

# Unit 1

# What is Machine Learning?

---

EE-UY4563/ EL-GY6143: INTRODUCTION TO MACHINE LEARNING  
PROF. SUNDEEP RANGAN



# Learning Objectives

---

- ❑ Identify **data-driven learning** vs. **expert** or **domain knowledge**-based approaches
- ❑ Provide examples of machine learning used today
- ❑ Given a new problem, qualitatively describe how machine learning can be used
  - Formulate a potential machine learning task
  - Identify the data needed for the task
  - Identify objectives
- ❑ Classify a machine learning task:
  - Supervised vs. unsupervised, regression vs. classification
- ❑ For supervised learning, identify the predictors and target variables

# Outline

---

## What is Machine Learning?

### ☐ Types of machine learning algorithms

- Classification
- Regression
- Unsupervised learning

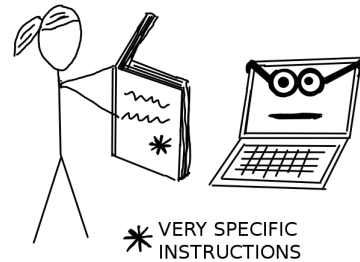
### ☐ Why the hype today?

### ☐ Some slides from:

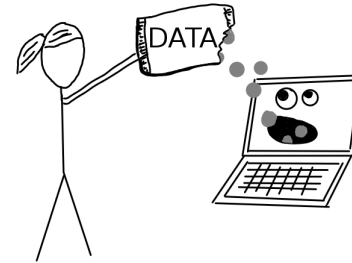
- A. Zisserman, “Machine Learning Introduction”
- Alpaydin, “Introduction to Machine Learning”

# What is Machine Learning?

Learn to improve algorithms from data.



Traditional approach  
Domain or expert knowledge



Machine Learning  
Data-driven

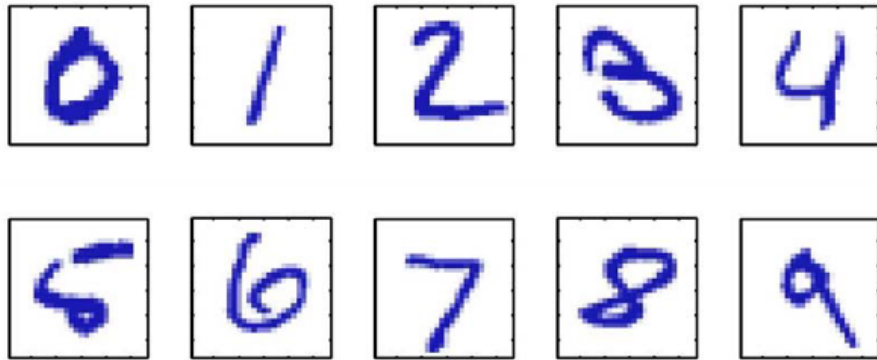
Image from Christoph Molnar,  
<https://christophm.github.io/interpretable-ml-book>

Why?

- Human expertise does not exist (ex: complex medical processes we don't fully understand)
- Humans are unable to explain their expertise (speech recognition)
- Solution change or adapt in time (routing on a computer network)

# Example 1: Digit Recognition

---



Images are 28 x 28 pixels

❑ **Problem:** Recognize a digit from the image

❑ **MNIST dataset challenge**

- Dataset developed in 1990s to spur AI research on a challenging problem for the time
- Data taken from census forms
- Became a classic benchmark for machine vision problems
- We will see this dataset extensively in this class

# Classical “Expert” Approach

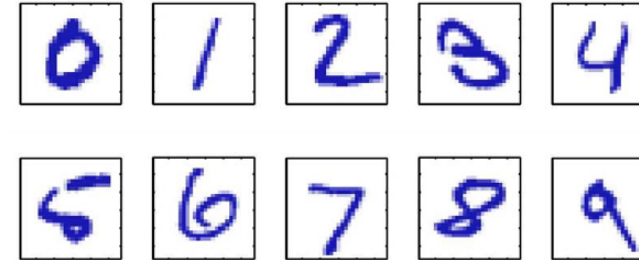
- ❑ **Idea:** Use your knowledge about digits
  - You are an “expert” since you can do the task

- ❑ Construct simple rules and code them

- ❑ **Expert rule** example: *“Image is a digit 7 if...”*:
  - There is a single horizontal line, and
  - There is a single vertical line

- ❑ Rule seems simple and reasonable

- ❑ But,...



Images are 28 x 28 pixels

```
def count_vert_lines(image):  
    ...  
def count_horiz_lines(image):  
    ...  
  
def classify(image):  
    ...  
    nv = count_vert_lines(image)  
    nh = count_horiz_lines(image)  
    ...  
  
    if (nv == 1) and (nh == 1):  
        digit = 7  
    ...  
  
    return digit
```

# Problems with Expert Rules



## ❑ Simple expert rule breaks down in practice

- Hard to define a “line” precisely
- Orientation, length, thickness, ...
- May be multiple lines...

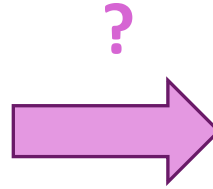
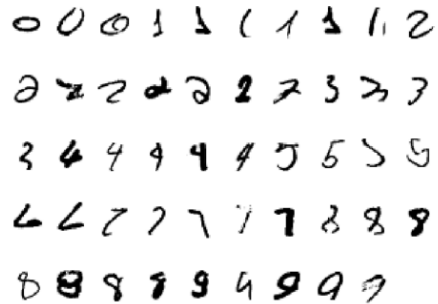
## ❑ General problem: We cannot easily code our knowledge

- We can do the task
- But, it is hard to translate to simple mathematical formula

```
def count_vert_lines(image):  
    ...  
def count_horiz_lines(image):  
    ...  
def classify(image):  
    ...  
    nv = count_vert_lines(image)  
    nh = count_horiz_lines(image)  
    ...  
  
    if (nv == 1) and (nh == 1):  
        digit = 7  
    ...  
  
    return digit
```

# ML Approach: Learn from Data

Training inputs images  $x_i$  (ex. 5000 ex per class)



Learned classifier  
 $f(x)$

Training output labels  $y_i \in \{0, 1, \dots, 9\}$

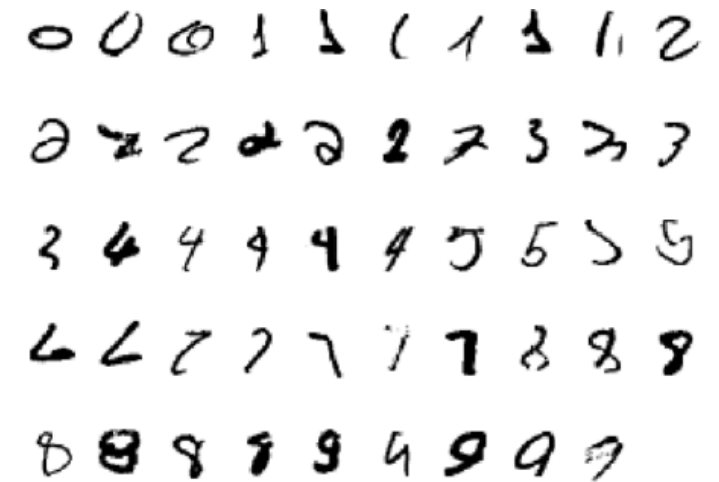
- ❑ Do not use your “expert” knowledge
- ❑ Learn the function from data!
- ❑ Supervised learning:
  - Get many **labeled examples**  $(x_i, y_i), i = 1, \dots, N$  (Called the training data)
  - Each example has an input  $x_i$  and output  $y_i$
  - **Learn** a **function**  $f(x)$  such that:  $f(x_i) = y_i$  for “most” training examples



# ML Approach Benefits and Challenges

## ❑ Learned systems do very well on image recognition problems

- On MNIST, [current systems](#) get <0.21% errors (as of 1/20/2018)
- Used widely in commercial systems today (e.g. OCR)
- Cannot match this performance with an expert system



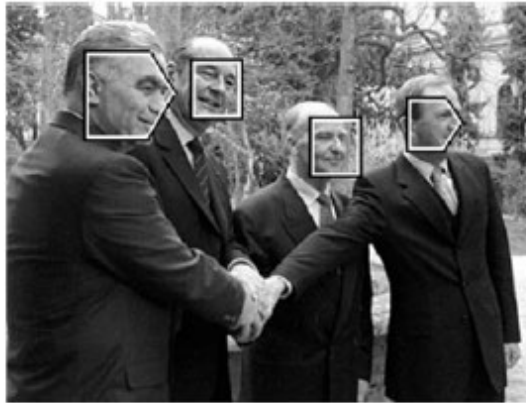
## ❑ But, there are challenges:

- How do we **acquire data**? Someone has to manually label examples.
- How do we **parametrize** a set of functions  $f(x)$  to search?
- How do we **fit** the function to data?
- If a function works on training example, will it **generalize** on new data?

## ❑ This is what you will learn in this class

# Example 2: Face Detection

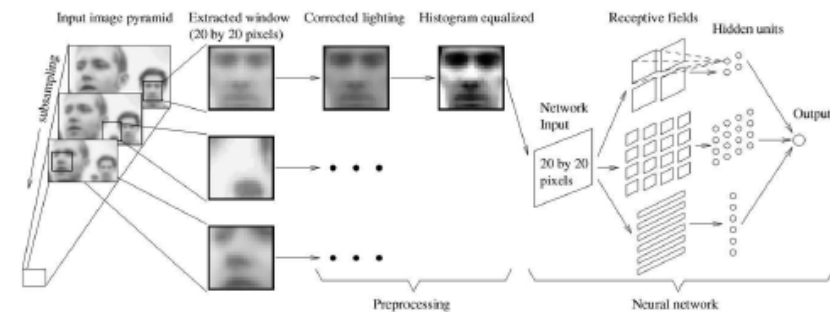
---



- ❑ **Problem:** For each image region, determine if face or non-face
- ❑ More challenging than digit recognition
  - Even harder to describe a face via “rules” in a robust way

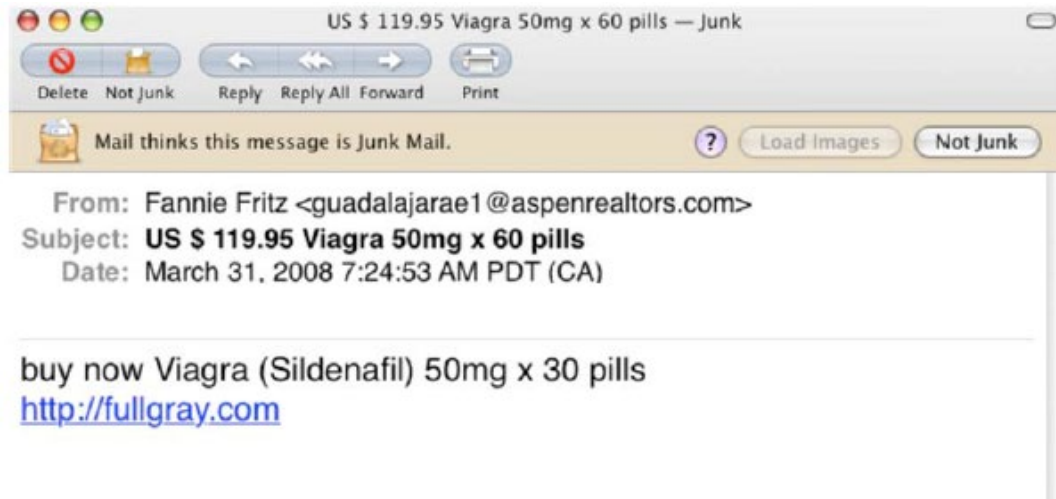
# Supervised Learning Approach

- ❑ **Data:** Get large number of face and non-face examples
- ❑ Typical early dataset
  - 5000 faces (all near frontal, vary age, race, gender, lighting)
  - $10^8$  non faces
  - Faces are normalized (scale, translation)
- ❑ **Learn** a classifier from a **class** of functions
  - Each function maps image to binary value “face” or “non-face”
  - Select function that works well on training data
  - For good performance, functions may be complex
  - Many **parameters**
- ❑ Many more datasets are available now:
  - See <http://www.face-rec.org/databases/>
  - You can use this for your project!



Rowley, Baluja and Kanade, 1998

# Example 3: Spam Detection



## ❑ Classification problem:

- Is email junk or not junk?

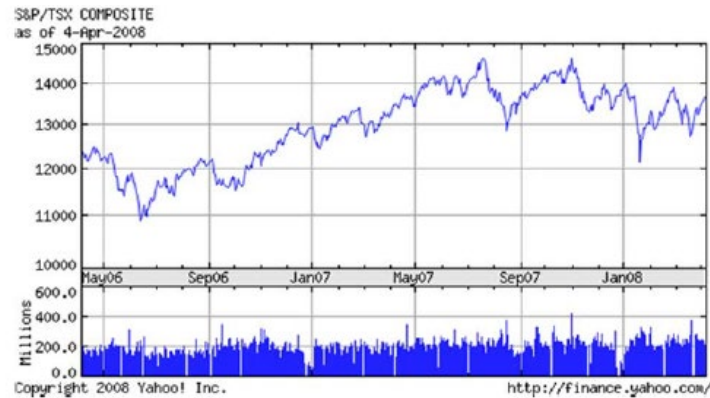
## ❑ For ML, must represent email numerically

- Common model: **bag of words**
- Enumerate all words,  $i = 1, \dots, N$
- Represent email via word count  
 $x_i = \text{num instances of word } i$

## ❑ Challenge:

- Very high-dimensional vector
- System must continue to adapt  
(keep up with spammers)

# Example 4: Stock Price Prediction



- ☐ Can you predict the price of a stock?
- ☐ What variables would you use?
- ☐ What is a non-machine learning approach?

# Machine Learning in Many Fields

---

- ❑ **Retail:** Market basket analysis, Customer relationship management (CRM)
- ❑ **Finance:** Credit scoring, fraud detection
- ❑ **Manufacturing:** Control, robotics, troubleshooting
- ❑ **Medicine:** Medical diagnosis
- ❑ **Telecommunications:** Spam filters, intrusion detection
- ❑ **Bioinformatics:** Motifs, alignment
- ❑ **Web mining:** Search engines
- ❑ ...

# In-Class Exercise 1

For each of the proposed algorithms below, indicate whether the use a machine learning (i.e. data driven) approach or not (e.g. expert or domain knowledge)

Num	Algorithm	ML Approach?	
		Yes	No
1	A robot determines its route in a room using a shortest path algorithm combined with data on the obstacle locations.		
2	You predict the weather tomorrow using data on how whether has changed in the past.		
3	A computer program playing poker decides to fold or not fold in a game by calculating the probability that its poker hand is the best.		
4	A program estimates whether a customer will purchase a product from sales records of past customers and their attributes.		

# Outline

---

❑ What is Machine Learning?

 Types of machine learning algorithms

- Classification
- Regression
- Unsupervised learning
- Reinforcement learning

❑ Why the hype today?

❑ Some slides from:

- A. Zisserman, “Machine Learning Introduction”
- Alpaydin, “Introduction to Machine Learning”



# Classification

## ❑ Supervised learning

- Learn mapping from features  $x$  to target  $y$

## ❑ Classification:

- Target is discrete. One of a finite number of values
- Ex: Binary  $y \in \{0,1\}$

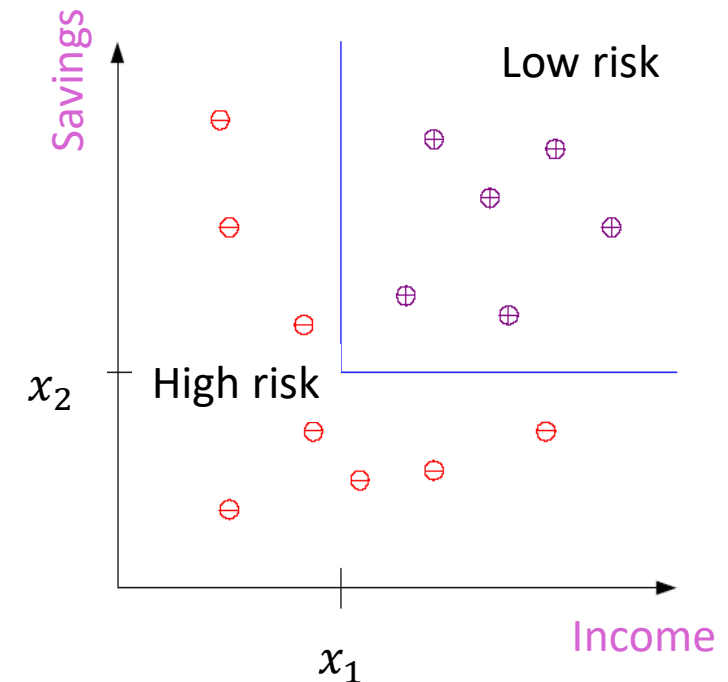
## ❑ Example: Credit assessment

- Target: customer is high-risk or low-risk
- Features: income & saving  $x = (x_1, x_2)$

## ❑ Learn a function from features to target

- Use past training data
- Need to get this data

## ❑ The function on the right is an example of a decision tree.

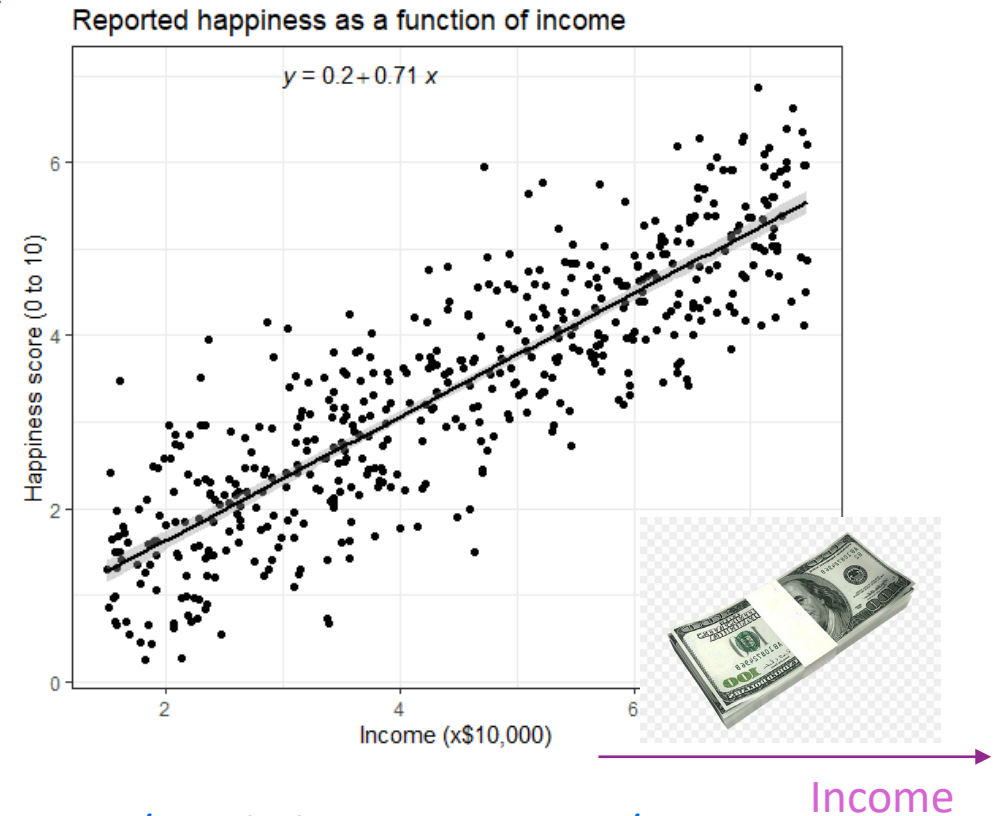


# Regression

- ❑ Also supervised learning
- ❑ Predicting a **continuous-valued** target
- ❑ **Example:**
  - Predict  $y$  = happiness score (e.g. from surveys)
  - From  $x$  = income, country, age, ...
  - Can use multiple predictors
- ❑ Assume some form of the mapping
  - Ex. Linear:  $y = \beta_0 + \beta_1 x$
  - Find parameters  $\beta_0, \beta_1$  from data



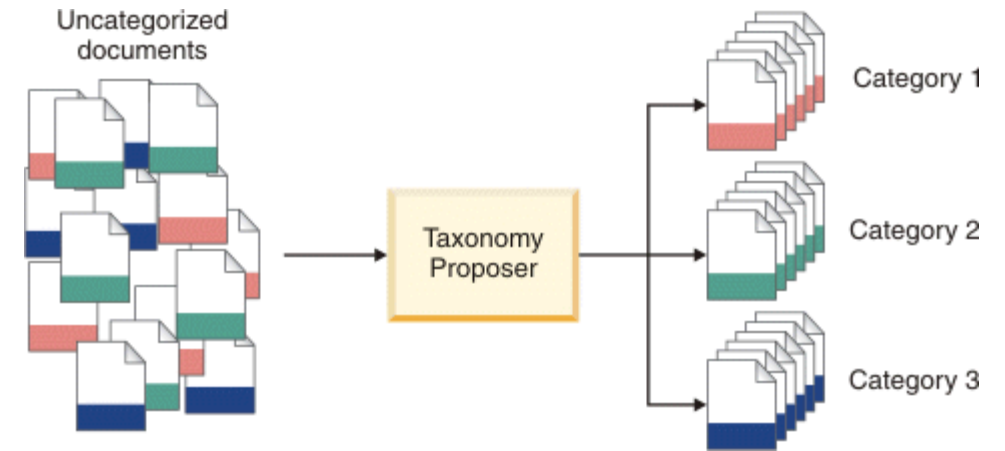
Happiness Score



<https://www.scribbr.com/statistics/simple-linear-regression/>

# Unsupervised Learning

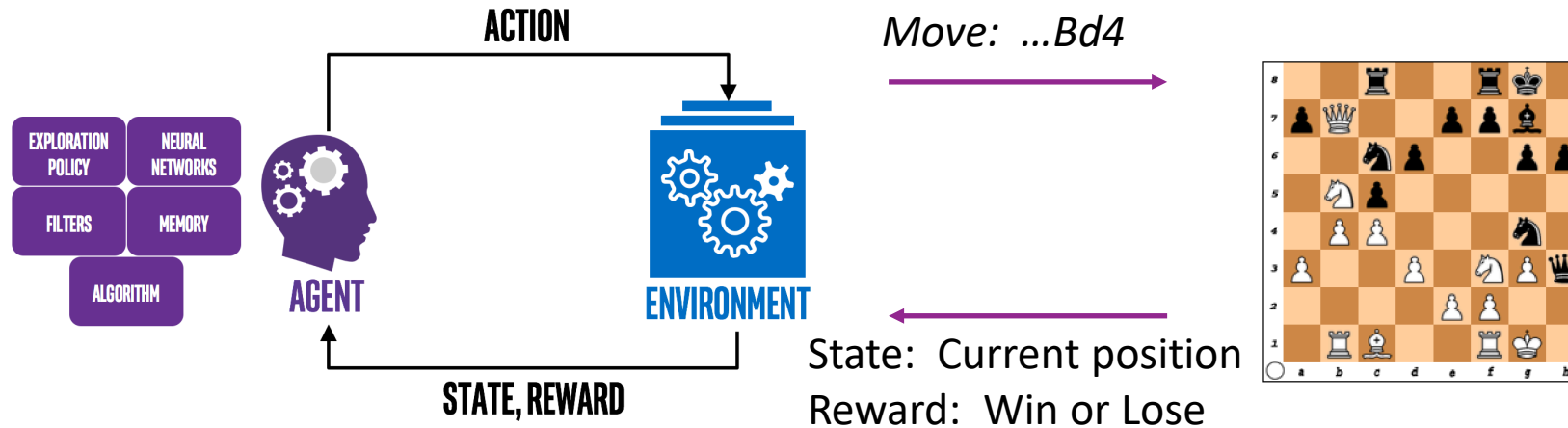
- ❑ Learning “what normally happens”
- ❑ No output
  - Just values  $x$ . No target  $y$
- ❑ Clustering: Grouping similar instances
- ❑ Example applications
  - Customer segmentation
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs



Example: Document classification

[http://www.ibm.com/support/knowledgecenter/SSBRAM\\_8.7.0/com.ibm.classify.ccenter.doc/c\\_WBG\\_Taxonomy\\_Proposer.htm](http://www.ibm.com/support/knowledgecenter/SSBRAM_8.7.0/com.ibm.classify.ccenter.doc/c_WBG_Taxonomy_Proposer.htm)

# Reinforcement Learning



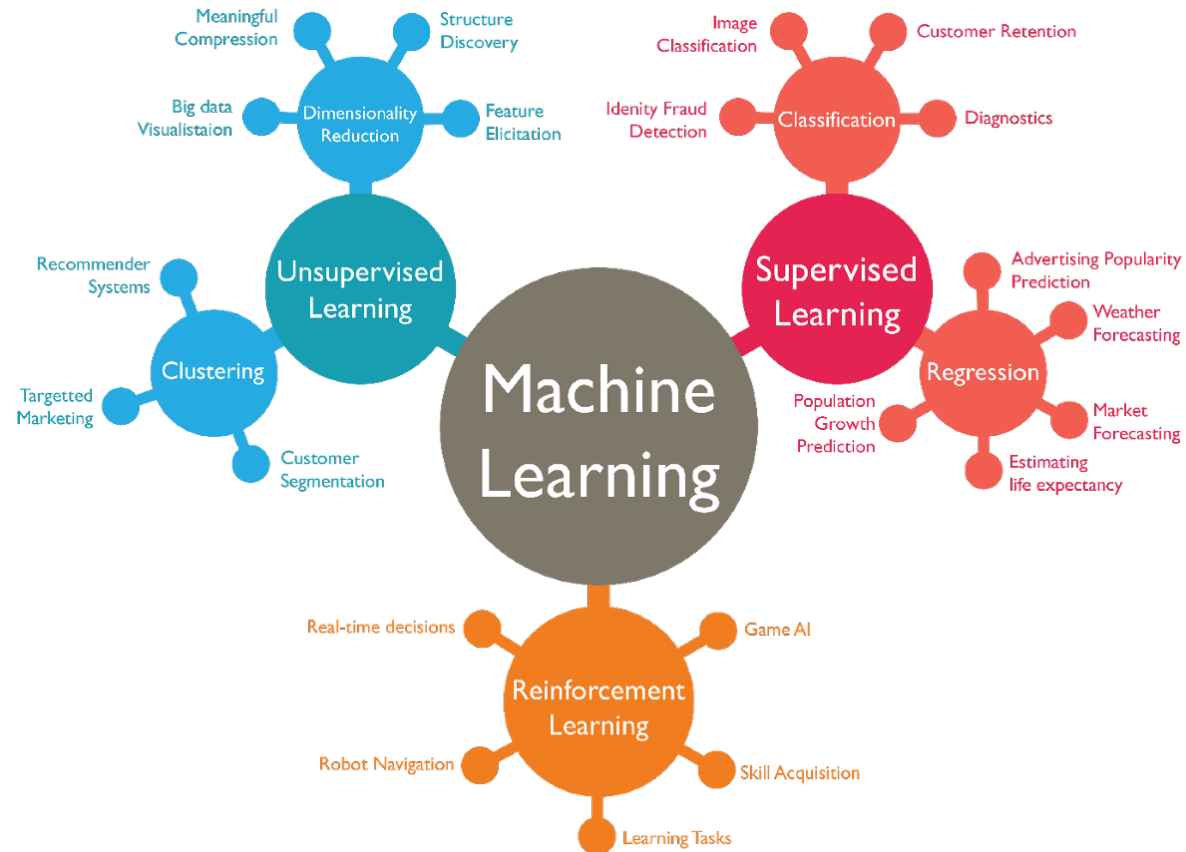
❑ Agent learns to make **actions** that interact with an **environment** to maximize a **reward**

- Agent typically acts in a **closed loop system**

❑ Key tradeoffs:

- **Exploitation** (Learn from past actions) vs. **exploration** (try new choices)
- **Credit assignment**: Which actions in the past led to the current reward?

# Types of Machine Learning



<https://www.7wdata.be/visualization/types-of-machine-learning-algorithms-2/>

# In-Class Exercise 2

For each machine learning problem below (Problem 1 to 5), determine which type of ML algorithm would be best:

- A. Supervised learning: Classification
- B. Supervised learning: Regression
- C. Unsupervised learning
- D. Reinforcement learning

For supervised learning problems, state possible predictors and target (There is no single correct solution).

Num	ML Problem	Algorithm: A to D
1	Estimate the increase in sales from attributes of an advertising campaign.	
2	Predict if a tissue sample is cancerous or not from an image of the tissue.	
3	Train a computer to steer a car from camera data. For training, you have recorded the steering actions of an expert human driver along with the camera data that the human saw.	
4	Train a computer to steer a car from camera data. In this case, there is no labeled data. The computer must learn how the steering affects the motion of the car.	
5	Classify survey data into groups with similar responses.	

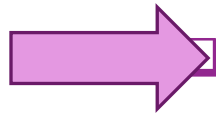
# Outline

---

❑ What is Machine Learning?

❑ Types of machine learning algorithms

- Classification
- Regression
- Unsupervised learning

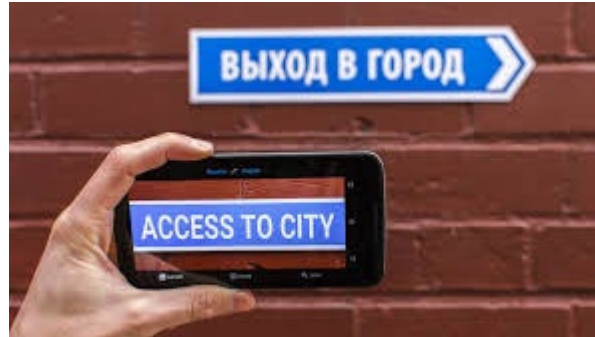
 ❑ Why the hype today?

❑ Some slides from:

- A. Zisserman, “Machine Learning Introduction”
- Alpaydin, “Introduction to Machine Learning”

# What ML is Doing Today?

- ❑ Autonomous driving
- ❑ Jeopardy
- ❑ Very difficult games: Alpha Go
- ❑ Machine translation
- ❑ Many, many others...





# Why Now?

---

- ❑ Machine learning is an old field
  - Much of the pioneering statistical work dates to the 1950s
- ❑ So what is new now?
- ❑ Big Data:
  - Massive storage. Large data centers
  - Massive connectivity
  - Sources of data from Internet and elsewhere
- ❑ Computational advances
  - Distributed machines, clusters
  - GPUs and hardware



Google Tensor Processing Unit (TPU)

# Exercise

---

- ❑ Break into small groups
- ❑ Take a field that interests you:
  - Ex. Driving a car, social networks, recommend a movie to watch, ...
- ❑ Identify a specific task that can be done with machine learning
  - What is the objective of the task?
  - What is the data you need?
  - What type of ML problem is this? Classification, regression, ...
  - How would your approach compare to an expert-driven method?