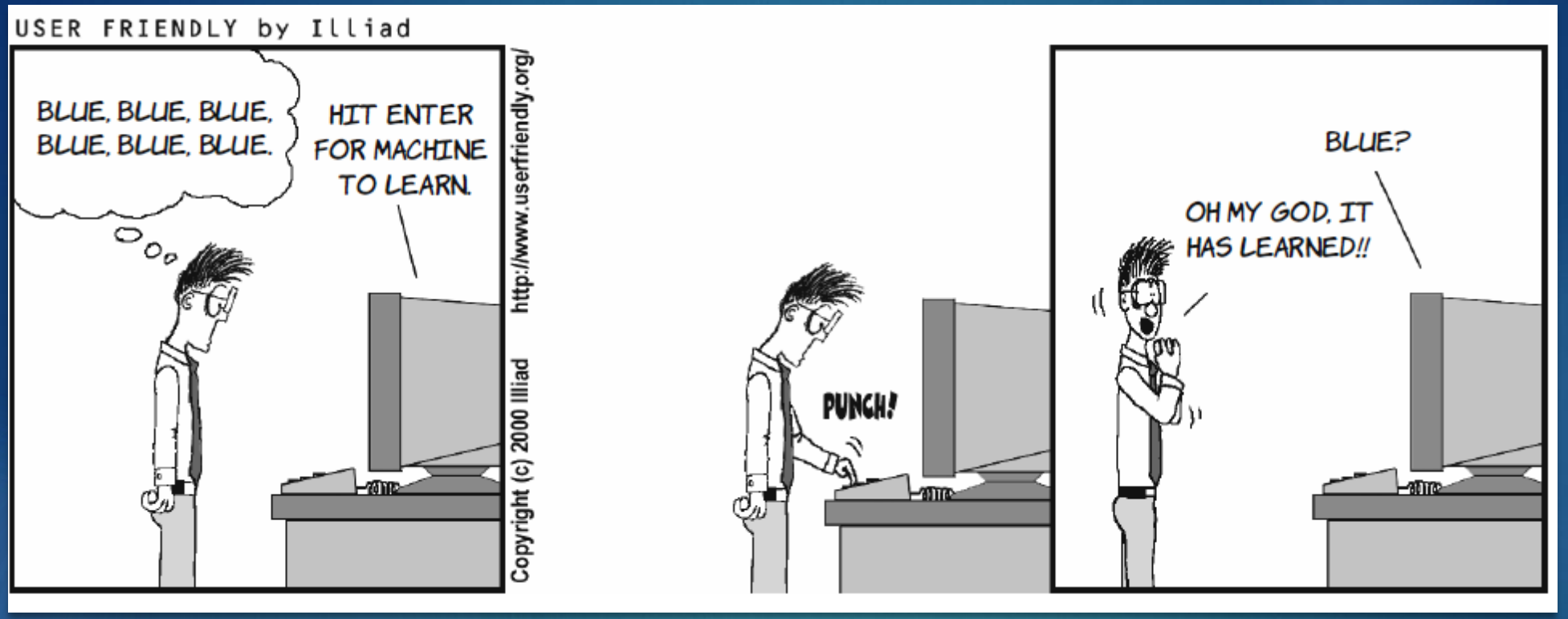
COMP-2704: Supervised Machine Learning

WEEK 1



Chapter 1: What is Machine Learning?

It is common sense, except done by a computer.

Machine learning is everywhere

- ML is now widely used in software, business, and research.
- This stems from recent (2012) advances in algorithms coupled with advances in computer hardware (GPUs).
- There is still much room for growth.

Q: Where do we see machine learning in use today?

Applications of machine learning

- Recommendation systems
- Image recognition
- Processing text for sentiment analysis
- Self-driving cars
- Spam recognition
- Medical diagnoses
- …

What exactly is machine learning?

Q: What is artificial intelligence?

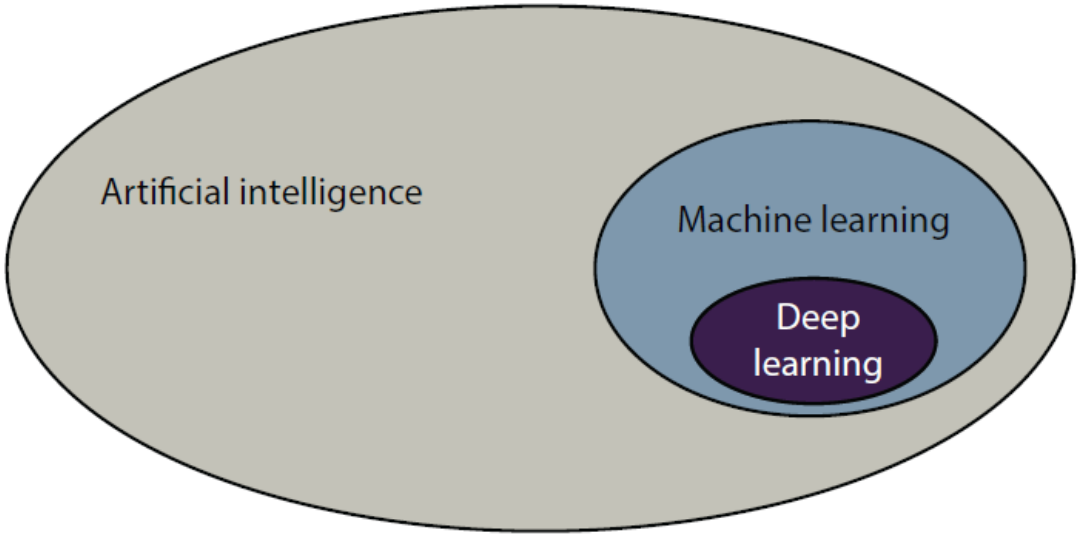
- The set of all tasks in which a computer can make decisions.

Q: What is machine learning?

- The set of all tasks in which a computer can make decisions *based on data*.

Q: What is deep learning?

- The field of machine learning that uses certain objects called neural networks.



How do humans think?

Q: Will the temperature be below -20 °C at noon tomorrow?

Q: How did you determine your answer?

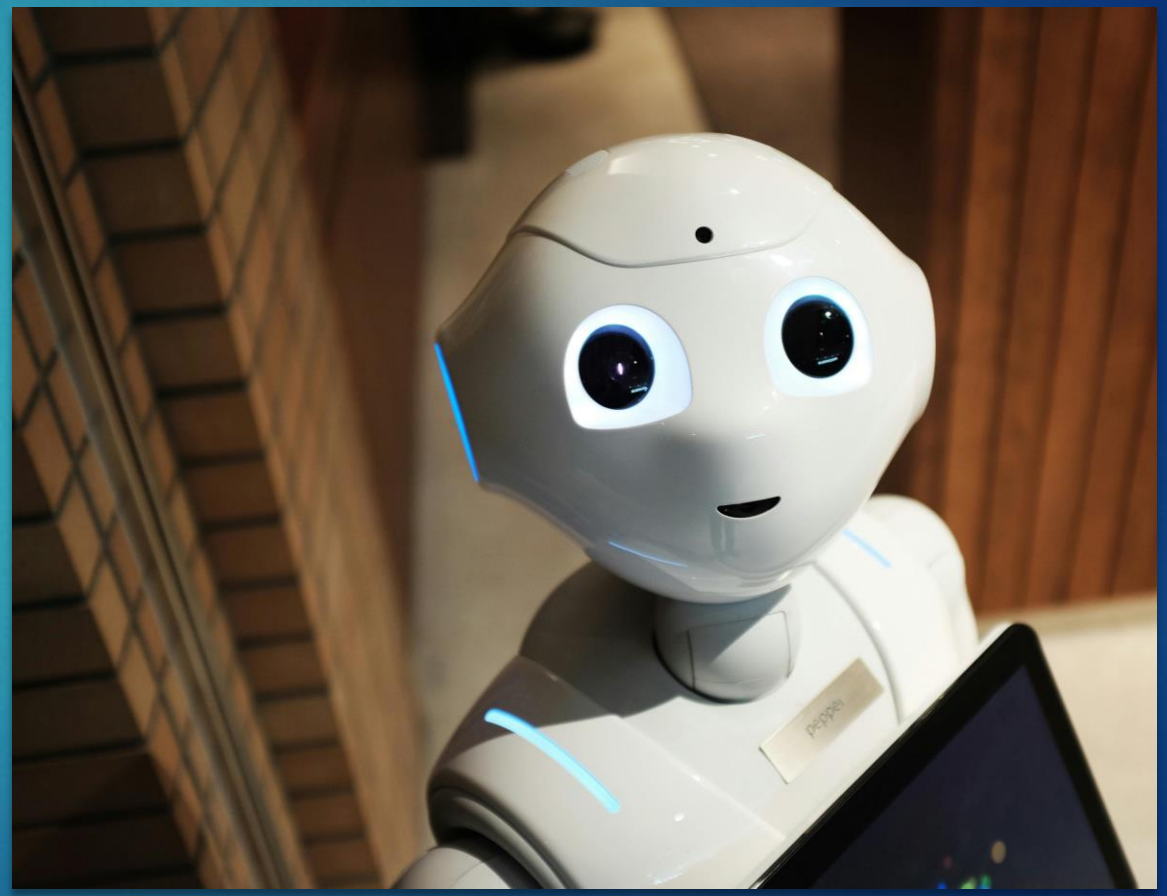
- *Remember*: temperature over last few days.
- *Formulate*: find average temperature over last few days.
- *Predict*: use average temperature as prediction.

How do computers think?

- In the past, computers were told exactly what to do by the code: *e.g.*

$$T_4 = \frac{1}{3}T_1 + \frac{1}{3}T_2 + \frac{1}{3}T_3$$

- In the ML approach, computers:
  - *Remember*: information is stored as data.
  - *Formulate*: process data to determine best equation.
  - *Predict*: use equation to make prediction.



Machine learning lingo

Model: A set of rules that represent our data and can be used to make predictions.

Algorithm: A procedure, or a set of steps, used to solve a problem or perform a computation.

- A ML *algorithm* processes data to create a *model*.

Spam filter examples

- Let us consider the specific example of a spam filter.
- A couple of definitions:
  - **Spam** is a common term for a junk or unwanted email, such as chain letters, promotions, and so on.
  - **Ham** is the term software developers use for non-spam email.
- The goal of the ML model for this use case is to predict whether an email is *spam* or *ham*.
- An algorithm creates the model by *training* it on data from past emails to the user.
- Email software uses the model to predict whether each incoming email is ham or spam, moving those predicted as spam to the junk folder.

Example 1

We have a friend, Bob, who sends a lot of emails; some are important, but others are chain letters. It is Saturday, and we just got an email from Bob.

Remember: 6 of the last 10 emails from Bob are spam.

Formulate: Most emails from Bob are spam.

Predict: Spam!

Example 2

Remember:

Day	Label
Monday	Ham
Tuesday	Ham
Saturday	Spam
Sunday	Spam
Sunday	Spam
Wednesday	Ham
Friday	Ham
Saturday	Spam
Tuesday	Ham
Thursday	Ham

Formulate:

If an email from Bob comes on a weekday, predict ham. Otherwise, predict spam.

It is Saturday, and we just got an email from Bob.

Predict: Spam!

Example 3

Remember:

Size	Label
1 kb	Ham
2 kb	Ham
16 kb	Spam
20 kb	Spam
18 kb	Spam
3 kb	Ham
5 kb	Ham
25 kb	Spam
1 kb	Ham
3 kb	Ham

Formulate:

If size > 10, predict spam
else predict ham.

We just got a 19 kb email from Bob.

Predict: Spam!

Example 4

Remember:

Day	Size	Label
Monday	1 kb	Ham
Tuesday	2 kb	Ham
Saturday	16 kb	Spam
Sunday	20 kb	Spam
Sunday	18 kb	Spam
Wednesday	3 kb	Ham
Friday	5 kb	Ham
Saturday	25 kb	Spam
Tuesday	1 kb	Ham
Thursday	3 kb	Ham

Formulate:

```
if (size > 10) or (day is on weekend)
        predict spam
else
        predict ham
```

It is Saturday, and we just got a 19 kb email from Bob.

Predict: Spam!

Example 5

Remember:

Day	Size	Label
Monday	1 kb	Ham
Tuesday	2 kb	Ham
Saturday	16 kb	Spam
Sunday	20 kb	Spam
Sunday	18 kb	Spam
Wednesday	3 kb	Ham
Friday	5 kb	Ham
Saturday	25 kb	Spam
Tuesday	1 kb	Ham
Thursday	3 kb	Ham

Formulate:

if day is a weekday
    if size > 15, predict spam
    else, predict spam
if day is on weekend
    if size > 5, predict spam
    else, predict ham

It is Saturday, and we just got a 19 kb email from Bob.

Predict: Spam!

Example 6

Remember:

Day	Size	Label
0	1 kb	Ham
1	2 kb	Ham
5	16 kb	Spam
6	20 kb	Spam
6	18 kb	Spam
2	3 kb	Ham
4	5 kb	Ham
5	25 kb	Spam
1	1 kb	Ham
3	3 kb	Ham

Map days to numbers: Monday -> 0, Tuesday -> 1, …

Formulate:

if day + size > 12, predict spam
else, predict ham

It is Saturday (6), and we just got a 19 kb email from Bob.

Predict: Spam!

Example 7

Remember:

- A ***feature*** is any property or characteristic of the data that the model can use to make predictions.

Q: What other features, should we consider?

- Spelling mistakes, sender, occurrence of 'buy' or 'win', …

Formulate:

if (more than one spelling mistake)
or (size > 10)
or (sender not in contact list)
or ('buy' or 'win' is present)
predict spam

else predict ham

We just got a 19 kb email from Bob with three spelling mistakes and a chance to win a prize.

Predict: Spam!

Example 8

Formulate:

if (size) + 10 (number of spelling mistakes) –
(number of appearances of the word ‘mom’) +
4 (number of appearances of the word ‘buy’) > 10
predict spam

else predict ham.

We just got a 19 kb email from Bob with three spelling mistakes and the word ‘buy’ appears twice.

Predict: Spam!

Which model is best?

- Whichever model generalizes to new data the best.
- We will see how to choose the best model as we go through specific examples.
- The main idea is to test models on data they have not seen before to see how well they perform.