**Coursera Data Science Course 5 Week 3: Reproducible Research**

**1. Communicating Results**

-Results of data analyses are presented in oral forms  
-Specificity etc

Hiearchy of Info  
-Title/ Author List  
-Abstract  
-Body / Results  
-Supplementary Materials / gory details  
-Code/Data/ really gory details

Email Presentation  
-Subject Line  
 -Minimum; include one  
 -Summarize findings in one sentence  
-Email body  
 -A brief description of the problem; recall what was proposed and executed; summarize findings/results; 1-2 paragraphs  
 -if action needs to be taken as a result of this presentation, suggest some options and make them as concrete as possible  
 -if questions need to be addressed, try to make them yes/no  
-Attachment(s)  
 -R Markdown file  
 -knitr report  
 -stay concise; don’t spit out pages of code  
-Links to Supplementary Materials  
 -Code, GitHub etc

**RPub**

-easy web publishing with R markdown

**2. Reproducible Research Checklist**

**Start with Good Science**  
-Coherent, focused question simplifies many problems  
-Working with good collaborators reinforces good practices  
-Something that’s interesting to you will motivate good habits

**Don’t: do things by hand:**Editing spreadsheets of data to “clean it up”  
-Removing outliers  
-QA/QC  
-Validating  
Editing tables or figures (e.g. rounding, formatting)  
Downloading data from a website (clicking links in a web browser)  
Moving data around your computer; splitting reformatting data file  
“We’re just going to do this once…” // it’s tempting to not have this automated because you’re doing it once

Things done by hand need to be precisely documented (so you can do it, but it’s also a lot harder than it sounds)

**DON’T: Point and Click**  
-Many data processing / statistical analysis packages have graphical user interfaces (GUI)  
-GUIs are convenient / intuitive but it’s difficult to reproduce  
-Some GUIs produce a log file or script that includes equivalent commands that are saved later  
-in general, be care with data analysis software that is highly interactive, but might not be reproducible then

**DO: Teach a Computer**  
-if something needs to be done as part of your analysis / investigation, try to teach your computer to do it (even if you need to do it once)  
-in order to give your computer instructions, you need to write down exactly what you nmean to do and how it should be done  
-teach a computer almost guarantees reproducibility

For example: by hand , you can go to a link and download it and save the zip file on your computer etc OR you can just use R:  
**download.file(URL, fileName)**everything is specified with one command and is executed by R! ☺

**DO: Use Some Version Control**  
-slow things down  
-add changes to small chunks  
-track / tag snapshots, revert to old versions  
-software like github etc make it easy to publish results  
// there’s a history of how you get to things if you use GitHub etc

**DO: Keep Track of your Software Environment**  
-if the project is complex, the software and computing environment can be critical  
**Computer Architecture:** CPU (Intel, AMD, ARM), GPUs  
**Operating System:** Windows, Mac OS, Linux / Unix  
**Software toolchain:** Compilers, interpreters, command shell, programming languages (C, Perl, Python, etc), database backends, data analysis software  
**Supporting software / infrastructure:** Libraries, R packages, dependencies  
**External dependencies:** Websites, data repos, remote databases, software repo  
**Version numbers**

**sessionInfo()** tells you everything about the packages platforms etc

**DON’T: Save Output**  
-Avoid saving data analysis output (table, figures, summaries etc), except perhaps temporarily for efficiency purposes  
-If a stray output file cannot be easily connected with the means by which it was created, then it is not reproducible  
-Save the data + code that generated the output, rather than the output itself  
-intermediate files are okay as long as there is clear documentation of how they were created

**DO: Set your seed**  
-random number generators generate pseudo-random numbers based on an initial seed (a number or set of numbers)  
**set.seed()** allows random numbers to be exactly reproducible  
remember to always set the seed

**DO:** **Think about the Entire Pipeline**  
-process is just as important as the results

**3. Evidence-based Data Analysis**

Recall Replication vs Reproducibility

Reproducibility gets us:  
-transparency  
-data availability  
-software / methods availability  
-improved transfer of knowledge

But we don’t know the validity / correctness of the analysis

Something can be reproducible and still wrong.

-Addresses the most “downstream” aspect of the research process – post-publication

Ex: Asthma – cause and effects, medications, but there is environmental intervention  
Scientific Dissemination – the pipeline – can have editors judgment etc, but it’s important to realize that there is something we can do at the most upstream part of the process.

**Who Reproduces Research?**For reproducibility to be effective as a means to check validity, someone needs to do something  
-Re-run analysis, check results match  
-Check the code for bugs/error  
-Try alternative approaches; check sensitivity  
-Need for someone to do something is inherited from traditional notion of replication  
-Who is “someone” and what are their goals?

Reproducibility does bring transparency  
But can we trust this analysis?  
Reproducibility are very downstream

**Evidence-based Data Analysis**-most data analyses involve stringing together many different tools and methods  
-some methods may be standard for a given field, but others are often applied ad hoc  
-we should apply thoroughly studied, mutually agreed upon methods

-Create analytic pipelines from evidence-based components  
-a Deterministic Statistical Machine  
-analysis with “transparent box”  
-analogous to a pre-specified clinical trial protocol

Ex: Air Pollution  
-we can encode everything we’ve found in a single package

**DSM Modules for Time Series Studies for Air Pollution**  
1. Check for outliers, high leverage, overdispersion  
2. Do NOT fill in missing data  
3. Model selection: Estimate degrees of freedom to adjust for unmeasured confounders  
 -Other aspects of model not as critical  
4. Multiple lag analysis  
5. Sensitivity analysis wrt  
 -Unmeasured confounder adjustments  
 -Influential points

Have multiple DSMs

**Cochrane Collaboration**  
-a curated library of data analysis

**Caching Computation**LaTeX/R -> Local/Remote -> PDF

Cacher package in R  
-evaluates code and stores results in data base  
-distribution purposes

Cacher used as an author  
-Parse the R source file; creates necessarily cache

Cacher as a reader

It’s basically kind of like a repo