Statistical Learning

Lecture 01a

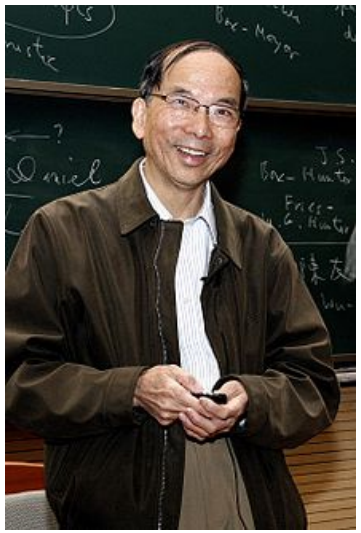
ANU - RSFAS - AHW

Last Updated: Tue Feb 22 14:57:58 2022

What is Statistics?

What is Statistics?

- Statistics is the **science of learning from data**.



- Professor Jeff Wu in November 1997, gave a talk for his appointment to the H. C. Carver Professorship at the University of Michigan titled:

Statistics = Data Science?

http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

American Statistical Association States . . .

Statistics is a big, important and growing field. In fact, it's a science . . . the science of learning from data. And as data has become more prevalent and important in our world, so has the field of statistics.

John Tukey (1915 - 2000)

"The best thing about being a statistician is that you get to play in everyone's backyard.' ' — J. Tukey

– coined the terms 'bit' and 'software'.

Backyards that I Play In

- Assessing uncertainty in weather predication → Atmospheric Science.
- Developing a 'Health' index for streams → Environmental Science.
- Developing a 'Health' (socio-economic) index for countries → Economics, Political Science.
- Statistical models for game theoretic data → Political Science, Economics.
- Statistical models for network data → Sociology, Political Science, Economics, Biology.
- Statistical inference for computer simulation models → Social Science, Biology
- Statistical inference for human activity spaces → Social Science

American Statistical Association

- This is Statistics (https://thisisstatistics.org)
- Why You Need to Study Statistics (https://youtu.be/wV0Ks7aS7YI)
- Statisticians Making A Difference (https://youtu.be/_EnoTvnx2gQ)
- Employment profiles (https://thisisstatistics.org/jobs-in-statistics/)

Stats + Stories (https://statsandstories.net)

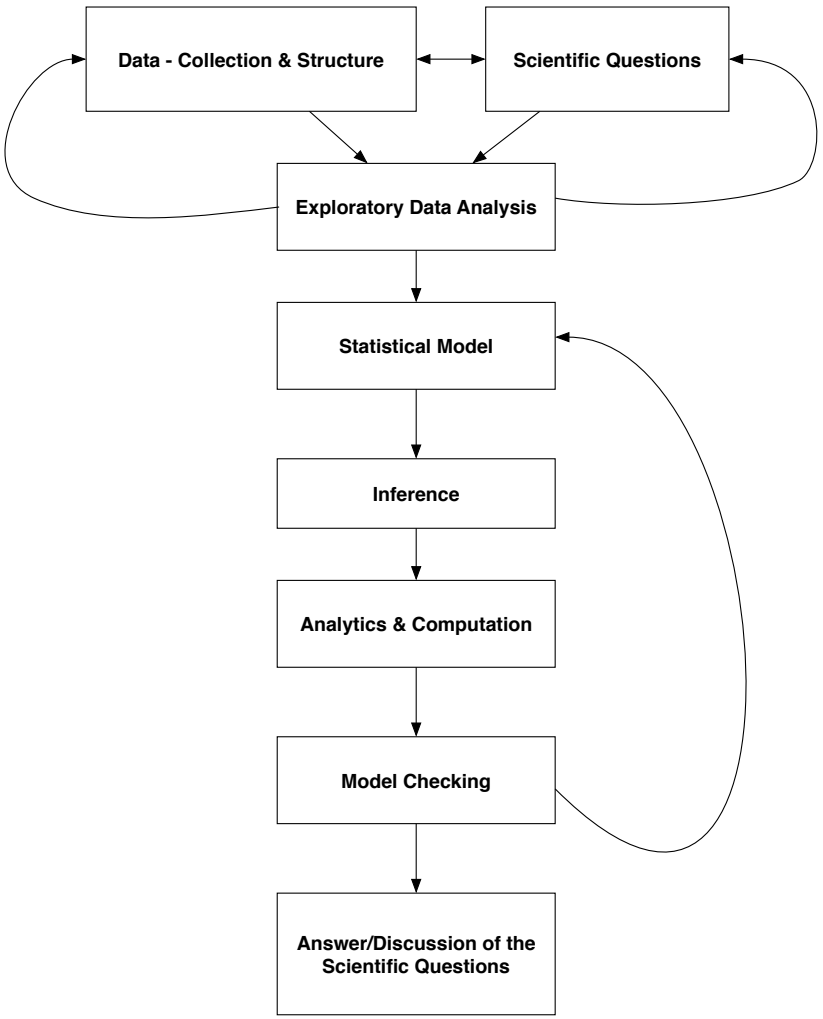
- Protecting Seaside Cities from Possible Future Impacts of Climate Change

https://statsandstories.net/environment1/2018/7/30/protecting-seaside-cities-from-possible-future-impacts-of-climate-change

More Data Journalism

- Nate Silver - Five Thirty Eight (https://fivethirtyeight.com)
- Hans Rosling - Gapminder (https://www.gapminder.org)
- 3 Statistics Lessons from a Summer as a Science Journalist (https://thisisstatistics.org/lessons-from-science-journalist/)

Thoughts on Statistics & Science



Course Description

This course introduces students to the techniques of **Statistical Learning** $\Rightarrow$ **Learning from Data!** This course has a focus on the idea of prediction.
We will examine:

- linear regression
- classification techniques
- resampling methods (e.g., the bootstrap)
- regularisation methods
- tree based methods
- unsupervised learning techniques (e.g., clustering)
- others . . .

Format

- Lectures:
  - Wednesday 10:00 - 11:30 (live on campus - will be recorded)
  - Thursday 4:00 - 5:30 (live on Zoom - will be recorded)
- Different material will be covered in both classes.
- Occasionally, a live session will be replaced with a video. A note will be placed on Wattle prior to this occurring.
- Tutorials (starting in the second week)
  - Three choice - choose only one
    - Live on campus
    - Online live via Zoom
    - Pre-recorded

Prescribed Texts

- **G. James, D. Witten, T. Hastie, and R. Tibshirani**
  An Introduction to Statistical Learning with Applications in R
  Springer
  - This text is freely available here:
    https://www.statlearning.com
- **T. Hastie, R. Tibshirani, and J. Friedman**
  The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition)
  Springer
  - This text is freely available here:
    http://statweb.stanford.edu/~tibs/ElemStatLearn/

Recommended Text

- **N. Silver**
  The Signal and the Noise: The Art and Science of Prediction
  Allen Lane Publishing

Assesments

- Assignment 1 (10%)
- Assignment 2 (30%)
- Project (60%)

Assignments

- There will be two take-home problem sets.
- Students should attempt all of the questions, showing appropriate mathematical and computational details, as well as discuss results.
- Solutions should be properly written-up. It is suggested that this is done using R-markdown within RStudio.

Project

- This compulsory project is designed to apply many of the statistical learning ideas you have been introduced throughout the course and requires the use of R to analyse real data.
- In addition, students will engage in a prediction competition, based on a withheld test data set.
- Students are required to communicate their findings in a formal written report using R-markdown within RStudio.

- A good resource is the **ModernDive** by Chester Ismay and Albert Y. Kim (https://ismayc.github.io/moderndiver-book/index.html).
- Another resource: **R for Data Science** by Hadley Wickham and Garrett Grolemund (https://r4ds.had.co.nz/index.html).

- For the class, you will use R markdown. R Markdown is a "quick" authoring tool where you can use an easy-to-write plain text format and combine that with R code to create a single, **reproducible** document.

- If you also have downloaded LaTex, then you can generally also include LaTex in your R Markdown (this is generally a good idea to do be able to do).

- For some cheat-sheets to R and RStudio see (https://www.rstudio.com/resources/cheatsheets/).

- Dr. Hadley Wickham has a nice video on getting your data into "R" (https://rstudio.com/resources/webinars/getting-data-into-r/)

- Other webinars that may be useful (https://www.rstudio.com/resources/webinars/)

John Tukey - Importance of Visualization

- PRIM~9 (Picturing, Rotation, Isolation, Masking in up to 9 dimensions)
  https://youtu.be/B7XoW2qiFUA