BST 219
Core Principles of Data Science

Lecture 21: Maps continued and Introduction to Machine Learning
November 12, 2024

Recipe of the Day!
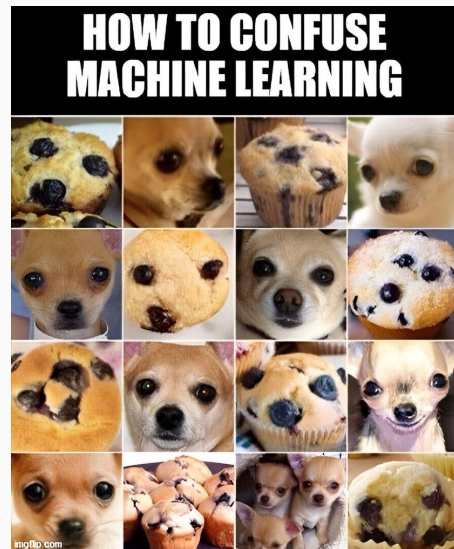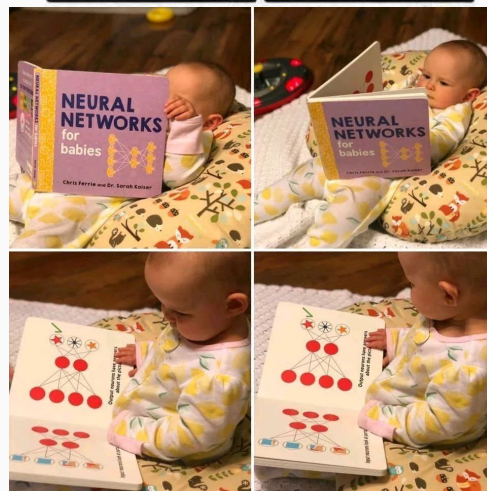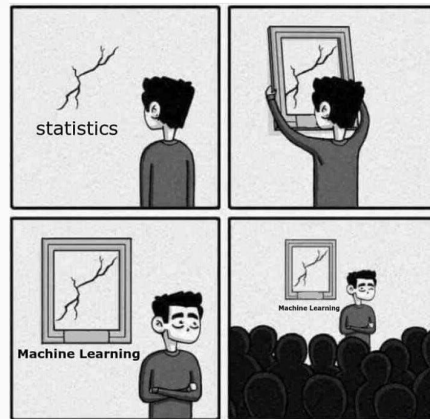
Crockpot Creamy Coconut Chicken Tikka Masala

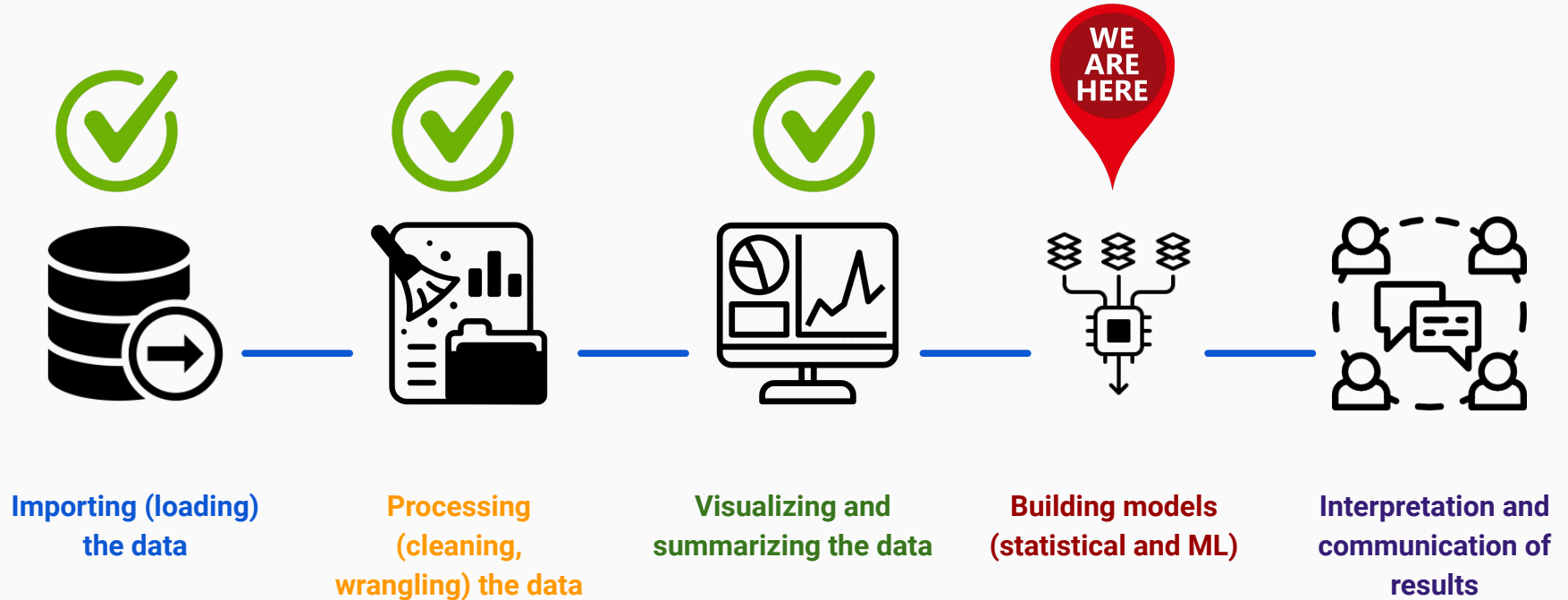**November in Kendall Square**

Agenda

- Announcements
  - Lab this week!
  - Midterm 11/8 - 11/17
  - Project preference due 11/12 (today!)
    - [Fill out this form](#)

- Continue maps

- Introduction to Machine Learning

Special Announcements

- Heather's **11/12** office hour will be moved to **12-1pm**

- The 11/14 lecture will be moved to **11/13, 12:30-2pm via Zoom**

- Lecture on **11/26** will be held via **Zoom**
  - Will be on a special topic that will be stand alone and not part of an assignment

- Heather's office hour on **11/26** will be **Zoom only**

- The TFs will not hold office hours the week of Thanksgiving (the week of 11/25)

Where we are in the data science pipeline

Importing (loading) the data

Processing (cleaning, wrangling) the data

Visualizing and summarizing the data

Building models (statistical and ML)

Interpretation and communication of results

WE ARE HERE

Machine Learning

To help us understand how a machine "learns", let's try to guess the function that takes x as an input and generates y as an output. The first values I have are x = 3 and y = 6.

```
          ?
  x    ───→  ▭  ───→   y
  3                    6
```

Machine Learning

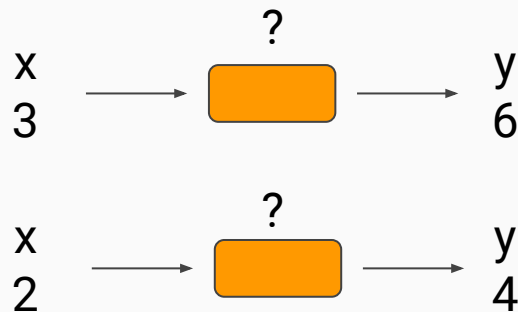To help us understand how a machine "learns", let's try to guess the function that takes x as an input and generates y as an output. The first values I have are x = 3 and y = 6.

$$x \quad \longrightarrow \quad \boxed{?} \quad \longrightarrow \quad y$$

x
3

?

y
6

x+3
2x
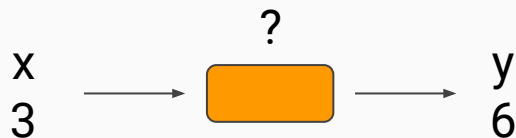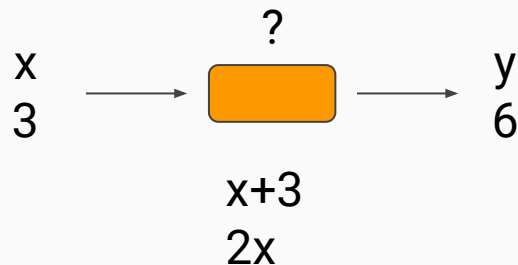
A couple of options for the function are y = x + 3 and y = 2x

Machine Learning

Since we have multiple options, let's look at another x and y combination to see if that will give us more information on which function is the correct one. Now I have x = 2 and y = 4.
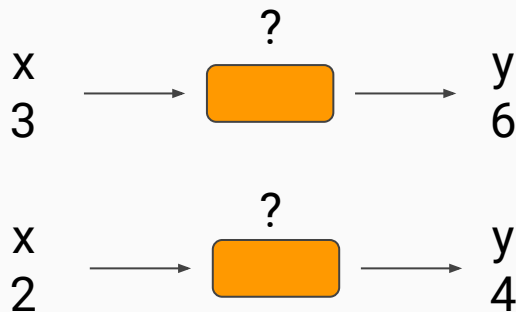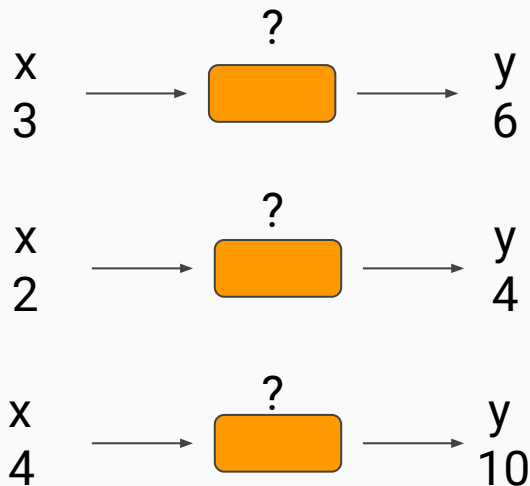
Since we have multiple options, let's look at another x and y combination to see if that will give us more information on which function is the correct one. Now I have x = 2 and y = 4.
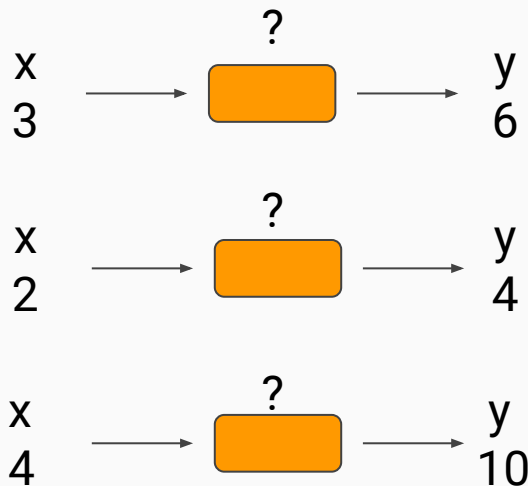

The function y = 2x would work for both!

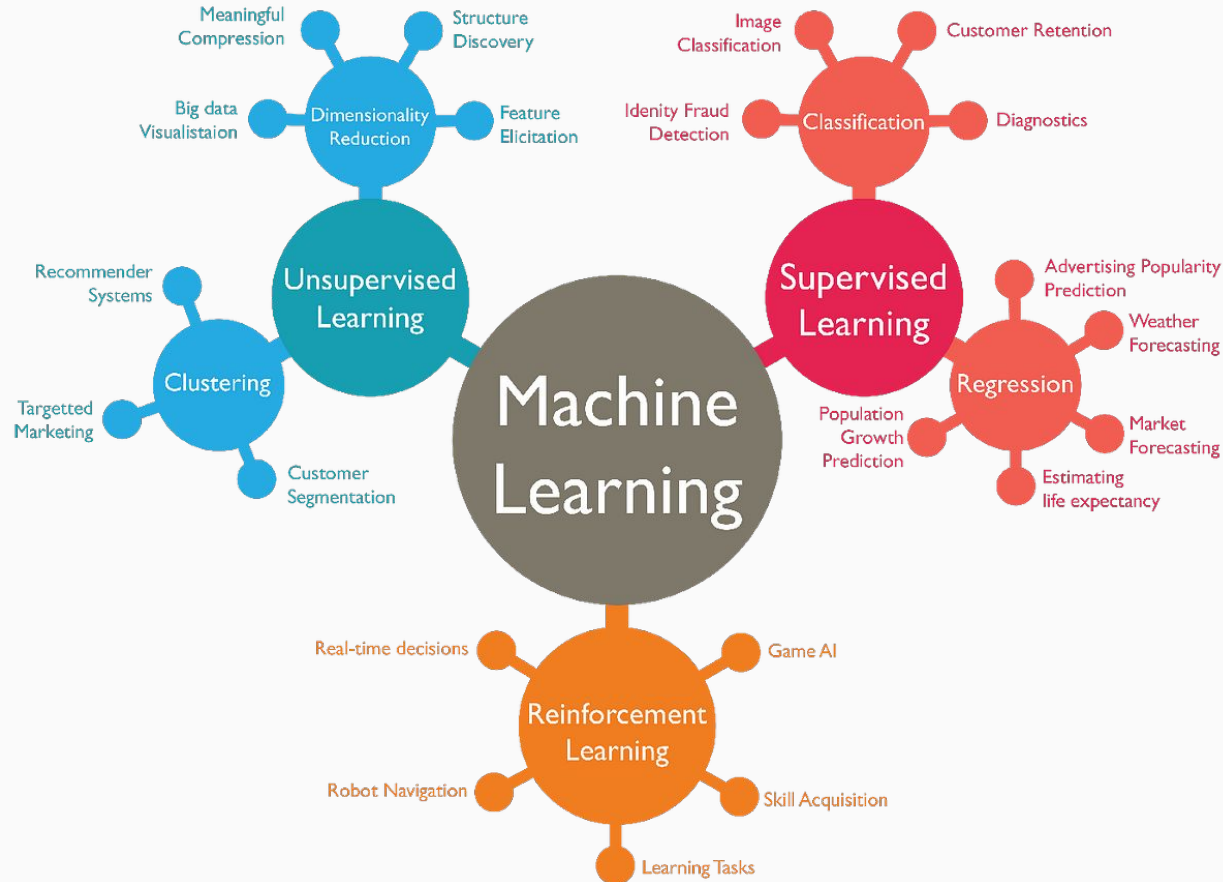But what if I have more data? Now I have x = 4 and y = 10. Our function y = 2x doesn't work anymore…

Machine Learning

But what if I have more data? Now I have x = 4 and y = 10. Our function y = 2x doesn't work anymore...

Now the function is something more complicated and I need more data to keep updating my guess for the function - this is a type of machine learning!
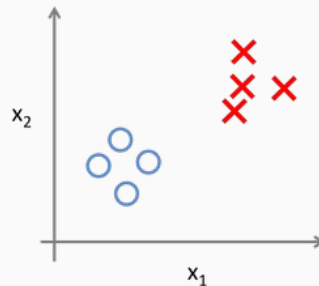
Machine Learning
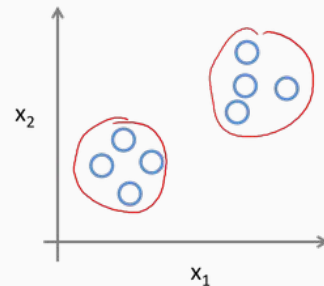
Machine Learning

Supervised Learning

A machine learning technique that uses labeled datasets to train algorithms to recognize patterns and predict outcomes. The algorithm is "supervised" because it's given information to help it learn.
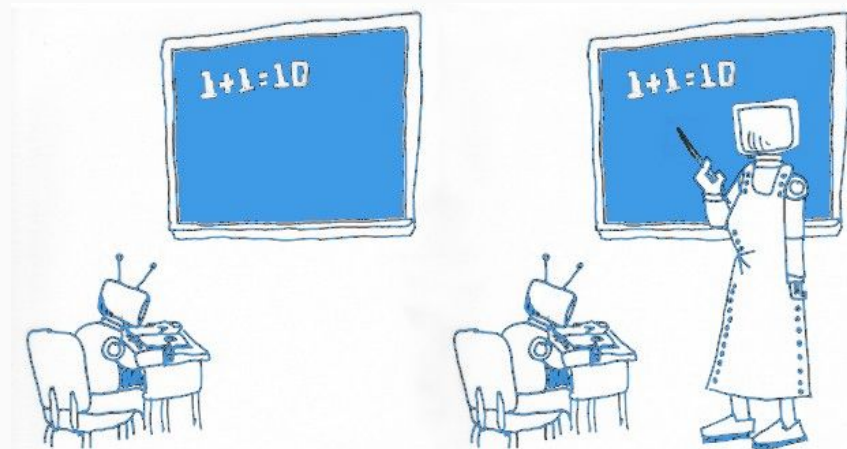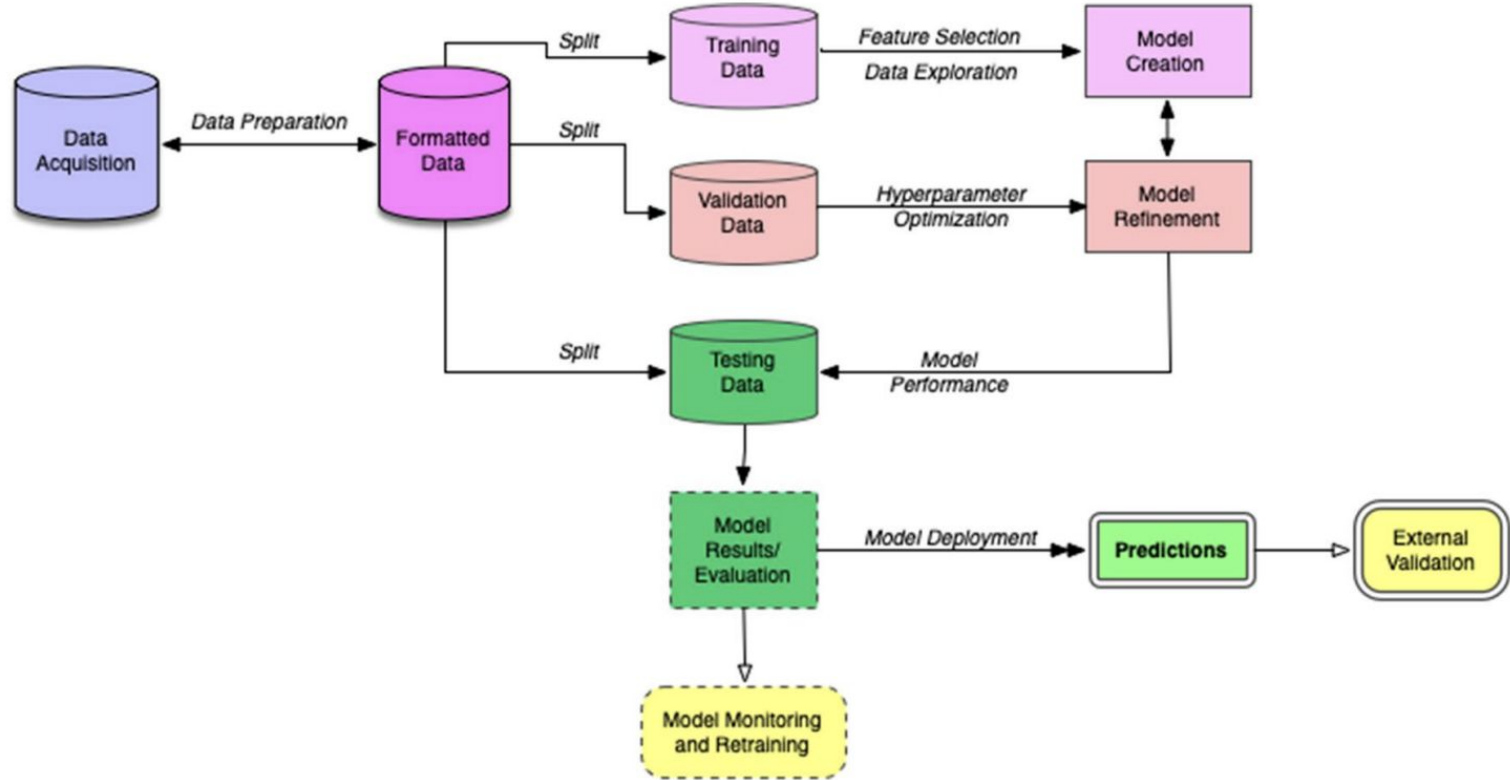
Unsupervised Learning

A machine learning technique that uses algorithms to analyze and learn patterns from unlabeled data without human intervention.

The Machine Learning Workflow

Train, Validation, and Test Sets

Available Data

Training	Testing
	(holdout sample)

New Available Data

Training	Validation	Testing
	(validation holdout sample)	(testing holdout sample)