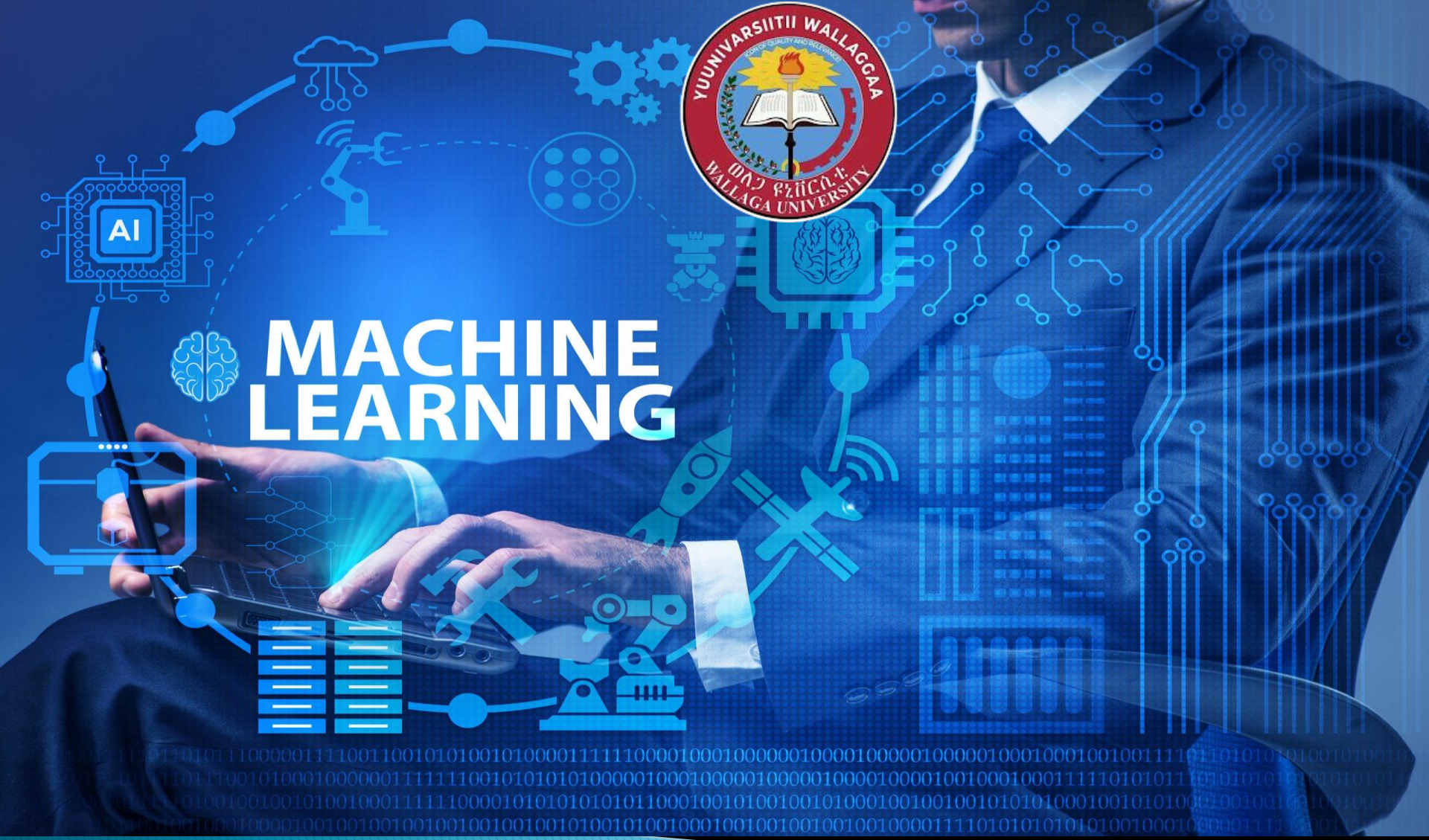


By: Dr. Etana Fikadu(PhD)
Computer Science

Foundations of Machine Learning

Outline

- Introduction to Machine Learning
  - What is machine learning?
  - Why machine learning
  - How machine learning relates to other fields
- Data, Types of Data, Datasets , and Applications in Machine Learning
- Machine Learning : Models and methods
  - Supervised
  - Unsupervised
  - Semi-supervised
  - Reinforcement Learning

» Introduction to Machine Learning

By Dr. Etana F. (PhD)

Intelligence

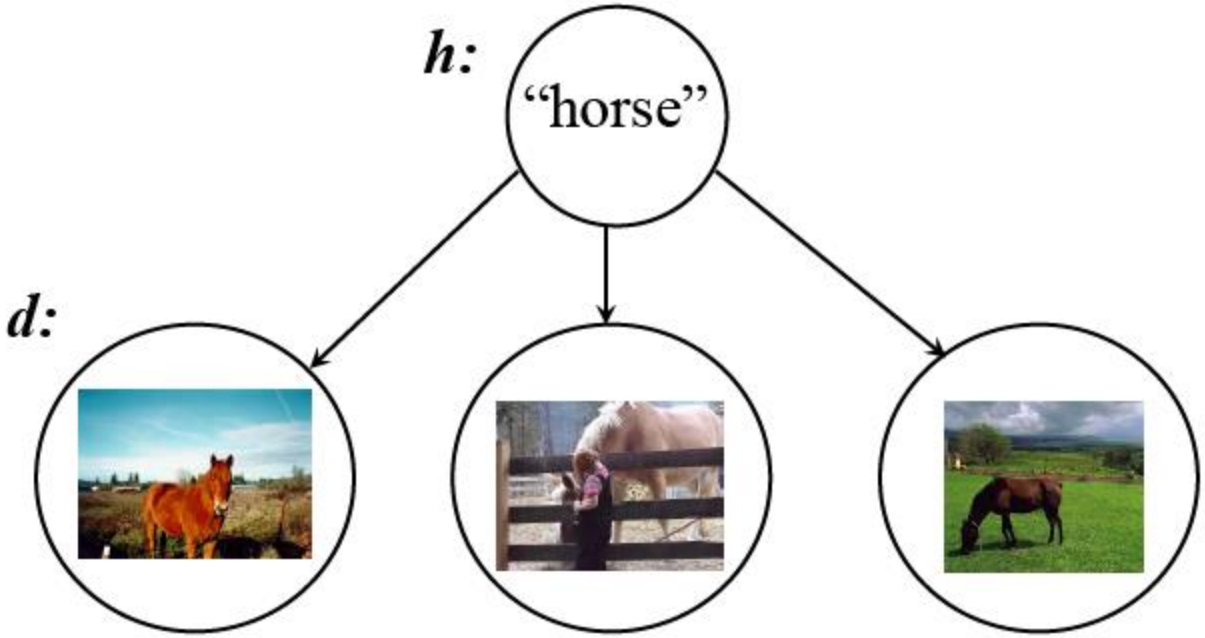
- Ability for abstract thought, understanding, communication, reasoning, planning, emotional intelligence, problem solving, **learning**
- The ability to learn and/or adapt is generally considered a hallmark of intelligence

Learning and Machine Learning

- ``*Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task(s) drawn from the same population more efficiently and more effectively the next time."--Herbert Simon*
- Machine Learning is concerned with the development of algorithms and techniques that allow computers to *learn*.

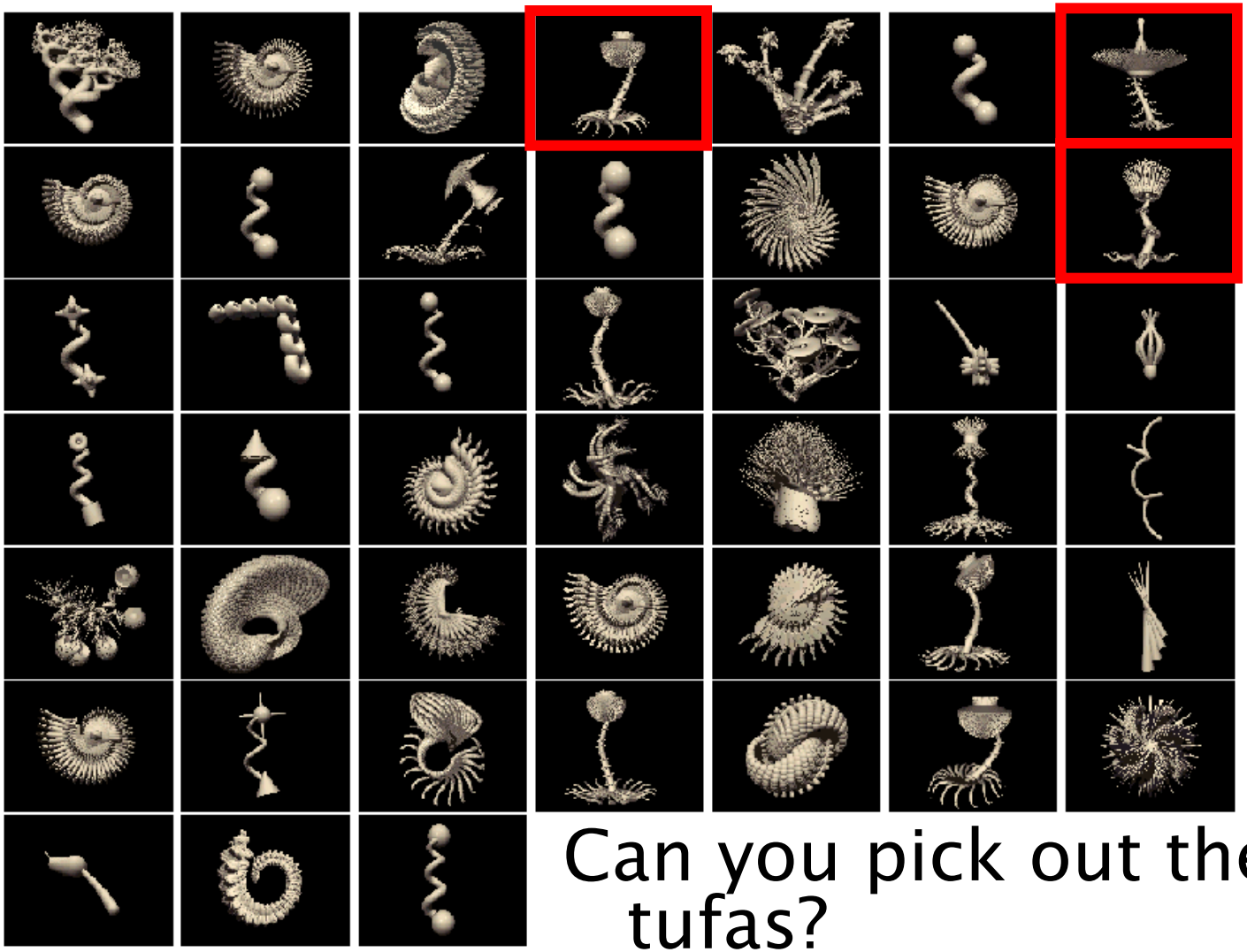
Machine Learning

- "Machine learning studies the process of constructing abstractions (features, concepts, functions, relations and ways of acting) automatically from data."



E.g.: Learning concepts and words

"tufa"



"tufa"

"tufa"

Can you pick out the tufas?

By Dr. Etana F. (PhD) Source: *Josh Tenenba*

Why Machine Learning ?

- Human expertise does not exist (e.g. *Martian exploration*)
- Humans cannot explain their expertise or reduce it to a rule set, or their explanation is incomplete and needs tuning (e.g. *speech recognition*)
- Situation changing in time (e.g. *spam/junk email*)
- Humans are expensive to train up (e.g. *zipcode recognition*)
- There are large amounts of data (e.g. *discover astronomical objects*)

Data, Types of Data, Datasets , and Applications in Machine Learning

By Dr. Etana F. (PhD)

Data Data Everywhere …

Google

YouTube

- Library of Congress text database of ~20 TB
- AT&T 323 TB, 1.9 trillion phone call records.
- World of Warcraft utilizes 1.3 PB of storage to maintain its game.
- Avatar movie reported to have taken over 1 PB of local storage at *WetaDigital for the rendering of the 3D CGI effects.*
- Google processes ~24 PB of data per day.
- YouTube: 24 hours of video uploaded every minute. More video is uploaded in 60 days than all 3 major US networks created in 60 years. According to *Cisco, internet video will generate over 18 EB of traffic per month in 2013.*

What is Data?

- Raw information used by machine learning models.
- Can be numbers, text, images, audio, or sensor data.
- ML learns patterns from examples, not rules.

Structured Data

- Organized into rows and columns.

Examples

- Student information (Name, Age, Score)
- Hospital data (Patient ID, Diagnosis, Temperature, BP)
- Bank records (Transaction Amount, Balance)

Age	Blood Pressure	Diabetes
45	134/85	Yes
67	150/90	No

Used in:

- Classification
- Regression
- Prediction models

Unstructured Data

Data that does **not follow a fixed table format**.

Examples

- ✓ Text (tweets, reviews, emails)
- ✓ Images (X-rays, faces)
- ✓ Audio (speech recordings)
- ✓ Video (lectures)

Used in:

- ➢ NLP (Natural Language Processing)
- ➢ Computer Vision
- ➢ Speech recognition

Numerical Data

Numerical: Continuous (weight), Discrete (#children).

Categorical Data

- Nominal(unordered) :-gender: Male/Female

  -Color: Red/Blue/Green

- Ordinal (ordered) :-education level: High school < Bachelor < Master).

Time-Series Data

Data collected over time at regular intervals.

Examples:

- Stock prices
- Weather tracking
- Heart-rate monitor readings

Text Data

Sentences, reviews, comments, documents.

Example:

- ✓ *"The service was very good and fast."* → Sentiment = Positive

Image Data

Pictures stored as pixels.

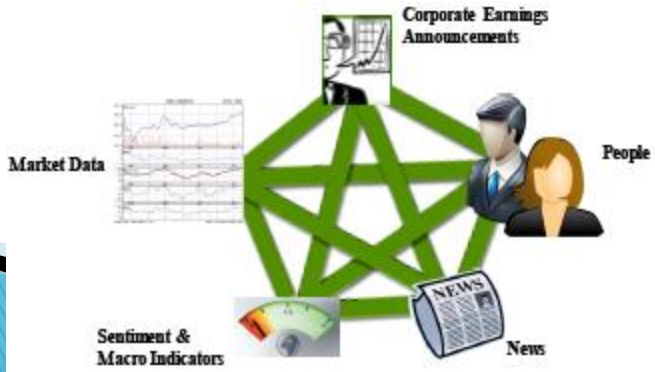
Example tasks:

- ➢ Disease detection from X-rays
- ➢ Face recognition
- ➢ Object detection

Machine Learning to the rescue

- Machine Learning is one of the front-line technologies to handle Information Overload
- **Business**
  - Mining correlations, trends, spatio-temporal predictions.
  - Efficient supply chain management.
  - Opinion mining and sentiment analysis.
  - Recommender systems.

amazon

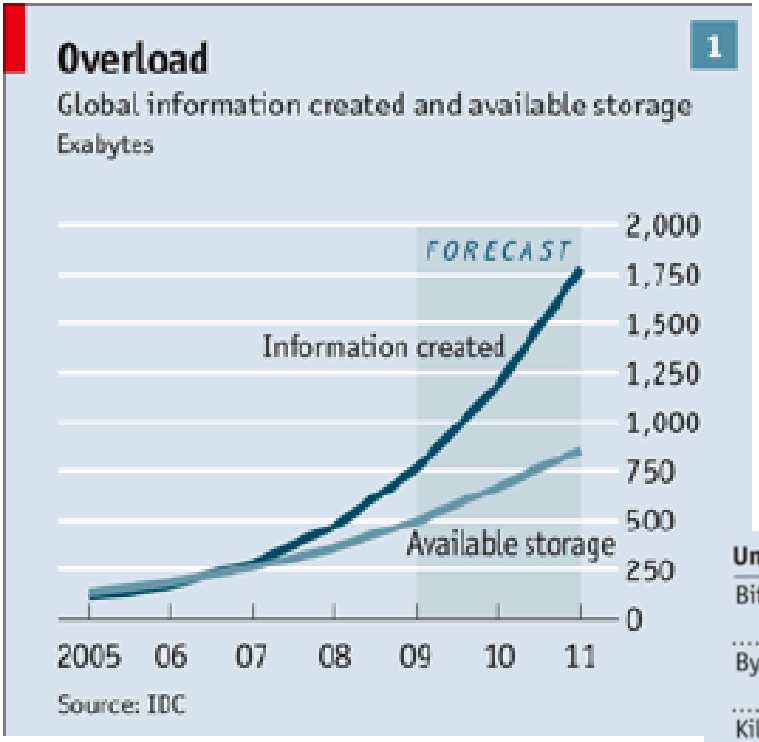


ebay market data reseller

Walmart Save money. Live better.

By Dr. Etana F. (PhD)

Information Overload

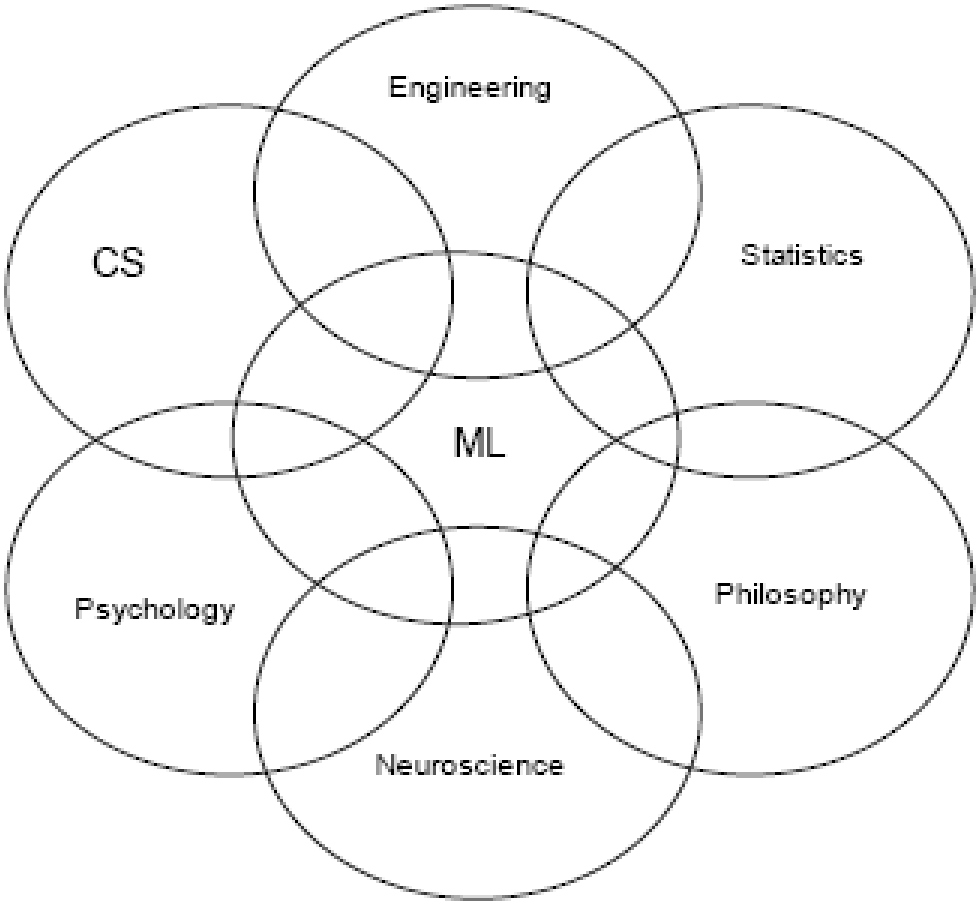


Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Fields related to Machine Learning



By Dr. Etana F. (PhD)

Fields related to Machine Learning

- Artificial Intelligence: computational intelligence
- Data Mining: searching through large volumes of data
- Neural Networks: neural/brain inspired methods
- Signal Processing: signals, video, speech, image
- Pattern Recognition: labeling data
- Robotics: building autonomous robots

Application of Machine Learning

- ML has become an essential tool across all engineering disciplines, transforming design, manufacturing, and maintenance processes by enabling systems to learn from massive amounts of data.

Manufacturing & Industrial Engineering

The primary goal is improving efficiency, quality, and uptime in production environments.

Application	ML Example	Explanation	Visualization
Predictive Maintenance (PdM)	Regression (e.g., Random Forests, RNNs)	Predicting the Remaining Useful Life (RUL) of equipment (e.g., pumps, motors, robotic arms) based on vibration, temperature, and current sensor data. This shifts maintenance from reactive or scheduled to just-in-time.	Time Series Plot: A graph showing the sensor value trending over time with a model-predicted line representing the RUL countdown.
Quality Control	Computer Vision (CNNs)	Automatically detecting microscopic defects (e.g., cracks, scratches, misalignments) in manufactured components (like circuit boards or metal parts) using high-resolution camera images.	Bounding Box Visualization: An image of a part with a red bounding box overlaid on the exact location of the detected defect.
Process Optimization • • •	Reinforcement Learning (RL)	Training an autonomous agent to control complex processes (like chemical mixing or furnace temperature) by optimizing controls to maximize yield or minimize energy usage in real-time.	Reward History Chart: A plot showing the agent's cumulative reward increasing over training episodes, indicating learning progress.

Cont….

Deep Blue and the chess Challenge

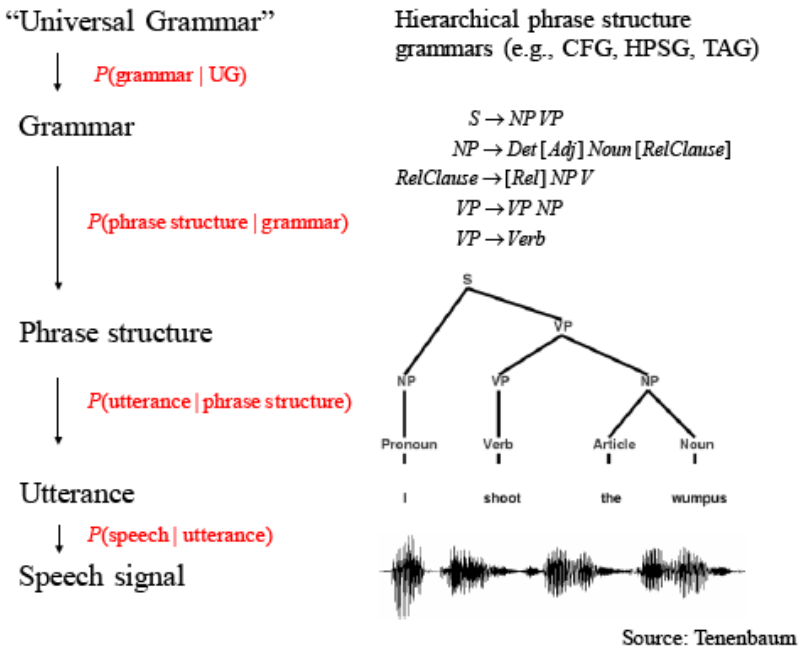
Source:
RoboCup web site

RoboCup

Online Poker

By Dr. Etana F. (PhD)

Application of Machine Learning



Source: Tenenbaum

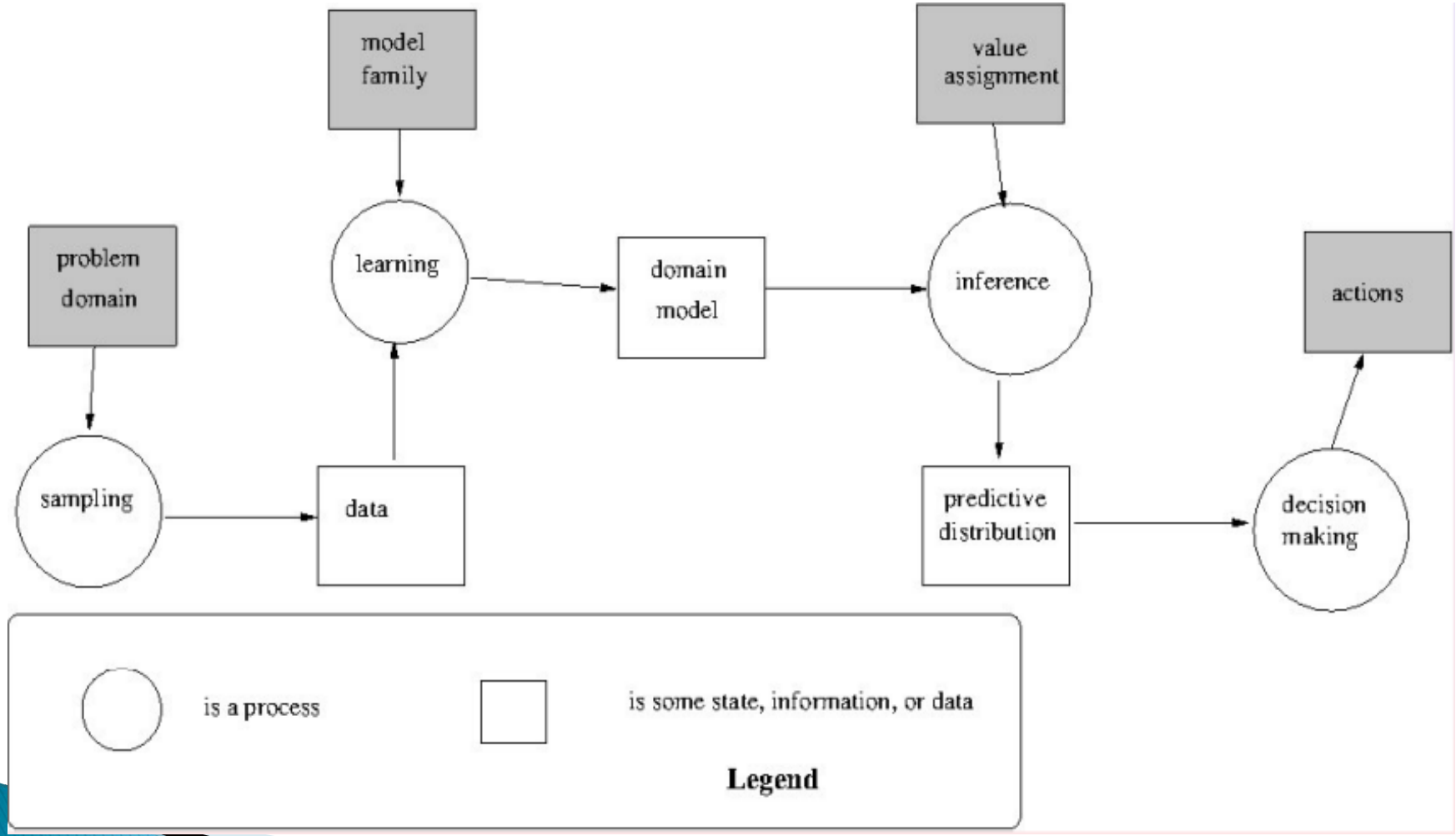
- Application in speech and Natural Language processing
- Probabilistic Context Free Grammars

- Graphical Models
- Social network graph analysis, causality analysis

Machine Learning Models and Methods

By Dr. Etana F. (PhD)

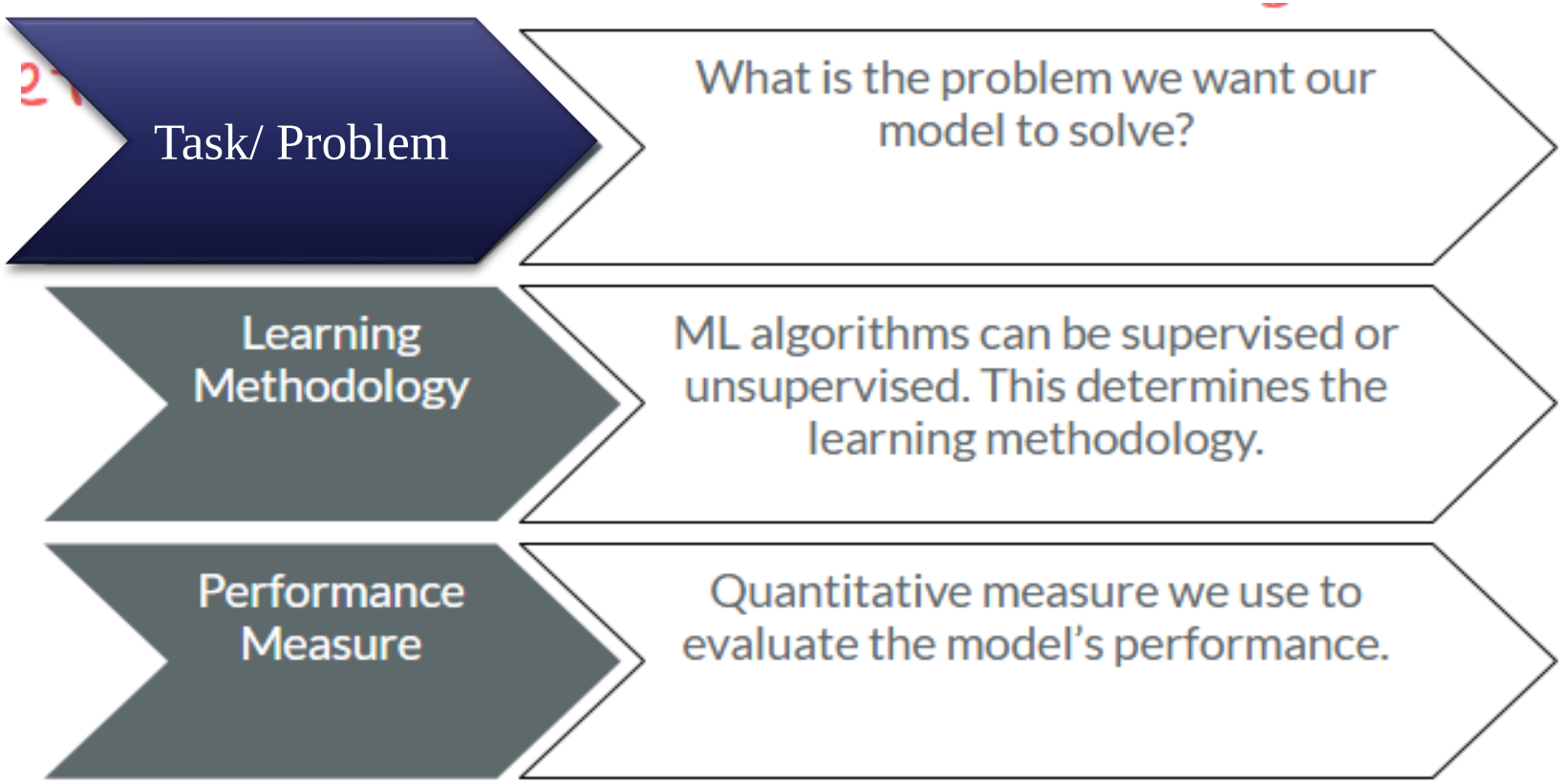
Machine Learning Process



By Dr. Etana F. (PhD)

Conti..

All ML models have 3 key components: a task, a performance measure and a learning methodology.



Source: Module 1: Introduction to Machine Learning - Google Slides

Learning (Formally)

- **Task**
  - To apply some machine learning method to the data obtained from a given domain (*Training Data*)
  - The domain has some characteristics, which we are trying to learn (*Model*)
- **Objective**
  - To minimise the error in prediction
- **Types of Learning**
  - Supervised Learning
  - Unsupervised Learning
  - Semi-Supervised Learning
  - Reinforcement Learning

Supervised Learning

- Classification / Regression problem
- Where some samples of data (Training data) with the correct class labels are provided.
  - i.e. Some correspondence between input (**X**) & output (**Y**) given
- Using knowledge from training data, the classifier/ regressor model is learnt
  - i.e. Learn some function **f : f(X) = Y**
- **f** may be probabilistic/deterministic
- Learning the model ≡ Fitting the parameters of model to minimise prediction error
- Model can then be tested on test-data

Regression

- Regression is a statistical measurement used in finance, investing, and other disciplines that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables or independent variable

Linear regression

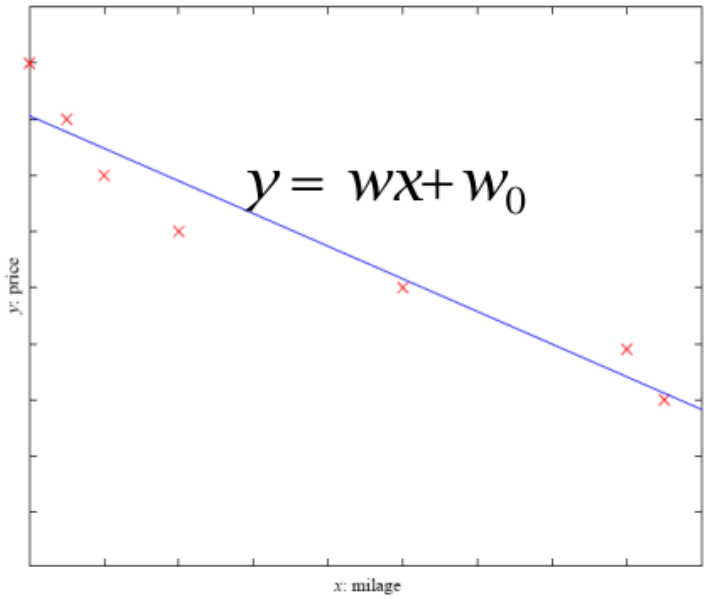
- Example: Price of a used car
- x : car attributes

  y : price

  $y = g\ (x,\theta)$

  $g\ (\ )$ model,

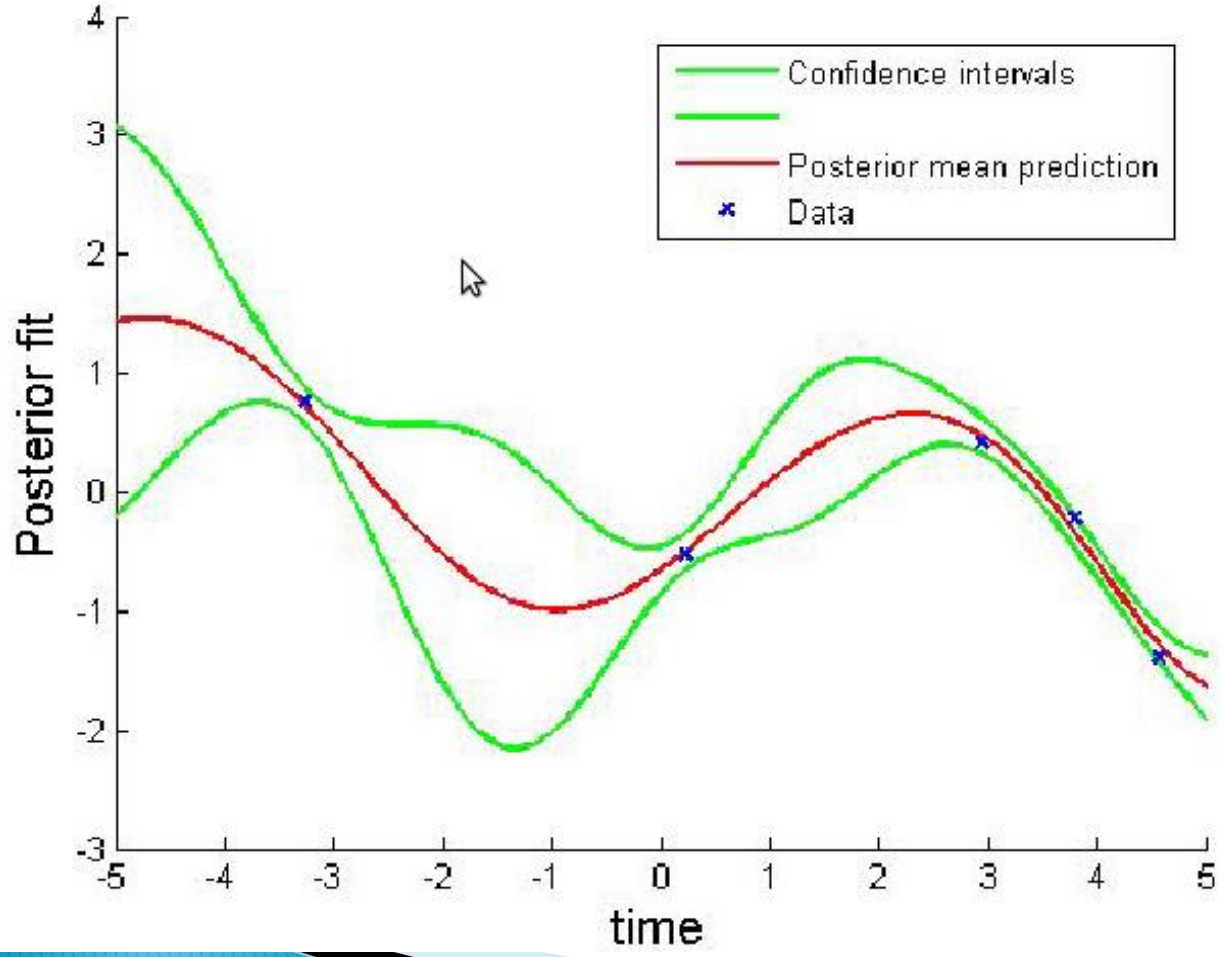
  $\theta = (w,w_0)$ parameters (slope and intercept)



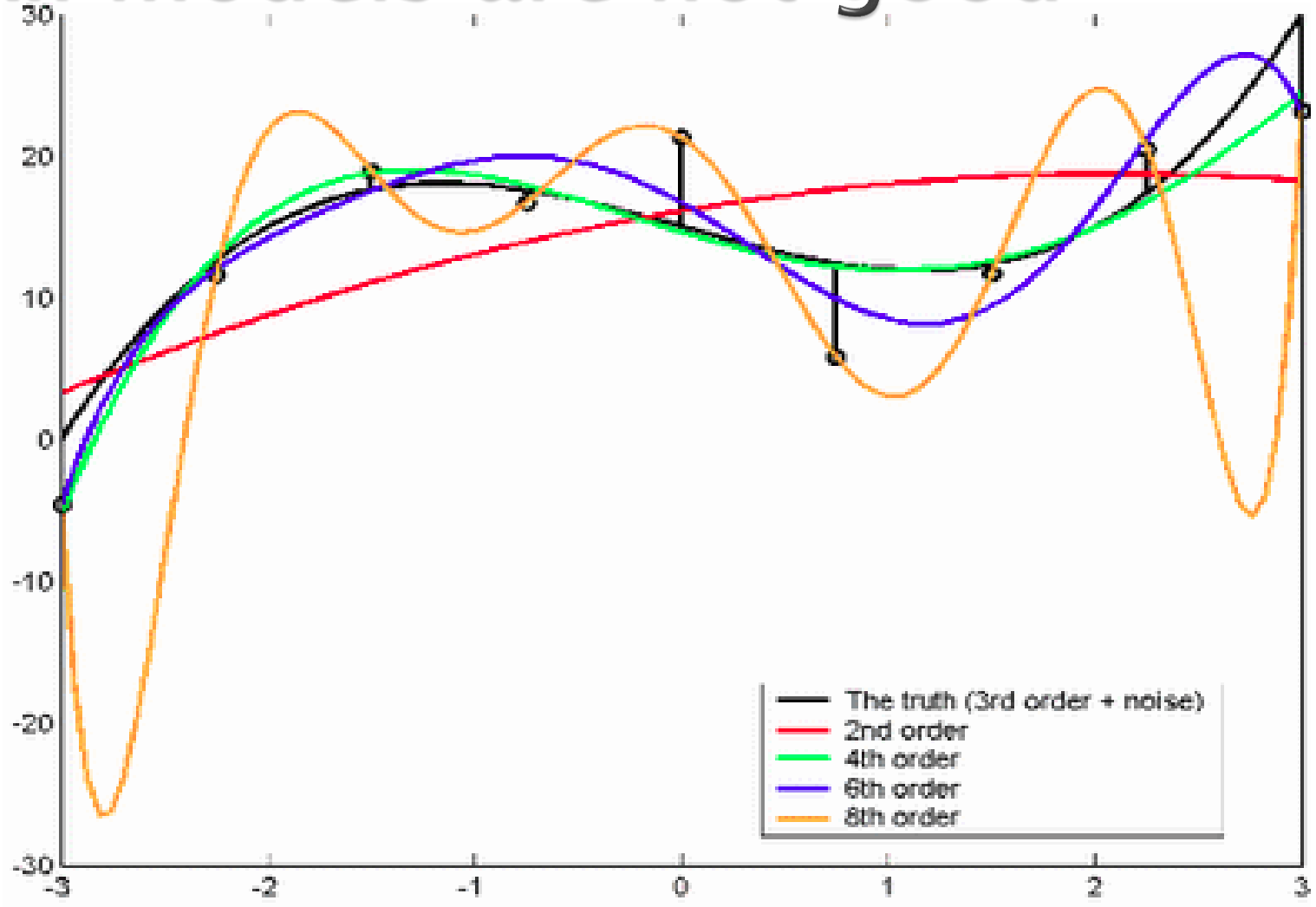
- Uses
  - Stock Prediction
  - Outlier detection

Regression

- Non Linear regression



All models are not good



- Constrain the parameters

By Dr. Etana F. (PhD)

Classification

Pattern recognition

- Glasses, beard, make-up, hair style
- Different handwriting styles.
- Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- From symptoms to illnesses
- Predict if a user clicks on an ad on the Internet.

Supervised Classification example

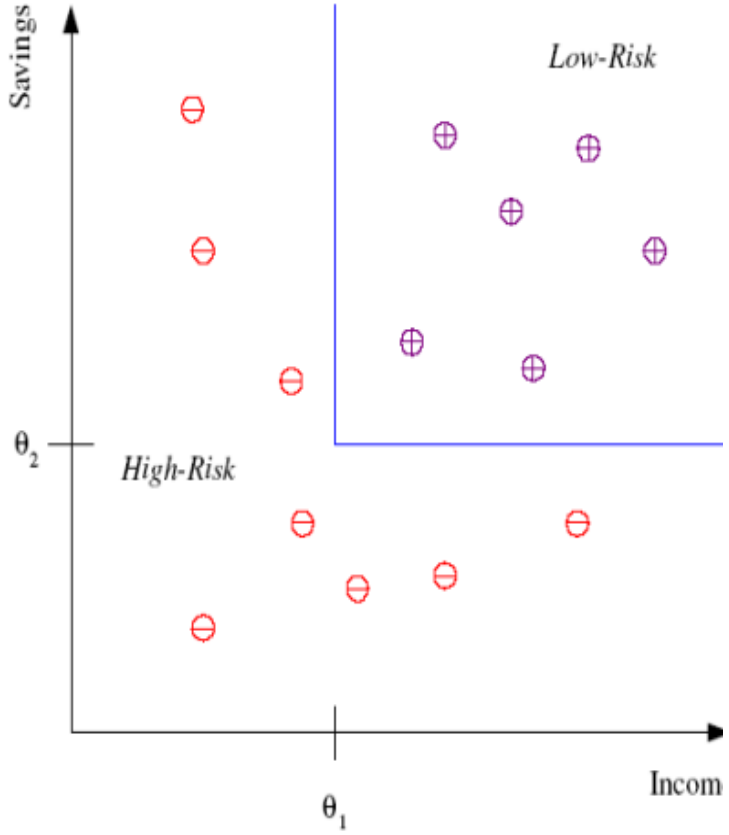
	f1	f2	f3	f4	Class label
d1					BearHead
d2					???
d3					LionHead

By Dr. Etana F. (PhD)

Source: LHI Animal Faces Dataset

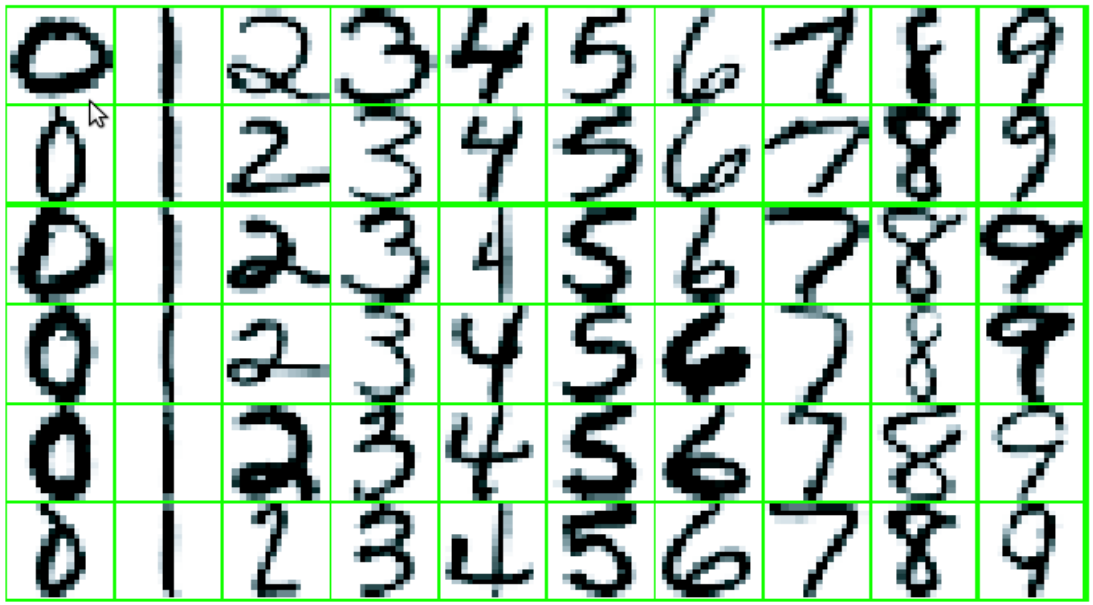
Classification

- Example:
  - Credit Scoring
- Goal:
  - Differentiating between high-risk and low-risk customers based on their income and savings
- Discriminant:
  - IF **income** > θ_1 AND **savings** > θ_2 THEN **low-risk** ELSE **high-risk**
- Discriminant is called '**hypothesis**'
- Input attribute space is called '**Feature Space**'
- Here Input data is 2-dimensional and the output is binary



Application

Example: Handwritten digit recognition for postal codes



Common Supervised Learning algorithms (Learning from Labeled Data)

Algorithm	Type	Explanation
Linear Regression	Regression	➢ Models the linear relationship between features and a continuous output by fitting a straight line or plane.
Logistic Regression	Classification	➢ Uses a sigmoid function to estimate the probability of a data point belonging to one of two classes.
K-Nearest Neighbors (KNN)	Classification/Regression	➢ Classifies a new point based on the majority class (or average value) of its 'K' closest data points in the feature space.
Support Vector Machine (SVM)	Classification/Regression	➢ Finds the optimal hyperplane that maximizes the margin (distance) between different classes.
Decision Tree	Classification/Regression	➢ Creates a flowchart-like structure where each node splits the data based on a feature test to reach a final decision (leaf).
Random Forest	Ensemble	➢ An ensemble method that combines the predictions of multiple individual Decision Trees to improve accuracy and control overfitting.
Naive Bayes	Classification	➢ A probabilistic algorithm based on Bayes' Theorem that assumes all features are independent of each other (the "naive" assumption).

Deep Learning (Using Neural Networks with Many Layers)

Algorithm	Type	Explanation
Convolutional Neural Network (CNN)	Feature Learning	✓ Uses convolutional layers to automatically learn spatial hierarchies of features from grid-like data, highly effective for image processing.
Recurrent Neural Network (RNN) / LSTM	Sequence Modeling	✓ Networks with loops that allow information from previous steps to persist, making them ideal for sequential data like text or time series. LSTMs solve the vanishing gradient problem in standard RNNs.
Transformer	Sequence Modeling	✓ Relies entirely on the Attention Mechanism to weigh the importance of different parts of the input sequence, forming the basis of most modern large language models (LLMs).

nature › scientific reports › articles › article

Article | Open access | Published: 16 May 2024

AI-based disease category prediction model symptoms from low-resource Ethiopian lang Afaan Oromo text

Etana Fikadu Dinsa, Mrinal Das & Teklu Urgessa Abebe

Scientific Reports **14**, Article number: 11233 (2024) | Cite this article

8531 Accesses | **9** Citations | Metrics

Home › Discover Applied Sciences › Article

MLP-SVM: a hybrid approach for improvin the performance of the classification mod for health-related documents from social media using multi-layer perceptron and support vector machine

Research | Open access | Published: 09 April 2025

Volume 7, article number 320, (2025) Cite this article

You have full access to this open access article

Download PDF

Home › Discover Computing › Article

A supervised learning approach for recommending medical specialists in t healthcare sector for the Afaan Oromo context

Research | Open access | Published: 24 April 2025

Volume 28, article number 46, (2025) Cite this article

You have full access to this open access article

Download PDF

Etana Fikadu, Mrinal Das, Teklu Urgessa & Krishnaraj Ramaswamy

Etana Fikadu, Teklu Urgessa & Mrinal Das

Home › Discover Applied Sciences › Article

Automatic categorization of medical documents in Afaan Oromo using ense machine learning techniques

Research | Open access | Published: 28 October 2024

Volume 6, article number 575, (2024) Cite this article

You have full access to this open access article

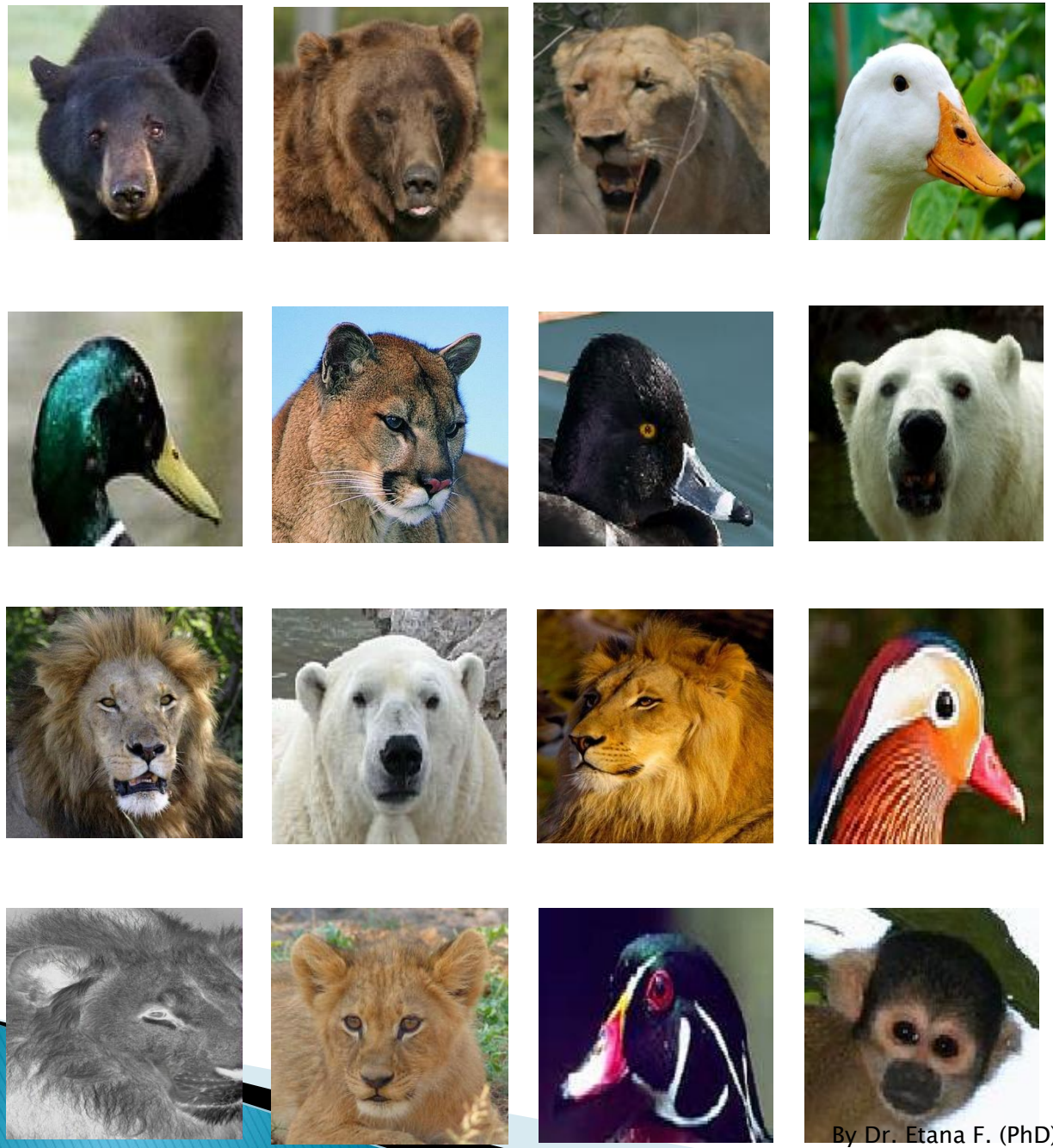
Download PDF

Etana Fikadu Dinsa, Mrinal Das, Teklu Urgessa Abebe & Krishnaraj Ramaswamy

By Dr. Etana F. (PhD)

Unsupervised Learning

- Labels may be too expensive to generate or may be completely unknown
- There is lots of training data but with no class labels assigned to it



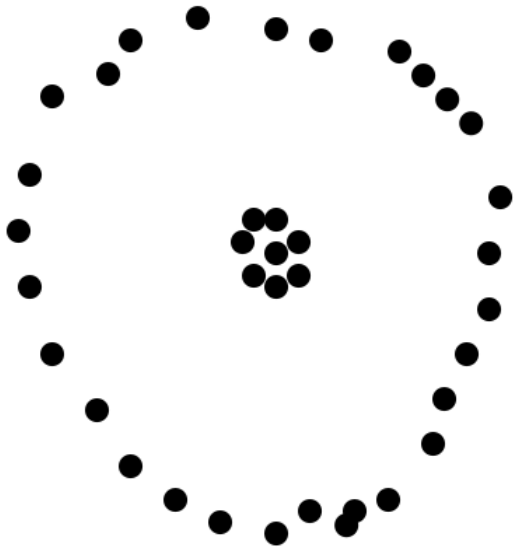
???

Source: LHI Animal Faces Dataset

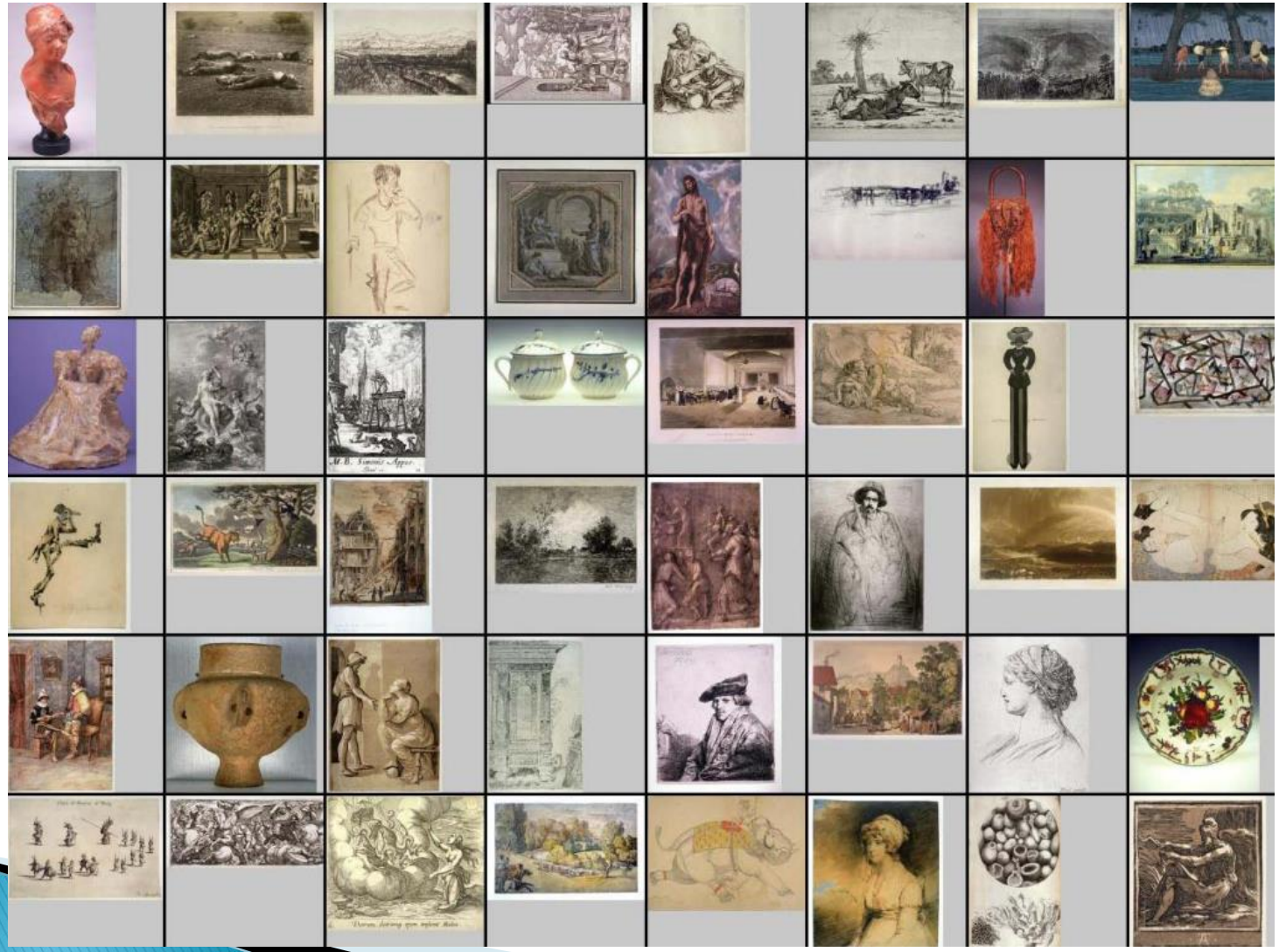
By Dr. Etana F. (PhD)

Unsupervised Learning

- For example clustering
- Clustering –
  - grouping similar objects
- Similar in which way?

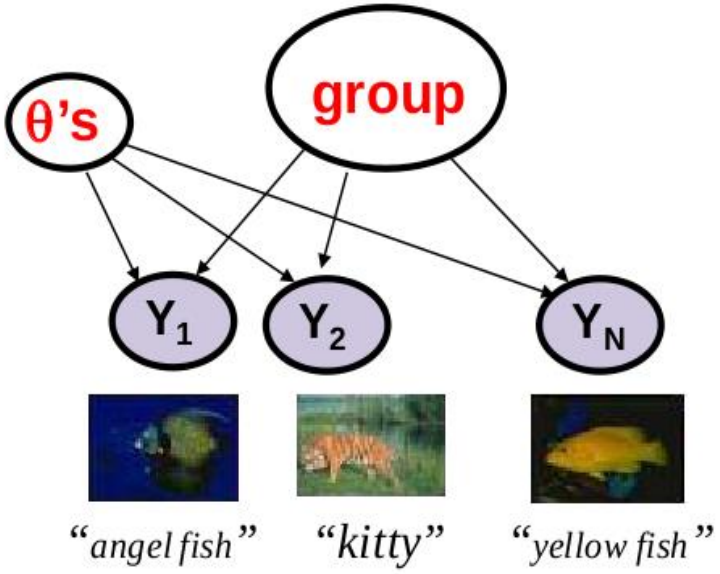


Clustering



By Dr. Etana F. (PhD)

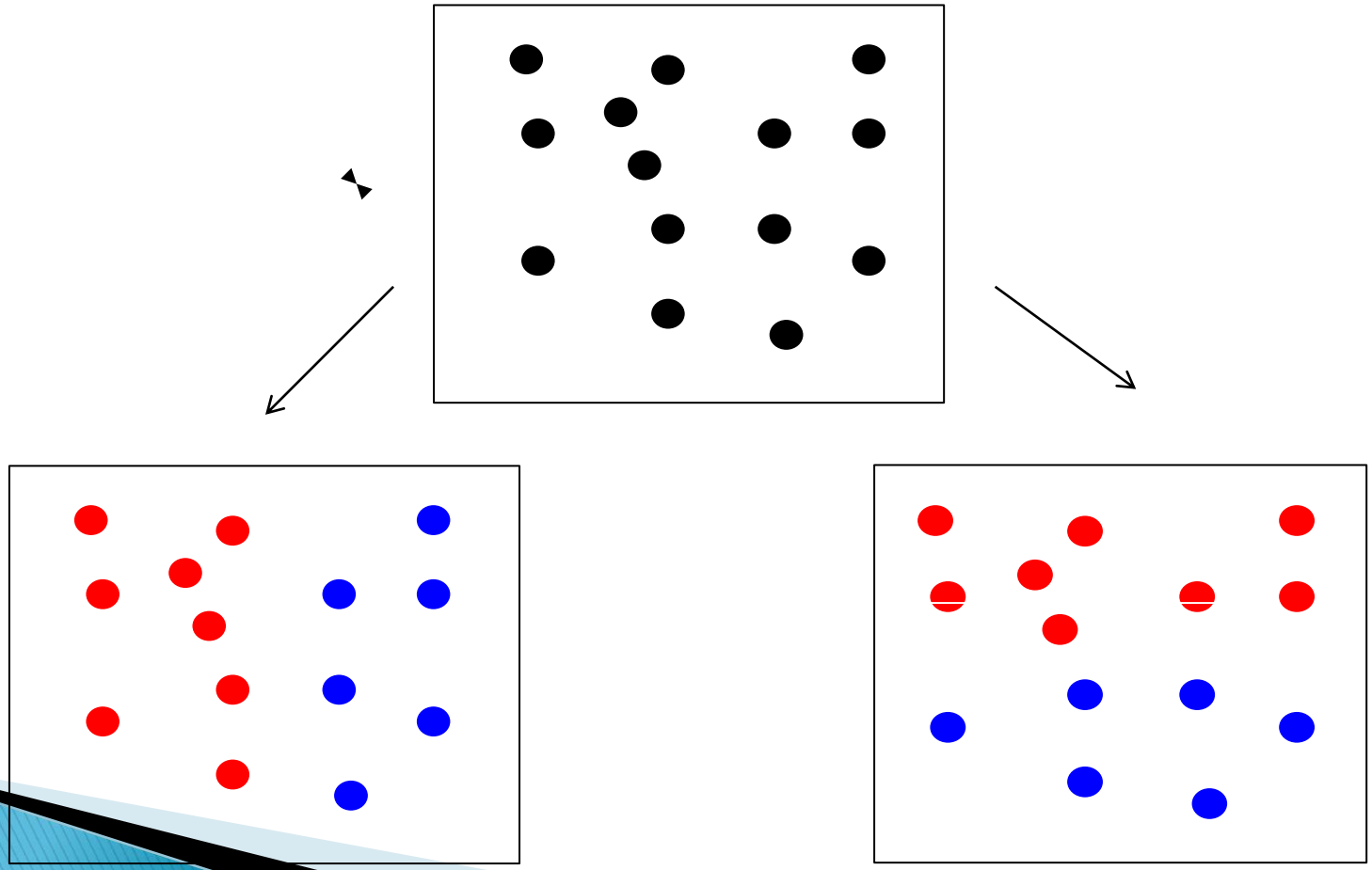
Clustering



By Dr. Etana F. (PhD)

Clustering Problems

- How to tell which type of clustering is desirable?



Common Unsupervised Learning

(Finding Patterns in Unlabeled Data)

Algorithm	Type	Explanation
K-Means Clustering	Clustering	➢ Partitions the data into K distinct clusters by minimizing the distance between points and their assigned cluster's centroid (mean).
DBSCAN	Clustering	➢ Groups together data points that are closely packed in a dense region, marking points in sparse areas as outliers.
Principal Component Analysis (PCA)	Dimensionality Reduction	➢ Transforms high-dimensional data into a lower-dimensional subspace (Principal Components) while retaining most of the data's variance.
Apriori	Association Rule Learning	➢ Identifies frequent itemsets in a dataset to discover strong association rules (e.g., "If X is bought, Y is also likely to be bought").

Published paper in this area

nature › scientific reports › articles › article

Article | Open access | Published: 30 December 2024

A topic modeling approach for analyzing and categorizing electronic healthcare documents in Afaan Oromo without label information

Etana Fikadu Dinsa ✉, Mrinal Das & Teklu Urgessa Abebe

Scientific Reports **14**, Article number: 32051 (2024) | Cite this article

3338 Accesses | **5** Citations | Metrics

Semi-Supervised Learning

- Supervised learning + Additional unlabeled data
- Unsupervised learning + Additional labeled data
- **Learning Algorithm:**
  - Start from the labeled data to build an initial classifier
  - Use the unlabeled data to enhance the model
- **Some Techniques:**
  - Co-Training: two or more learners can be trained using an independent set of different features
  - Or to model joint probability distribution of the features and labels

Reinforcement Learning

Topics:

Policies: what actions should an agent take in a particular situation

Utility estimation: how good is a state (→used by policy)

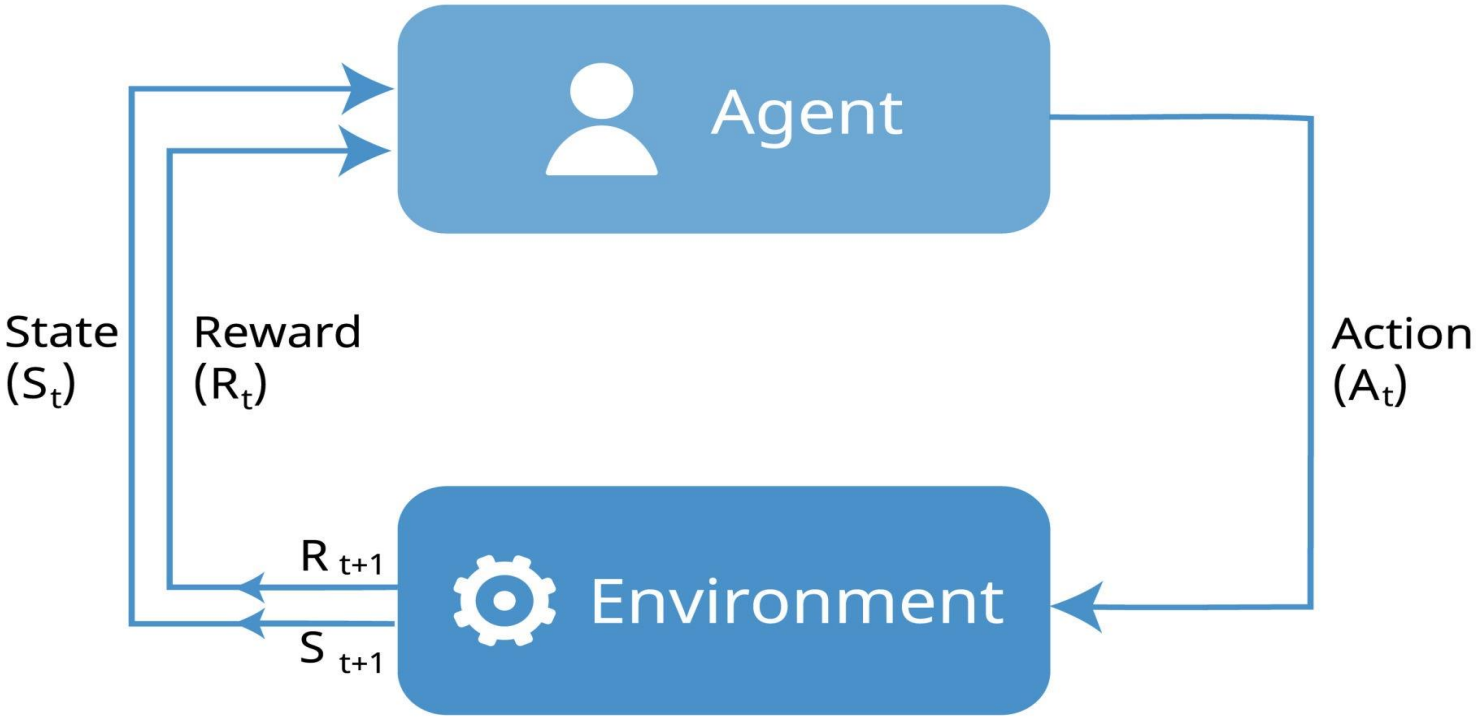
No supervised output but delayed reward

Credit assignment problem (what was responsible for the outcome)

Applications:

- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Reinforcement Learning

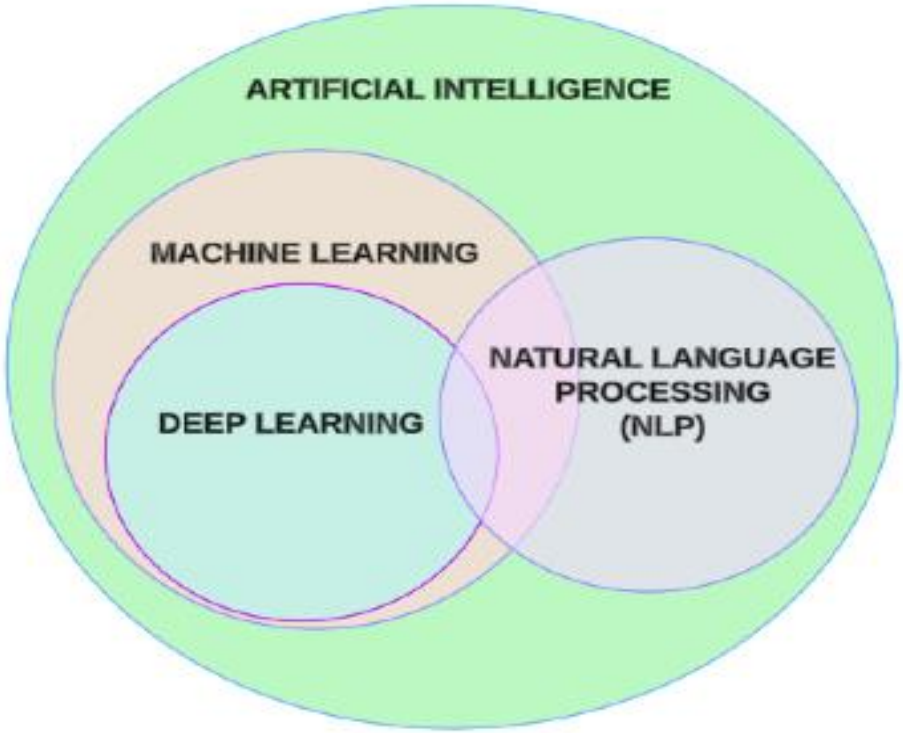


- This diagram shows how the **Agent** observes the **State** of the **Environment**, chooses an **Action**, and then receives a **Reward** (or penalty) and a new State, learning the best strategy over time.

How learning happens

Human	Machine
Memorize	k-Nearest Neighbours, Case/Example-based learning
Observe someone else, then repeat	Supervised Learning, Learning by Demonstration
Keep trying until it works (riding a bike)	Reinforcement Learning
20 Questions	Active Learning
Pattern matching (faces, voices, languages)	Pattern Recognition
Guess that current trend will continue (stock market, real estate prices)	Regression

Over all Summary



High level diagram to show the relationship between AI, ML, DL and NLP

Thanks

By Dr. Etana F. (PhD)