01 - Introduction to ST 597

ST 597 | Spring 2017
University of Alabama

01-Intro.pdf

Preliminaries

What is Analytics?

Syllabus

Course Tools

About

Last Things

Preliminaries

The Course

- This is ST 597: Introduction to Data Analytics
- Me: Dr. Porter
- Meet the TA: Huan Li

About you

Fill out a notecard with the following information:

1. Your name (with pronunciation hints)
2. Degree, Major and expected graduation date
3. Summer plans (intern, research, etc.). What industry or topics?
4. List coding/programming experience. Language and level (Scale 1-5, 5 highest).
5. Why are you taking this course?
6. Is there anything specific you want to learn?
7. 3 interesting things about you (to help me remember you)

Course Webpage

- Course material (including syllabus) can be found at: https://mdporter.github.io/ST597/
- Some material (e.g., solutions) will be posted on blackboard

What is Analytics?

Famous Quotes

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H. G. Wells

Not many executives are information-literate. They know how to get data. But most still have to learn how to use data.

Peter Drucker

The greatest value of a picture is when it forces us to notice what we never expected to see.

John Tukey

Analytics Revolution

Data Analytics

- This course will provide you with an introduction to **Data Analytics** (using the R software).
- There are lots of buzzwords and opinions about definitions.
- Some topics we cover may be described as:
  - Analytics
  - Data Science
  - Statistical/Machine Learning
  - Exploratory Data Analysis (EDA)
  - Statistical Computing
  - Data Mining
  - Big Data
- Mix of data analysis, computing, and statistics

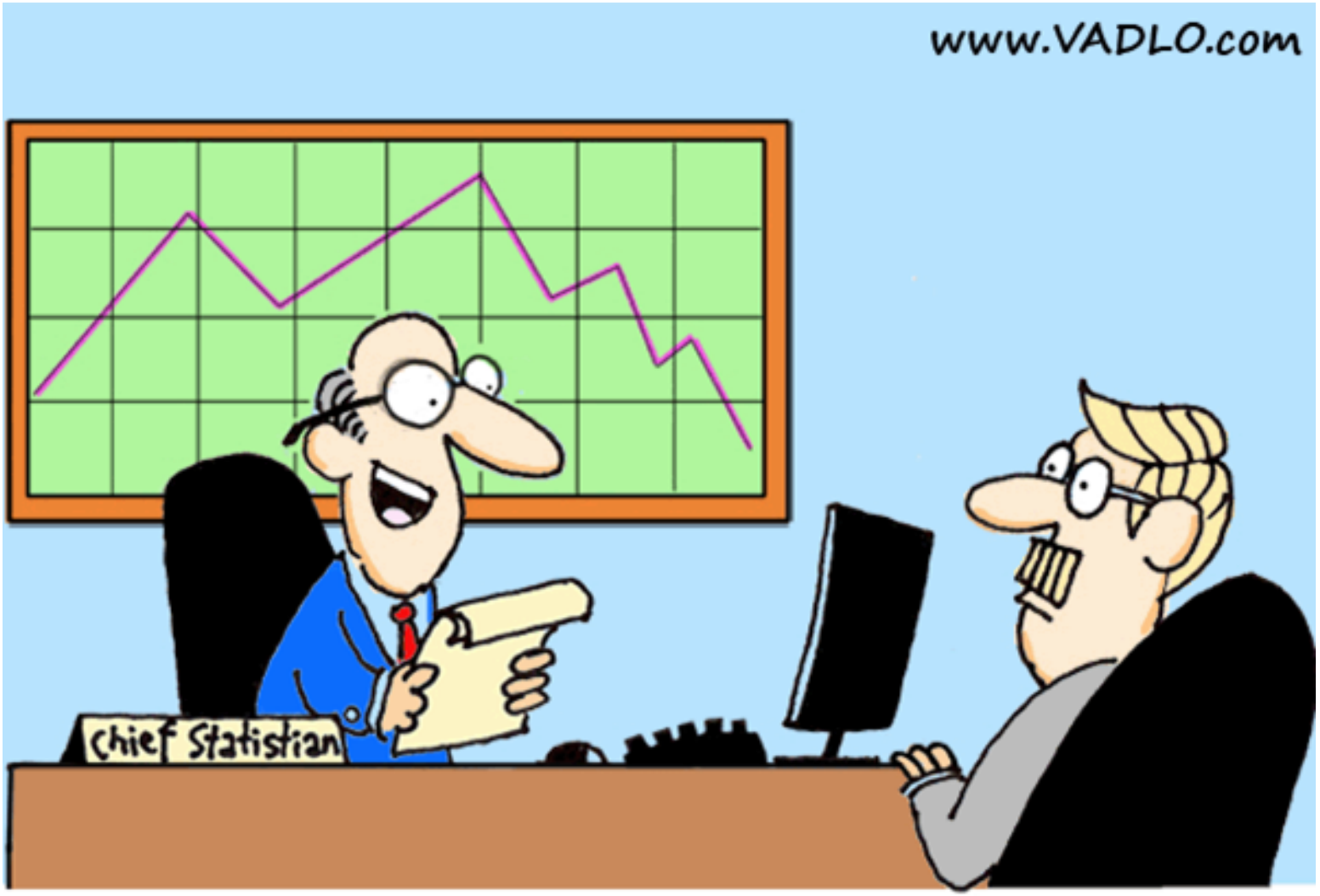
Statistics and Data Science

- Some *branches of* or *interfaces to* Statistics:
  - bio-statistics
  - econometrics
  - chemometrics, envirometrics, psychometrics
  - *metrics (Science + Statistics)
- Data Science and Analytics may be developing into another branch of statistics
  - **Data Science & Analytics** = Computer Science + Statistics
  - **Business Analytics** = Business Discipline + IT + Statistics
  - Besides Big Data, computing and coding is a major part of these fields
- Video Jeff Wu suggests **modern** Statistics should be renamed Data Science
- Video Terry Speed talk on Big Data

Why Study Statistics

What is Statistics?

There are three kinds of lies: lies, damn lies, and statistics. (Mark Twain) -



"I can prove it or disprove it! What do you want me to do?"

What is Statistics?

Seriously,

> Statistics is the science of learning from data.
>
> *- American Statistical Association (ASA)*

- Statistics is at the foundation of analytics, big data, and data science (and many other quantitative fields).
  - You will find many statistical methods being *rediscovered* in other fields
  - This is probably why some statisticians are reluctant to embrace these new *fields*
- As such, Statistics is one of the most sought after skills in the 21st century and ranked as one of the best graduate degrees.

Analytics Categories

The Institute for Operations Research and the Management Sciences (INFORMS) has proposed three categories of Analytics:

1. **Descriptive analytics**
   - Prepares and analyzes historical data
   - Identifies patterns from samples for reporting of trends
2. **Predictive analytics**
   - Predicts future probabilities and trends
   - Finds relationships in data that may not be readily apparent with descriptive analysis
3. **Prescriptive analytics**
   - Evaluates and determines new ways to operate
   - Targets business objectives
   - Balances all constraints

Two Phases of Data Analytics

A better way to think about analytics is in terms of two phases:

1. **Exploratory Data Analysis**
   - The exploratory phase "isolates patterns and features of the data and reveals these forcefully to the analyst" (Hoaglin, Mosteller, and Tukey; 1983)
   - If a model is fit to the data, exploratory analysis finds patterns that represent deviations from the model.
   - These patterns lead the analyst to revise the model, and the process is repeated.
2. **Confirmatory Data Analysis**
   - In contrast, confirmatory data analysis "quantifies the extent to which [deviations from a model] could be expected to occur by chance" (Gelman; 2004)
   - Uses the traditional statistical tools of inference, significance, and confidence.

Exploratory and Confirmatory Analytics

- Exploratory data analysis is sometimes compared to detective work: it is the process of gathering evidence.
- Confirmatory data analysis is comparable to a court trial: it is the process of evaluating evidence.
- Exploratory analysis and confirmatory analysis "can -and should- proceed side by side" (Tukey; 1977).

Why learn stats

http://stattrak.amstat.org/2013/08/01/insurance/

I recently went on a recruiting trip to a top school with both a stats department and a very well recognized "analytics" program, but we only visited the stats department. That was intentional because we have found that the graduates we have seen from that analytics program didn't understand the fundamentals of statistics or modeling, and that's so important to what we do it was a deal breaker. They knew how to push the right buttons and in what order, but the conversation broke down when we asked them why they push those buttons, and what happens if they deviate from that script?

- *Andy Pulkstenis, program director of analytics for State Farm Insurance*

Students perspective

MBA switching to analytics

- Several decent comments in this post. Notice how there is no agreement on what each term does. Good news is there are jobs for people with your exact skills, no matter the technical level.

Syllabus

Syllabus

Let's check out the course syllabus:
https://mdporter.github.io/ST597/syllabus.html

- Course Info
- Course Pre-Reqs

Course Description

- This course is an introduction to (exploratory) data analytics using the free and open-source software R.
- You will learn about the basics of exploratory and descriptive data analysis.
- We will cover things like obtaining, cleaning, combining, and wrangling the data into a more usable form.
- We will learn how to break up a large dataset into manageable pieces and then use a variety of quantitative and visual tools to summarize and learn about it.
- The challenges of big data (e.g., size, streaming data, mixed variable types) will be addressed throughout the course.
- As an introductory course, focus will be on understanding basic concepts and how to implement them in R.

Coding is not optional

- You *will* need to write code
- Learning a new language is frustrating, but will be rewarding
- Languages: R
  - Optionally: RMarkdown (to create html or pdf)
- Software: R, RStudio

Textbooks and Software

- Textbooks
- Software

Course Assessment

- Homework with DataCamp
  - I will use your email address from blackboard, so look for email
  - Homework posted at `https://mdporter.github.io/ST597/homework.html`
  - Start homework early - you will get more out of lecture
  - Note: 1st HW due in a week
- In-Class Participation
- Midterm Exam
  - similar to in-class problems
- Group Project

More Syllabus

- Course Outline
  - Schedule (https://mdporter.github.io/ST597/)
  - Read **before** class. And try the exercises.
    - Don't forget that the homework overlaps too!
- Academic Misconduct

Any questions about the syllabus?

Course Tools

What is R?

- https://www.r-project.org/about.html

Your Turn

(If using lab computer, skip step 1)

1. Download R: Go to `http://cran.r-project.org/` and click on your OS
   - Install **base** (for windows)
   - Choose your version for Mac
   - If you use linux, you don't need me to tell you what to do
   - Detailed instructions here
2. Open R
3. Use R to find:
   - $7 + 3$
   - $10 - 13$
   - $\pi * 3^2$
   - log(1)
   - Try this in R:

```
plot(x=-10:10, y=(-10:10)^2, type='l')
```

Who uses R?

- Core language of almost all Statisticians
- Top data science and analytics tool
  - 49% of Data Scientists use R
  - #5 overall programming language
  - Overall popularity
- R is forefront of many on-line courses
  - DataCamp
  - Coursera: Data Science Specialization
- Many companies using R
  - Microsoft
  - Google
  - New York Times
  - Airbnb
  - General Mills, LexisNexis Risk Solutions, Novartis
  - and many, many others

Why use R?

- $115K average salary for R users (Dice.com 2014 survey)
- R is Good for Business
- R is Popular
- R is free!
- R is the **best** program for interactive data analysis
- R can detect credit card fraud at 1M transactions/second
- R can run on multiple platforms (Windows, Mac, Linux)
- R is open-source
  - Companies can use and modify
    - R used at Microsoft
    - R incorporated into SQL server 2016
  - You can find out what the code is actually doing
- SAS (and many other programs) allows you to run R code
- We will do a textual analysis of business analytics postings which will reveal the growing demand of R

What can R do?

- R is a programming language so you can get it to do all sorts of things
  - but its core strength is *data analysis*
- Some basic functionality:
  https://www.r-project.org/about.html
- But strength of R is in its **packages**
  - Contributed by users
  - Over 12,000 R packages
- Task Views http://cran.r-project.org/web/views/
- Reproducible research

RStudio

- RStudio is an integrated development environment (IDE) for R
  - It is also free, open source, and cross-platform!
  - Download RStudio
  - Install after R

http://vimeo.com/97166163

- We will do all "coding" in RStudio
  - R
  - Rmarkdown
- RStudio also facilitates Shiny for interactive visualizations: http://shiny.rstudio.com/gallery/
- Note: RStudio is not R. But it facilitates the use of R (and many other things).

Open RStudio if you have it loaded

RMarkdown

- RMarkdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R.
  - RMarkdown is included with RStudio; there is nothing extra to install
- It combines the core syntax of markdown (an easy-to-write plain text format) with embedded R code chunks that are run so their output can be included in the final document.
- RMarkdown documents are fully reproducible (they can be automatically regenerated whenever underlying R code or data changes).
- RMarkdown cheatsheet

Final Class Details

- Use Lab Computers or Bring laptop to class (with R and Rstudio loaded)
  - We will be doing lots of in class examples
  - Ensure you can access lab computer info (if necessary)
- Class is extremely cumulative
  - If you can't come to class, spend time with the lecture notes and talk to someone who attended
  - If you are falling behind, catch up as soon as possible
- The first few weeks can be frustrating; stick with it and you will start getting the hang of it
  - Use google!
- Expected time commitment (weekly):
  - 2.5 hrs in class, 3-6 hrs homework, 1-2 hrs reading/practicing
  - Total of 6.5-10.5 hrs/week

About

About your instructor

- 1994-1998: Purdue University
  - B.S. Industrial Engineering
- 1998-2001: Chicago, IL - Sanford Markers
  - Engr/Maintenance
- 2001-2003: Vanderbilt University
  - M.S. Systems Engr
- 2003-2006: University of Virginia
  - PhD Sys and Info Engr
- 2006-2008: North Carolina State University & SAMSI
  - Postdoc
- 2008-2013: Spadac/GeoEye/DigitalGlobe
  - Principal Research Scientist
- 2013-Present: University of Alabama
  - Assistant Prof

My Family

About you

Fill out a notecard with the following information:

1. Your name (with pronunciation hints)
2. Degree, Major and expected graduation date
3. Summer plans (intern, research, etc.). What industry or topics?
4. List coding/programming experience. Language and level (Scale 1-5, 5 highest).
5. Why are you taking this course?
6. Is there anything specific you want to learn?
7. 3 interesting things about you (to help me remember you)

Last Things

ToDo

- Start DataCamp homework (look for email)
- Read R4DS 1 and 2
- Note any questions from reading and homework
- Get these RStudio Cheatsheets
  - Data Visualization
  - Data Wrangling
  - Base R
  - Regular Expressions

Interesting Reading

- Deliberate Practice: the making of an expert https://hbr.org/2007/07/the-making-of-an-expert

```
# Hello from R World
```