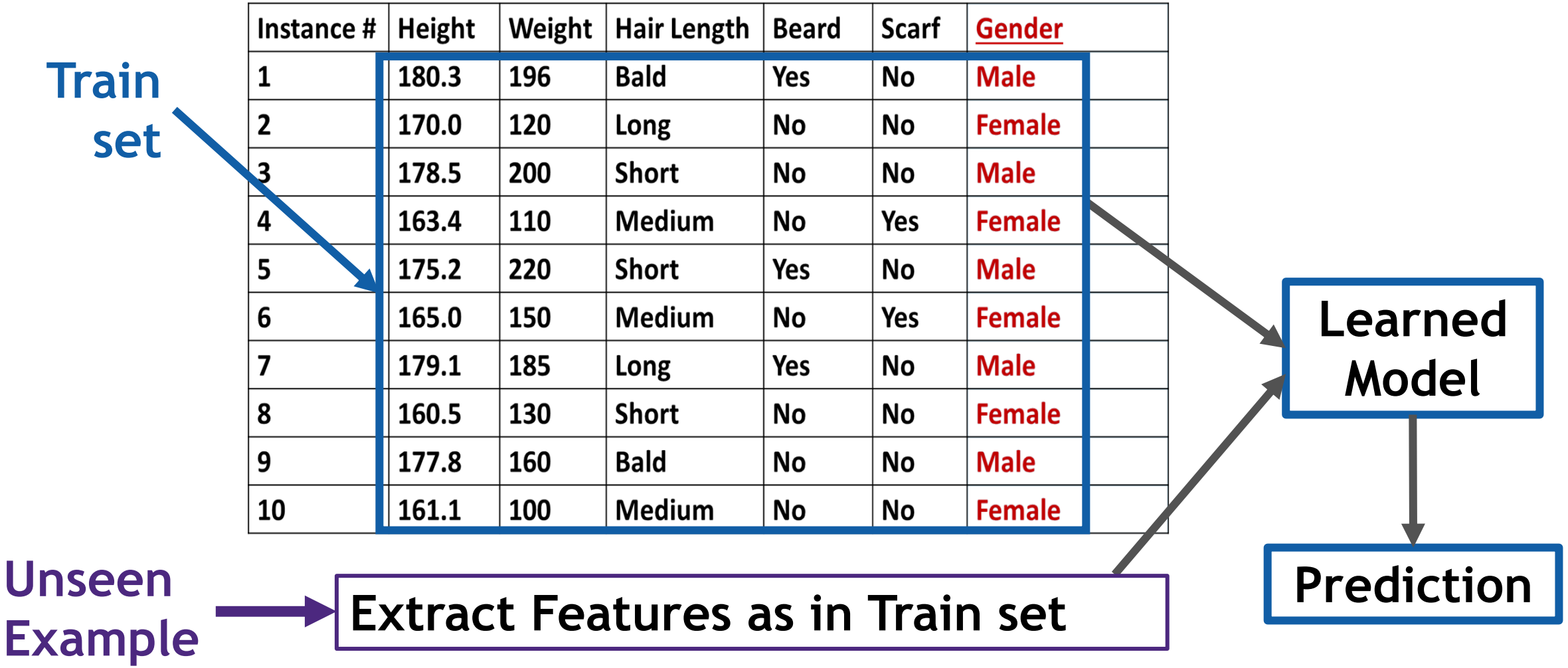
179.1	185	Long	Yes	No
160.5	130	Short	No	No
177.8	160	Bald	No	No
161.1	100	Medium	No	No



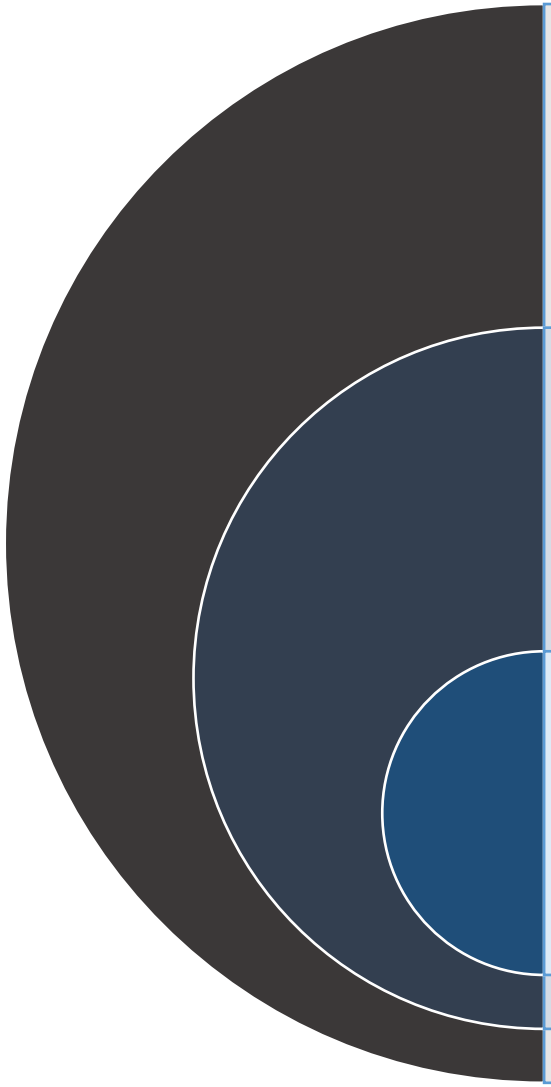
					Actual	Predicted	Accuracy
179.1	185	Long	Yes	No	Male	Male	1
160.5	130	Short	No	No	Female	Male	0
177.8	160	Bald	No	No	Male	Female	0
161.1	100	Medium	No	No	Female	Female	1

Accuracy = Correctly Classified / Total instances = 2 / 4 = 0.50 or 50%

Application Phase



Lecture Summary

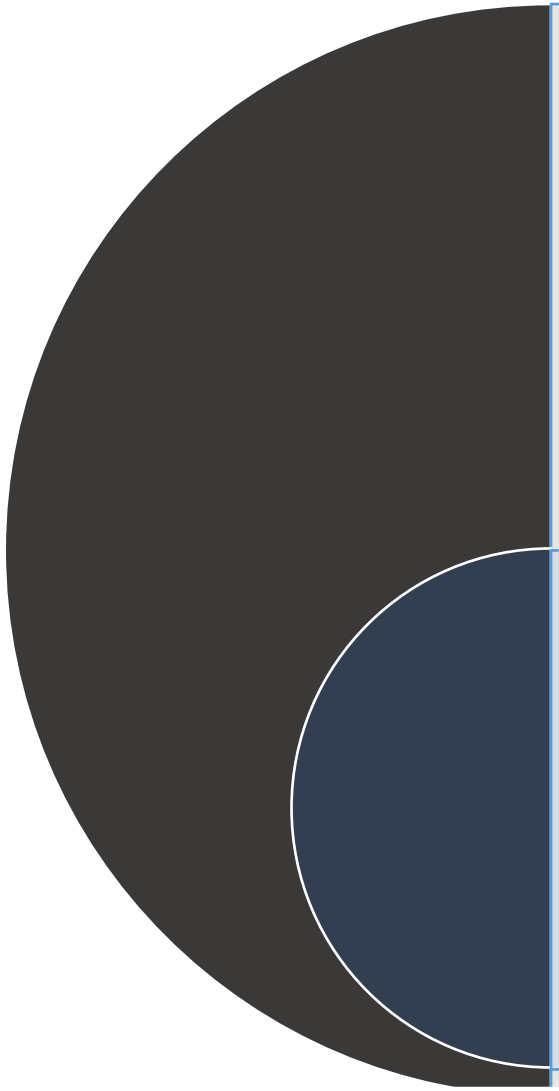


The goal of Machine Learning algorithms is to learn from existing data to make predictions on new (unseen) data

Majority of Machine Learning involves learning input-output functions i.e. learn from input to predict output

For any learning setting, we need large amount of data, high quality data and preferably balanced data

Lecture Summary



The three main phases of machine learning are:

(1) Training – learn from training data and output a model, (2) Testing – evaluate the performance of the model on test data and (3) Application – deploy the trained model in the real world

Machine learning can also be summarized in the following equation:

Data = Model + Error