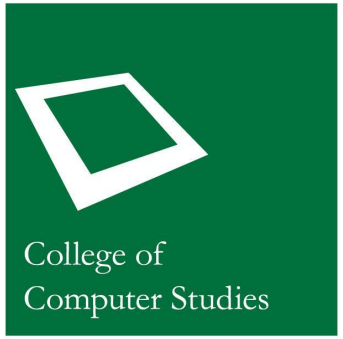
Machine Learning Overview

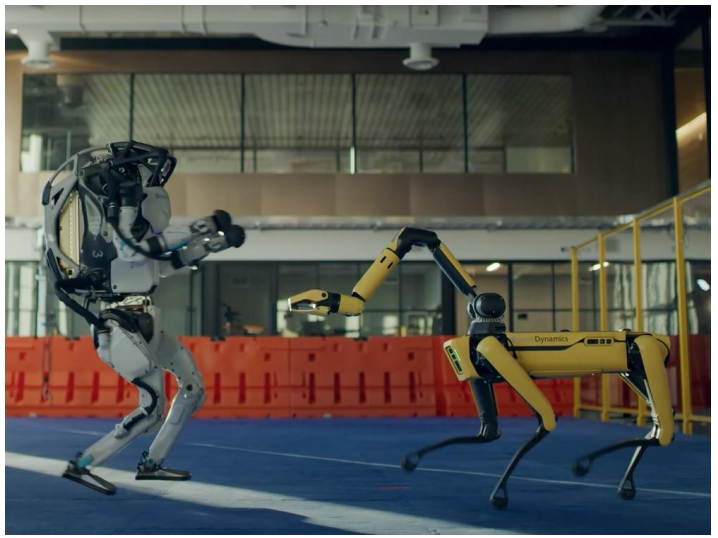
Original Slides by:
Courtney Anne Ngo
Daniel Stanley Tan, PhD
Arren Antioquia

Updated (AY 2023 – 2024 T3) by:
Thomas James Tiam-Lee, PhD



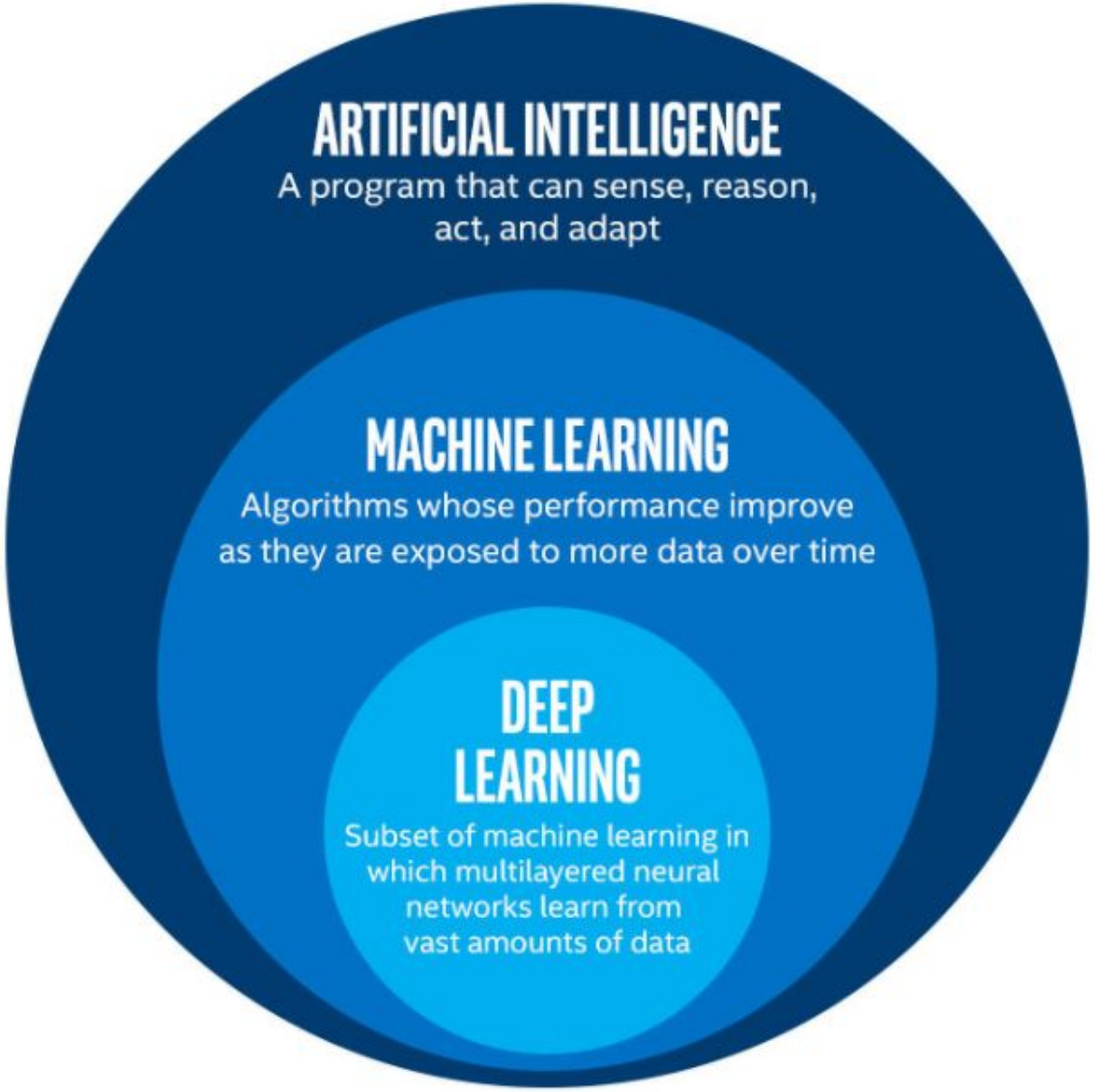
Artificial Intelligence

- A field of research in computer science that focuses on **making computers exhibit intelligent behavior**.



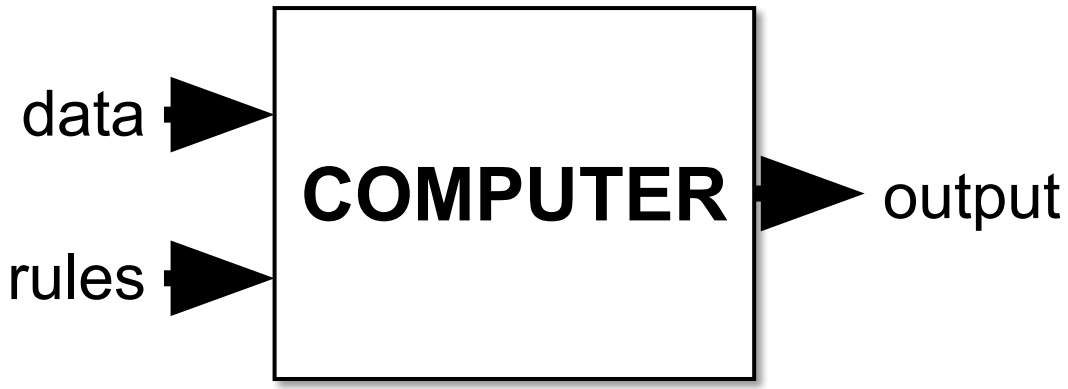
Machine Learning

- A sub-field of artificial intelligence where the rules are automatically **inferred from data**.

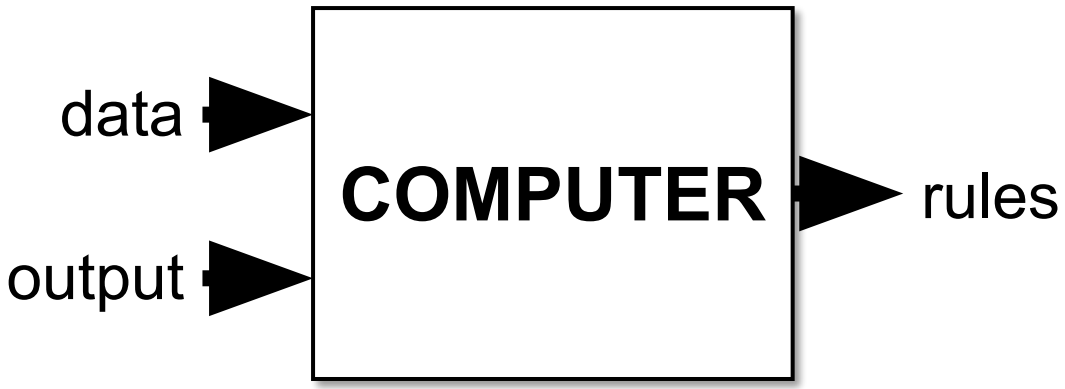


Intellectual Traditions in AI

Symbolic AI
(Traditional AI)

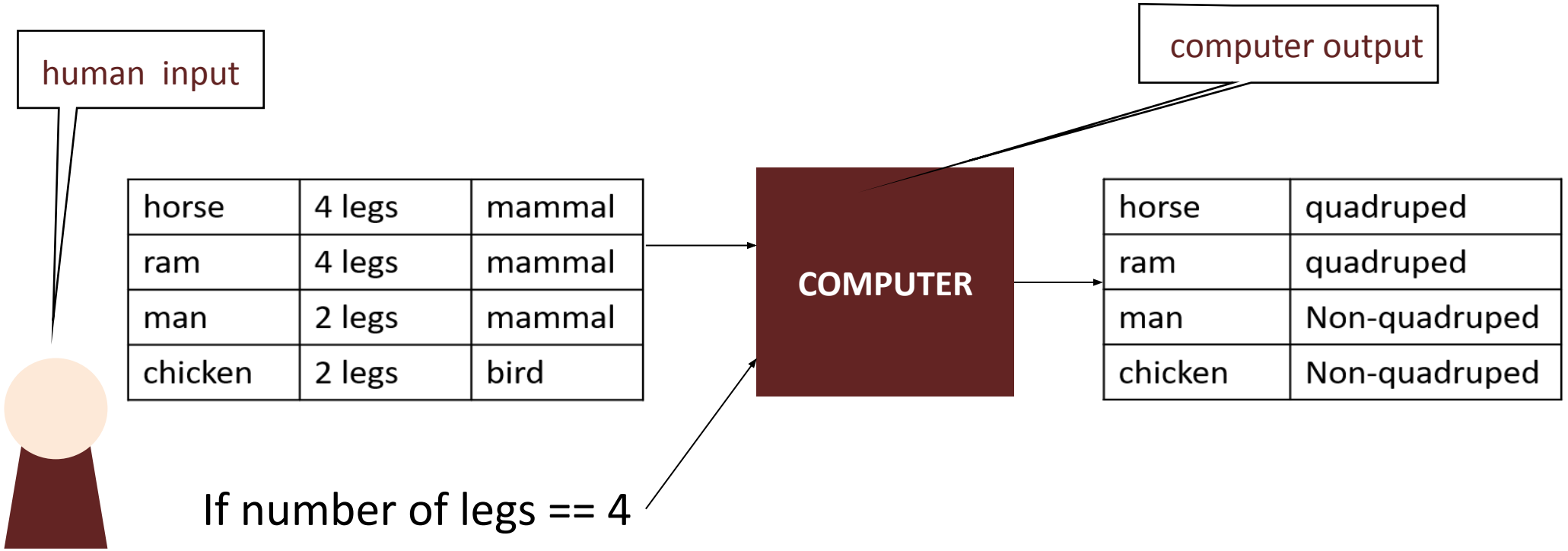


Neural AI
(Machine Learning)



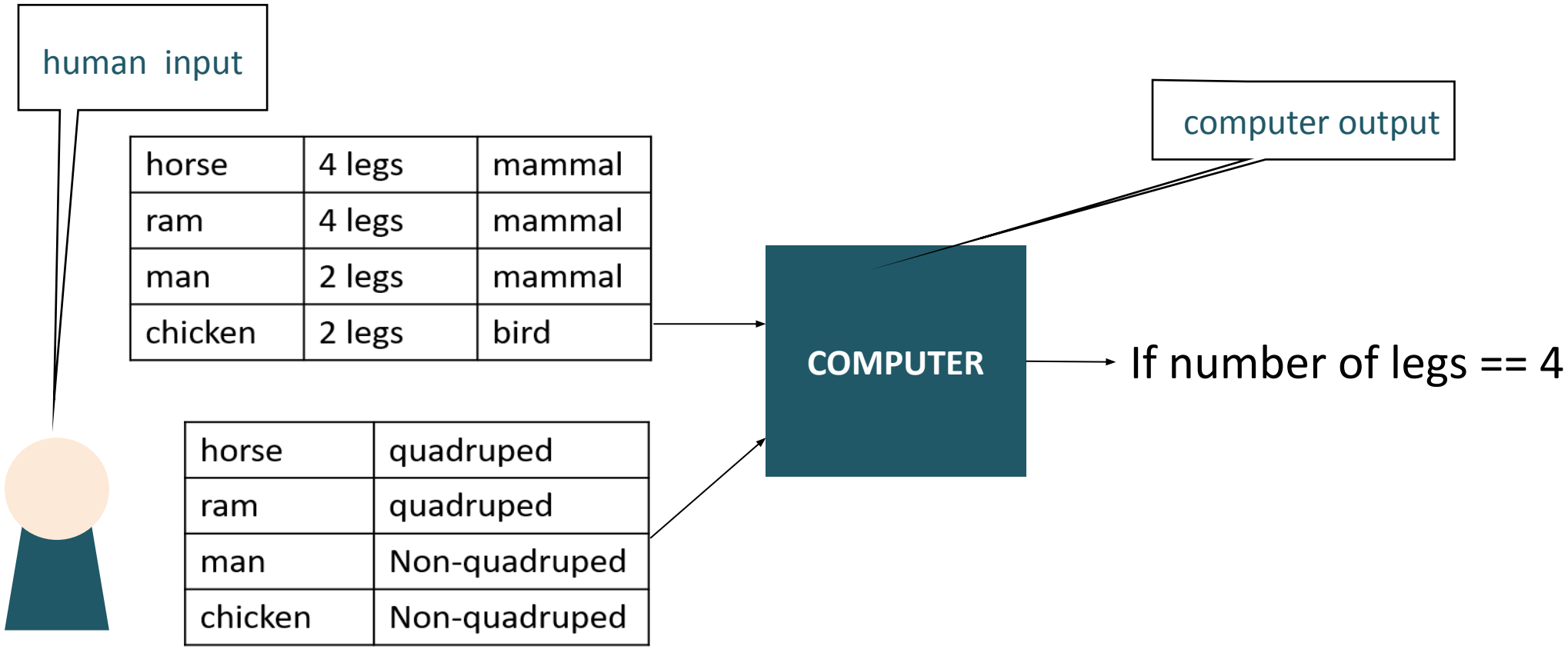
Symbolic / Traditional AI

- Determine if an animal is ***quadruped***.



Machine Learning

- Determine if an animal is ***quadruped***.

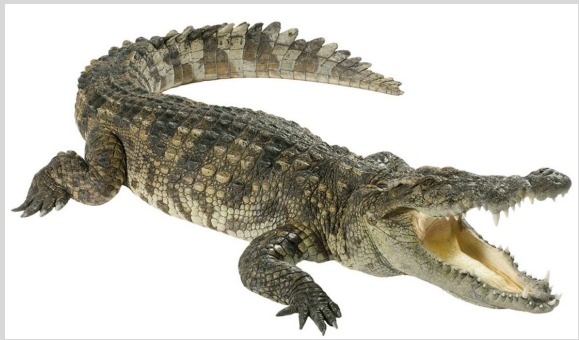


Short Game

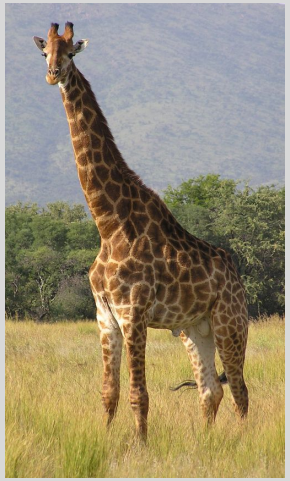
Challenge

- Determine if a horse is **acerous** or **non-acerous**.

ACEROUS



NON-ACEROUS



The horse is **acerous**.

Acerous (n).

1. Having no antennae
2. Having no horns

Source: Collins Dictionary

Traditional AI Vs. Machine Learning

- A computer that translates from Filipino to English

TRADITIONAL AI

- Encode the grammatical rules mathematically
- Parse the sentence, match it to the appropriate rules, then translate to English

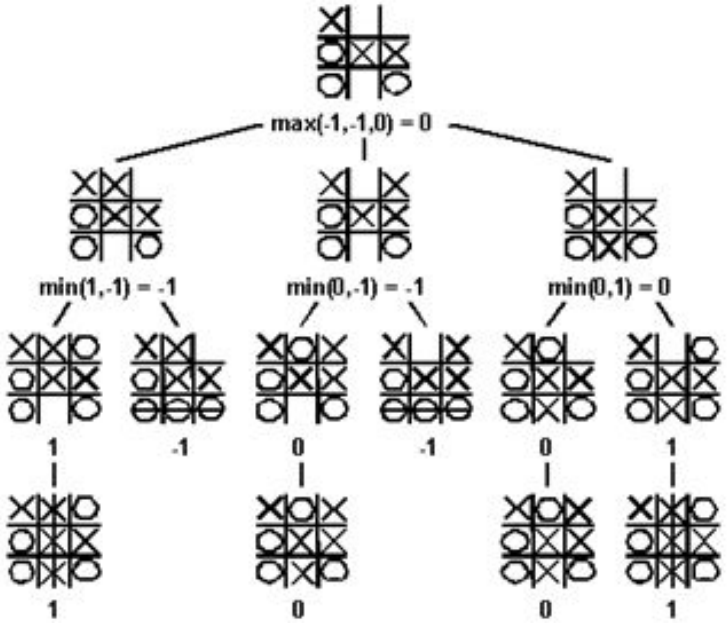
MACHINE LEARNING

- Feed lots of sentences and the corresponding translations.
- Capture the statistical patterns.
- Use the patterns to predict the translation for new text.

Traditional AI Vs. Machine Learning

- A computer that plays tic-tac-toe well

TRADITIONAL AI



- Generate the game tree
- Search for the best move

MACHINE LEARNING

- Feed lots of tic-tac-toe games to the model.
- Learn the patterns on which moves yield the best outcomes.

Types of Machine Learning Tasks

Supervised Learning

Unsupervised Learning

Reinforcement Learning

STINTSY will mostly focus on this

Supervised Learning

- The model tries to **predict a label** (also called target or output).
- There is a “correct answer” to every example.
- You can measure if the model is good by comparing its predictions to the “correct answer”

Supervised Learning

Regression
(predicting a numerical value)

- Predict the age of a person
- Predict the temperature for a given day
- Forecast the stock price of a company on a given day

Classification
(predicting a categorical value)

- Predict if a person has COVID or not ***(binary classification)***
- Detect if an email is spam ***(binary classification)***
- Classify the type of flower ***(multiclass classification)***
- Recognize a handwritten digit ***(multiclass classification)***

Unsupervised Learning

- There is **no label**. No “correct answer” for each example in the data.
- The goal is to simply find patterns from the data.
- Requires interpretation or validation of results.

Reinforcement Learning

- The model learns by interacting with the environment and **receiving rewards or penalties for its actions**.
- Will not be covered in STINTSY.

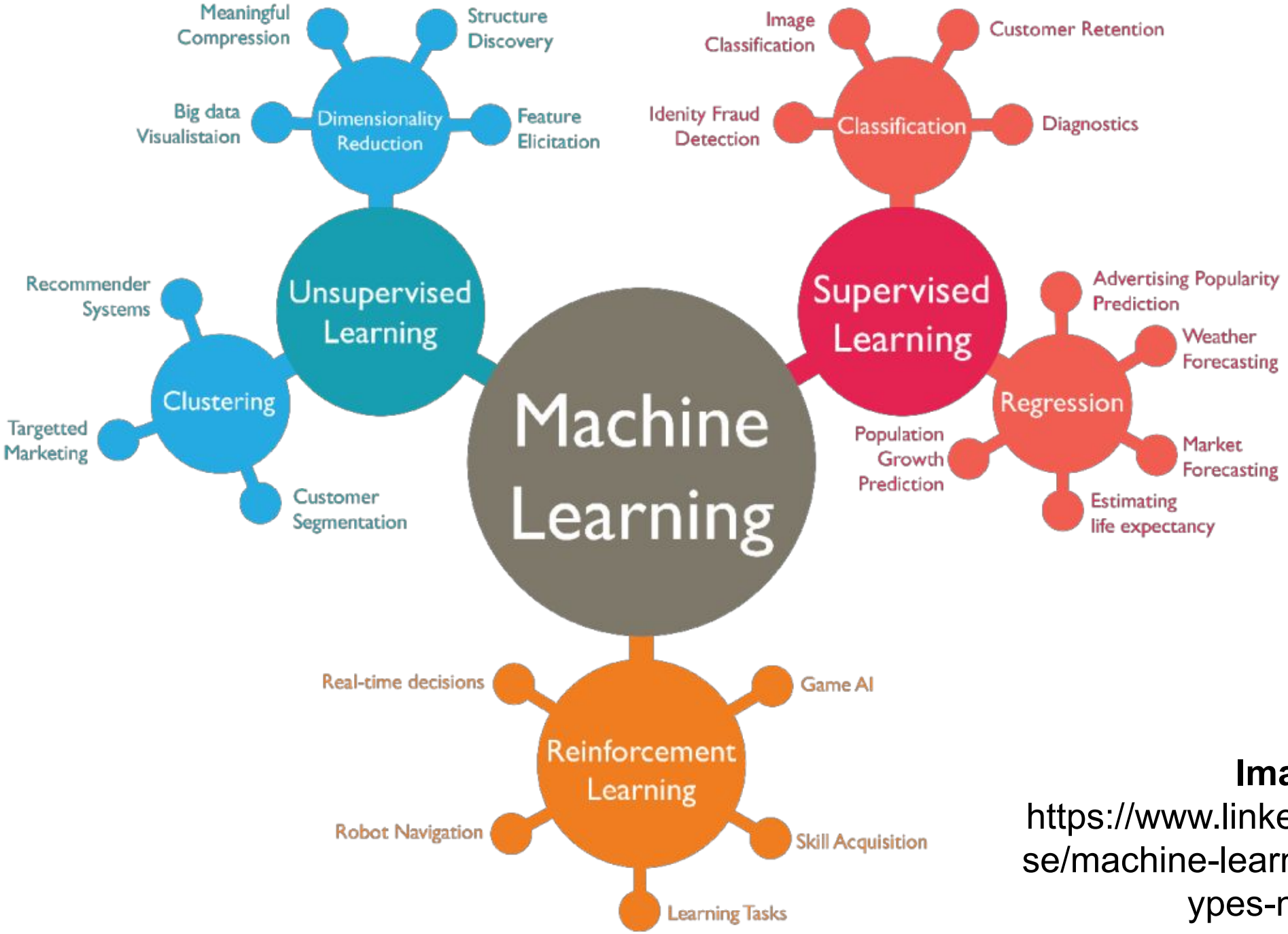


Image source:
https://www.linkedin.com/pulse/machine-learning-model-types-navjot-singh/

Why Machine Learning?

- It turns out that while symbolic AI is good at a lot of intelligent tasks, it **has a lot of limitations when it comes to other tasks**.

> *"The main lesson of thirty-five years of AI research is that the **hard problems are easy and the easy problems are hard**. The mental abilities of a four-year-old that we take for granted – recognizing a face, lifting a pencil, walking across a room, answering a question – in fact solve some of the hardest engineering problems ever conceived..."*
>
> Steven Pinker, *The Language Instinct (1994)*

Machine Learning Terms

- **Dataset:** collection of data instances where the model will “learn” from.

	index	sepallength	sepalwidth	petallength	petalwidth	class
0	0	5.0	3.2	1.2	0.2	Iris-setosa
1	1	6.7	3.1	4.4	1.4	Iris-versicolor
2	2	5.7	2.8	4.5	1.3	Iris-versicolor
3	3	7.7	3.0	6.1	2.3	Iris-virginica
4	4	6.7	3.1	5.6	2.4	Iris-virginica
5	5	4.7	3.2	1.3	0.2	Iris-setosa
6	6	6.4	2.7	5.3	1.9	Iris-virginica

Machine Learning Terms

- **Instance:** a single object / row in the dataset

	index	sepallength	sepalwidth	petallength	petalwidth	class
0	0	5.0	3.2	1.2	0.2	Iris-setosa
1	1	6.7	3.1	4.4	1.4	Iris-versicolor
2	2	5.7	2.8	4.5	1.3	Iris-versicolor
3	3	7.7	3.0	6.1	2.3	Iris-virginica
4	4	6.7	3.1	5.6	2.4	Iris-virginica
5	5	4.7	3.2	1.3	0.2	Iris-setosa
6	6	6.4	2.7	5.3	1.9	Iris-virginica

Machine Learning Terms

- **Label:** the target variable that is being predicted, i.e., the model is “learning” how to predict this variable (supervised learning only).

	index	sepallength	sepalwidth	petallength	petalwidth	class
0	0	5.0	3.2	1.2	0.2	Iris-setosa
1	1	6.7	3.1	4.4	1.4	Iris-versicolor
2	2	5.7	2.8	4.5	1.3	Iris-versicolor
3	3	7.7	3.0	6.1	2.3	Iris-virginica
4	4	6.7	3.1	5.6	2.4	Iris-virginica
5	5	4.7	3.2	1.3	0.2	Iris-setosa
6	6	6.4	2.7	5.3	1.9	Iris-virginica

Machine Learning Terms

- **Classes:** The list of possible values for the label (applicable to classification tasks). In this case: **{iris-setosa, iris-versicolor, iris-virginica}**

	index	sepallength	sepalwidth	petallength	petalwidth	class
0	0	5.0	3.2	1.2	0.2	Iris-setosa
1	1	6.7	3.1	4.4	1.4	Iris-versicolor
2	2	5.7	2.8	4.5	1.3	Iris-versicolor
3	3	7.7	3.0	6.1	2.3	Iris-virginica
4	4	6.7	3.1	5.6	2.4	Iris-virginica
5	5	4.7	3.2	1.3	0.2	Iris-setosa
6	6	6.4	2.7	5.3	1.9	Iris-virginica

Machine Learning Terms

- **Features:** The variables that will be considered when “learning” the rules / making the prediction

	index	sepallength	sepalwidth	petallength	petalwidth	class
0	0	5.0	3.2	1.2	0.2	Iris-setosa
1	1	6.7	3.1	4.4	1.4	Iris-versicolor
2	2	5.7	2.8	4.5	1.3	Iris-versicolor
3	3	7.7	3.0	6.1	2.3	Iris-virginica
4	4	6.7	3.1	5.6	2.4	Iris-virginica
5	5	4.7	3.2	1.3	0.2	Iris-setosa
6	6	6.4	2.7	5.3	1.9	Iris-virginica

Machine Learning Task Formulation

- **Goal:** We want to make a system that can estimate the price of a house.

- Can we use traditional AI for this?

Machine Learning Task Formulation

- **Supervised machine learning approach:** we can collect data on existing houses.

House ID	Location	Number of Bedrooms	Number of Bathrooms	Land Area (sq m)	House Area (sq m)	House Price (PHP)	Year Built
1	Quezon City	3	2	150	120	5,500,000	2010
2	Makati City	2	1	80	60	8,000,000	2015
3	Pasig City	4	3	200	180	6,800,000	2005
4	Taguig City	1	1	50	40	3,200,000	2020
5	Mandaluyong City	3	2	120	100	7,200,000	2018
6	Paranaque City	5	4	300	250	12,000,000	2000
7	Las Pinas City	2	1	70	55	4,500,000	2013

⋮

+ 1,000 more houses

Machine Learning Task Formulation

Instance: a single house

House ID	Location	Number of Bedrooms	Number of Bathrooms	Land Area (sq m)	House Area (sq m)	House Price (PHP)	Year Built
1	Quezon City	3	2	150	120	5,500,000	2010
2	Makati City	2	1	80	60	8,000,000	2015
3	Pasig City	4	3	200	180	6,800,000	2005
4	Taguig City	1	1	50	40	3,200,000	2020
5	Mandaluyong City	3	2	120	100	7,200,000	2018
6	Paranaque City	5	4	300	250	12,000,000	2000
7	Las Pinas City	2	1	70	55	4,500,000	2013

⋮

+ 1,000 more houses

Machine Learning Task Formulation

Label: what we want to predict

House ID	Location	Number of Bedrooms	Number of Bathrooms	Land Area (sq m)	House Area (sq m)	House Price (PHP)	Year Built
1	Quezon City	3	2	150	120	5,500,000	2010
2	Makati City	2	1	80	60	8,000,000	2015
3	Pasig City	4	3	200	180	6,800,000	2005
4	Taguig City	1	1	50	40	3,200,000	2020
5	Mandaluyong City	3	2	120	100	7,200,000	2018
6	Paranaque City	5	4	300	250	12,000,000	2000
7	Las Pinas City	2	1	70	55	4,500,000	2013

⋮

+ 1,000 more houses

Machine Learning Task Formulation

Features: What our predictions are based on

House ID	Location	Number of Bedrooms	Number of Bathrooms	Land Area (sq m)	House Area (sq m)	House Price (PHP)	Year Built
1	Quezon City	3	2	150	120	5,500,000	2010
2	Makati City	2	1	80	60	8,000,000	2015
3	Pasig City	4	3	200	180	6,800,000	2005
4	Taguig City	1	1	50	40	3,200,000	2020
5	Mandaluyong City	3	2	120	100	7,200,000	2018
6	Paranaque City	5	4	300	250	12,000,000	2000
7	Las Pinas City	2	1	70	55	4,500,000	2013

⋮

+ 1,000 more houses

Machine Learning Task Formulation

Features: What our predictions are based on

House ID	Location	Number of Bedrooms	Number of Bathrooms	Land Area (sq m)	House Area (sq m)	House Price (PHP)	Year Built
1	Quezon City	3	2	150	120	5,500,000	2010
2	Makati City	2	1	80	60	8,000,000	2015
3	Pasig City	4	3	200	180	6,800,000	2005
4	Taguig City	1	1	50	40	3,200,000	2020
5	Mandaluyong City	3	2	120	100	7,200,000	2018
6	Paranaque City	5	4	300	250	12,000,000	2000
7	Las Pinas City	2	1	70	55	4,500,000	2013

⋮

+ 1,000 more houses

Should we include this as a feature?

Machine Learning Task Formulation

Supervised machine learning task general format:

Predict ***<label>*** of a ***<instance>***
given the ***<features>***

Machine Learning Task Formulation

Supervised machine learning task general format:

Predict ***the price*** of a ***single house*** given the ***location, number of bedrooms, number of bathrooms, land area, house area, and the year it was built***.

This is a **regression** task!

Machine Learning Task Formulation

Features: What our predictions are based on

House ID	Location	Number of Bedrooms	Number of Bathrooms	Land Area (sq m)	House Area (sq m)	House Price (PHP)	Year Built
1	Quezon City	3	2	150	120	5,500,000	2010
2	Makati City	2	1	80	60	8,000,000	2015
3	Pasig City	4	3	200	180	6,800,000	2005
4	Taguig City	1	1	50	40	3,200,000	2020
5	Mandaluyong City	3	2	120	100	7,200,000	2018
6	Paranaque City	5	4	300	250	12,000,000	2000
7	Las Pinas City	2	1	70	55	4,500,000	2013

⋮

+ 1,000 more houses

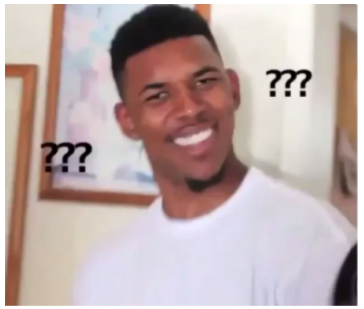
Should we include this as a feature?

Feature Selection Matters!

- If you put many features that are not connected to the label, models will generally have a harder time “learning” the patterns.

Predict the **favorite ice cream flavor** of a **person** given **their height, shoe size, and the number of siblings**.

Machine Learning Model:



- **Remember:** the features you use will also be the inputs in making a prediction!

Machine Learning Task Formulation

- **Goal:** I want to make a system that can automatically detect if an animal image is a photo of a mammal, reptile, or bird.

Machine Learning Task Formulation

- We can collect lots of data…

bird

bird

mammal

reptile

mammal

reptile

mammal

bird

+ 1,000 more images

Machine Learning Task Formulation

- We can collect lots of data…

Instance: A single image


bird


bird


mammal


reptile


mammal


reptile


mammal


bird

+ 1,000 more images

Machine Learning Task Formulation

- We can collect lots of data…

Label: mammal, bird, or reptile

bird

bird

mammal

reptile

mammal

reptile

mammal

bird

+ 1,000 more images

Machine Learning Task Formulation

- We can collect lots of data…

What are the features?

bird

bird

mammal

reptile

mammal

reptile

mammal

bird

+ 1,000 more images

Machine Learning Task Formulation

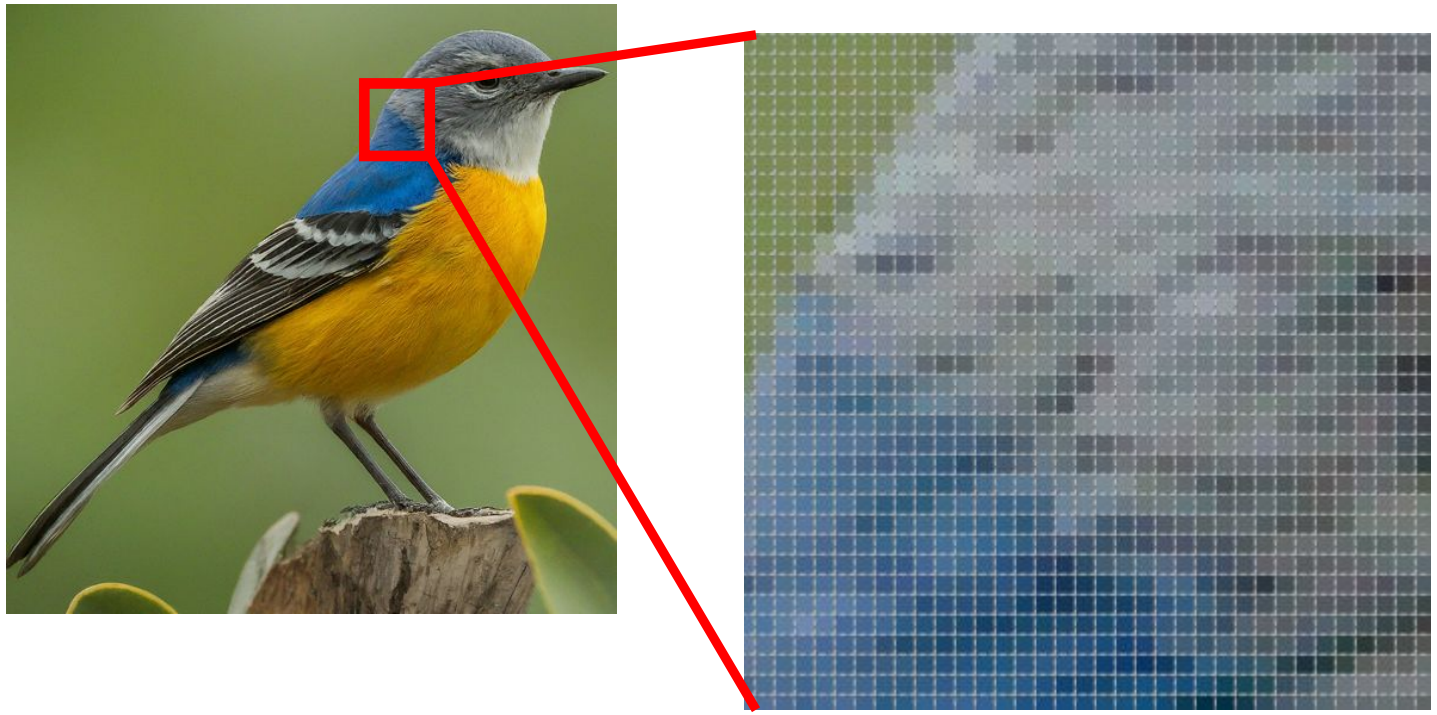
It would be nice if we have the following information…

- **Number of legs**
- **Presence of fur**
- **Presence of scales**
- **Presence of beak**
- **Head size ratio**

and so on…

But these are impossible for the computer to extract just from the image!

Machine Learning Task Formulation



- The computer sees an image as a 2D array of pixels
- Each pixel is represented by 3 numbers (RGB channels)
- Each of these numbers can be a feature!

Assuming we have 100 x 100 images, how many features do we have in total?

Machine Learning Task Formulation

Supervised machine learning task general format:

Given an ***animal image***, Predict if it is a ***mammal, reptile, or bird*** given the ***colors of each pixel in the image***.

This is a **multi-class classification** task!

What are Features, Really?

- You can think of the features as the **numerical representation** of a single instance.

[City, # bedrooms, # bathrooms, house area, lot area, year built]

```
[3, 3, 2, 150, 120, 2010]
```

Feature vector, i.e., the numerical representation of this house

6 features = 6-dimensional vector

What are Features, Really?

100 x 100 image

[1st pixel R, 1st pixel G, 1st pixel B, 2nd pixel R, 2nd pixel G, 2nd pixel B, …, last pixel R, last pixel G, last pixel B]

```
[72, 133, 79, 72, 134, 60, …, 73, 130, 61]
```

30,000 features = 30,000-dimensional vector

Is this the only way?

An Alternative Representation

100 x 100 image

[Average R value across all pixels, Average G value across all pixels, Average B value across all pixels]

```
[21, 159, 50]
```

3 features = 3-dimensional vector

This representation attempts to capture the dominant color channel in the image.

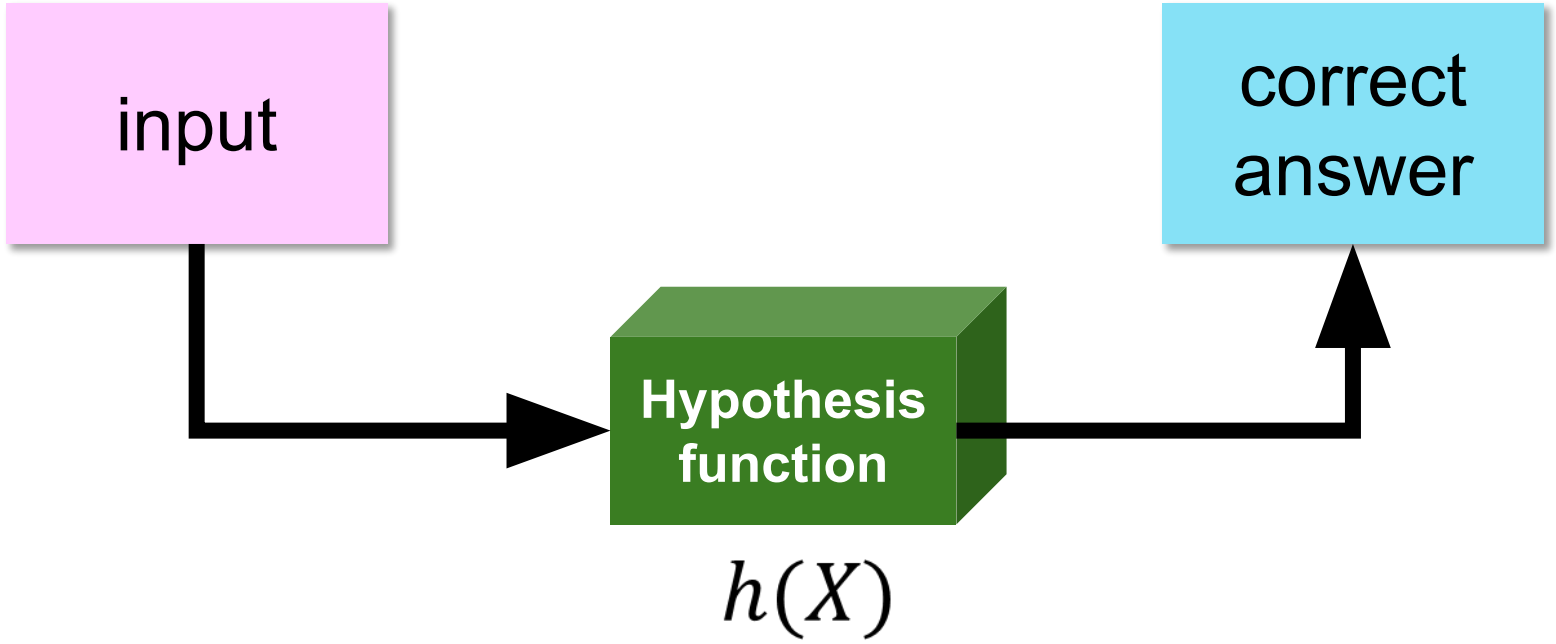
How will this affect the machine learning model?

Notations

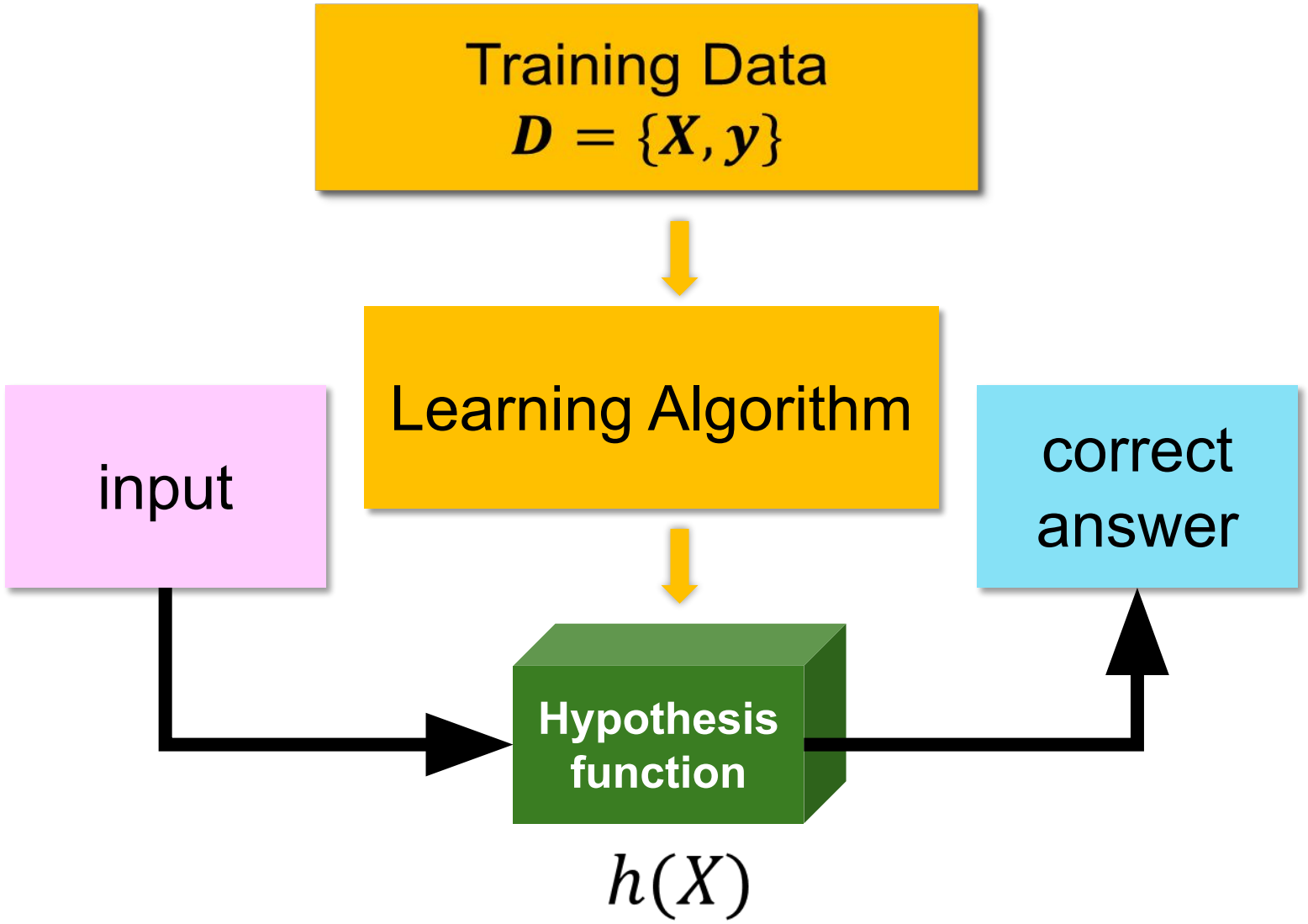
- D: the dataset (a subset of the population)
- X: feature matrix
- y: labels

	index	sepallength	sepalwidth	petallength	petalwidth	class
0	0	5.0	3.2	1.2	0.2	Iris-setosa
1	1	6.7	3.1	4.4	1.4	Iris-versicolor
2	2	5.7	2.8	4.5	1.3	Iris-versicolor
3	3	7.7	3.0	6.1	2.3	Iris-virginica
4	4	6.7	3.1	5.6	2.4	Iris-virginica
5	5	4.7	3.2	1.3	0.2	Iris-setosa
6	6	6.4	2.7	5.3	1.9	Iris-virginica

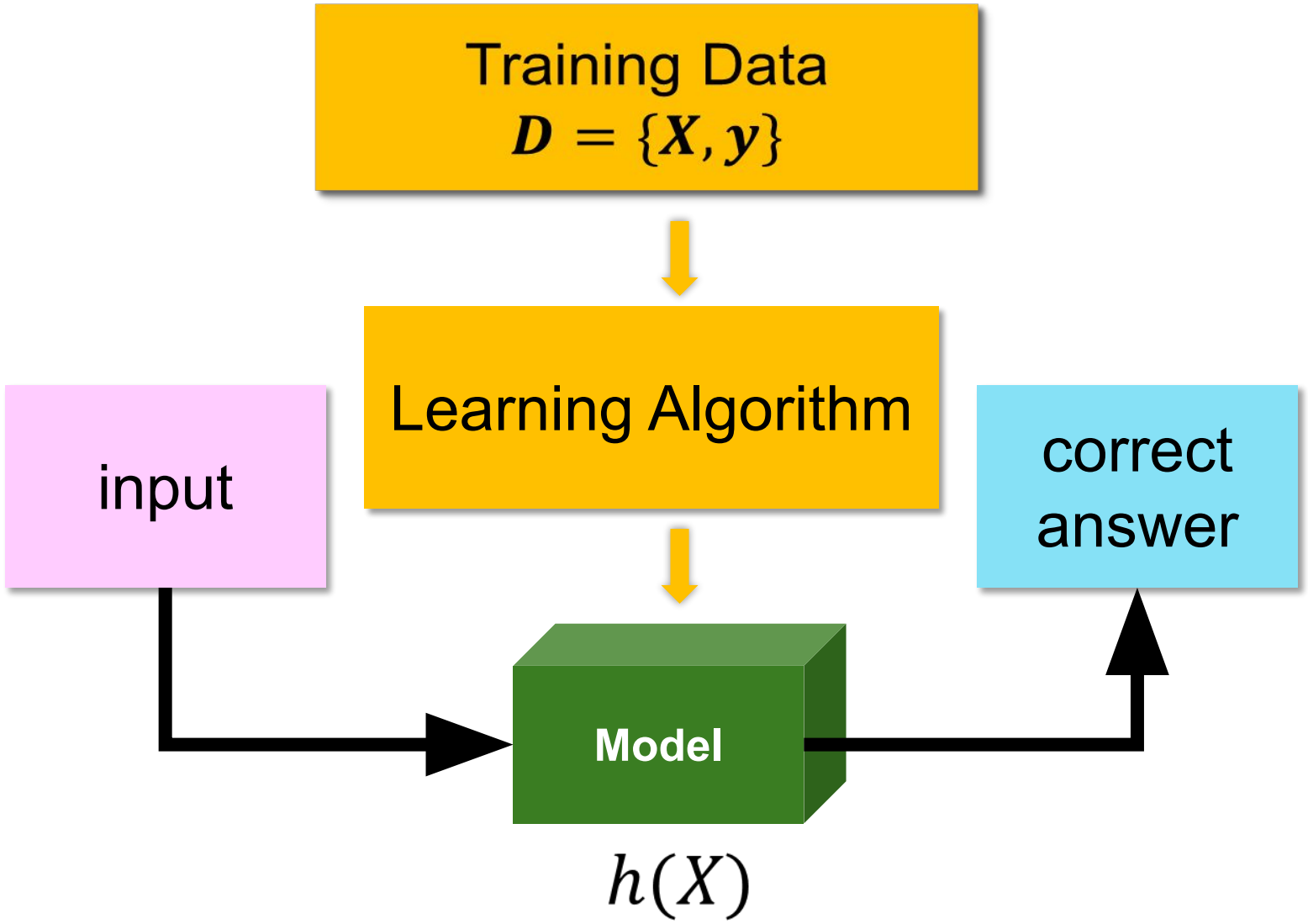
Training a Supervised ML Model



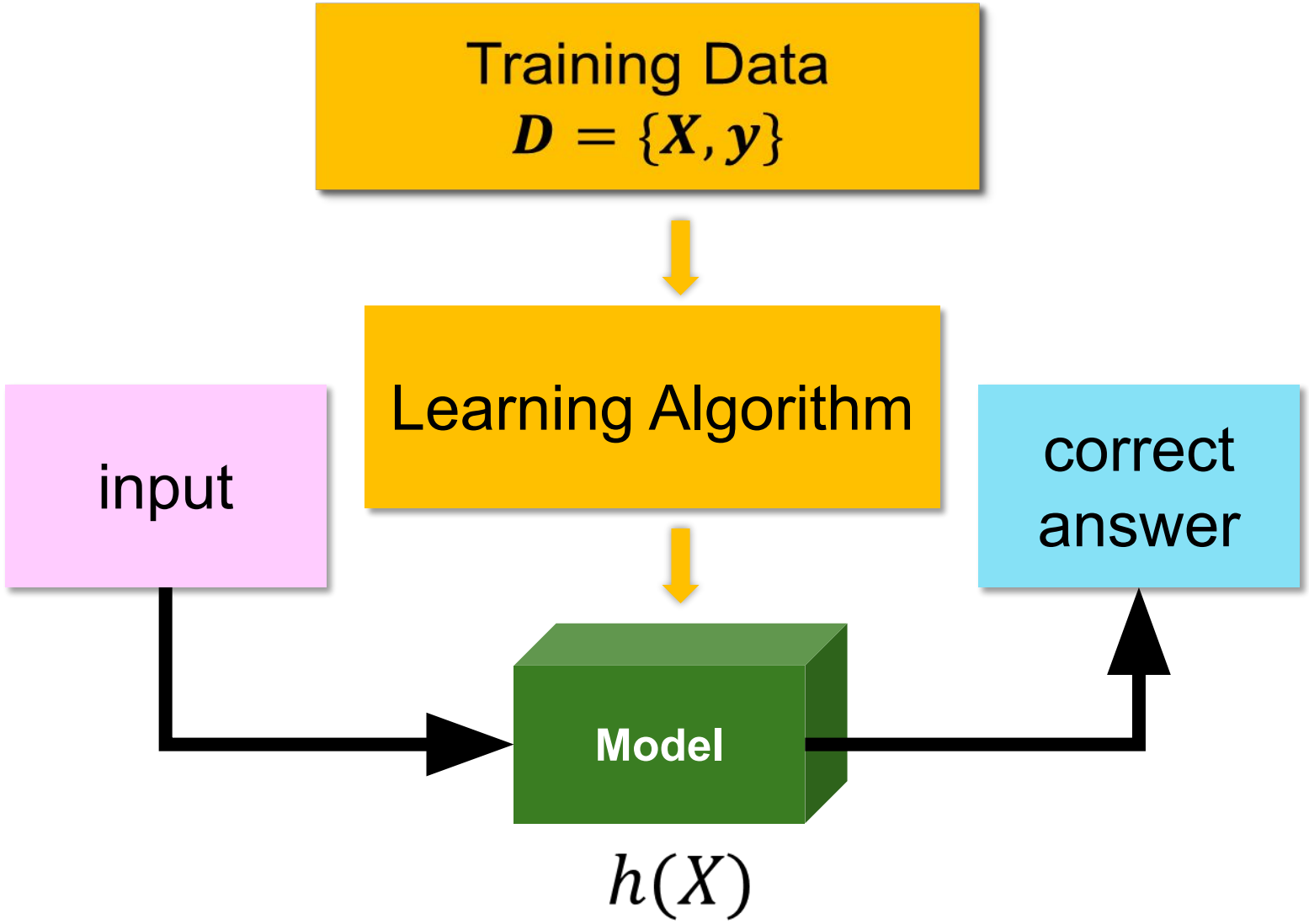
Training a Supervised ML Model



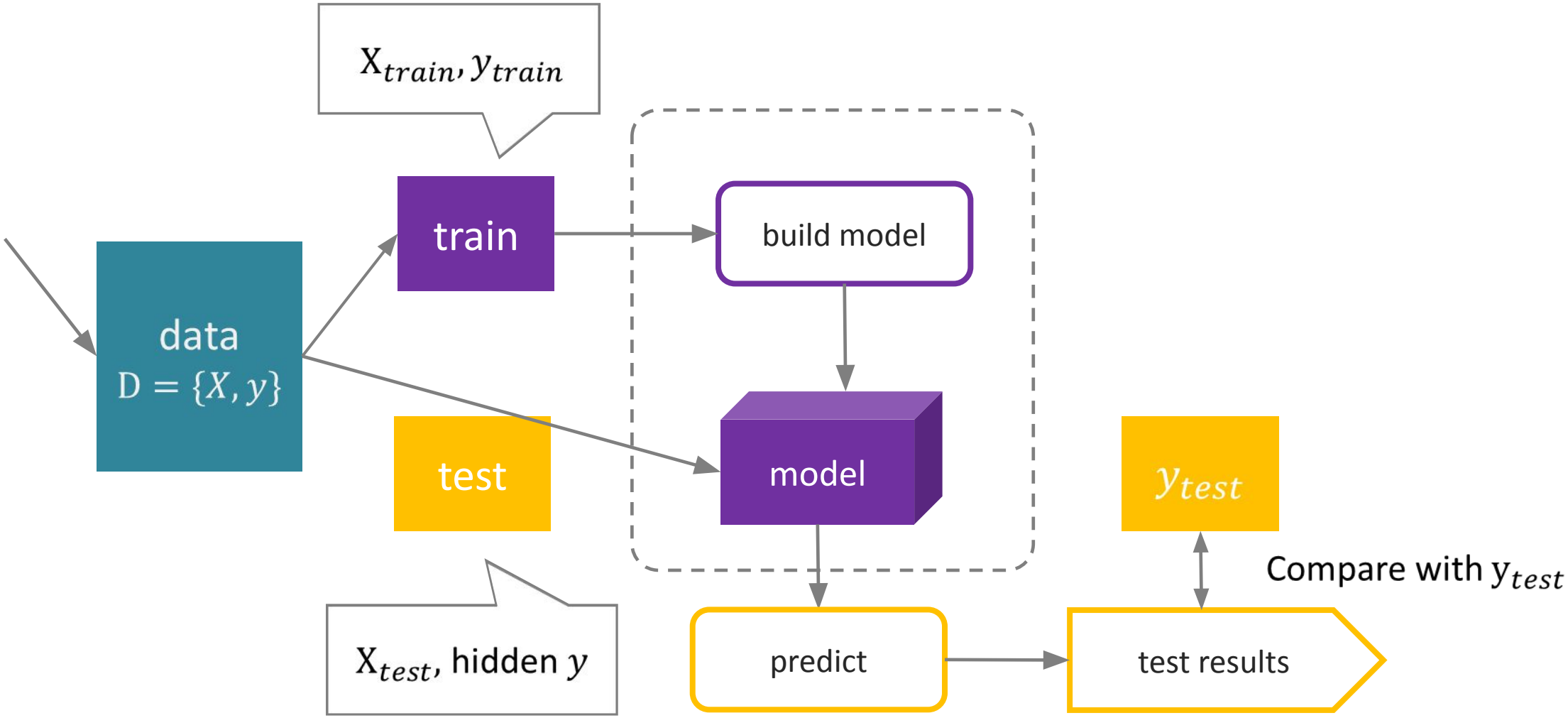
Training a Supervised ML Model



How to Know If Model is Good?



Evaluating a Supervised ML Model



Why Split the Train and Test Set?

- The model “learned” the patterns based on the training data.
- If the same data was used to “test” the model, then the result will likely be good, but biased.
- Therefore, we need to test the performance of the model on **data it has not yet seen before**.

Basic Supervised ML Pipeline

- Collect data.
- Preprocess data (exploratory data analysis, cleaning, etc.)
- Identify features and label.
- Split data into training set and test set.
- Build and fine-tune model from the training set.
- Run the test set on the model to measure its performance.
- Iterate as needed.

Basic Unsupervised ML Pipeline

- Collect data.
- Preprocess data (exploratory data analysis, cleaning, etc.)
- Identify features.
- Build and fine-tune model from the dataset.
- Perform expert interpretation and validation on results.
- Iterate as needed.