Contents

1 Overview

1.1 Data Science

Working with data in a scientific way that will produce new and reproducible insight

This course is an introduction to the

- key ideas behind working with data in a scientific way
- tools that will allow you to execute on a data analytic strategy, from raw data in a database to a completed report

---

1.1.1 Why do DS

DS is being able to push through a lot of the difficulties that you have when you're dealing with either large or messy or poor data. It includes

- collecting the data
- clean them up
- and then building new announced techniques that explore new information about that data

---

What is the key challenge in DS?

You are interested in answering questions with data and are in a situation where

- either you really don't have enough data to answer the question that you're interested in, and you have to go out and try to search for it, find it on the web, or find it in other places.
- or you are overwhelmed with a surplus of data and you have to filter out all of the irrelevant information to try to narrow in on your question.

In summary, you can work on problems where

- the data aren't always clean and nice and easy to handle
- the questions that we want to answer are complicated and you have to break them down into parts
- you are passionate about trying to get the right answer so that you can help people in human health, science progress, industry, finance, etc.

From The Economist, The data deluge.

Over the last several years

- data has become much, much cheaper to collect.
- It's much easier to store.
- And there's so many free computing tools out there, that you can actually do something with this entire data deluge that is assaulting all different areas of science and business.

From McKinsey Global Institute, Big data, The next frontier.

The other term that comes into play now is **big data** which is a sort of a new frontier: we have data in areas that we didn't used to have that data. For example, now

- we have access to information about GPS coords from cars from everybody in the entire world
- it is possible to sequence everybody's genome.

1.1.2 Statistical Learning

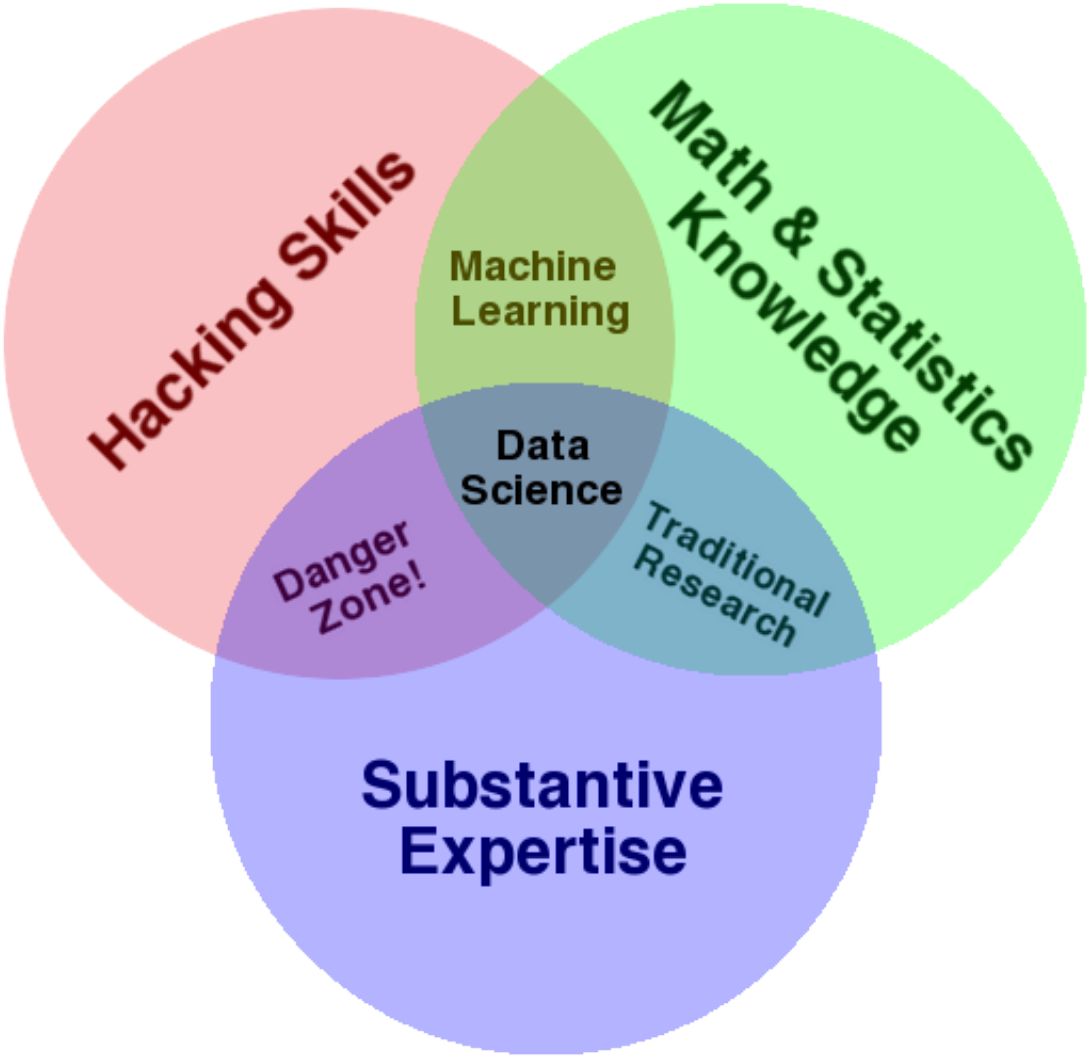
This DS track will have a statistical bend.

Machine learning and data analytics can be synthetized by **Statistical Learning**. Why?

Statistics is the science of learning from data.

t's very rare that you'll get a data set where all of the answers are really clear, and there's no uncertainty.

In any case where there is uncertainty, that's where statistics comes and plays a role.



From Drew Conway.

1.1.3 Why R

- It is free
- It has a comprehensive set of packages for
  - data access
  - data cleaning
  - analysis
  - data reporting
- It has one of the best development enviroments - Rstudio
- It has an amazing ecosystem of developers
- Packages are easy to install and "play nicely together"

1.2 Data Analytics

1.2.1 What a data scientist do

- Define a question of interest
- Identify the ideal data set
- Determining if/what data is accessible
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction or modeling
- Interpret results, challenging them.
- Synthese/write up results
- Create reproducible code
- Share results with other people

1.2.2 Getting and cleaning data

- Raw versus tidy data
- How to download files
- Read in data from a very large number of different sources
- Merge, reshape, summarize data

Raw data

- The original source of the data, often hard to use for data analysis
- Data analysis includes processing

Processed data

- Data ready for analysis
- Processing can include merging, subsetting, transforming into the nice tidy data set that people can use.

1.2.3 Exploratory data analysis

- Exploratory techniques for summarizing data are typically applied before formal modeling commences and can help inform the development of more complex statistical models. They are also important for eliminating or sharpening potential hypotheses that can be addressed by the data.
- Exploratory graphs
- Plotting systems in R: base, lattice, ggplot2

1.3 Machine Learning

- Grew out of work in AI
- New capability for computers

Examples:

- Database mining
  - Large datasets from growth of automation/web. E.g., Web click data, medical records, biology, engineering
- Applications can't program by hand.
  - E.g., Autonomous helicopter, handwriting recognition, most of Natural Language Processing (NLP), Computer Vision.
- Self-customizing programs
  - E.g., Amazon, Netflix product recommendations
- Understanding human learning (brain, real AI).

1.3.1 ML definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

1. Classifying emails as spam or not spam.
2. Watching you label emails as spam or not spam.
3. The number (or fraction) of emails correctly classified as spam/not spam.
4. None of the above—this is not a machine learning problem.

1.3.2 ML algorithms

- Supervised learning
- Unsupervised learning

Others: Reinforcement learning, recommender systems.

1.3.3 Examples

1.3.3.1 Regression: Predicting house prices

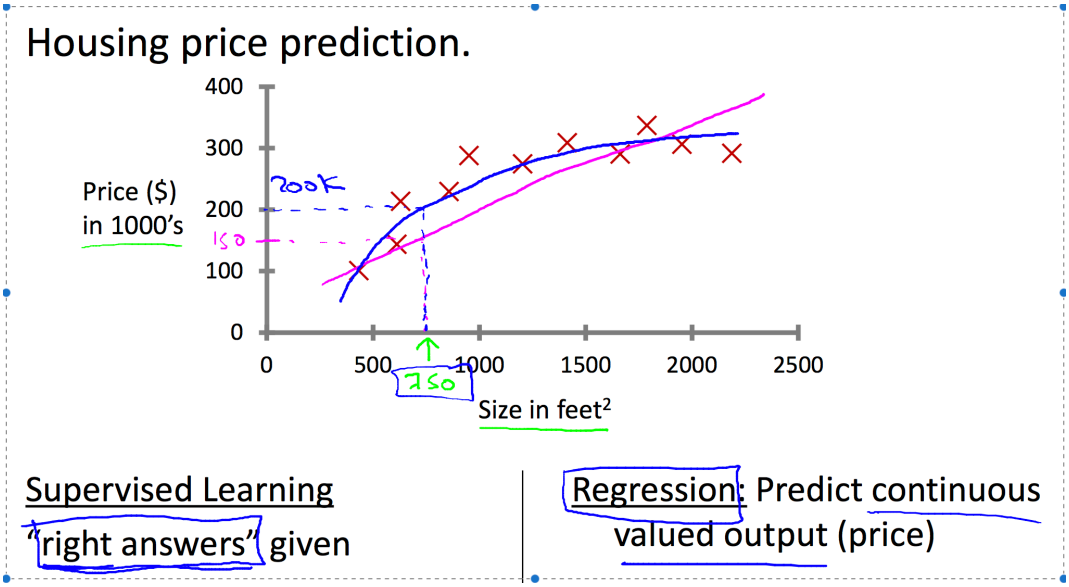
See slides

1.3.3.2 Classification: Analysing sentiments and others

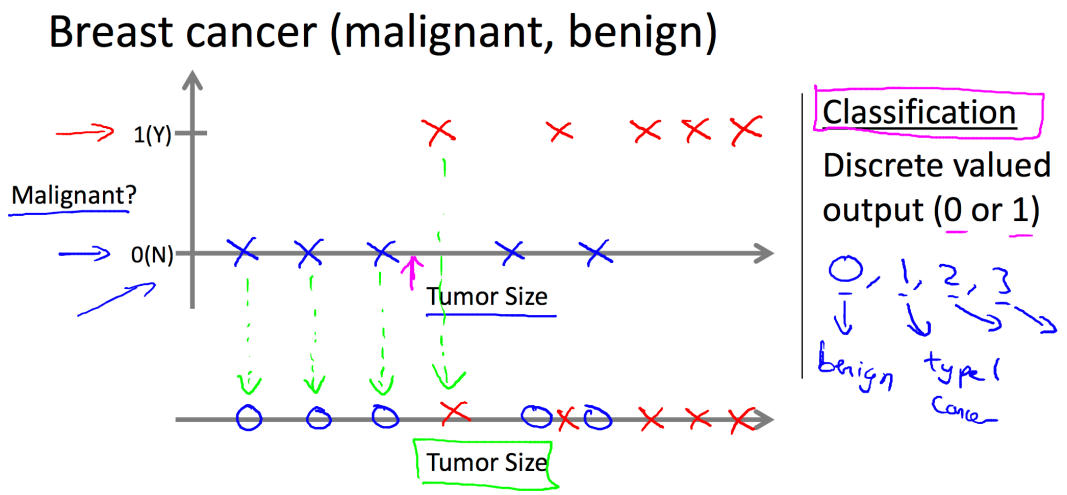
See slides

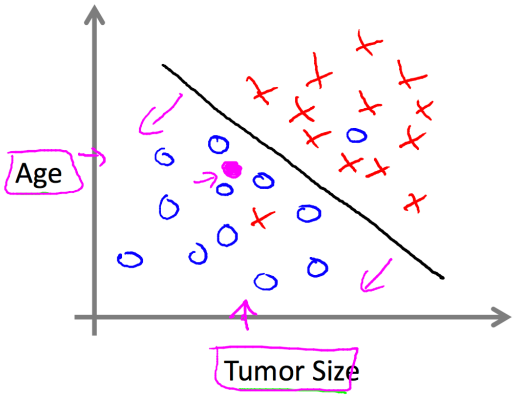
1.3.3.3 Clustering: document retrieval and others

See slides



From Ng.

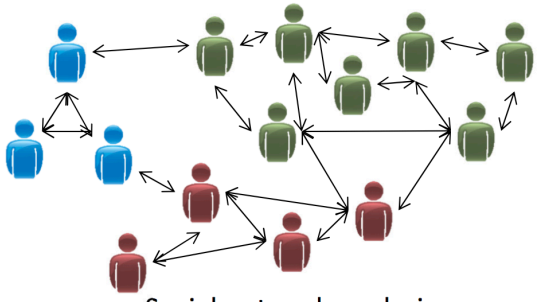




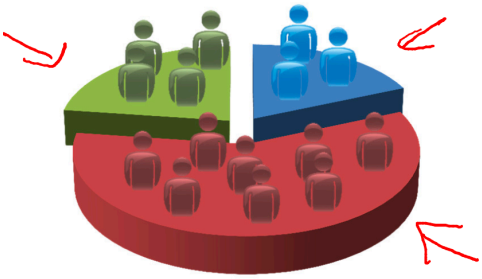
- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- ...

From Ng.

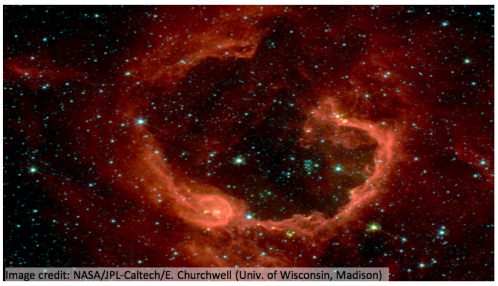
Organize computing clusters



Social network analysis



Market segmentation



Astronomical data analysis

From Ng.

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

1. Treat both as classification problems.
2. Treat problem 1 as a classif. pr., problem 2 as a regress. pr..
3. Treat problem 1 as a regress. pr., problem 2 as a classif. pr..
4. Treat both as regression problems.

Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

1. Given email labeled as spam/not spam, learn a spam filter.
2. Given a set of news articles found on the web, group them into set of articles about the same story.
3. Given a database of customer data, automatically discover market segments and group customers into different market segments.
4. Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

1.4 Statistical learning

1.4.1 The supervised learning problem

Suppose we observe

- an outcome measurement Y (also called dependent variable, response, target).
  - In the **regression** problem, Y is quantitative (e.g price, blood pressure).
  - In the **classification** problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- a vector of p predictor measurements $\mathbf{X}$ (also called inputs, regressors, covariates, features, independent variables).
- We have *training* data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.
- (Of course), we believe that there is a relationship between Y and at least one of the X's.

Objectives

On the basis of the training data we would like to:

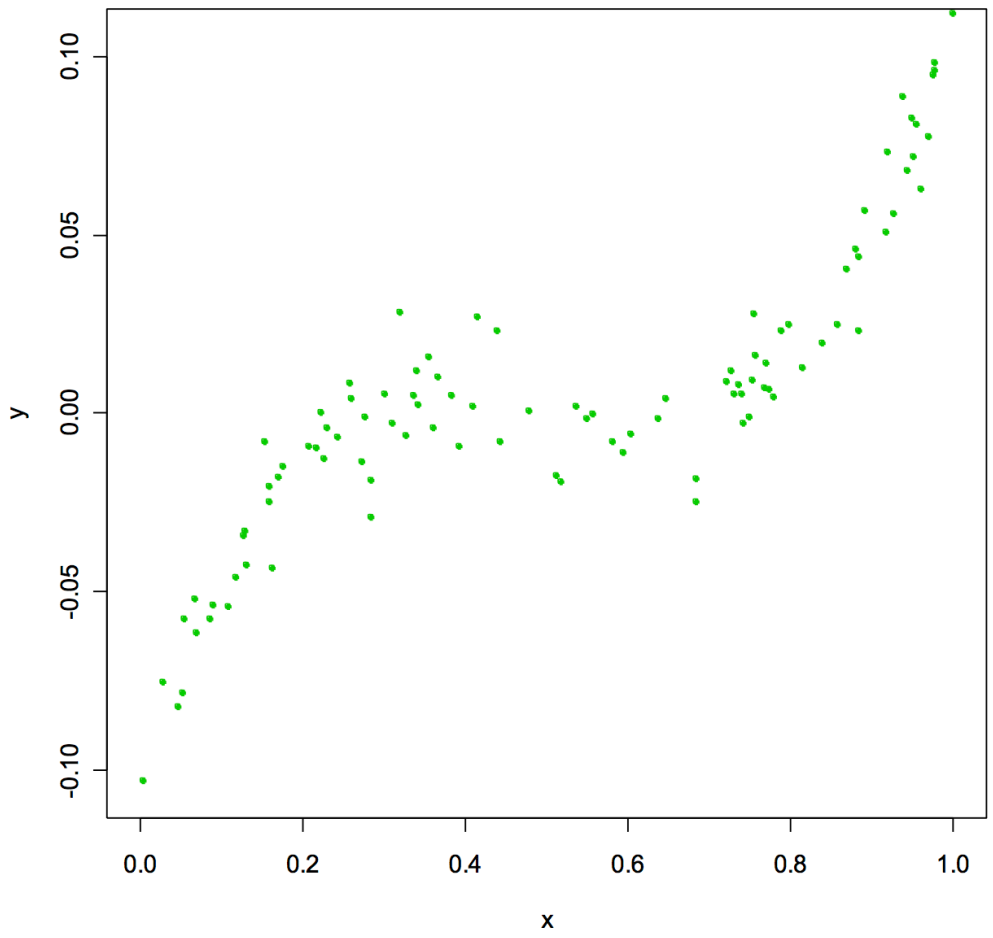
- Accurately predict unseen *test* cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

We can model the relationship as

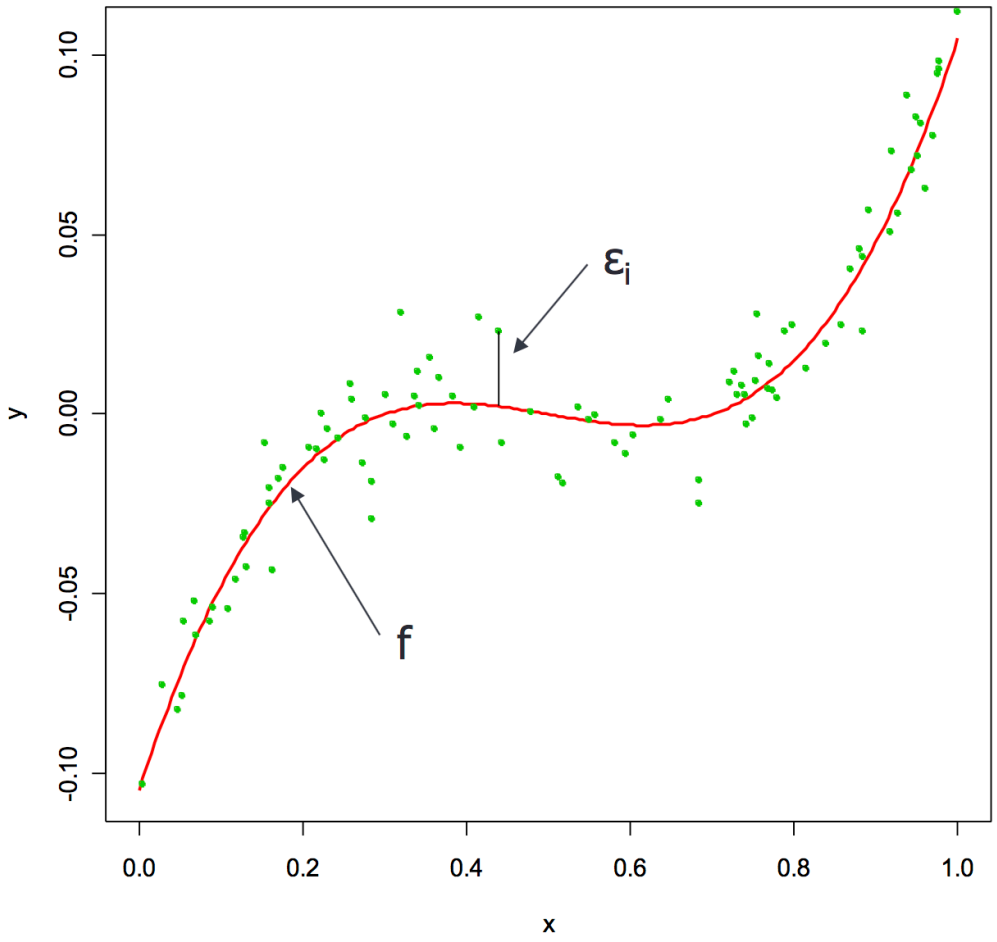
$$Y_i = f(\mathbf{X}_i) + \epsilon_i$$

where f is an unknown function and ϵ is a random error (with mean 0).

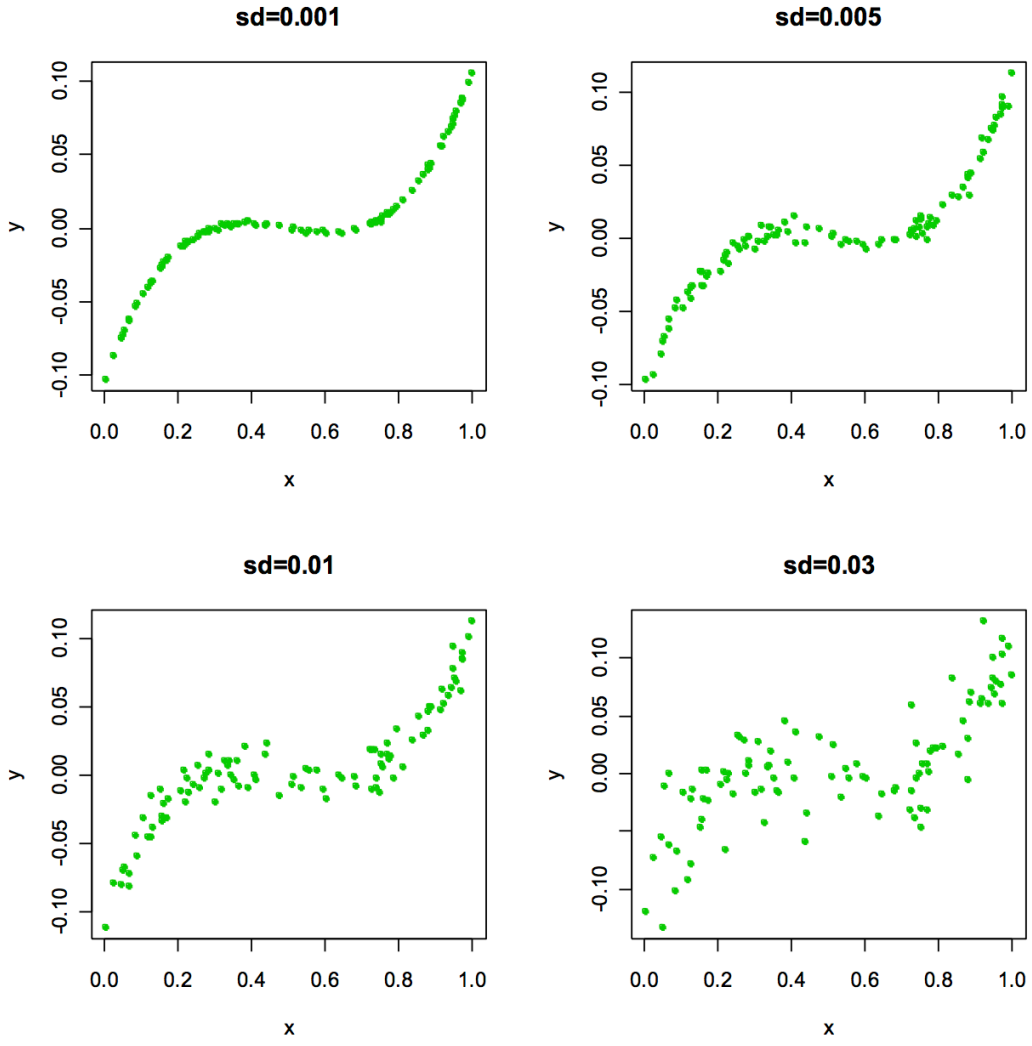
A simple example



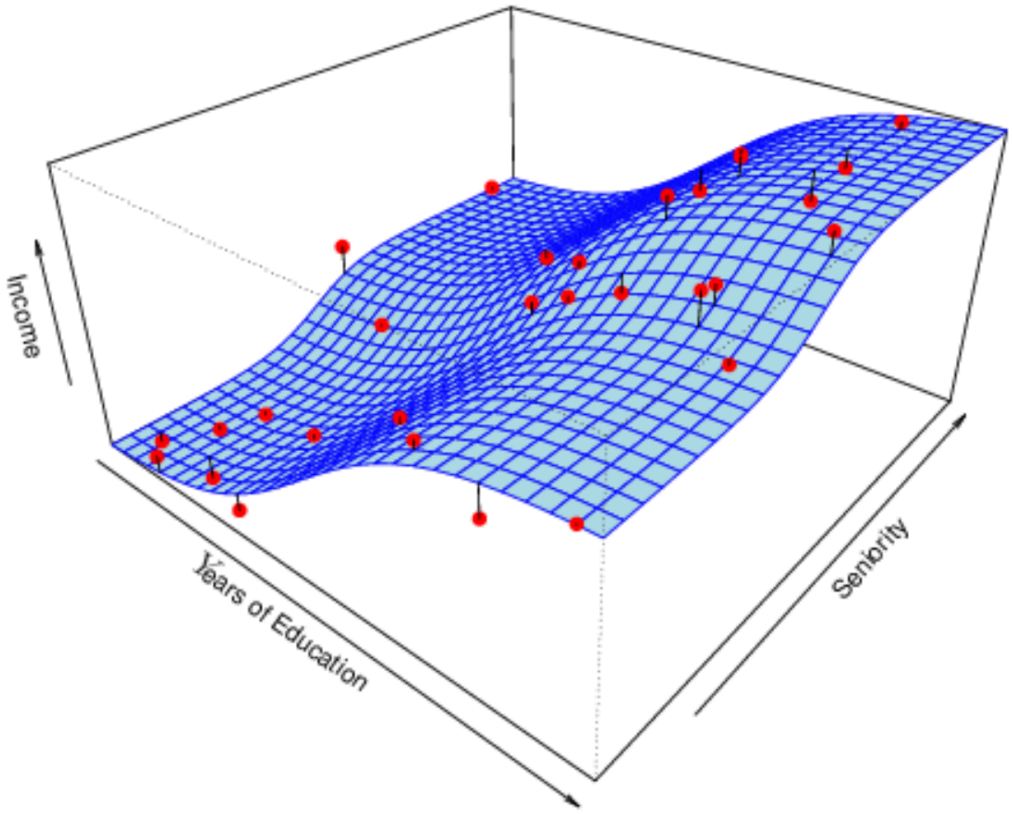
From Al Sharif.



Estimating f gets more difficult as uncertainty (error size) increases.



Another example



From Al Sharif.

Statistical learning refers to using the data to "learn" f.

Why do we care about estimating f?

- For prediction: make accurate predictions for Y based on a new value of $\mathbf{X}$.
- For inference: know the type of relationship between Y and $\mathbf{X}$.

For example,

- Direct mailing prediction:
  - Interested in predicting how much money an individual will donate based on observations from $90,000$ people on which we have recorded over 400 different characteristics.
  - Don't care too much about each individual characteristic.
  - Just want to know: For a given individual should I send out a mailing?
- Establish the relationship between salary and demographic variables in population survey data.
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

Some ideas

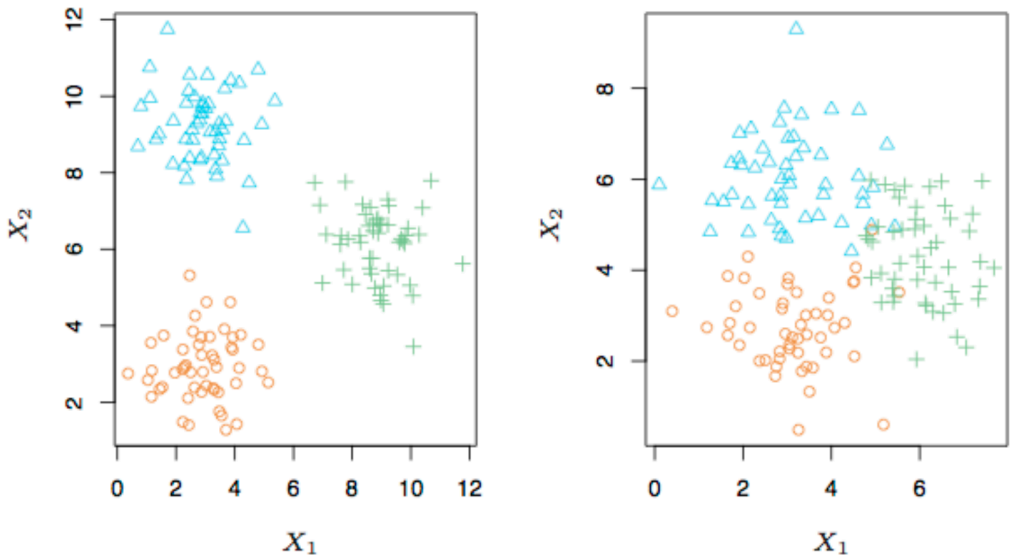
- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.

- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working (simpler methods often perform as well as fancier ones!)
- Statistical learning is a fundamental ingredient in the training of a modern **data scientist**.

1.4.2 The unsupervised learning problem

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- can be useful as a pre-processing step for supervised learning.

A common example is market segmentation where we try to divide potential customers into groups based on their characteristics.



Supervised (statistical) learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs.

In unsupervised there are inputs (but no supervising output) from which we can learn structure.

1.4.3 Statistical learning vs Machine learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap - both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, and precision and uncertainty.

- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.