**Coursera Course 5 Week 1: Reproducible Research**

**What is Reproducible Research?**

Replication  
-ultimate standard for strengthening scientific evidence is replication of findings and conducting studies with independent  
 -investigators  
 -data  
 -analytical methods  
 -laboratories  
 -instruments  
-replication is important  
BUT  
-some studies can’t be replicated because of resources restraint  
**Reproducible Research**- make analytic data and code available so that others may reproduce findings

Replication vs Nothing (middle ground)

**Why do we need Reproducible Research?**-new technologies increasing data collection throughput; data are more complex and extremely high dimensional  
-existing databases can be merged into new “megadatbases”  
-computing power is greatly increased, allowing more sophisticated analyses  
-for every field “X” there is a field “Computational X” // biology etc

Ex: Reproducible air pollution and health  
-estimating small (but important) health effects in the presence of much strong signals  
-results inform substantial policy decisions, affect many stakeholders  
i.e. EPA  
-complex statistical methods are needed and subjected to intense scrutiny  
🡪 iHAAPS

**Research Pipeline**-a flow chart to show from process to analyze to present etc

**The IOM Report  
Data/metadata** used to develop test should be made publicly available  
**Computer code** and fully specified computational procedures should be made sustainably available  
Computer code should have **ALL steps of computational analysis, including data preprocessing steps.**

**We need:**Analytic data are available  
Analytic code are available  
Documentation of code and data   
Standard means of distribution

Authors  
-want to make their research reproducible  
-want tools for RR to make their lives easier

Readers  
-want to reproduce this research

**Challenges**  
-authors must undertake considerable effort to put data on the web  
-readers must download data and piece them together  
-readers don’t have the same resources  
-need more tools

SO…

Authors throw thins on the web, journal supplementary materials etc, biology ICPSR

Readers – just download the data and try to figure it out by hand

**Literate (Statistical) Programming**-an article is a stream of text and code  
-analysis code is divided into text and code “chunks”  
-each code loads data / compuptes results  
-presentation code formats results  
-article text explains whats going on  
“weaved” and “tangled”

We will need a documentation language for humans and a programming language for machines. (LaTeX and R)

-difficult to learn for LaTeX, don’t have much tools

-knitr allows you to mix other programming language with R, can also use Markdown, HTML

**Scripting your Analysis**WRITE EVERYTHING DOWN ☺R Script – you can even just use the text editor that’s built in.

**Structure of a Data Analysis**  
define question  
define ideal data set  
determine what data you can access  
obtain the data  
clean the data

Key Challenge: Problems versus sufficient information to solve them

**1. Define a Question**-> 1. Statistical methods development  
-> 2. Danger zone!!!  
-> 3. Proper data analysis  
-always think of the science behind the data, not just data to statistics

**Example:**  
General Q: Can I automatically detect emails that are SPAM that are not?  
Concrete: Can I use quantitative characteristics of the emails to classify them as SPAM/HAM?

**2. Define the ideal data set**

-The data set may depend on your goal  
 - Descriptive – a whole population  
 - Exploratory – a random sample with many variables measured  
 - Inferential – the right population, randomly sampled  
 - Predictive – a training and test data set from the same population  
 - Causal – data from a randomized study  
 - Mechanistic – data about all components of a system

Ex: Google Data Centers – a population of emails

**3. Determine what data you can access**

-sometimes you can find data free on the web  
-other times you may need to buy the data  
-be sure to respect the terms of use  
-you might need to make your own data

Collection of SPAM

**4. Obtain the data**-new data  
-reference source  
-polite emails  
-record URL and time accessed if its from the internet

Kernlab -> spam

**5. Clean the data**  
-raw data often needs to be processed  
-if it is pre-processed, make sure you understand how  
-understand the source of the data  
-may need re-formating – record these steps  
-ARE THE DATA GOOD ENOUGH?

Check out Kernlab

**More steps in a data analysis**Exploratory data analysis  
Statistical prediction/modeling  
Interpret results  
Challenge results  
Synthesize/write up results  
Create reproducible code

**Subsampling our data set**  
-part of data set to build our model, part of data set to predict how good our data set actually is  
-use a coin flip to split the data in half

**Exploratory data analysis**  
summaries of data  
check for missing data  
create exploratory plots  
perform exploratory analyses

**Names(trainSpam)  
table(trainSpam$type)  
plot(trainSpam$capitalAve ~ tramSpam$type)   
plot(log10(trainSpam$capitalAve + 1) ~ trainSpam$type)** // shows the data to be less skewed  
**plot(log10(trainSpam[, 1:4] + 1))**   
Cluster: **hCluster = hclust(dis(t(trainSpam[, 1:57]))) or hClusterUpdated = hclust(dist(t(log10(trainSpam[, 1:55] + 1))))  
plot(hClusterUpdated)**

**Statistical prediction/modeling**  
-should be informed by the results of your exploratory analysis  
exact methods should depend on questions of interest  
-transformations should be done when necessary  
-record uncertainty

Ex: Logistic Regression

**Interpret Results**Uses the appropriate language  
 - describes  
 - correlates with/associated with  
 - leads to/causes  
 - predicts  
Give an explanation  
Interpret coefficients  
Interpret measures of uncertainty

Ex:  
Fraction of characters that are dollar signs can be used to predict if an email is spam  
6.6% dollar signs = spam  
Error rate was 22.4%

**Challenge results**CHALLENGE ALL STEPS:  
Questions, Data source, processing, analysis, conclusions  
-Challenge measures of uncertainty  
-Challenge choices of terms to include in models  
Think of potential alternative analyses

**Synthesize/write-up results**  
-lead with the question  
-summarize the analyses into the story  
-don’t include every analysis, include it  
 - if it is needed for the story  
 - if it is needed to address a challenge  
-order the analyses according to the story  
-include “pretty” figures that contribute to the story

Our example:  
-Lead with the question – the spam questions:  
Describe the approach  
-How to collect data, explore relationships  
-Choose logistic model by cross validation – 78% test set accuracy  
Interpret results  
-number of dollar signs seems reasonable  
Challenge results  
-78% is meh, why logistical regression etc

**Document all code and make sure it’s reproducible**

**Organize a Data Analysis**

Data  
- Raw  
- Processed

Figures  
- Exploratory figures  
- Final figures

R code  
- Raw / unused scripts  
- Final scripts  
- R Markdown files

Text  
- README files  
- Text of analysis / report

**Raw data**  
-should be stored in your analysis folder  
-would want to put it in a “nice” text editor or whatever it might be  
-if accessed from web, include URL, description, and date access in README  
-add raw data to repository if the data is not too huge

**Processed data**  
-should be named so its easy to see which script generated the data  
-processing script – processed data mapping should occur in the README  
-processed data should be tidy

**Exploratory figures**  
-figures made during the course of your analysis, not necessarily part of your final report  
-do NOT need to be pretty, need to be usable enough

**Final Figures**  
-usually a small subset of the original figures  
-axes/colors set to make the figure clear  
-possibly multiple panels

**Raw scripts**-may be less commented (but comments help you!)  
-may be multiple versions  
-may include analyses that are later discarded

**Final scripts**  
-clearly commented  
 -> small comments liberally  
 -> big blocks of comments  
-include processing details  
-only analyses that appear in the final write-up

**R markdown files**  
-used to generate reproducible report

R markdown can replace README, but should contain step-by-step instructions for analysis

**Text of the document**include a title, introduction (motivation) , methods (statistics used), results (including measures of uncertainty), and conclusions (including potential problems)  
should tell a story  
should not include every analysis performed  
references for statistical methods are important