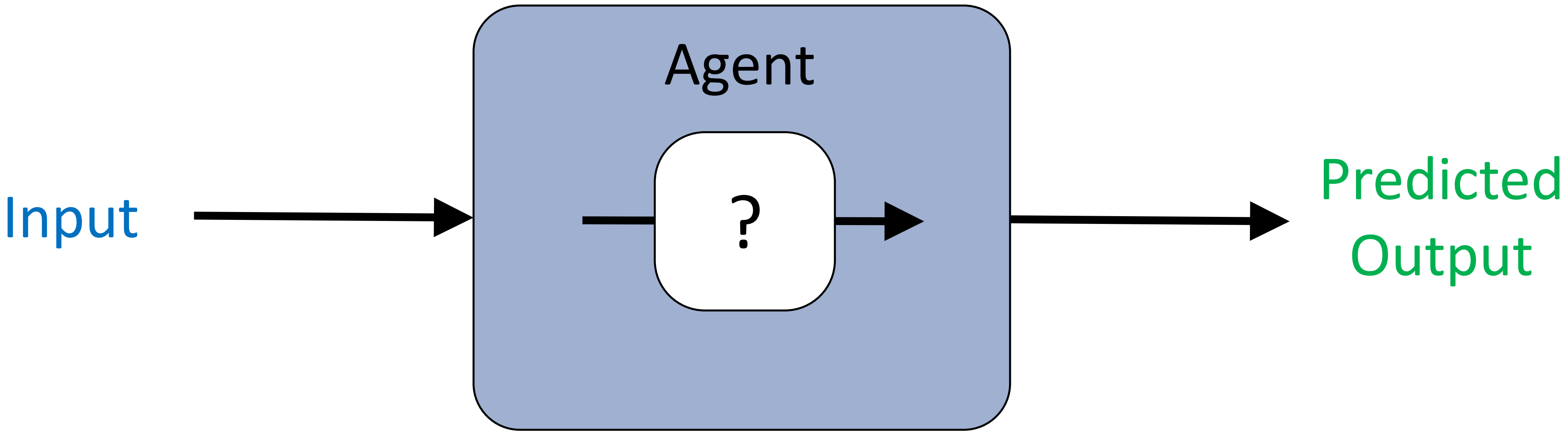
10-315
Machine Learning Problem Formulation

Instructor: Pat Virtue

Agent: Simple Input/Output Task



Task Input and Output

Input	Task	Output
Petal measurements	Iris classification	Category
Time of day	Traffic prediction	Traffic Volume
Image	Image classification	Category
Image	Image denoising	Image
Text	Text to image generation	Image
???	Face generation	Image

Today

ML Problem Formulation

- Task input and output
- Task, Performance, Experience
- Data and notation
- Supervised Learning
  - Classification and Regression
- Unsupervised Learning

ML Training and Models

- Nearest Neighbor
- Linear
- Neuron

ML Problem Formulation

Machine Learning Problem Formulation

Three components <*T,P,E*>:

1. Task, *T*
2. Performance measure, *P*
3. Experience, *E*

Definition of learning:

A computer program **learns** if its performance at tasks in *T*, as measured by *P*, improves with experience *E*

Definition from (Mitchell, 1997)

Machine Learning Problem Formulation

Notation

Task

Formalize the task as a mapping from input to output

$$h(x) \rightarrow \hat{y}$$

Experience

Data! Task experience examples will usually be pairs: (input, measured output)

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$$

Performance measure

Objective function that gives a single numerical value representing how well the system performs for a given dataset

- Classification: error rate
- Regression: mean squared error

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y^{(i)} \neq \hat{y}^{(i)})$$

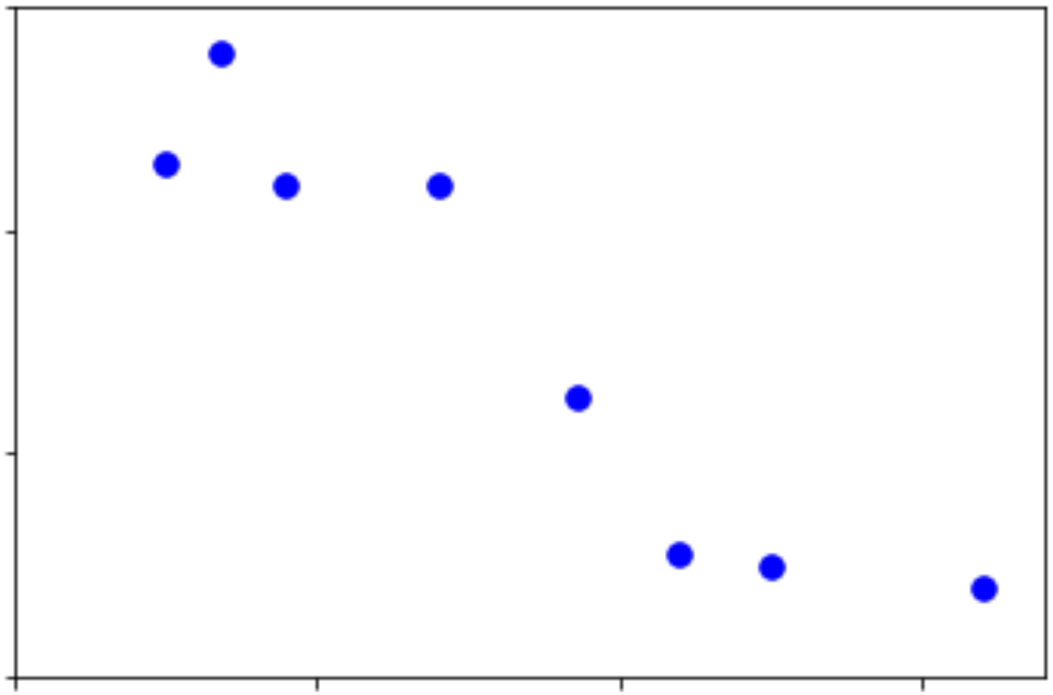
$$\frac{1}{N}\sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2$$

Experience: Data and Notation

Example Dataset: Selling My Car

Example

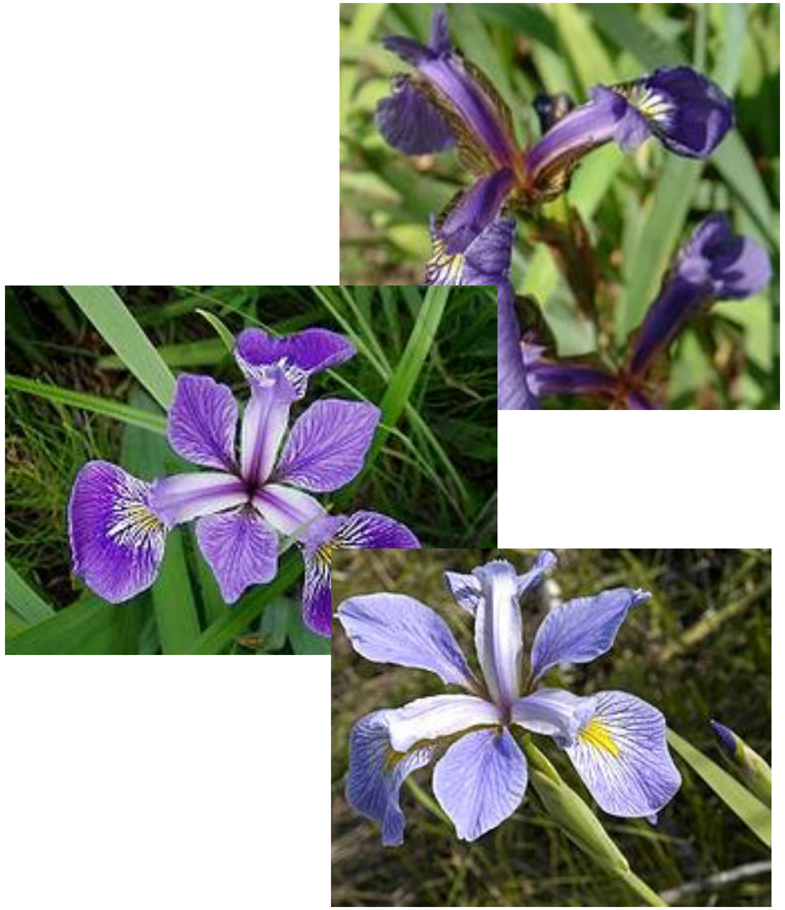
Trying to see how much I should sell my car for. Looking up data from car websites, I find the mileage for a set of cars and the selling price for each car.



Example Dataset: Fisher Iris Dataset

Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8



Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Example Dataset: Fisher Iris Dataset

```
from sklearn import datasets

iris = datasets.load_iris()
X = iris.data
y = iris.target
```

Assume samples in data are i.i.d.

Dataset notation

$$\mathcal{D} = \left\{\left(y^{(i)}, \mathbf{x}^{(i)}\right)\right\}_{i=1}^{N}$$

$$= \left\{\left(y^{(i)}, x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}\right)\right\}_{i=1}^{N}$$

Linear algebra can represent all data

$\mathbf{y} \in \{0,1,2\}^N$

$X \in \mathbb{R}^{N \times 4}$ (design matrix)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Example Dataset: Fisher Iris Dataset

```
from sklearn import datasets

iris = datasets.load_iris()
X = iris.data
y = iris.target
```

Assume samples in data are i.i.d.

Dataset notation

$$\mathcal{D} = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^{N}$$

$$= \{(y^{(i)}, x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)})\}_{i=1}^{N}$$

Data point $i = 6$: $(y^{(6)}, \mathbf{x}^{(6)})$

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Example Dataset: Fisher Iris Dataset

```
from sklearn import datasets

iris = datasets.load_iris()
X = iris.data
y = iris.target
```

Assume samples in data are i.i.d.

Dataset notation

$$\mathcal{D} = \left\{\left(y^{(i)}, \mathbf{x}^{(i)}\right)\right\}_{i=1}^{N}$$

$$= \left\{\left(y^{(i)}, x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}\right)\right\}_{i=1}^{N}$$

Linear algebra can represent all data

$\mathbf{y} \in \{0,1,2\}^N$

$X \in \mathbb{R}^{N\times 4}$ (design matrix)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

ML Data : Supervised vs Unsupervised

Supervised training data:

Pairs of input and output

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Unupervised training data:

No output data (i.e., no answers!)

$$\mathcal{D} = \{(\mathbf{x}^{(i)} \quad)\}_{i=1}^{N}$$

Sepal Length	Sepal Width	Petal Length	Petal Width
4.3	3.0	1.1	0.1
4.9	3.6	1.4	0.1
5.3	3.7	1.5	0.2
4.9	2.4	3.3	1.0
5.7	2.8	4.1	1.3
6.3	3.3	4.7	1.6
5.9	3.0	5.1	1.8

ML Data : Supervised vs Unsupervised

Supervised training data:

Pairs of input and output

$$\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$$

Unupervised training data:

No output data (i.e., no answers!)

$$\mathcal{D} = \left\{\left(\mathbf{x}^{(i)} \quad\quad \right)\right\}_{i=1}^{N}$$

ML Tasks: Supervised Learning

Supervised learning: Pairs of input and output in training data

$$\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N} \qquad h(\mathbf{x}) \to \hat{y}$$

Classification

- Output labels
- $y \in \mathcal{Y}$, where $\mathcal{Y}$ is discrete and order of values has no meaning

Regression

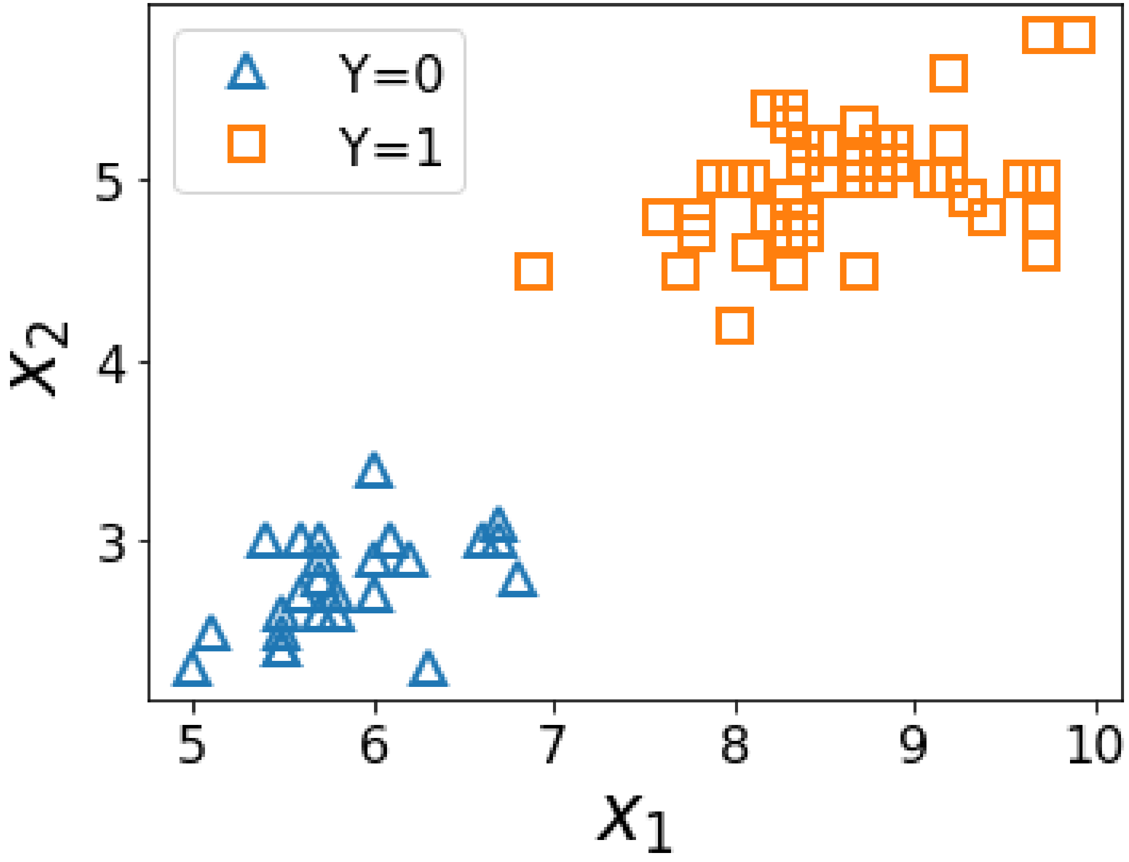
- Output values
- $y \in \mathcal{Y}$, where $\mathcal{Y}$ is usually continuous, order of values has meaning

Task: Classification

ML Task: Classification

Predict species label from first two input measurements

$h(\mathbf{x}) \rightarrow \hat{y}$



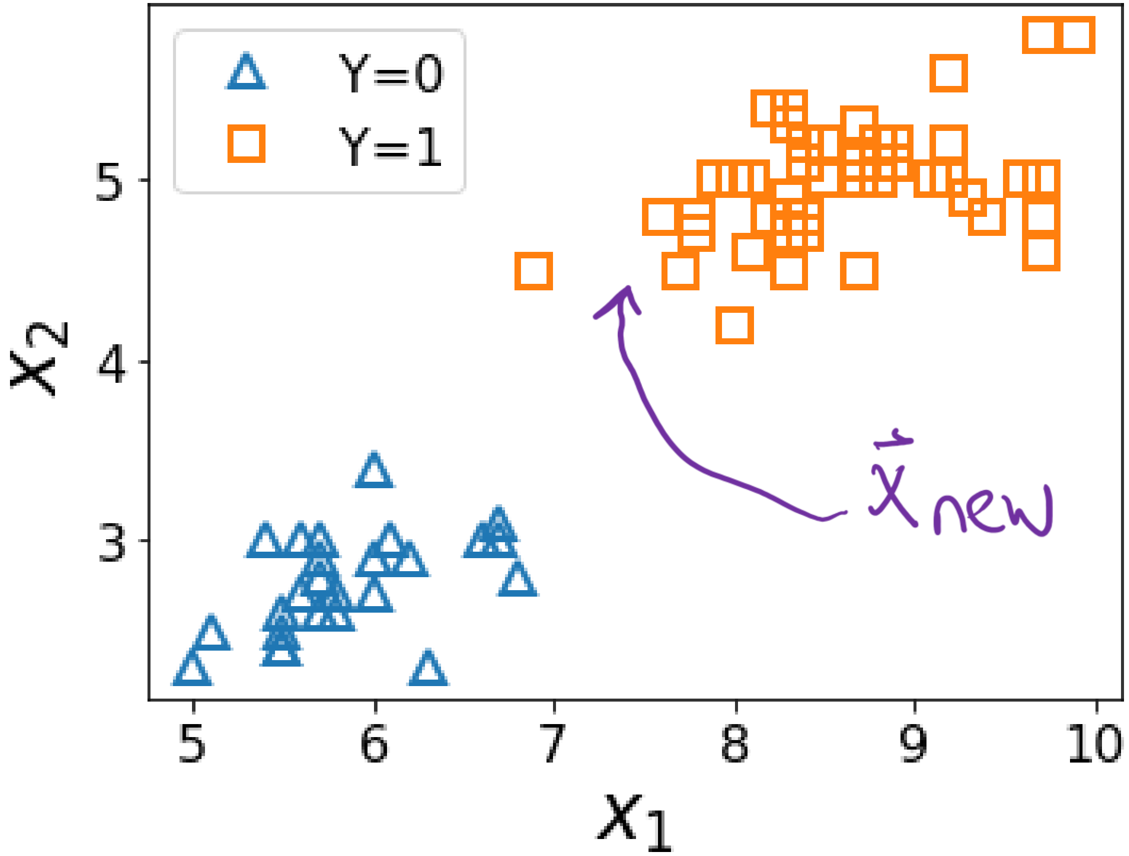
Species	Sepal Length	Sepal Width
0	4.3	3.0
0	4.9	3.6
0	5.3	3.7
1	4.9	2.4
1	5.7	2.8
1	6.3	3.3

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

ML Task: Classification

Nearest neighbor classification algorithm

Find $\mathbf{x}^{(i)}$ closest to $\mathbf{x}_{new}$. Then return it's label $y^{(i)}$.



Species	Sepal Length	Sepal Width
0	4.3	3.0
0	4.9	3.6
0	5.3	3.7
1	4.9	2.4
1	5.7	2.8
1	6.3	3.3

$\vec{x}_{new}$

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Classification

Notation alert: Indicator function

$$\mathbb{I}(z) = \mathbf{1}(z) = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Iris data example

$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, where $\mathbf{x}^{(i)} \in \mathbb{R}^4$, $y^{(i)} \in \{0, 1, 2\}$

Predict species label from input measurements

$$h(\mathbf{x}) \to \hat{y}$$

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Performance measure?

Classification error rate

- Fraction of times $y \neq \hat{y}$ in a given dataset
- $\frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y^{(i)} \neq \hat{y}^{(i)})$

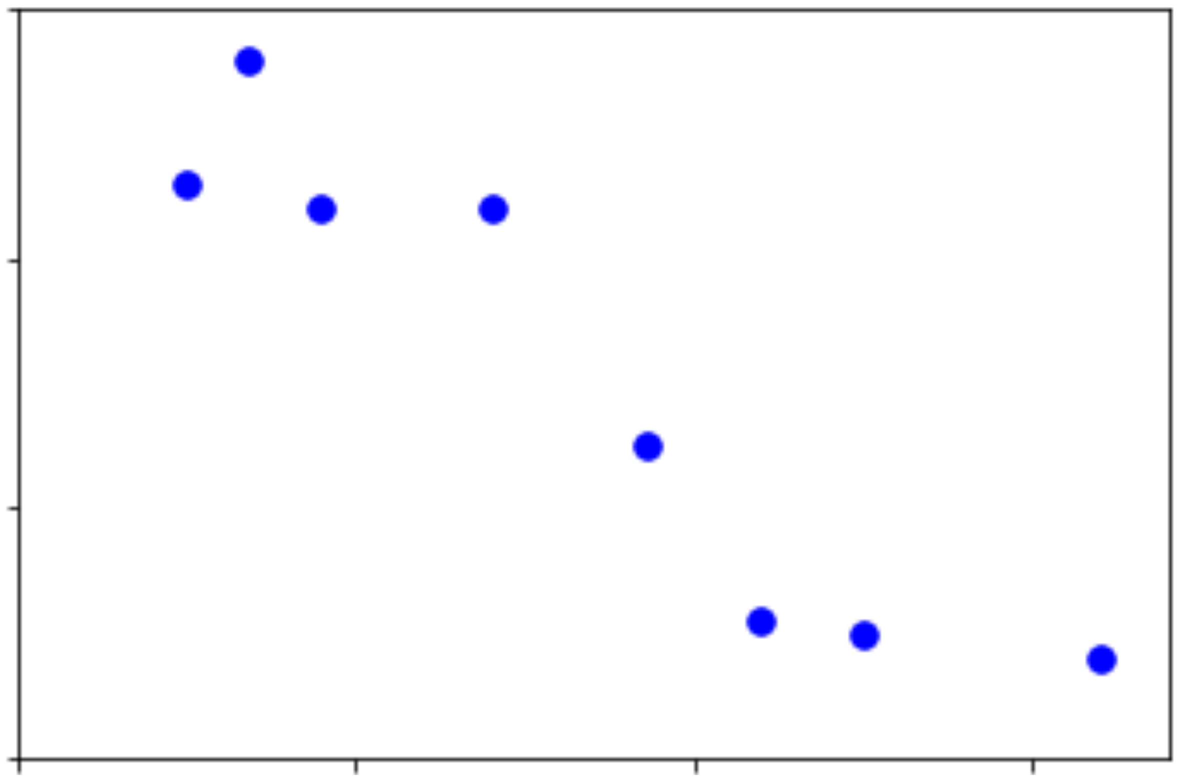
Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

ML Task: Regression

Regression: learning a model to predict a numerical output (but not numbers that just represent categories, that would be classification)

Example

Trying to see how much I should sell my car for. Looking up data from car websites, I find the mileage for a set of cars and the selling price for each car.



Unsupervised Tasks

ML Tasks

Unsupervised learning

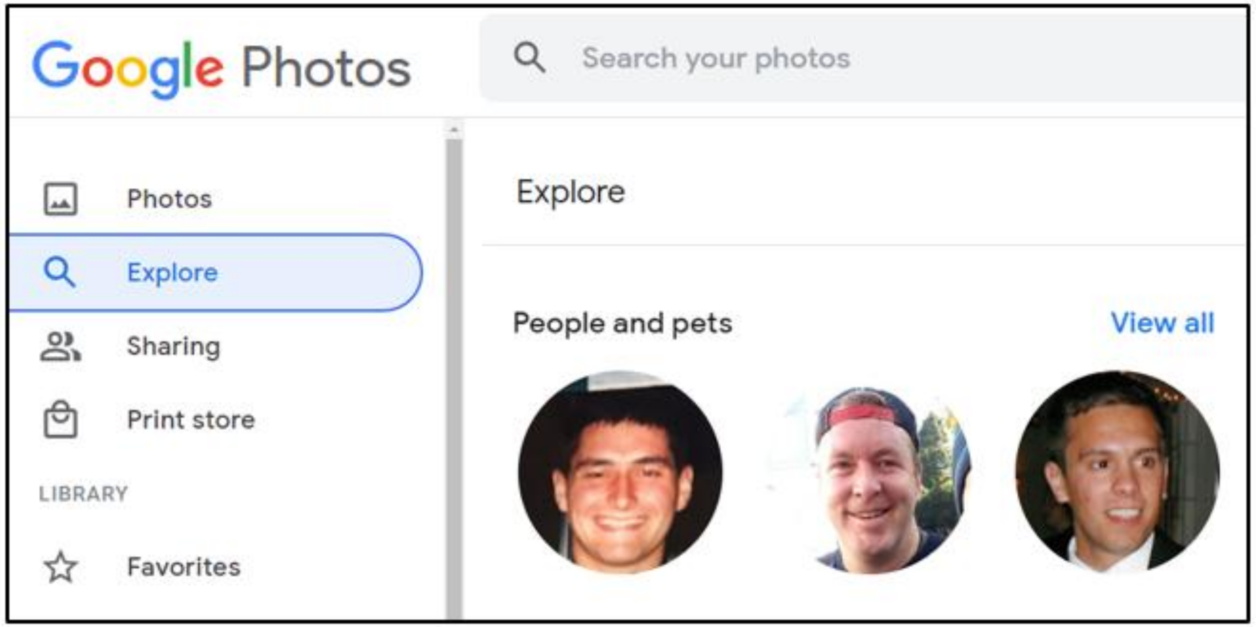
$$\mathcal{D} = \{\ \mathbf{x}^{(i)}\ \}_{i=1}^{N} \qquad h(\mathbf{x}) \rightarrow ???$$

- Training data has no output values
- Tasks can vary
- Often used to organize data for future (minimally) supervised learning

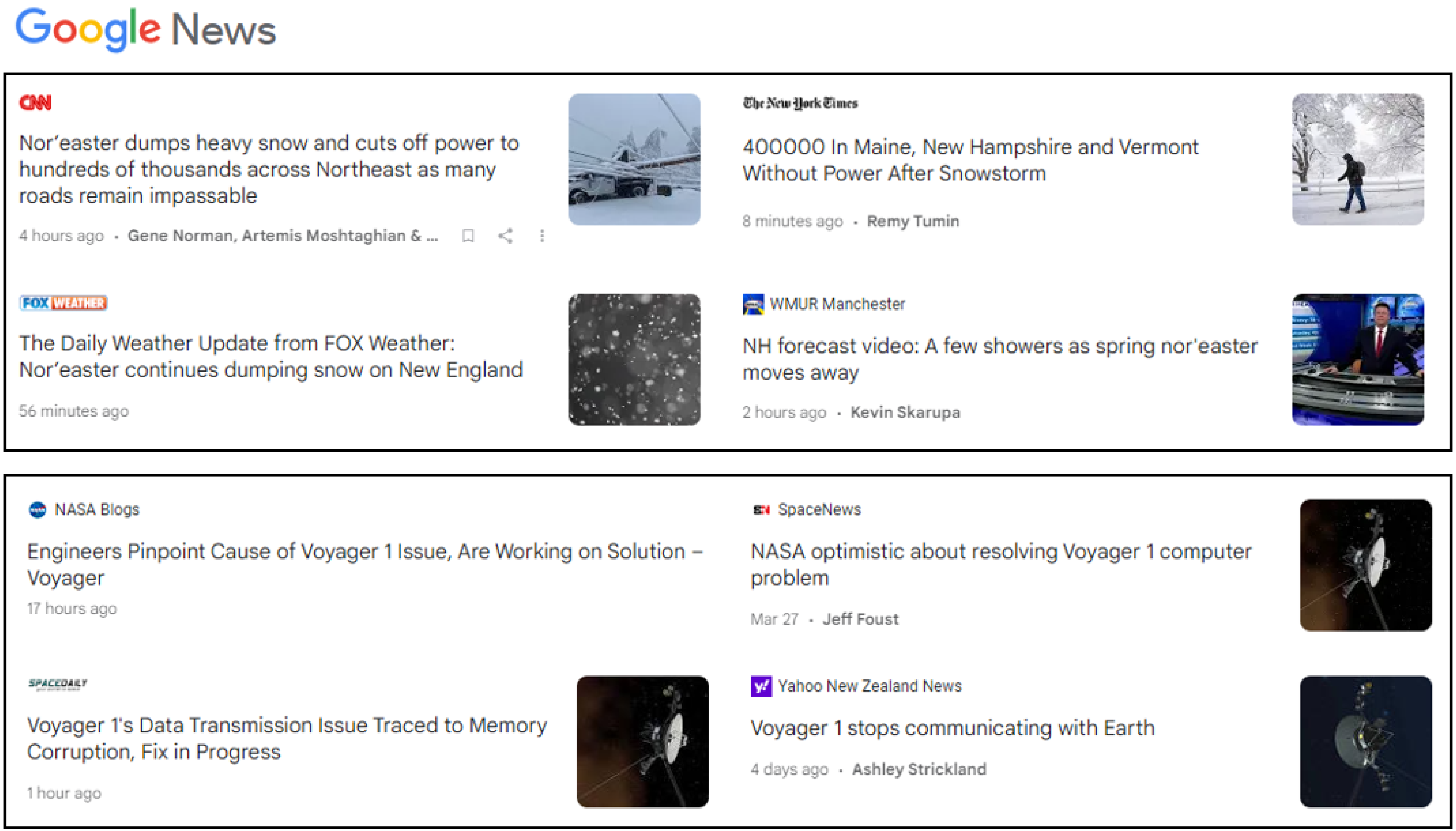
Task: Unsupervised Face Generation

https://thispersondoesnotexist.com/

Tasks: Unsupervised Clustering (Photos)



Tasks: Unsupervised Clustering (News)



ML Tasks

Unsupervised learning

$$\mathcal{D} = \{ \mathbf{x}^{(i)} \}_{i=1}^{N} \qquad h(\mathbf{x}) \rightarrow ???$$

- Training data has no output values
- Tasks can vary
- Often used to organize data for future (minimally) supervised learning

Example: Unsupervised autoencoder → Random image generation

$$\mathbf{x} \rightarrow \boxed{h(\mathbf{x})} \rightarrow \hat{\mathbf{x}}$$

$$\mathbf{x} \rightarrow \boxed{f(\mathbf{x}) \rightarrow \mathbf{z} \rightarrow g(\mathbf{z})} \rightarrow \hat{\mathbf{x}} \qquad \mathbf{z} \rightarrow \boxed{g(\mathbf{z})} \rightarrow \hat{\mathbf{x}}$$

ML Tasks

Unsupervised learning

$$\mathcal{D} = \{ \ \mathbf{x}^{(i)} \ \}_{i=1}^{N} \qquad h(\mathbf{x}) \rightarrow ???$$

- Training data has no output values
- Tasks can vary
- Often used to organize data for future (minimally) supervised learning

Example: Text Generation

Vocab pause

Task

- Prediction
- Inference
- Hypothesis function
- Classification
- Regression

Experience/Data

Input

- Input feature
- Measurement
- Attribute

Output

- Target
- Class/category/label
- True output
- Measured output
- Predicted output

Supervised

Unsupervised

Performance Measure

- Objective function

Classification

- Error rate
- Accuracy rate

Regression

- Mean squared error

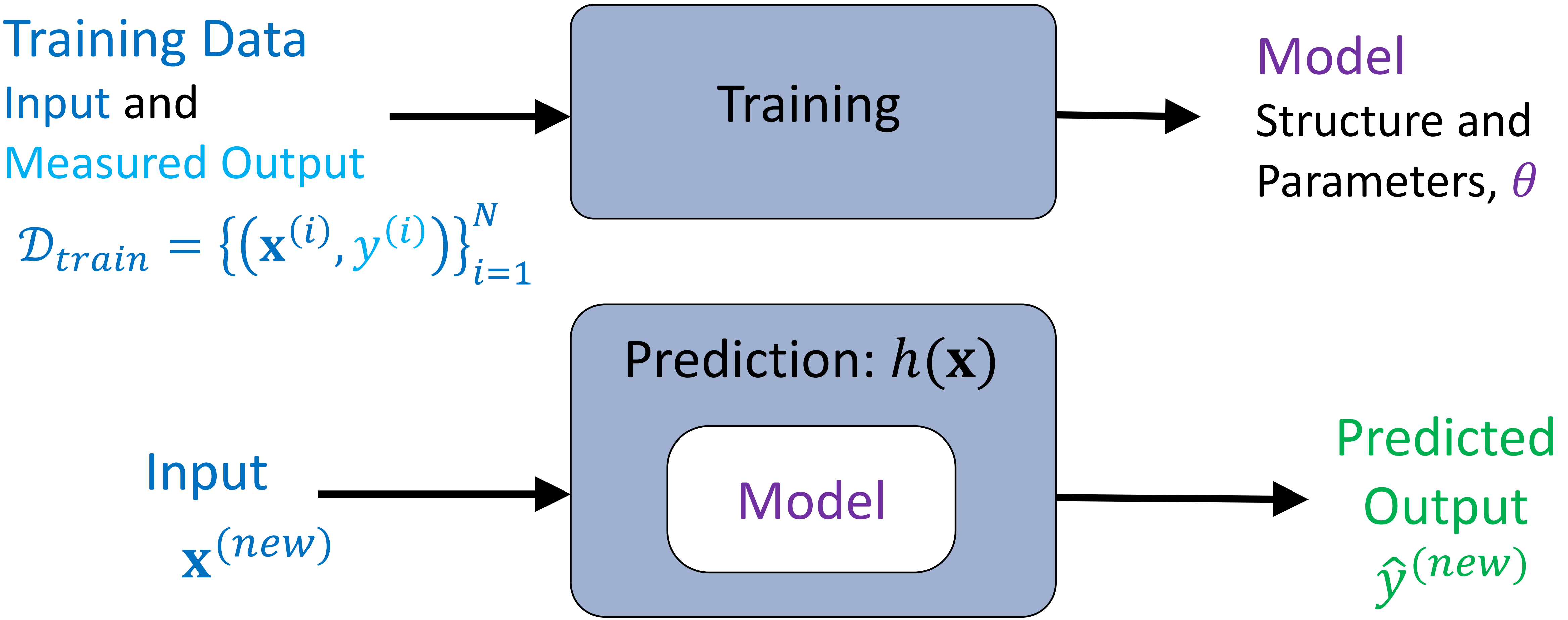
Training

- Model
- Model structure
- Model parameters

Training and ML Models

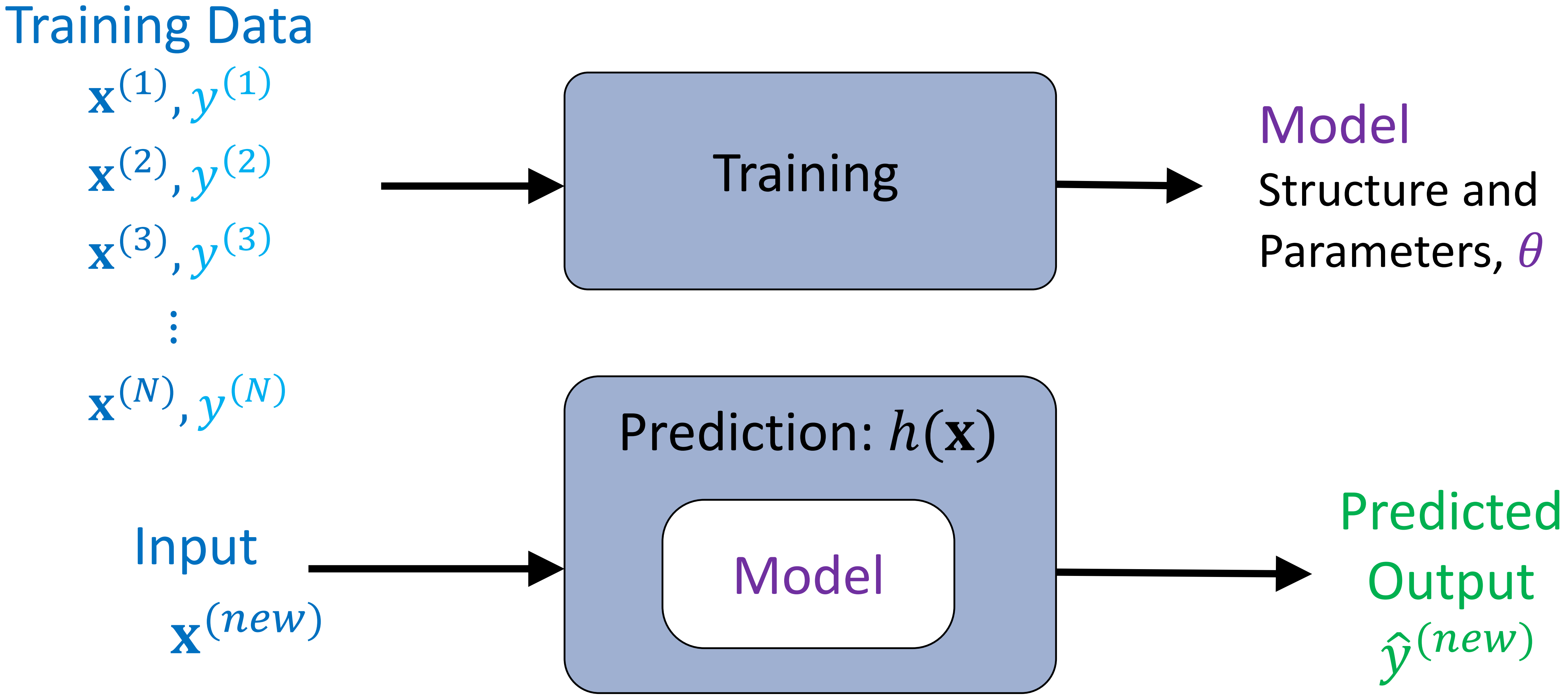
Machine Learning

Using (training) data to learn a model that we'll later use for prediction



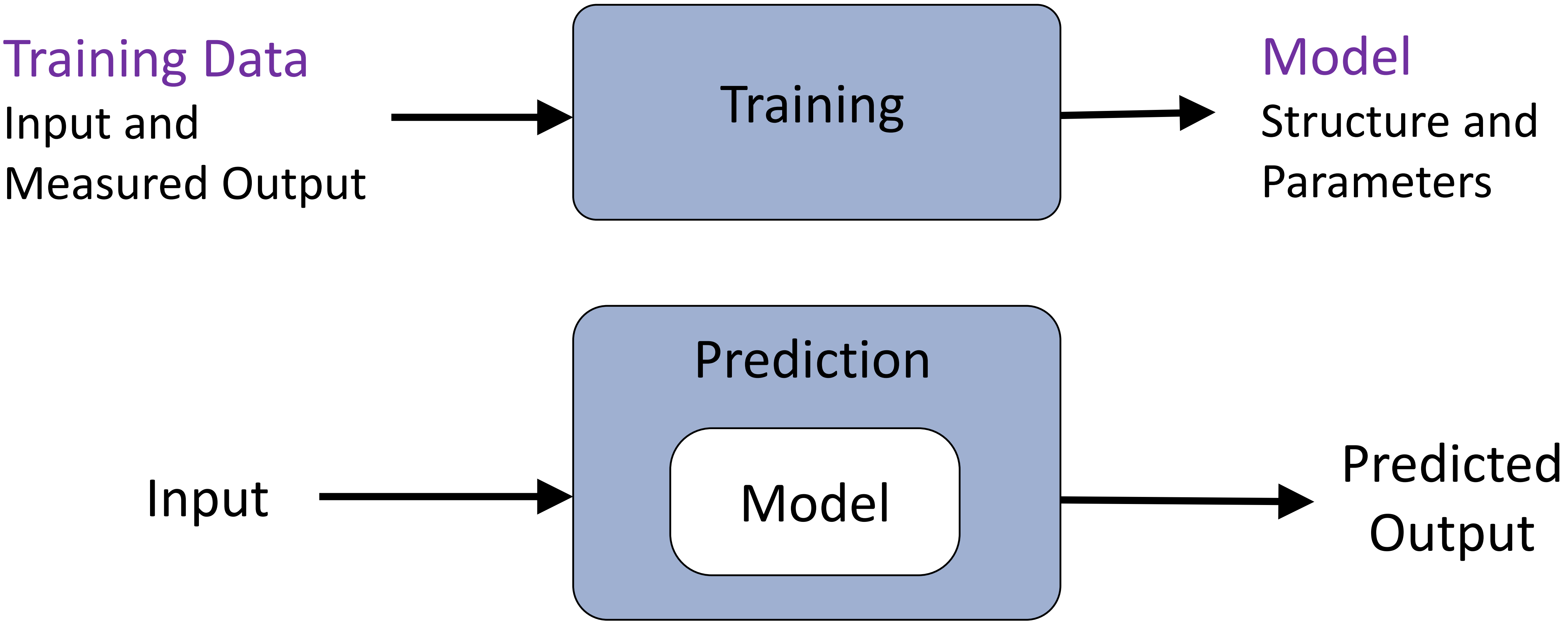
Machine Learning

Using (training) data to learn a model that we'll later use for prediction



Machine Learning

Using (training) data to learn a model that we'll later use for prediction



Classification

Iris data example

$$\mathcal{D} = \left\{\left(y^{(i)}, \mathbf{x}^{(i)}\right)\right\}_{i=1}^{N}, \text{ where } \mathbf{x}^{(i)} \in \mathbb{R}^4,\ y^{(i)} \in \{0, 1, 2\}$$

Predict species label from input measurements

$$h(\mathbf{x}) \rightarrow \hat{y}$$

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Performance measure?

Classification error rate

- Fraction of times $y \neq \hat{y}$ in a given dataset

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Classification

Notation alert: Indicator function

$$\mathbb{I}(z) = \mathbf{1}(z) = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Iris data example

$$\mathcal{D} = \left\{\left(y^{(i)}, \mathbf{x}^{(i)}\right)\right\}_{i=1}^{N}, \text{ where } \mathbf{x}^{(i)} \in \mathbb{R}^4,\, y^{(i)} \in \{0, 1, 2\}$$

Predict species label from input measurements

$$h(\mathbf{x}) \rightarrow \hat{y}$$

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Performance measure?

Classification error rate

- Fraction of times $y \neq \hat{y}$ in a given dataset
- $\frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(y^{(i)} \neq \hat{y}^{(i)}\right)$

Images and full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

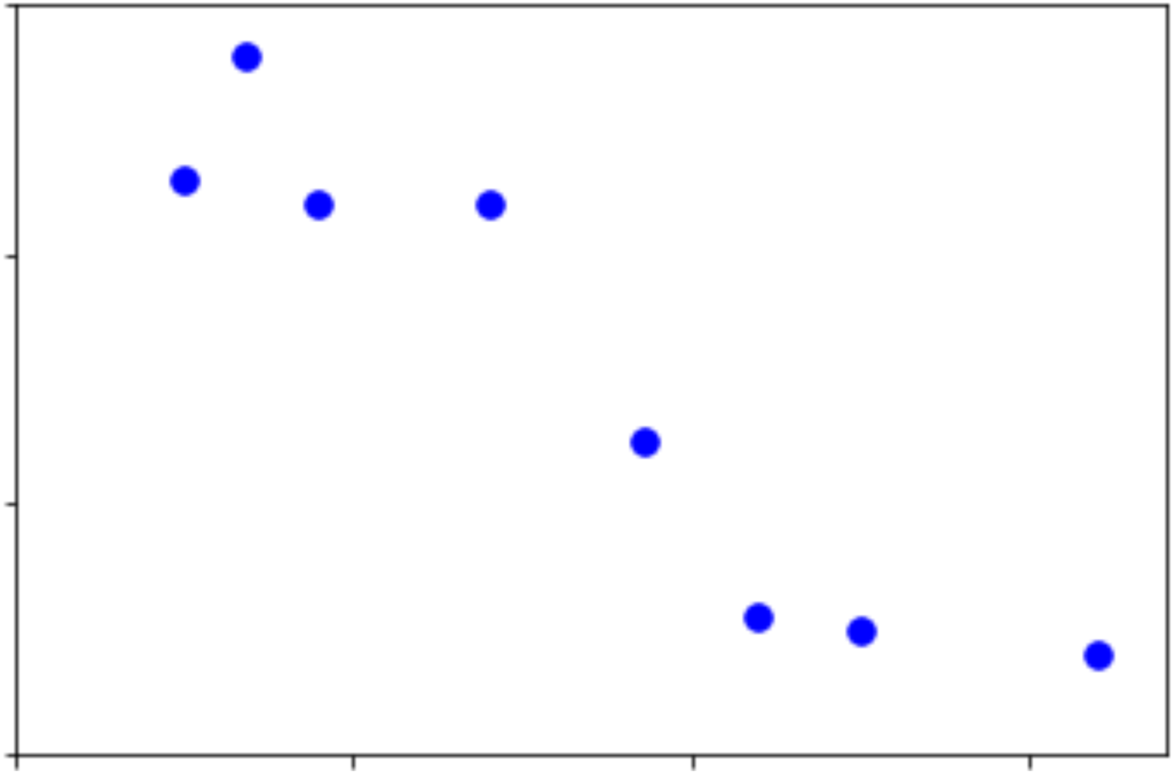
Regression Model

Regression: learning a model to predict a numerical output (but not numbers that just represent categories, that would be classification)

Model: Linear

Structure:

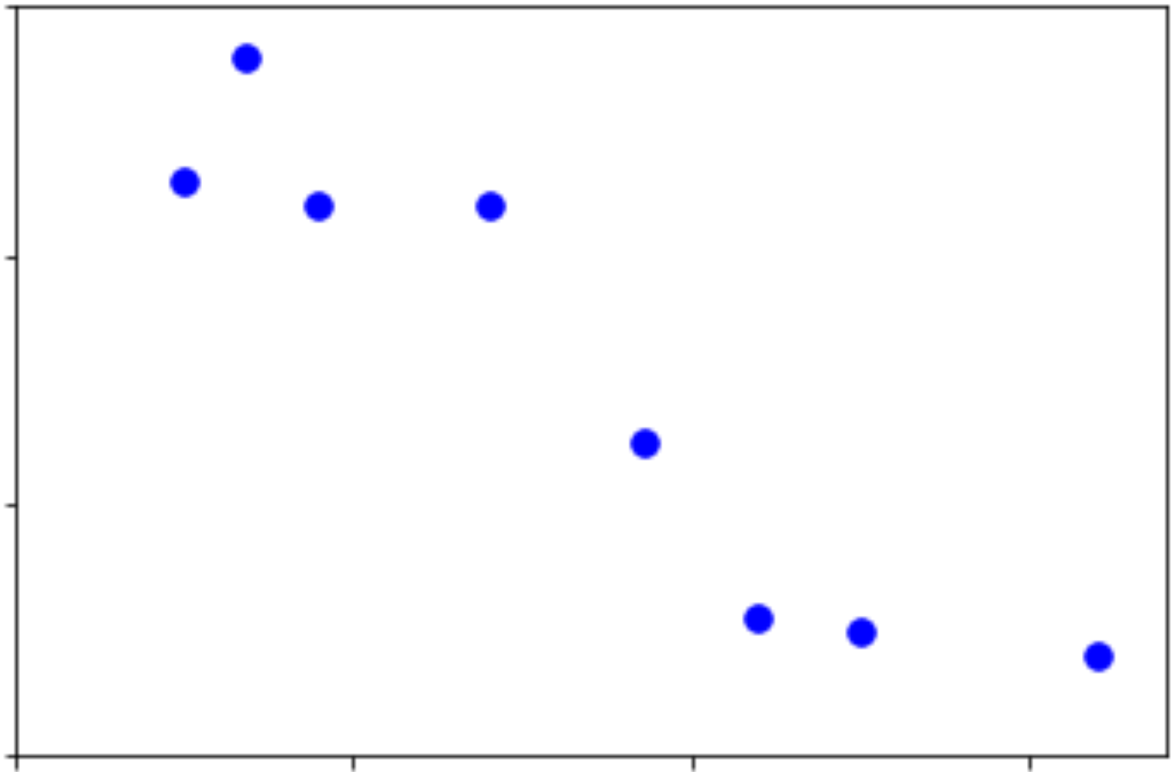
Parameters:



Regression Model

Regression: learning a model to predict a numerical output (but not numbers that just represent categories, that would be classification)

How do we know which model and model parameters are best?



Regression Model

Regression: learning a model to predict a numerical output (but not numbers that just represent categories, that would be classification)

How do we know which model and model parameters are best?