Lecture 1: Course logistics, homework 0

STATS 202: Data mining and analysis

Guenther Walther
September 23, 2019

(Slide credits: Sergio Bacallado)

Syllabus

- **Videos:** Every lecture will be recorded by SCPD.

Syllabus

- **Videos:** Every lecture will be recorded by SCPD.
- **Email policy:** Please use the Piazza site for most questions. For administrative issues that only concern you, email a TA or the instructor.

Syllabus

- **Videos:** Every lecture will be recorded by SCPD.
- **Email policy:** Please use the Piazza site for most questions. For administrative issues that only concern you, email a TA or the instructor.
- **Class website:** stats202.stanford.edu.

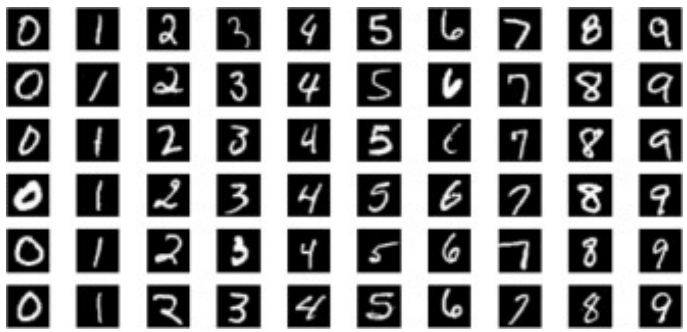
Prediction challenges

In a typical prediction challenge, you are given a training set of labeled datapoints.

Prediction challenges

In a typical prediction challenge, you are given a training set of labeled datapoints.

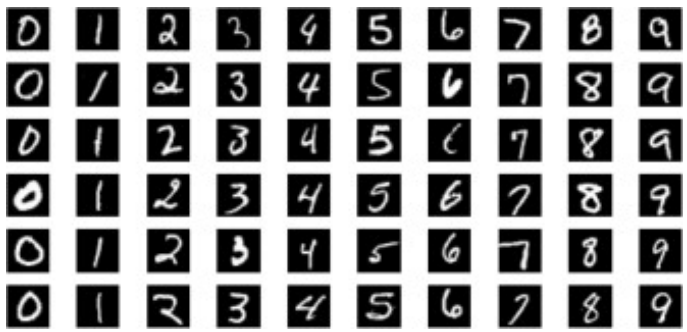
As an example, the MNIST dataset is a library of handwritten digits, labeled from 0 to 9.



Prediction challenges

In a typical prediction challenge, you are given a training set of labeled datapoints.

As an example, the MNIST dataset is a library of handwritten digits, labeled from 0 to 9.

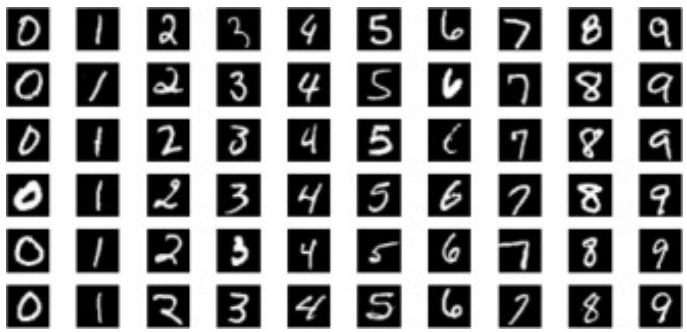


You are also given a test set of unlabeled datapoints.

Prediction challenges

In a typical prediction challenge, you are given a training set of labeled datapoints.

As an example, the MNIST dataset is a library of handwritten digits, labeled from 0 to 9.



You are also given a test set of unlabeled datapoints.

Your job is to assign a label (digit) to each image in the test set as accurately as possible.

The Netflix prize

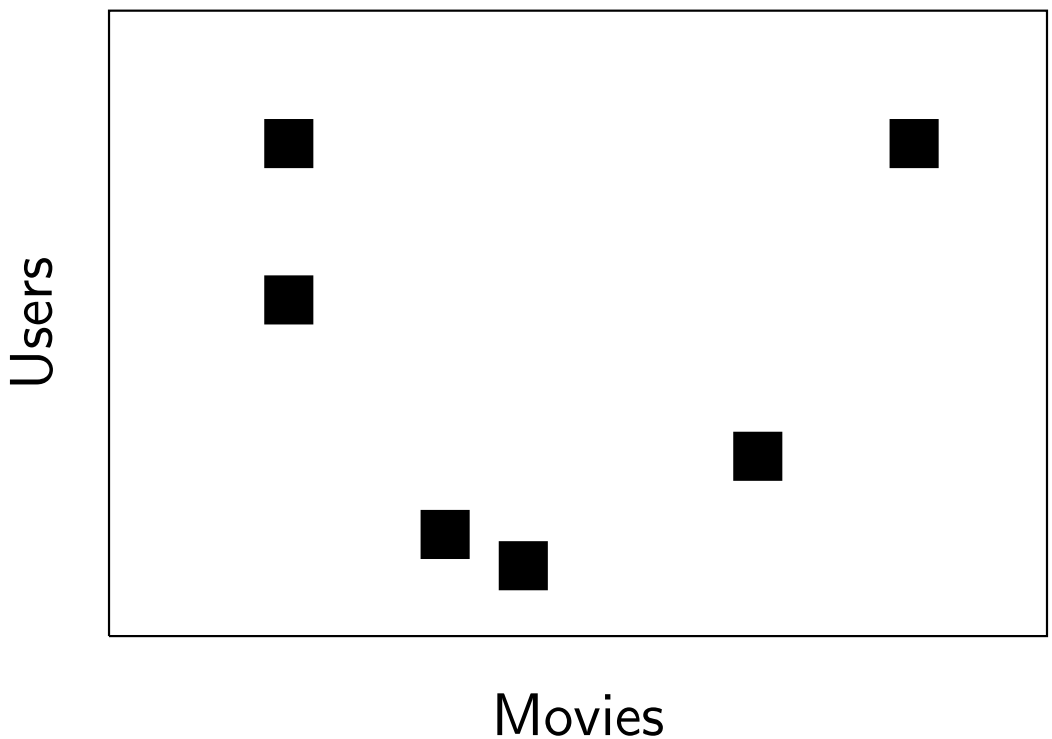
Netflix popularized prediction challenges by organizing an open, blind contest to improve its recommendation system.

The prize was $1 million.

The Netflix prize

Netflix popularized prediction challenges by organizing an open, blind contest to improve its recommendation system.

The prize was $1 million.

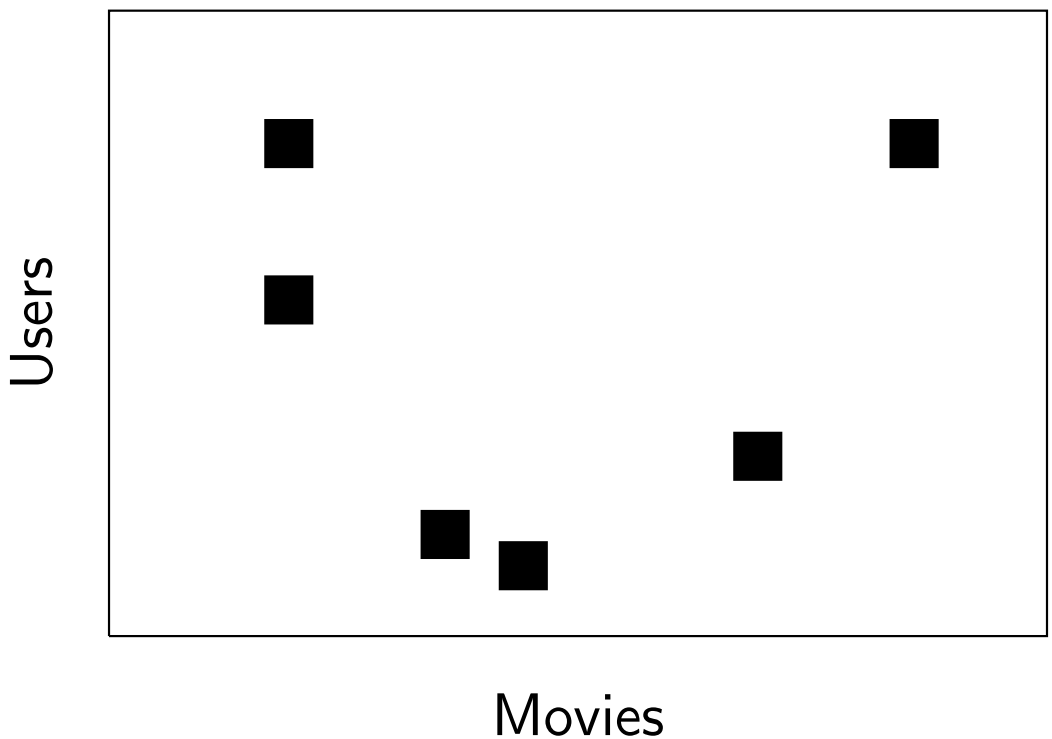


Only 1% of the 100 million possible user-movie rating pairs were revealed

The Netflix prize

Netflix popularized prediction challenges by organizing an open, blind contest to improve its recommendation system.

The prize was $1 million.



The challenge was to predict the unseen ratings

Kaggle

Company founded in 2010.

Business model:

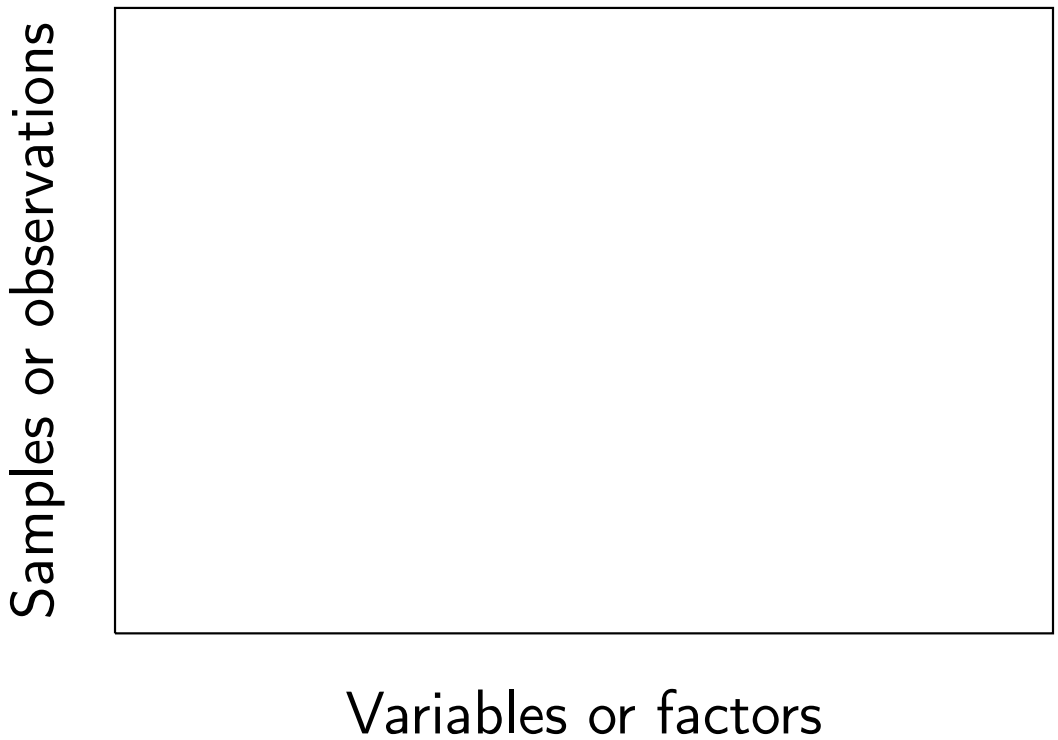
- Organize prediction competitions hosted online.
- Offer companies consulting services from Kaggle "masters".

Kaggle-in-class is a competition engine offered to degree-granting institutions for free. Stats 202 was the first class to use it!

This quarter you will team up with up to three students and participate in a Kaggle challenge TBD.

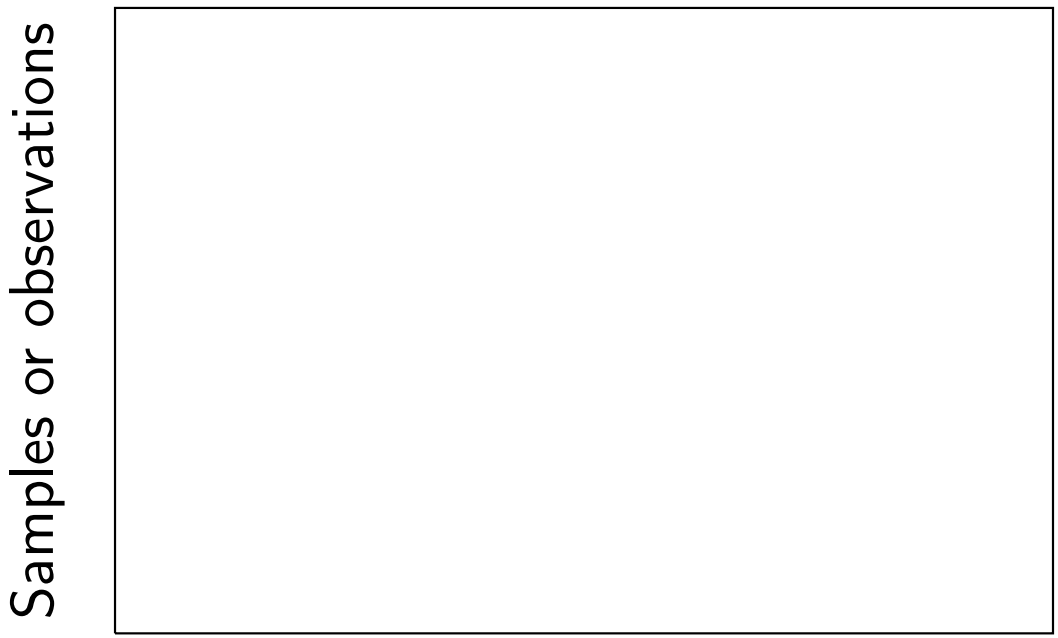
Supervised vs. unsupervised learning

In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:



Supervised vs. unsupervised learning

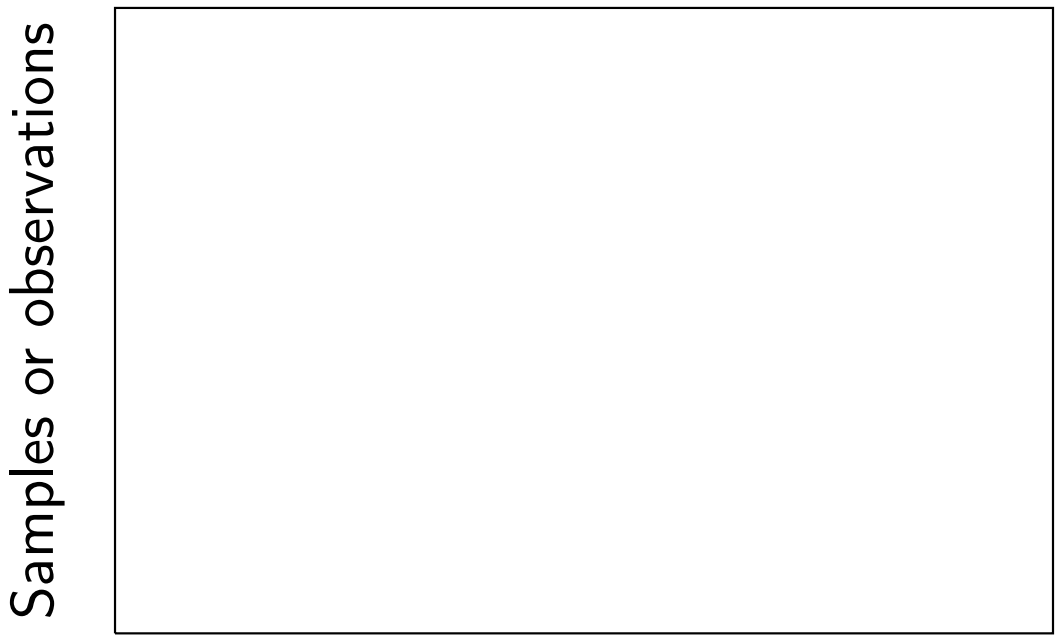
In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:



Quantitative variables, eg. weight, height, number of children, ...

Supervised vs. unsupervised learning

In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:



Qualitative variables, eg. college major, profession, gender, ...

Supervised vs. unsupervised learning

In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:

Our goal is to:

- Find meaningful relationships between the variables.

Supervised vs. unsupervised learning

In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:

Our goal is to:

- Find meaningful relationships between the variables.
- Find low-dimensional representations of the data which make it easy to visualize the variables.

Supervised vs. unsupervised learning

In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:

Our goal is to:

- Find meaningful relationships between the variables.
- Find low-dimensional representations of the data which make it easy to visualize the variables.
- Find meaningful groupings of the data.

Supervised vs. unsupervised learning

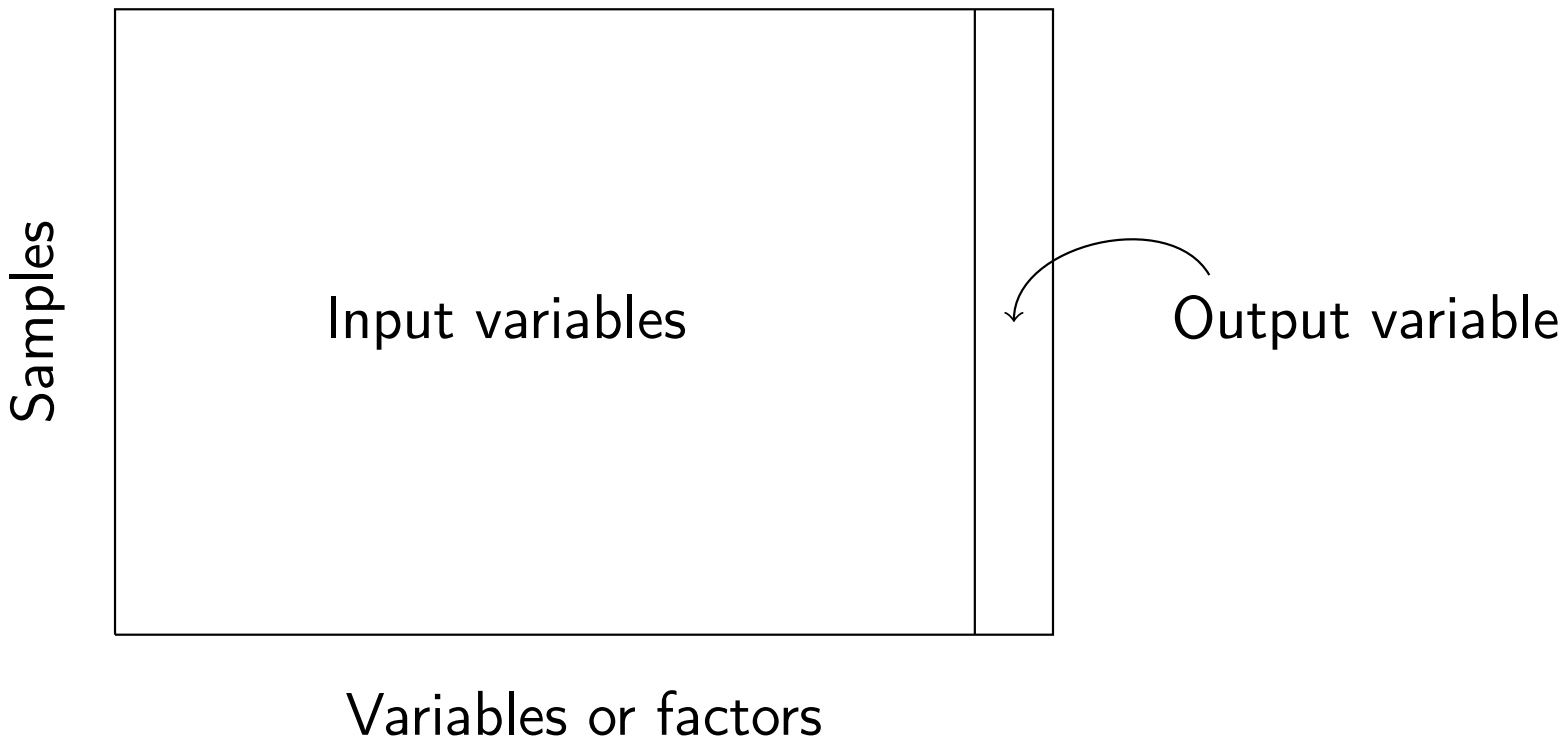
In **unsupervised learning** we seek to understand the relationships between variables in a data matrix:

Our goal is to:

- Find meaningful relationships between the variables. Correlation analysis.
- Find low-dimensional representations of the data which make it easy to visualize the variables. PCA, ICA, isomap, locally linear embeddings, etc.
- Find meaningful groupings of the data. Clustering.

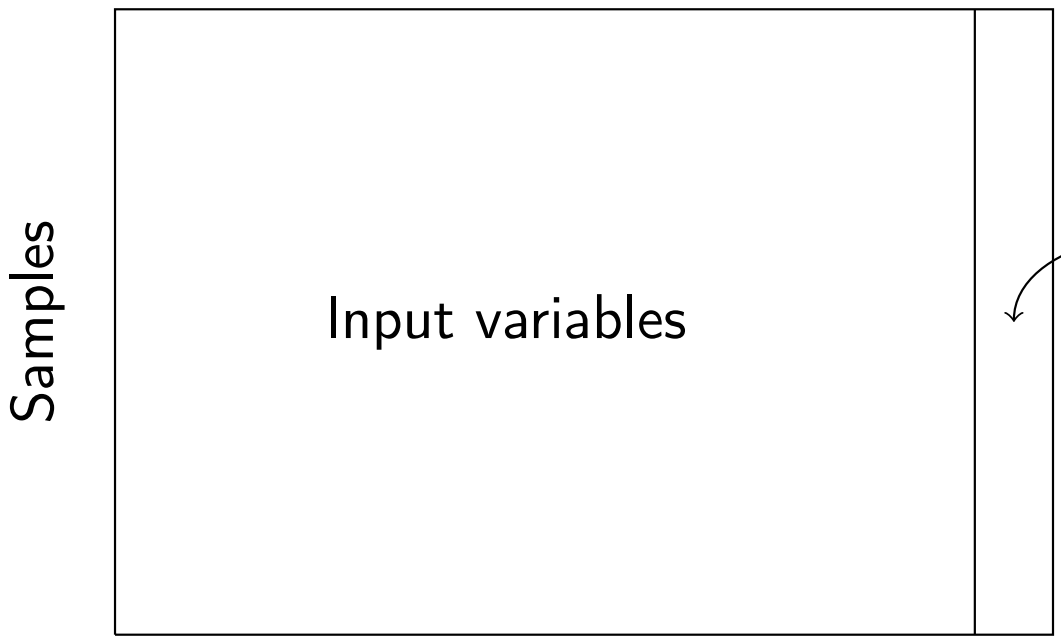
Supervised vs. unsupervised learning

In **supervised learning**, there are *input* variables, and *output* variables:



Supervised vs. unsupervised learning

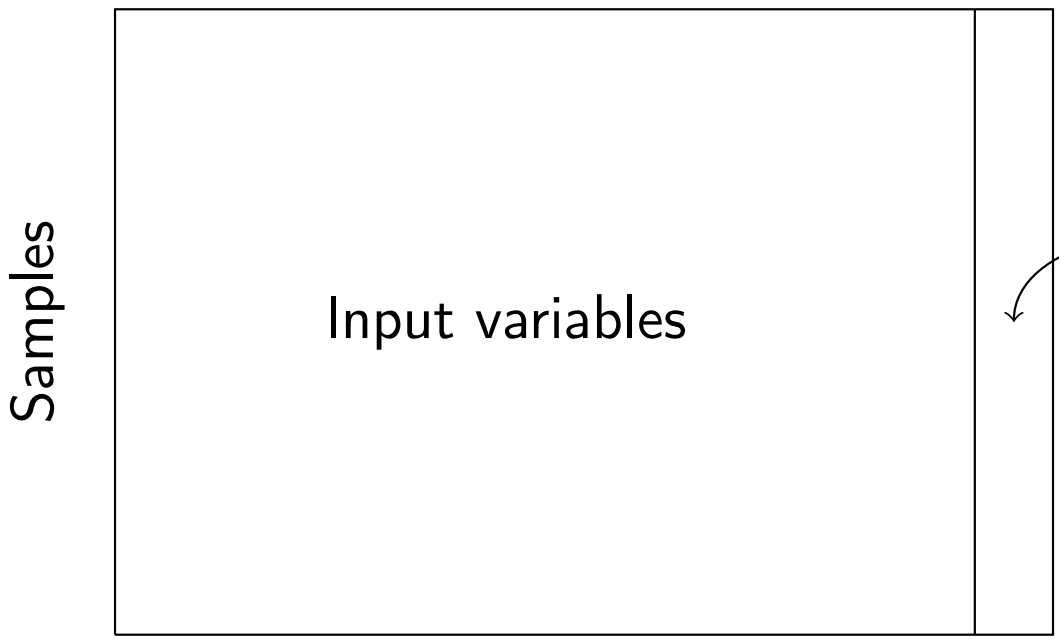
In **supervised learning**, there are *input* variables, and *output* variables:



If quantitative, we say this is a regression problem

Supervised vs. unsupervised learning

In **supervised learning**, there are *input* variables, and *output* variables:



If qualitative (categorical), we say this is a classification problem

Supervised vs. unsupervised learning

In **supervised learning**, there are *input* variables, and *output* variables:

If X is the vector of inputs for a particular sample. The output variable is modeled by:

$$Y = f(X) + \underbrace{\varepsilon}_{\text{Random error}}$$

Supervised vs. unsupervised learning

In **supervised learning**, there are *input* variables, and *output* variables:

If X is the vector of inputs for a particular sample. The output variable is modeled by:

$$Y = f(X) + \underbrace{\varepsilon}_{\text{Random error}}$$

The function f captures the systematic relationship between X and Y. f is fixed and unknown.

Supervised vs. unsupervised learning

In **supervised learning**, there are *input* variables, and *output* variables:

If X is the vector of inputs for a particular sample. The output variable is modeled by:

$$Y = f(X) + \underbrace{\varepsilon}_{\text{Random error}}$$

The function f captures the systematic relationship between X and Y. f is fixed and unknown.

ϵ represents the unpredictable "noise" in the problem.

Supervised vs. unsupervised learning

In **supervised learning**, there are *input* variables, and *output* variables:

If X is the vector of inputs for a particular sample. The output variable is modeled by:

$$Y = f(X) + \underbrace{\varepsilon}_{\text{Random error}}$$

The function f captures the systematic relationship between X and Y. f is fixed and unknown.

ϵ represents the unpredictable "noise" in the problem.

Our goal is to learn the function f, using a set of training samples.

Supervised vs. unsupervised learning

$$Y = f(X) + \underbrace{\varepsilon}_{\text{Random error}}$$

Motivations:

- **Prediction:** Useful when the input variable is readily available, but the output variable is not.

  Example: Predict stock prices next month using data from last year.

Supervised vs. unsupervised learning

$$Y = f(X) + \underbrace{\varepsilon}_{\text{Random error}}$$

Motivations:

- **Prediction:** Useful when the input variable is readily available, but the output variable is not.
- **Inference:** A model for f can help us understand the structure of the data — which variables influence the output, and which don't? What is the relationship between each variable and the output, e.g. linear, non-linear?

  Example: What is the influence of genetic variations on the incidence of heart disease.