

# Lecture 1: Course logistics, homework 0

STATS 202: Data mining and analysis

Jonathan Taylor, 9/24

Slide credits: Sergio Bacallado

September 24, 2018

# 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 the course staff mailing list:

`stats202-aut1819-staff@lists.stanford.edu`

# Syllabus

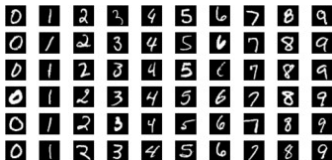
- ▶ **Videos:** Every lecture will be recorded by SCPD.
- ▶ **Email policy:** Please use the Piazza site for most questions. For administrative issues that only concern you, email the course staff mailing list:

`stats202-aut1819-staff@lists.stanford.edu`

- ▶ **Class website:** `stats202.stanford.edu`. If you are auditing the class (not registered on Axess), email us your SUNet ID in order to gain access to the lectures and homework.

## Prediction challenges

The MNIST dataset is a library of handwritten digits.



## Prediction challenges

The MNIST dataset is a library of handwritten digits.



In a prediction challenge, you are given a **training** set of images of handwritten digits, which are labeled from 0 to 9.

## Prediction challenges

The MNIST dataset is a library of handwritten digits.



In a prediction challenge, you are given a **training** set of images of handwritten digits, which are labeled from 0 to 9.

You are also given a **test** set of handwritten digits, which are not identified.

## Prediction challenges

The MNIST dataset is a library of handwritten digits.



In a prediction challenge, you are given a **training** set of images of handwritten digits, which are labeled from 0 to 9.

You are also given a **test** set of handwritten digits, which are not identified.

Your job is to assign a digit to each image in the test set.



## The Netflix prize

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

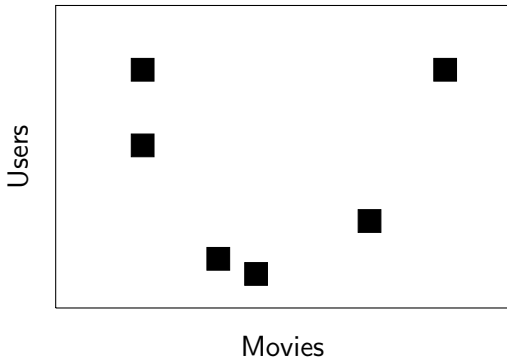
**The prize was \$1 million.**



## The Netflix prize

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

**The prize was \$1 million.**

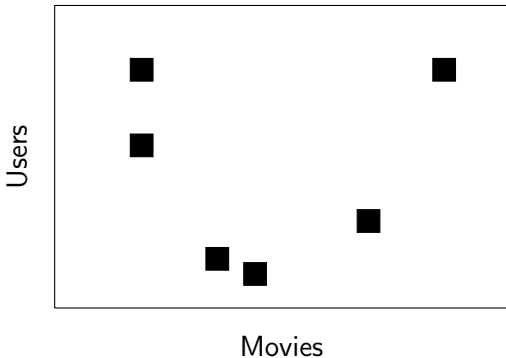


Some rankings were hidden in the training data

## The Netflix prize

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

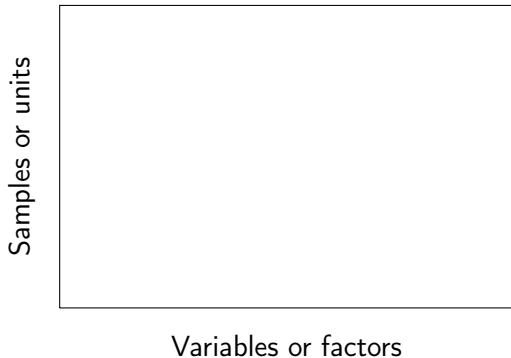
**The prize was \$1 million.**



The challenge was to predict those rankings

## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:



## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:



Variables or factors

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

## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:



Variables or factors

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

## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:

Our goal is to:

- ▶ Find meaningful relationships between the variables or units.

## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:

Our goal is to:

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



## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:

Our goal is to:

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

## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:

Our goal is to:

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

## Supervised vs. unsupervised learning

In **unsupervised learning** we start with a data matrix:

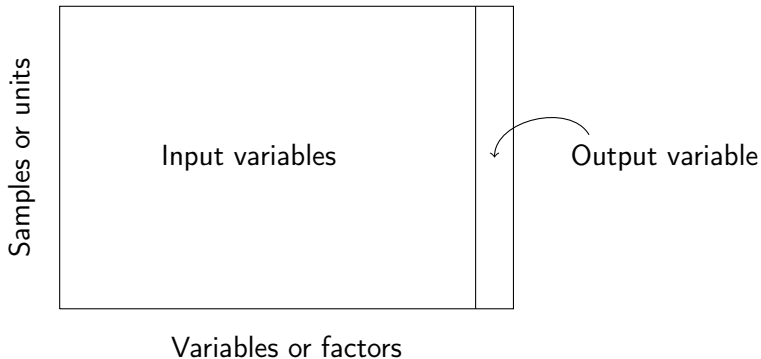
Our goal is to:

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

Unsupervised learning is also known in Statistics as **exploratory data analysis**.

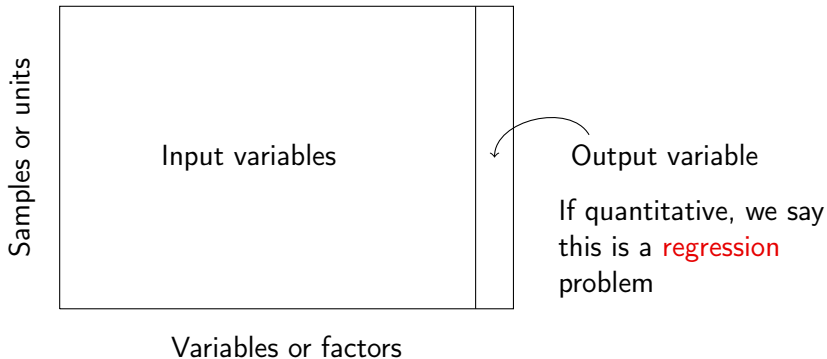
## Supervised vs. unsupervised learning

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



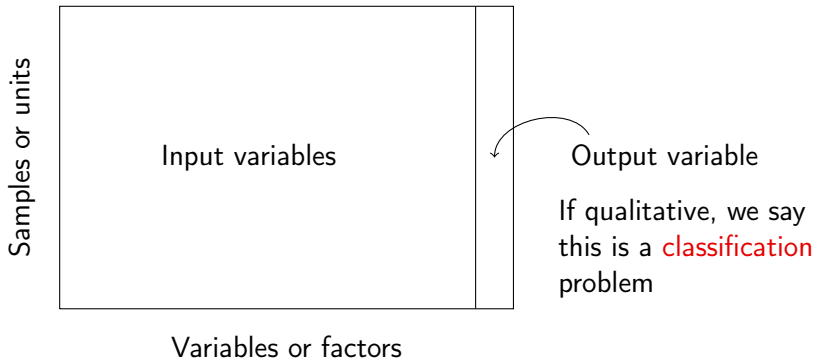
## Supervised vs. unsupervised learning

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



## 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  $X$  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:

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

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.

# Kaggle

## **Business model:**

- ▶ Organize prediction competitions hosted online.
- ▶ Offer companies consulting services from Kaggle “stars”.

# Kaggle

## **Business model:**

- ▶ Organize prediction competitions hosted online.
- ▶ Offer companies consulting services from Kaggle “stars”.

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

## A sample Kaggle challenge

Help out San Francisco's foremost Baroque ensemble bring in subscriptions!

## A sample Kaggle challenge

Help out San Francisco's foremost Baroque ensemble bring in subscriptions!

- ▶ **Option 1:** Using Philharmonia's database of subscriptions and single ticket sales, including information about concerts, and patrons, predict who will subscribe for the 2014-2015 season.

## A sample Kaggle challenge

Help out San Francisco's foremost Baroque ensemble bring in subscriptions!

- ▶ **Option 1:** Using Philharmonia's database of subscriptions and single ticket sales, including information about concerts, and patrons, predict who will subscribe for the 2014-2015 season.
- ▶ **Option 2:** Create an interactive visualization of Philharmonia's database using the R package Shiny.

## A sample Kaggle challenge

Help out San Francisco's foremost Baroque ensemble bring in subscriptions!

- ▶ **Option 1:** Using Philharmonia's database of subscriptions and single ticket sales, including information about concerts, and patrons, predict who will subscribe for the 2014-2015 season.
- ▶ **Option 2:** Create an interactive visualization of Philharmonia's database using the R package Shiny.

This year's competition coming soon!