Intro to R

Fuzzy Rogers
Research Computing Administrator
Materials Research Laboratory (MRL)
Center for Scientific Computing (CSC)
fuz@mrl.ucsb.edu
MRL 2066B

Sharon Solis
Research Computing Consultant
Enterprise Technology Services (ETS)
Center for Scientific Computing (CSC)
swsolis@ucsb.edu
Elings Hall 3229

Paul Weakliem
CNSI Research Computing Support
California NanoSystems Institute (CNSI)
Center for Scientific Computing (CSC)
weakliem@cnsi.ucsb.edu
Elings Hall 3231

Pre-class Instructions:

- Download R: http://cran.stat.ucla.edu/ (closest mirror)
- Download RStudio: https://www.rstudio.com/products/rstudio/download/
- Download Kaggle dataset and R code: Box download

What is R?

- R is a programming language and free software environment for statistical computing and graphics
- R was initially written by **Robert Gentleman** and **Ross Ihaka**—also known as “R & R” of the Statistics Department of the University of Auckland in 1992.

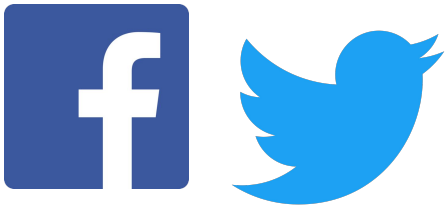
THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

What is R?

- R was inspired by the S environment which has been principally developed by John Chambers in 1976, while at Bell Labs.

http://thegrantlab.org/bio3d/79-static-content/85-why-use-r

Who uses R?



https://www.listendata.com/2016/12/companies-using-r.html

Why Use R?

- Powerful, state-of-the-art
- Used by professional statisticians
- Lot of documentation (StackOverflow)
- Freely available for Unix, Windows & Mac
- Extendable, with numerous add-on packages available.
- R produces publication quality graphics.

http://thegrantlab.org/bio3d/79-static-content/85-why-use-r

What is RStudio?

Integrated development environment (IDE)

- Console
- Syntax-highlighting editor that supports direct code execution
- Tools for plotting, history, debugging and workspace management
- Pretty!

Titanic Data Set

- https://www.kaggle.com/c/titanic
- Overview data set - titanic
- What problem are we solving?
- What is machine learning modelling?
-
- What is the train vs test set?
    - Different data, not duplicate
- Why combine the data sets?
- Can we predict who survived the Titanic?

https://www.kaggle.com/c/titanic/data

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Our Variables

- **pclass**: A proxy for socio-economic status (SES)
  1st = Upper
  2nd = Middle
  3rd = Lower
- **age**: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- **sibsp**: The dataset defines family relations in this way...
  Sibling = brother, sister, stepbrother, stepsister
  Spouse = husband, wife (mistresses and fiancés were ignored)
- **parch**: The dataset defines family relations in this way...
  Parent = mother, father
  Child = daughter, son, stepdaughter, stepson
  Some children travelled only with a nanny, therefore parch=0 for them.

https://www.kaggle.com/c/titanic/data

What is a Script?

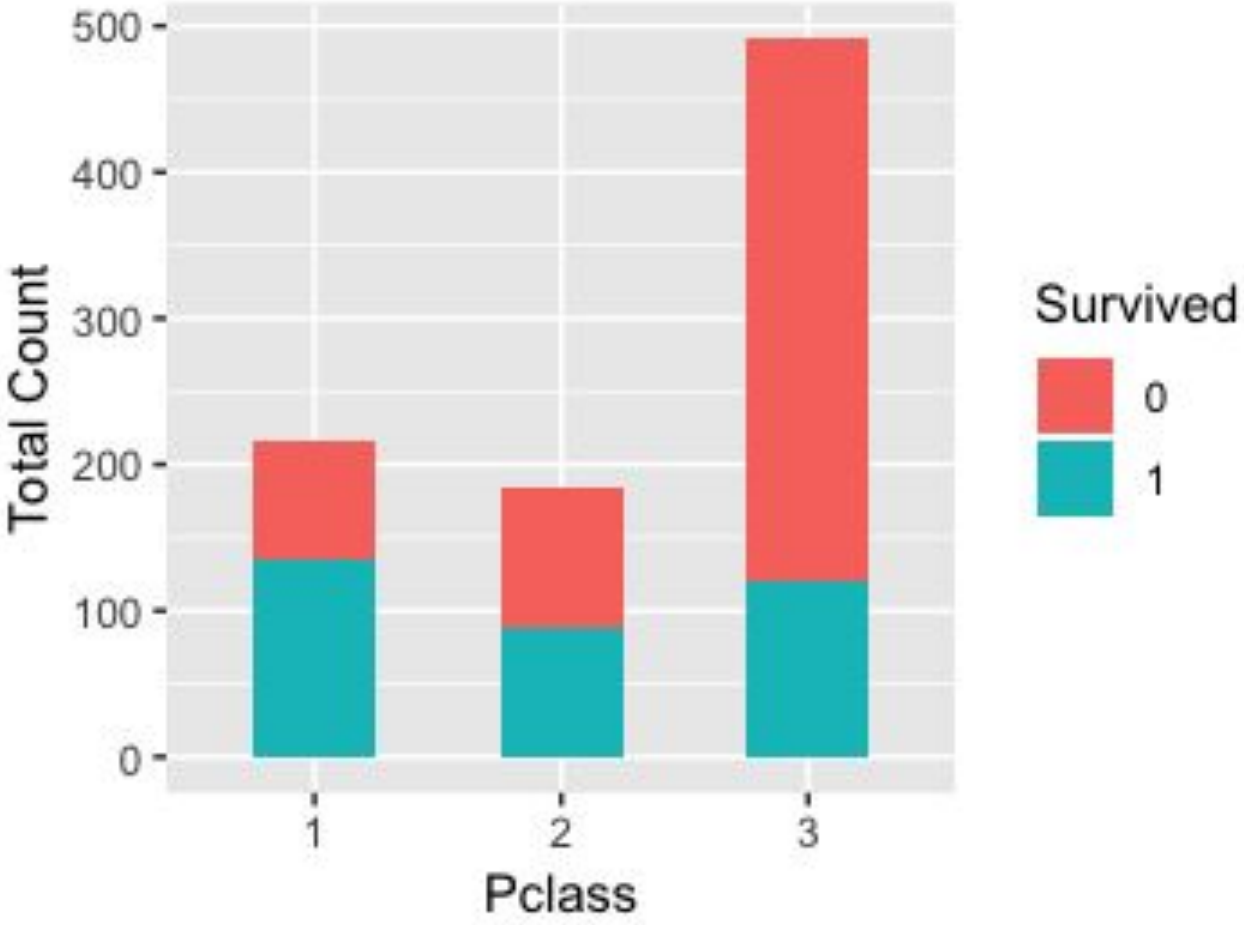
- How to run code
- Save yourself work!
- Don’t need to type over and over again
- Move easily between machines

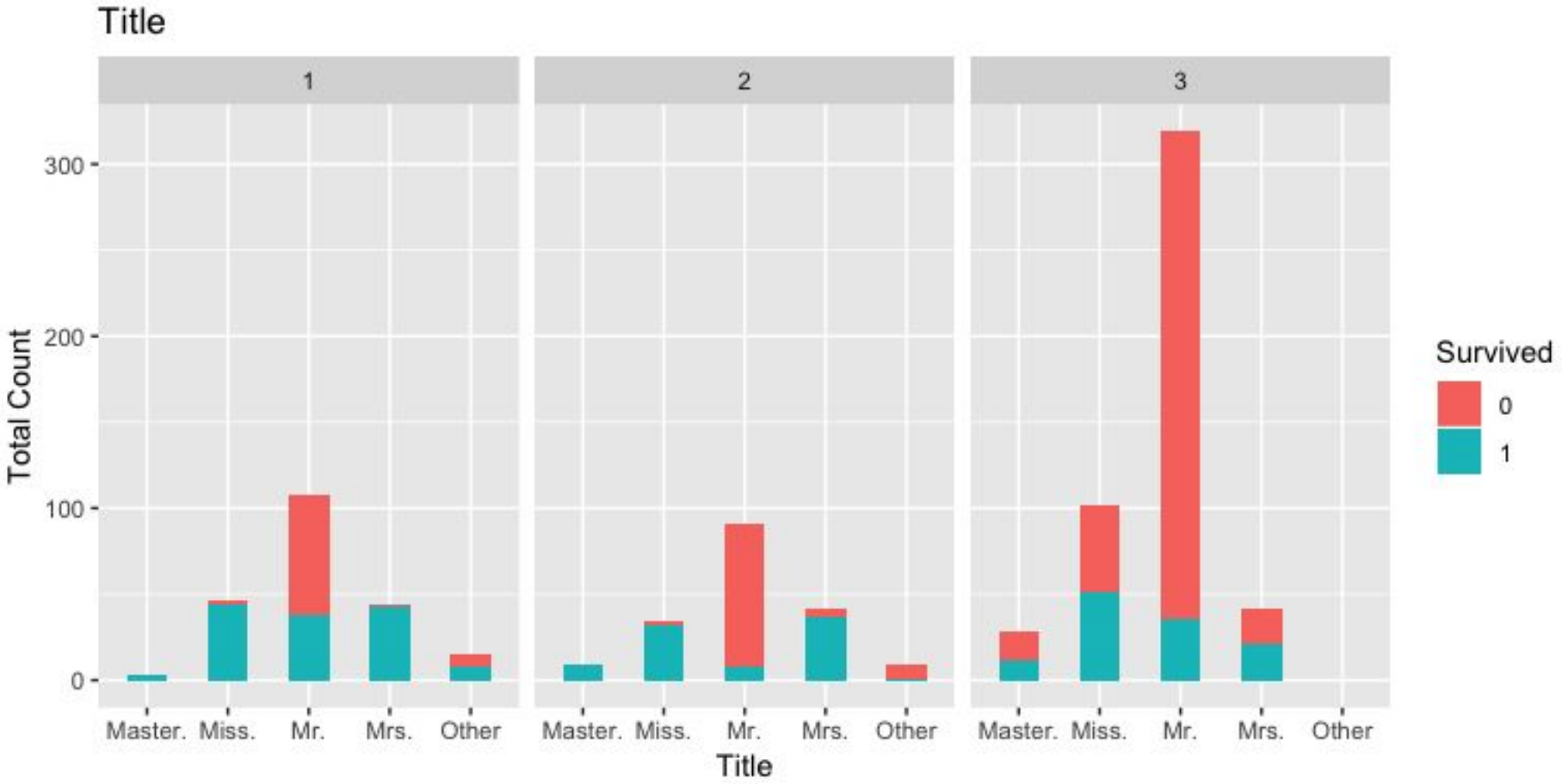
How to Read a File

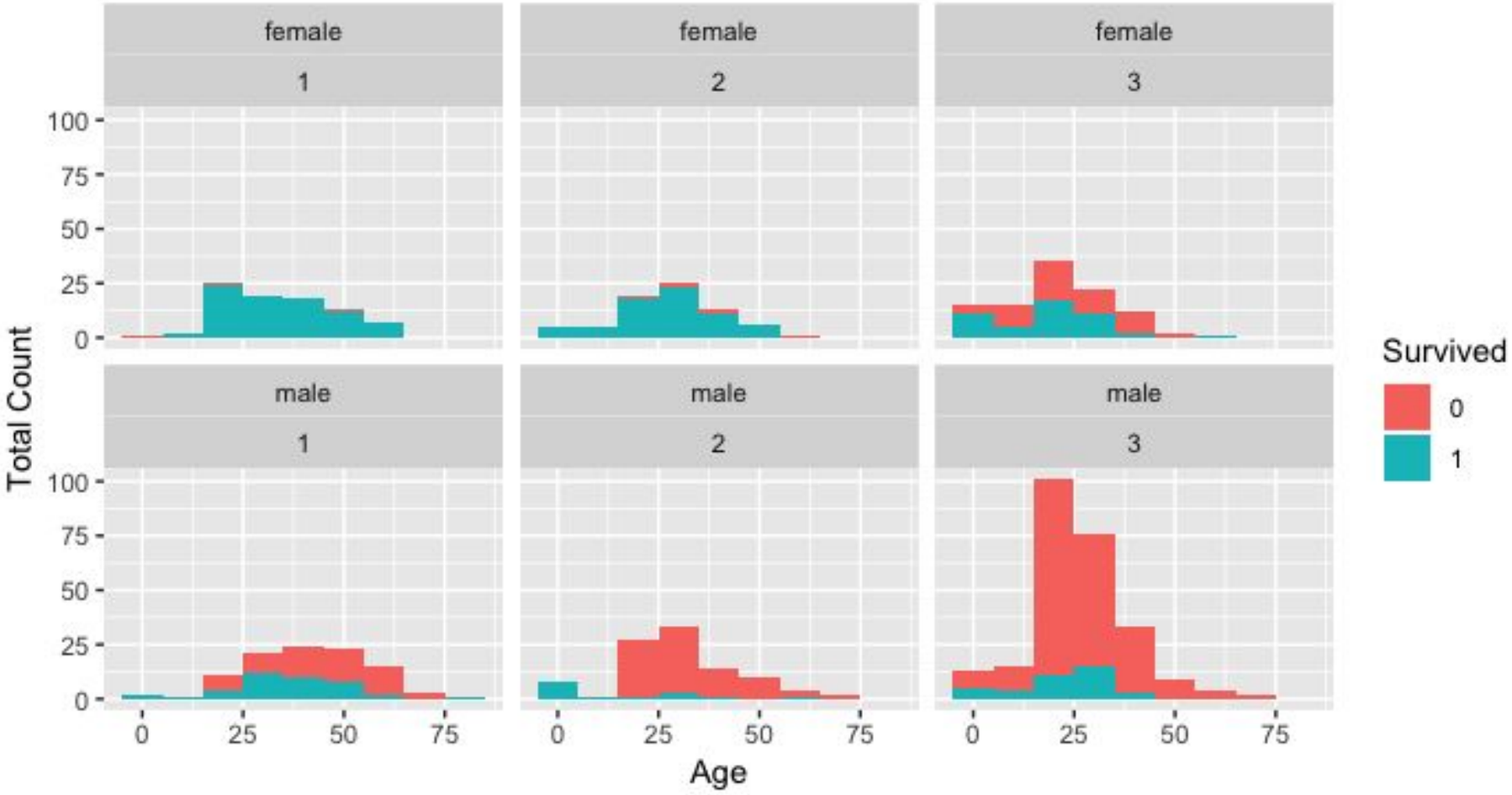
- Remember to set working directory

How to Get Help

- E.g., ?factor
- Help window
- StackOverflow


500
400
300
200
100
0
Total Count
1
2
3
Pclass
Survived
0
1


Title
1
2
3
300
200
100
0
Total Count
Master.
Miss.
Mr.
Mrs.
Other
Title
Survived
0
1


female
1
female
2
female
3
male
1
male
2
male
3
Total Count
100
75
50
25
0
0
25
50
75
Age
Survived
0
1

Installing Packages

- What are packages
  - libraries
- How to install packages
- Fftw
- Ggplot2
- Other examples
- Swirl

Using R on a Cluster

- Use R not RStudio on the cluster
- Make sure your R code runs from start to end on your own machine
- Perform tests on your computer first
- A simple script (text file) can be used to submit to the queue:

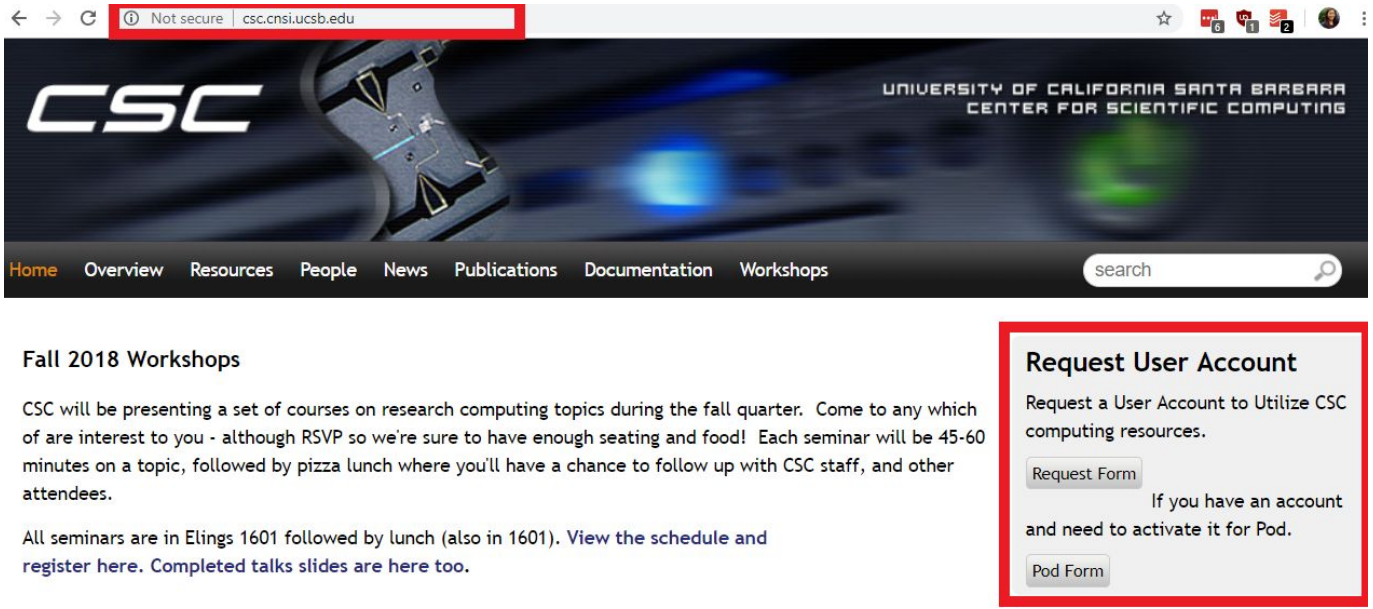
```
#!/bin/bash -l
#Serial (1 core on one node) job...
#SBATCH --nodes=1 --ntasks-per-node=1
cd $SLURM_SUBMIT_DIR
Rscript --vanilla example.R
```

Using R on the Cluster

```
#!/bin/bash -l
#Serial (1 core on one node) job...
#SBATCH --nodes=1 --ntasks-per-node=1
cd $SLURM_SUBMIT_DIR
Rscript --vanilla example.R
```

#!/bin/bash : the shell you are using
#Serial (1 core on one node) job : serial job, not parallel
#SBATCH --nodes=1 --ntasks-per-node=1 : Asking for one node and one task per node
cd $SLURM_SUBMIT_DIR : change directory to the one where job is submitted from
Rscript --vanilla example.R : Run your example.R code

How to Request a User Account



What Else Can You Do with R?

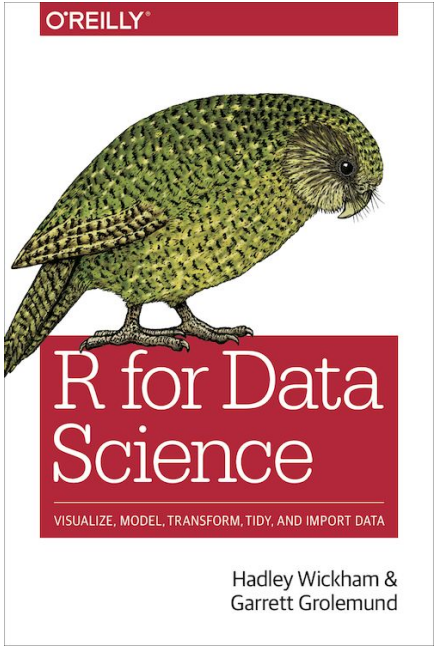
- Predictive modeling
- Machine Learning
- Statistical Analysis
- Economic forecasting
- Predict financial market changes
- Data visualization
- Semantic clustering

Swirl

- R package that leads you through an interactive tutorial to learn R
- Interactive within console

How to Learn More about R

- Online Tutorials:
  - Coursera, DataCamp, YouTube
  - Lynda.com (available to UCSB employees, including student employees)
- Swirl
  - (RStudio package, interactive tutorial within console)
- Stackoverflow
  - (a great forum of questions and answers about computer programming)
- One-on-One Consultation
  - Center for Scientific Computing (Elings Hall 3229)
  - Collaboratory
- Books
  - R for Data Science by Hadley Wickham and, Garrett Grolemund



Contact Us

csc.cnsi.ucsb.edu

Sharon Solis
Research Computing Consultant
Enterprise Technology Services (ETS)
Center for Scientific Computing (CSC)
swsolis@ucsb.edu
Elings Hall 3229

Fuzzy Rogers
Research Computing Administrator
Materials Research Laboratory (MRL)
Center for Scientific Computing (CSC)
fuz@mrl.ucsb.edu
MRL 2066B

Paul Weakliem
CNSI Research Computing Support
California NanoSystems Institute (CNSI)
Center for Scientific Computing (CSC)
weakliem@cnsi.ucsb.edu
Elings Hall 3231