Two Different Skillsets

1. Statistics

2. Programming

Computational

```
#subset consumptionData into mixed diet treatments only
#(mixed diet BOX #'s are multiples of 5)
mixedData<-consumptionData[consumptionData$BOX %% 5 == 0,]
mixedData$delta<-mixedData$ADD_WT - mixedData$REM_WT
mixedData$rate<-mixedData$delta/mixedData$days

#reshape to get species-specific consumption rate table
mixed.summary<-ddply(mixedData,.(BOX,SP_CODE),function(x){
  data.frame(CONSUMPTION_RATE=mean(x$rate,na.rm=TRUE))
  })

#############fit linear models (not including consumption--see below)
LMtestChange <- lm(formula = TEST_CHANGE ~ TREATMENT,
                   data = expData, na.action = na.omit)
```

Coding is power

Computational

```
#subset consumptionData into mixed diet treatments only (mixed diet BOX #'s ar
mixedData<-consumptionData[consumptionData$BOX %% 5 == 0,]
mixedData$delta<-mixedData$ADD_WT - mixedData$REM_WT
mixedData$rate<-mixedData$delta/mixedData$days

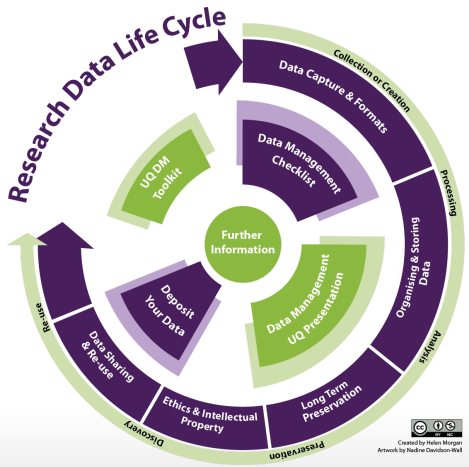
#reshape to get species-specific consumption rate table
mixed.summary<-ddply(mixedData,.(BOX,SP_CODE),function(x){
  data.frame(CONSUMPTION_RATE=mean(x$rate,na.rm=TRUE))
  })

#############fit linear models (not including consumption--see below)
LMtestChange <- lm(formula = TEST_CHANGE ~ TREATMENT, data = expData, na.actio
```

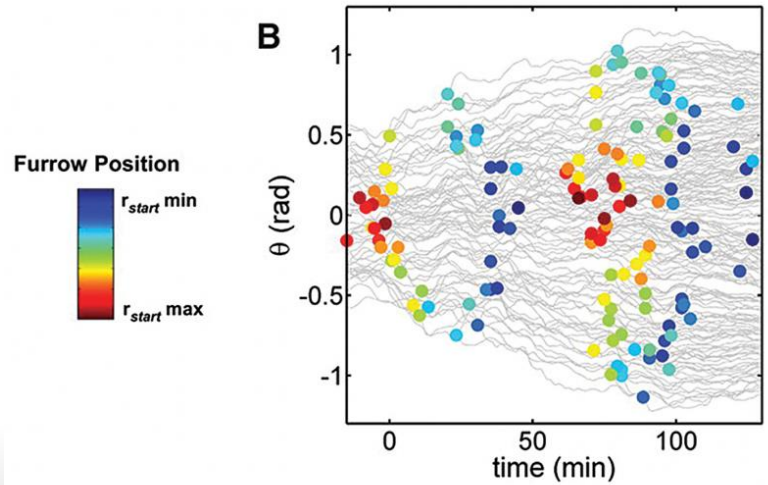
Repeatable Research

Data (acquisition)

Data (maintenance)



Analysis (visual)



Analysis (model)

```
##
## Call:
## lm(formula = shoots ~ treatment.genotypes, data = eelgrass)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27.47 -10.72  -1.30   8.96  35.70
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            30.66       5.32    5.76  2.7e-06 ***
## treatment.genotypes     4.63       1.40    3.31   0.0024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.4 on 30 degrees of freedom
## Multiple R-squared:  0.267,  Adjusted R-squared:  0.243
## F-statistic: 10.9 on 1 and 30 DF,  p-value: 0.00245
```


for Biology

for Biology

incremental: true ## SCIENCE FIRST!

- What is your model(s)?
- THEN decide on statistical approach
- Can you get data to paramaterize that model?
- How does biology inform your modeled results?

Course Goals

1. Learn how to think about your research in a systematic way to design efficient observational & experimental studies.
2. Understand how to get the most bang for your buck from your data.
3. Make you effective collaborators with statisticians.
4. Make you comfortable enough to learn and grow beyond this class.

Who are You?

1. Name
2. Lab
3. Brief research description
4. Why are you here?

Outline for Today

1. What are we doing here?
2. Who are we?
3. How will this course work?
4. A Philosophy of answering scientific questions with data
5. R!

Lecture/Lab/Labinar?

- I will yammer on
- R lab will be part of class
- Notes available at http://byrneslab.net/biol607
- Slide source available at http://github.com/jebyrnes/biol607

Yes, Lectures are Coded

R Presentations + R Markdown via knitr

```
<!--
##Lecture 1: Introduction

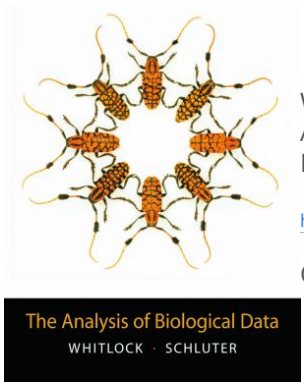
# Changelog
#
-->
Introduction to An Introduction to Computational Data Analysis for Biology
==========================================================
author: Jarrett Byrnes
date: 9/2/2014
css: 607slides.css
```{r echo=FALSE}
source("./slidePrep.R")
```

Outline for Today
==
incremental: true

1. What are we doing here?
2. Who are we?
3. How will this course work?
4. A Philosophy of answering scientific questions with data
5. R!

Two Different Skillsets
==
```

Readings for Class: W&S

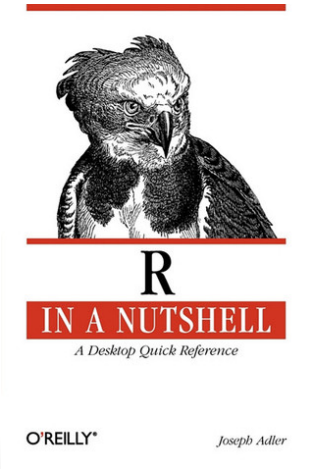


Whitlock, W.C. and Schluter, D. (2014) The Analysis of Biological Data, 2nd Edition. Roberts and Company Publishers

http://www.zoology.ubc.ca/~whitlock/ABD/teaching/index.html

Chapter 1 this week!

Readings for Class: Adler

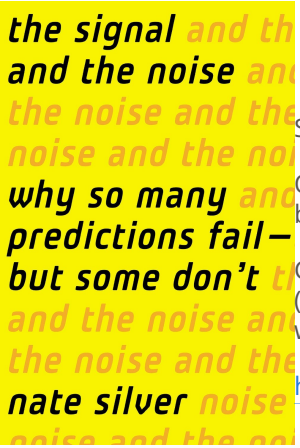


Adler, J. (2009) R in a Nutshell: A Desktop Quick Reference. O'Reilly.

Can be found online

Chapter 1 this week!

Think Big: Nate Silver



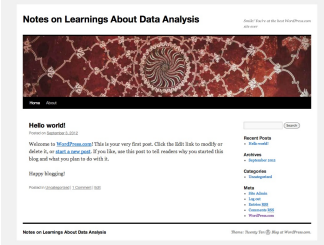
Silver, N. 2012. The Signal and the Noise.

One chapter a week to make you think bigger than this class

Got something else you'd rather read (The Lady Tasting Tea, Deb Mayo's writing, etc.) - Great!

http://fivethirtyeight.com/

Think Big: Blog



http://learningdata.wordpress.com/

Get a http://wordpress.com account NOW & email me

Signup for 3 dates at http://goo.gl/P0L8ls

Comment 3 times

10% of your grade

Can use class, Silver, or other blogs as points of inspiration

Problem Sets

- 40% of your grade
- "Adapted"" from Whitlock and Schluter
- Will often require R
- Complete them using knitr package and turn in pdfs
- http://yihui.name/knitr/ (and tutorial next class)

Midterm

- Advanced problem set
- Due Oct. 21st
- 20% of your grade

Final Project

- Topic of your choosing
  - Your data, public data, any data!
  - Make it dissertation relevant!
  - If part of submitted manuscript, I will retroactively raise your grade
- Dates
- Proposal Due Oct 14th
- Presentations on Dec 11th
- Paper due Dec 16th (but earlier fine!)
- 30% of your grade

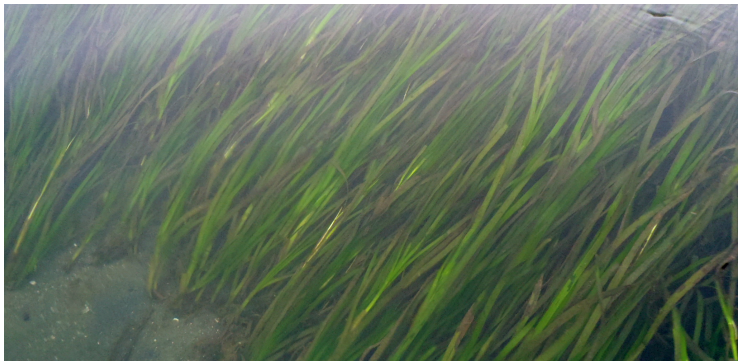
Outline for Today

1. What are we doing here?
2. Who are we?
3. How will this course work?
4. A Philosophy of answering scientific questions with data
5. R!

Our Approach to Data Analysis

Data from Reusch et al. 2005 PNAS

Start with a Question

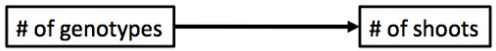


Does seagrass genetic diversity increase productivity?

Build an Understanding of the System

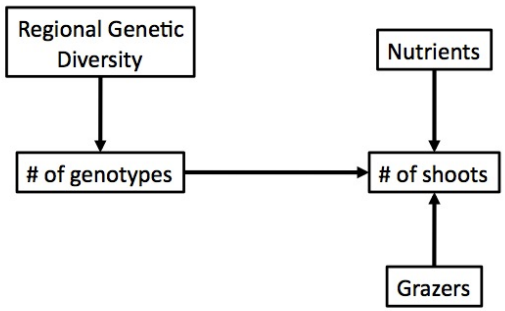
1. Literature
2. Observation
3. Disciplinary History

Construct a Causal Model of the System



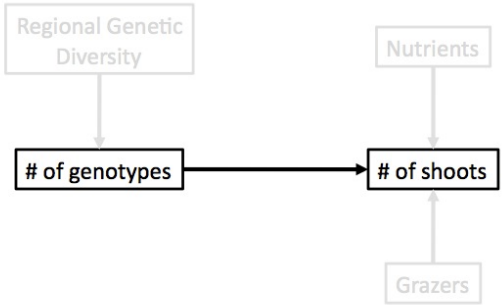
Start with your question

Construct a Causal Model of the System



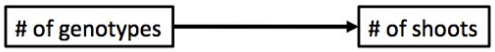
Big Picture

Construct a Causal Model of the System



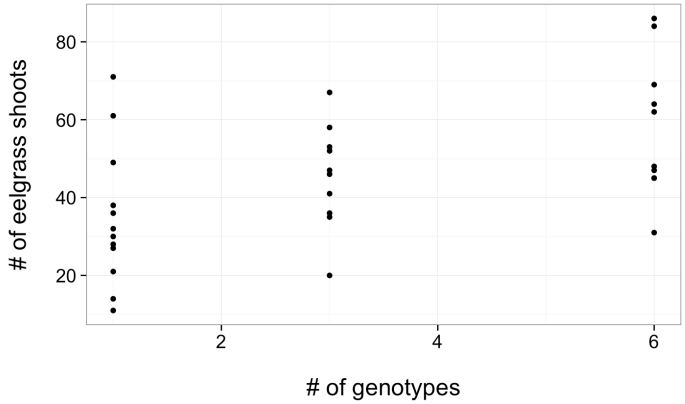
What can you isolate?

Collect the Data to Best Estimate & Test the Model}



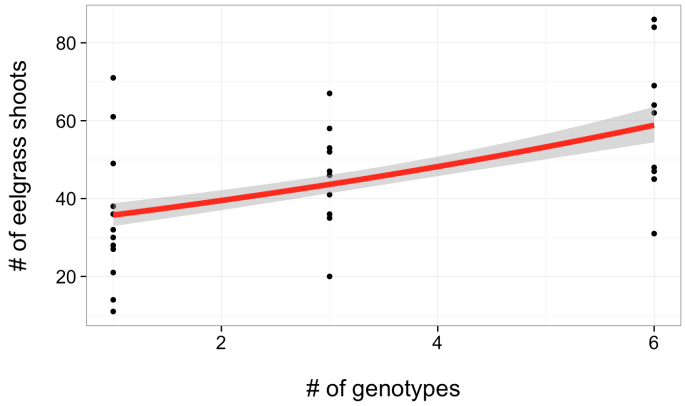
1 Genotype	3 Genotypes	6 Genotypes

Look at Your Data



Fit a model(s), chosen to suit this data!

Analysis!



Build Open Reproducible Research

Many Methods of Sharing Data, Methods, and Results Beyond Publication

1. GitHub - public code repository
2. FigShare - share key figures, get a doi
3. Blog - open 'notebook'
4. Dryad or Other Repository - post-publication data sharing

Outline for Today

1. What are we doing here?
2. Who are we?
3. How will this course work?
4. A Philosophy of answering scientific questions with data
5. R!

What is R?

A programming language uniquely developed for statistical analysis

Why R?

1.Free

1. Huge growing community
2. Packages to do almost anything
3. Makes reusable research easy
4. C-based language
5. Syntax naturally matches analytical thinking

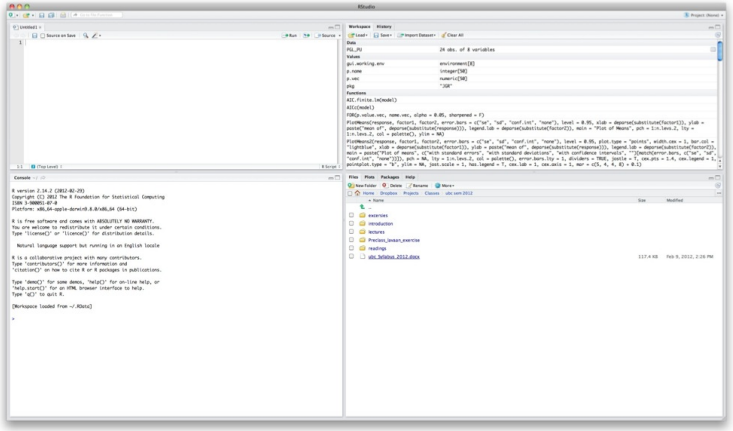
What is R Studio?

Let's Fire It Up!

Open R-Studio.

Don't have it? Download it from http://rstudio.org

What do you see?



The Console and Math

```
1+1
```

```
## [1] 2
```

You try - different mathematical operators

Everything is an Object

```
a.number<-1+1
```

```
a.number
```

```
## [1] 2
```

You try - what can you save as an object?

Note: Comment Your Code as You Write with

The text after # is not evaluated.

```
#This is going to be the number two
a.number<-1+1
```

```
#####----------

# You can get creative with comments to separate code blocks
# And write a lot, which is good practice

#####----------
```

Your comments tell readers - including yourself - what you are doing

Functions Work on Objects

```
sin(a.number)
```

```
## [1] 0.9093
```

How to get help for a function

```
?cos

help(cos)

??'cosine function'
```


Lots of Object Types - like Data!

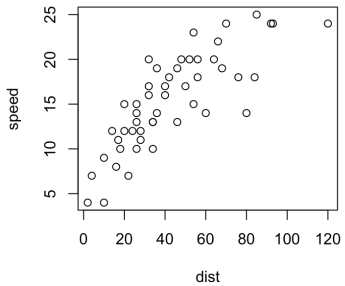
```
head(cars, n=3) #note the n= argument!
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

Try looking at all of cars and names(cars)

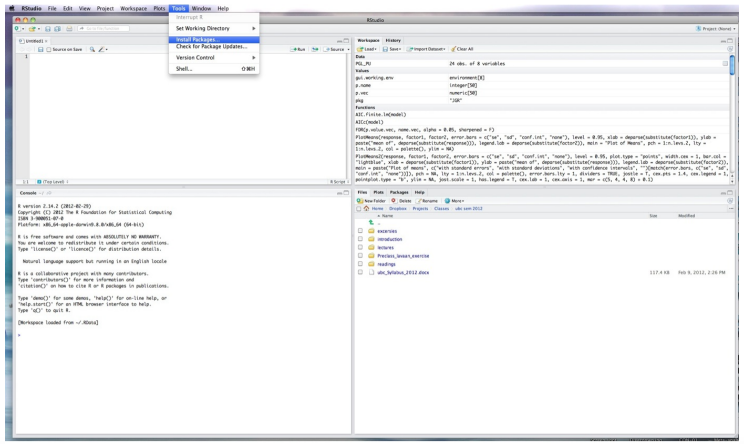
Graphics are a Snap

```
plot(speed ~ dist, data=cars)
```

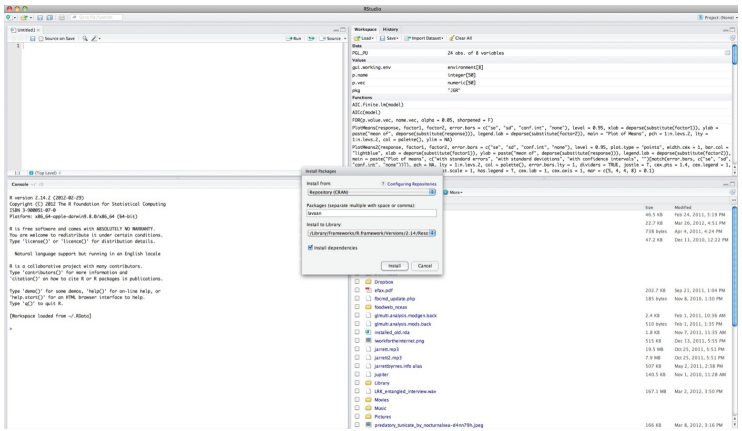


Look at ?plot to see other arguments to change appearance

Installing Packages



Installing Packages



Installing Packages

You can also install packages from the command line.

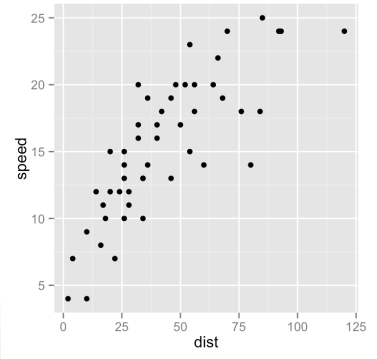
```
install.packages('ggplot2',
                 repos='http://cran.case.edu/',
                 dependencies=TRUE)
```

Using one of the above methods, install the package ggplot2 and its dependencies now.

Using a Package

```
library(ggplot2)

qplot(dist, speed, data=cars)
```



You Try It

1. Load ggplot2 and look at the mtcars data set

2. Look at the qplot help file & demos

3. Make two plots (ggplot or plot)