Supervised Learning:

Universe consists of everything

Stuff we don't know

Stuff we do know

Examples:
- Predict if the Coca-Cola stock will go up tomorrow.
- Predict if an email is spam or not.
- "" if a photo contains a human face.
- "" what a user said to a home assistant device (e.g. Alexa)

Goal: Learn $h$ from available data.

Ingredients:

- Labeled Data: $D = \{(x, y)\}_{i=1}^{n}$
- Hypothesis Class $H$:
  - a set of possible functions $h: \mathcal{X} \rightarrow \mathcal{Y}$
- Loss function $l: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_{+}$
- Algorithm $A$
- Data Scientist

Example:

- Convolutional neural networks
- Cross-entropy loss
- Stochastic Gradient Descent

Model/Program/Hypothesis $h(x) \rightarrow y$

Output $\hat{y} \uparrow$

label The thing we would like to predict (from $x$).

$Y = \mathbb{R}$: Regression

$Y = \{0,1\}$: Binary Classification

$Y = \{1, \ldots, M\}$: Multi-class classification
ML Stages and Concepts:

**Learning/Training**
- Use $\mathcal{A}$ to find $h: \mathcal{X} \rightarrow \{0,1\}$ on training data $\mathcal{D}$.
- Data used to find $h$ gives rise to the training loss:
  \[ \frac{1}{\mathcal{D}} \sum_{(x,y)} l(h(x), y) \]

**Inference/Testing**
- For some testing data $\mathcal{R}$, not in the training data, predict the label $y = h(x)$.
- Data used to evaluate $h$. Generalization loss:
  \[ \mathbb{E}_{(x,y) \sim \mathcal{P}} [l(h(x), y)] \]

### Training Data:
- $\frac{1}{\mathcal{D}} \sum_{(x,y)} l(h(x), y)$

### Testing Data:
- Typically you split your data $\approx 80/20$ into train/test.
  - Often people split into train/validation/test. Why?

### Typical ways to split your data:
- Uniformly at random
- By time (e.g., $\mathcal{A}_t \sim \mathcal{P}$)
- By patient/instance

### Standard rule: Simulate the test case
Never predict the past from the future?

### Assumptions:
- No free lunch theorem: You must make assumptions in order to learn.
- $\Rightarrow$ there is no single ML algorithm that works for all settings.
- Example assumptions:
  - Data is locally smooth
  - $P$ consists of natural image
  - $P$ is a mixture of Gaussians
  - Features are independent given the labels