STAT318/462 — Data Mining

Dr Gábor Erdélyi

University of Canterbury, Christchurch, New Zealand

Course developed by Dr B. Robertson. Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Organizational

- **Lecturer Term 3:** Dr Gábor Erdélyi.
- **Office Hours:** Erskine 704, Tuesdays 10am-11am or by appointment.

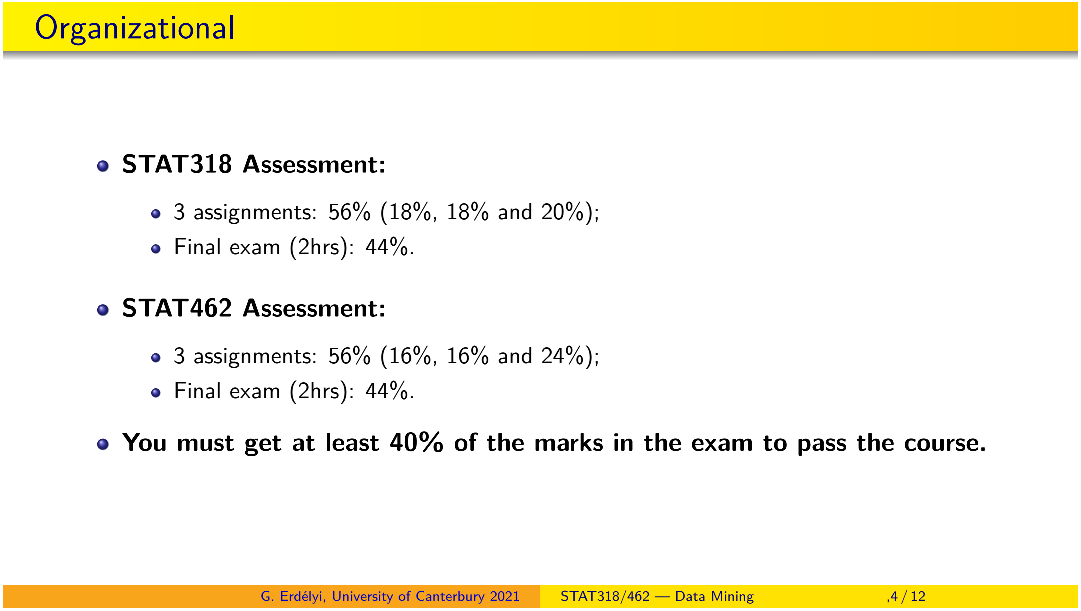
- **Course Co-ordinator/Lecturer Term 4:** Dr Varvara Vetrova.

- **Lectures (Echo360 recorded):**
  - Wednesdays 11am-12pm E7 Lecture Theatre;
  - Thursdays 1-2pm E5 Lecture Theatre.

- **Lecture slides:** on LEARN before lectures;
- **Lecture notes:** on LEARN after lectures.

Organizational

- **Tutors:**
  - Nicki Cartlidge;
  - Pooja Immattiparambil Baburaj;
  - Martin Nguyen;
  - Krzysztof Maliszewski.
- **Weekly labs/help sessions (starts in week 2):**
  - Mondays 3-4pm Ernest Rutherford 212 (Zoom livestream);
  - Tuesdays 1-2pm Ernest Rutherford 212;
  - Wednesdays 12-1pm Jack Erskine 035 Lab 2;
  - Wednesdays 3-4pm Jack Erskine 035 Lab 2 (Zoom livestream).



Organizational

- **STAT318 Assessment:**
  - 3 assignments: 56% (18%, 18% and 20%);
  - Final exam (2hrs): 44%.
- **STAT462 Assessment:**
  - 3 assignments: 56% (16%, 16% and 24%);
  - Final exam (2hrs): 44%.
- **You must get at least 40% of the marks in the exam to pass the course.**

- A student who narrowly fails to achieve 40% in the exam, but who performs very well in the other assessment, may be eligible for a pass in the course.
- You may do the assignments by yourself or with one other person from the same cohort (300-level students cannot work with 400-level students on the assignments). If you hand in a joint assignment, you will each be given the same mark.

Course Textbook

- G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning, with applications in R*.

  Available online free (pdf):

  http://www-bcf.usc.edu/∼gareth/ISL/

- T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*.

  Available online free (pdf):

  http://web.stanford.edu/∼hastie/ElemStatLearn/

The *Elements of Statistical Learning* is an excellent book on the subject, but requires a level of mathematical sophistication that is beyond the scope of this course. Students with sufficient backgrounds in mathematics may find this book useful to explain technical details that are not covered in this course.

An Introduction to Statistical Learning (ISL) is a simplified version of the *Elements of Statistical Learning*. The authors have removed much of the mathematics and keep things simple to make the subject accessible to a wider audience. They have done an excellent job and this course is pitched at the same level as the ISL text. The course follows ISL closely, but we do not have enough time to cover all of the topics in ISL.

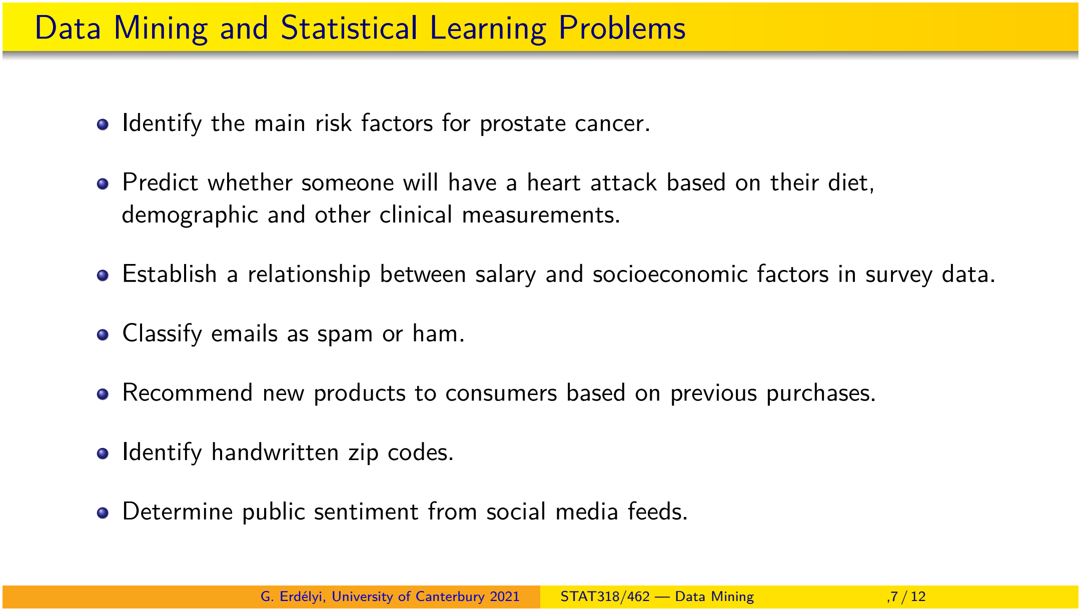
Course Objectives

1. Introduce statistical learning and data mining.
2. Introduce techniques for classification, regression, clustering and association analysis.
3. Understand the basic concepts and underlying assumptions of each technique and determine when they might be useful.
4. Implement various statistical learning techniques using *R* and real-world data.

1. Data Science, Machine Learning, Statistical Learning and Data Mining have much overlap. I will refer to the subject as statistical learning and data mining, but trendier names like ‘data science’ and ‘machine learning’ would also be appropriate for many topics in this course.

3. We will not be covering the technical/mathematical details of statistical learning methods in this course. Covering these details requires a level of mathematical sophistication that was not asked for in the course prereqs. At times we will delve into the details, but only when it is essential to better understand a method.

4. We will be implementing *R* functions rather than programming in *R*. This course does not assume prior knowledge of *R* and introductory *R* labs will be given.

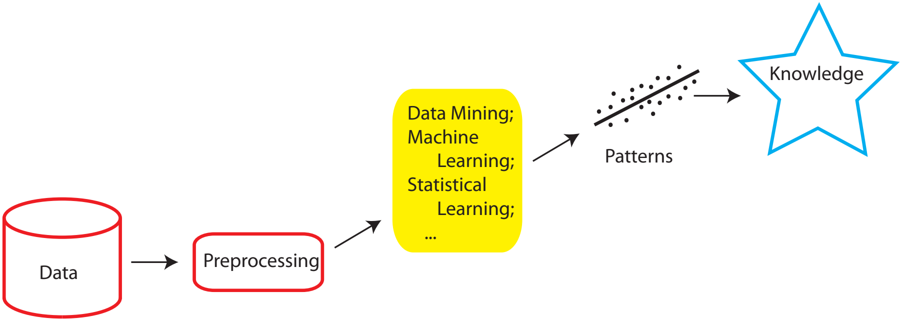


Data Mining and Statistical Learning Problems

- Identify the main risk factors for prostate cancer.
- Predict whether someone will have a heart attack based on their diet, demographic and other clinical measurements.
- Establish a relationship between salary and socioeconomic factors in survey data.
- Classify emails as spam or ham.
- Recommend new products to consumers based on previous purchases.
- Identify handwritten zip codes.
- Determine public sentiment from social media feeds.

There are many interesting statistical learning problems (some far more interesting than those listed above). The point here is that there are a variety of statistical learning problems including making predictions, looking for relationships, making inferences, classifying things, looking for interesting associations, ... There is no *best* method to tackle problems like these, so we will consider a variety methods for each type of problem.

Knowledge Discovery (Data Science?)



Data

Preprocessing

Data Mining; Machine Learning; Statistical Learning; ...

Patterns

Knowledge

Data storage and preprocessing (aka data wrangling) is a subject in its own right (we have a data wrangling course DATA201/422). This includes

- **Cleaning** (e.g. veracity — the accuracy and trustworthiness of the data)
- **Integration** (e.g. combining data from multiple sources/data types)
- **Reduction** (e.g. removing redundancy)
- **Transformation** (e.g. algorithm readable form)

Clearly, data wrangling is an incredibly important step in any learning process because when training a model, **rubbish in equals rubbish out**. This class considers fairly benign data sets that do not require much (if any) wrangling. This is possibly an unrealistic situation in practice, but means we can focus on understanding the algorithms/methods that are used in statistical learning. This course focuses on the last three steps of knowledge discovery.

Supervised Learning

- Response (outcome, target, dependent) variable Y and a vector of p predictor (input, feature, independent) variables X.

- We have a **training data set** of n observations (examples, instances) of the form

$$\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, n\},$$

where $\mathbf{x}_i$ is a vector of p predictor variables and y_i is the response value for the ith observation.

- If Y is quantitative, we have a **regression problem**. If Y takes values in a finite (unordered) set, we have a **classification problem**.

Consider trying to predict the *Sales* of a product based on the advertising budgets of three different media: *TV, Radio, Newspaper*.

- Response: $Y = \mathit{Sales}$
- Predictors: $X_1 = \mathit{TV}$, $X_2 = \mathit{Radio}$ and $X_3 = \mathit{Newspaper}$
- Predictor vector: $X = (X_1, X_2, X_3) = (\mathit{TV, Radio, Newspaper})$.

This is a regression problem because Y is quantitative. If we discretize *Sales* into *Low* and *High*, for example,

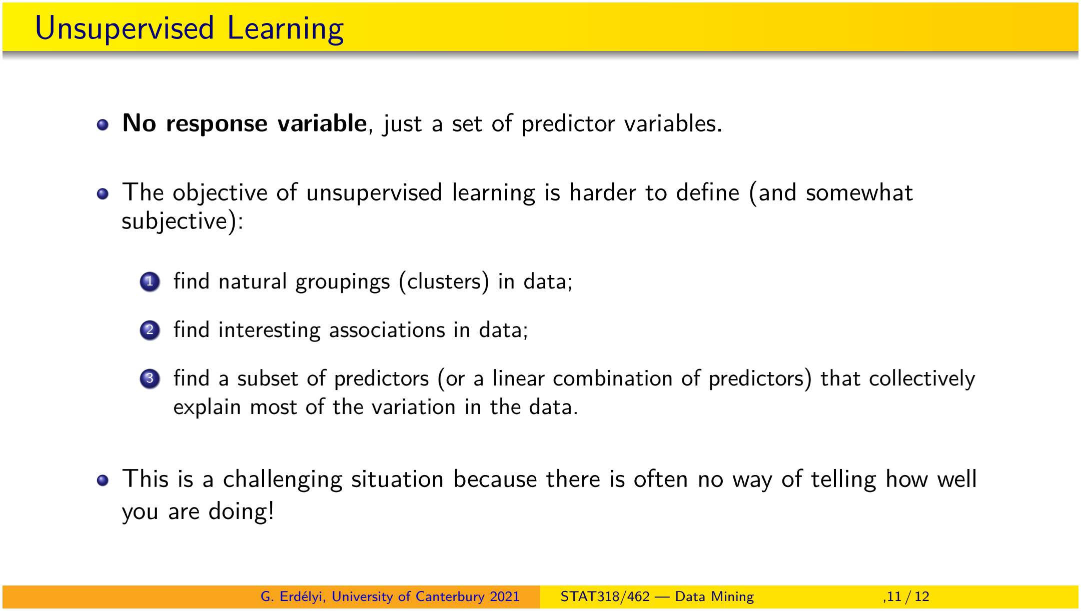
$$Y = \begin{cases} \mathit{Low} & \text{if } \mathit{Sales} < 10\text{k} \\ \mathit{High} & \text{otherwise,} \end{cases}$$

we would have a classification problem, where the goal is to predict whether sales are high or low (binary classification problem). Both regression and classification problems are considered in this course.

In supervised learning we:

1. **Predict Outcomes:** Use an observed $\mathbf{x}$ to predict an unobserved y;
2. **Make Inferences:** Understand how each predictor variable affects the response;
3. **Quantify Uncertainty:** Assess the quality of any predictions and/or inferences made.

1. It is usually difficult (or impossible) and/or expensive to measure the response variable directly and relatively easy to measure the predictors. Hence, we build a model to training data to make predictions. For example, if I spend \$50k on *TV* advertising, what is my expected *Sales*?
2. Should I reduce the amount I spend on *TV* advertising?
3. How confident am I about the predictions made by my model?

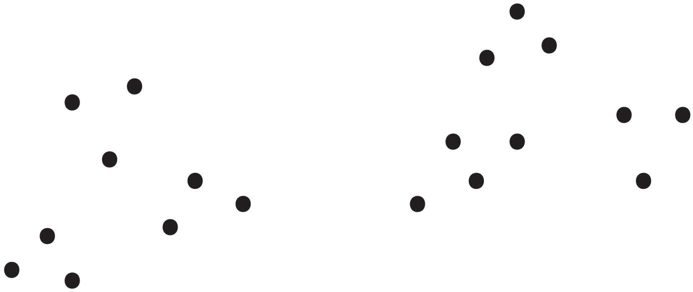


Unsupervised Learning

- **No response variable**, just a set of predictor variables.
- The objective of unsupervised learning is harder to define (and somewhat subjective):
  1. find natural groupings (clusters) in data;
  2. find interesting associations in data;
  3. find a subset of predictors (or a linear combination of predictors) that collectively explain most of the variation in the data.
- This is a challenging situation because there is often no way of telling how well you are doing!

The only unsupervised learning techniques considered in this course are cluster analysis and association analysis. We will not be considering, for example, dimensionality reduction and feature subset selection. Principal components analysis (PCA) is a popular dimensionality reduction method (see section 10.2 in the course textbook if you're interested), but we will not cover it here (it is covered in other courses we offer, for example, STAT315).

Unsupervised Learning: How many clusters?



The number of *natural* clusters in data is subjective. This data set could have 2,4 or 6 clusters, it really depends on how we define clusters. Different algorithms will tend to find different clusterings in data because their cluster definitions are different. More about this later in the course.