By the way…..

2 upcoming 1-day workshops in Davis

Saturday, 4th April Web Scraping & Web Services

Saturday, 25th April Visualization

Both 9.45 - 4.30 & free!

Apply via
http://datascience.ucdavis.edu/NSFWorkshops/

Sponsored by the National Science Foundation,
Data Science Initiative, Institute for Social Sciences.

Teaching Assistant

Nick Ulle (naulle@ucdavis.edu)

You’ll interact with him via office hours & Piazza.

Post (non-personal) questions to Piazza rather than to Nick or myself individually

You’ll get a faster response (2 of us).

Other students may reply even faster.

Answer & discussion will be beneficial for all students.

Course Sites

Course Web Site http://eeyore.ucdavis.edu/stat242

Piazza class discussion forum
http://piazza.com/ucdavis/spring2015/stat242

Git repository (with assignments, etc.)
https://bitbucket.org/duncantl/stat242_2015

Smartsite - for recording grades.
Please ensure they are correctly recorded and bring any anomalies to our attention ASAP.
Themes & Topics

Computational reasoning for working with data & simulation - especially “not-small” data.

Fundamentals of a Programming Language for data analysis, and basics of other languages

Different computational paradigms & technologies for data analysis, e.g., parallel computing, block operations.

Tools for your research and data analysis.

Developing Software (i.e. for others to use)

Themes & Topics ctd.

Exposure to some statistical methods not always seen in other courses (with a computational and heuristic focus)

Fundamental skills & technologies of data science.

Aspire to

More than just getting the computations done.

Do them elegantly, efficiently, flexibly by design and with reuse in mind.

Waiting List

Limited room capacity !

On Thursday 4th April, we are in 107 Cruess Hall. Split class into 2.

???

If you are considering dropping this course, please decide as soon as possible so others can take that spot.
There is a tension in this course because 57% of you have taken STA141, and 43% of you haven't.

We'll cover the core material in STA141 reasonably quickly.

It will be a good refresher for those of you who have already seen it, and we'll cover additional concepts in this startup period.

I strongly encourage you to use assignment 1 to go further than the material you have already seen and learn new related functionality and concepts.

Basics of R?
Vectors, and vectorized operations?
Writing functions?
lapply() & sapply()?
tapply()/by()/split()-lapply() - group by?
creating graphical displays?
text manipulation?
regular expressions?
UNIX shell, redirection (pipes)?
databases?

Review of the fundamentals of R
Exploratory Data Analysis and Visualization
Reading from Non-Standard Data Formats
Interfacing to C and the basics of the C language
Creating R packages
Object-oriented programming concepts and designing software
Scope, Non-Standard Evaluation, Formulas in R
Efficient computing - in R and beyond.

High-level Parallel Computing using multiple cores, multiple machines, cluster software
MapReduce & Hadoop/Spark concepts and related technologies (e.g., Hive, Pig)
When to use what approach - shell, databases, C, Hadoop, clusters, GPUs.
Using and interfacing to other languages, e.g., Java, Python
Version control (e.g., git)
Reproducible computations and dynamic documents (e.g., Sweave & knitr)
- Visualization - interactive, dynamic, Web-based approaches (HTML, JavaScript, D3, SVG, …, Client-Server, In-browser, Shiny)
- Text Processing and Natural Language Processing
Teaching Style

Semi “flipped” class

Spend first part of class discussing questions that you have from doing the assignment.

You need to be continuously working on the assignments for this to work and for you to learn from the process.

Discussion will also happen a lot on Piazza.

Other half - introduction to topics, collectively solving new problems - lots of interaction and participation!

Student Style - 4 I’s

I expect students to have 4 I’s

Interest

Independence

Initiative

Inquisitive

In other words, you should be self-motivated, self-starting, self-learning and aiming to excel.

This course involves “active learning” - i.e. you learn by doing.

Reflect on what you have learned once it is “complete”.

Grading

Assignments  75%

Probably 5 or 6 assignments (1 or 2 shorter ones)

Group project - 20%

3 in each group

Participation - 5%

In class, Piazza

HOWEVER, if there is any cheating, I will hold an in-class exam

Group Project

Get started now.

Be ambitious.

Consider it part of your portfolio, advertising your value and capabilities.
Possible Project Types

R package or python module (e.g. implementing a statistical method) for something other people may want to do/use.

Interface to some existing 3rd-party library.
Advanced visualization or dashboard for some dataset
Reproduce computations for computationally intensive paper, or reproduce a study on a different data set.
New data analysis.
Compare statistical methods on a rich data set, or via simulation.
Interface to Web service.

Three people per group.
The grade for your project is based on
ambition and novelty of the project
effective implementation & completion
quality of communication of the result.

For many of you, think of this as a capstone project and do a terrific job.
Submit a project “proposal” - in 2 weeks.
Goal, why, how and what you will create.

Homework 1

Provide me with your backgrounds and a picture of yourself
Connect your name and your face.
I hope to get to know you all & be able to adapt the course to your hopes & backgrounds.
Important information and process.
Do this today! Not for grade, but important.

Assignment 1

Analyze several years of results from a 10 mile run event - the Cherry Blossom race.
Males & Females, age, 5 mile and 10 mile times, name & home city and state.
Exploratory data analysis, visualization,
Data input, data “munging”, data matching.
Cross-sectional and longitudinal data analysis.
Assignment 1

Due Friday 10th, 4pm

Electronic version (via email and git) to me
dtemplelang@ucdavis.edu

Printed copy of report to Nick
physical mailbox in Statistics Department Office)

Maximum of 10 pages (excluding code)
print double-sided.

Code in 2 appendices.

Get started immediately.

Late Assignments

Start early.
Computing has lots of thorny details that take time.
Also, some computationally intensive assignments that take time to run.
Tell me ahead of time - not the day it is due.

Don't assume you can hand it in late.
Always submit something for each assignment.
I may give an individual more time.

2 Topics for Today & Thu

Introduction to git, bitbucket.org and accessing the class repository.

Example of data input & exploratory data analysis - EDA

Important Related Concepts

Automation
Reproducability
Provenance
Reuse
Extensibility