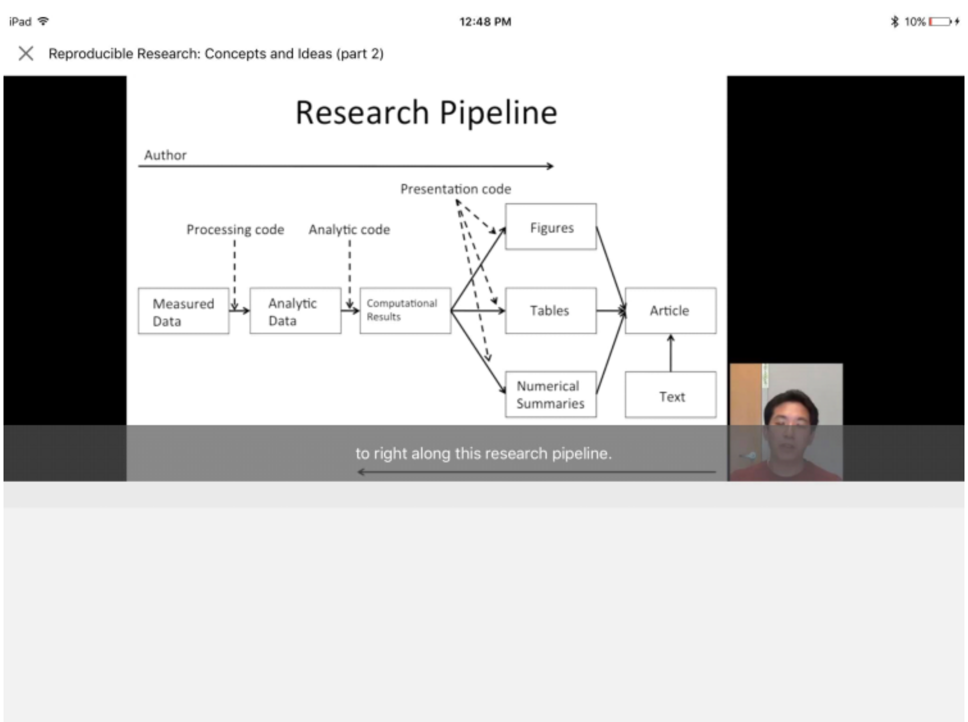
Concepts and Ideas

Thursday, August 11, 2016 12:35 PM

- Replication
  - Ultimate standard for strengthening scientific findings- constant replication of same results
  - Offers credibility
- Issues with replication
  - Difficult due to complexities of each study
  - Expensive timewise and moneywise
- Reproducible research
  - Middle ground: not full replication, but enough to be usable
- Data size and amount increased, thus we need to make sure we can replicate the collection/storing and analyses
- Research Pipeline:



- Recommendations for reproducibility:
  - Data/metadata available
  - Computer code available
  - Descriptions for steps of code
- What do we need
  - Analytic data avialbalbe
  - Analytic code available
  - Documentation of code
  - Standard means of distribution
- Reality
  - Authors just put stuff on web
  - Some supplementary materials via journal
  - Reader just try to piece it together
- Literate statistical programming
  - Article is a stream of text and code

- Documentation language (human readable)
- Programming language (machine readable)
- Put together to make an interpretable way to access data
- Different languages for this:
    - Knitr (uses markdown, R)
    - Sweave (LaTeX focused)

Scripting Analysis

Thursday, August 11, 2016 12:56 PM

Make everything via scripts, record everyhting on comp to make as reproducible as possible.

Structure of a Data Analysis

Thursday, August 11, 2016 1:00 PM

- Key challenge in data analysis
    - You will either have too much or too little data

1. Define a question
    a. Narrow down question, you know what data you need and what dimensions to reduce
    b. Choose data, pick specific stats method development and choice
    c. Start general, then make concrete
2. Define ideal data set
    a. Different types based on situation
    b. Descriptive (a whole population)
    c. Exploratory (random sample with many vars)
    d. Inferential (correct population but random samples)
    e. Predictive (training and test data set from same population)
    f. Causal (data from random study)
    g. Mechanistic (data of all components of study)
3. Determine data you can access
    a. Sometimes you can find free data on web
    b. Respect terms of use
4. Obtain the data
    a. Try to obtain raw data
    b. Reference the source
    c. Properly cite how getting data (maybe include time in case the data changes on public domain)
5. Clean data
    a. Raw data often needs to be processed
    b. If pre-processed understand how
    c. Understand source
6. Exploratory data analysis
    a. Look at summaries of data
    b. Check for missing data
    c. Create exploratory plots
    d. Perform exploratory analyses
7. Statistical prediction/modeling
    a. Informed by exploratory results
    b. Methods + data dependent on question
    c. **Uncertainty should be reported**
    d. Example steps:
        i. Create classifications of data
        ii. Get measure of uncertainty by calculating error rate of classification methods
8. Interpret results
    a. Objective, descriptive
        i. "predicts", "correlates with", "causes"
    b. Give potential explanations for correlations
    c. Acknowledge uncertainties
9. Challenge results
    a. Challenge all steps
        i. Question
        ii. Data source issues
        iii. Processing

        iv. Analysis
        v. Conclusions
    b. Challenge uncertainty
    c. Challenge model choices
    d. Think of potential alternative analyses/potentially utilize
    e. **You will be challenged inevitably, so recognize and challenge yourself first**
10. Synthesize/write-up result
    a. Lead with question
    b. Summarize analyses in story
    c. Don't include every analysis
        i. Only relevant for story
        ii. Or to address a foresought challenge
        iii. Still, remember all analyses that came into play to use later if potentially further challenged
    d. Strong figures to support
11. **CREATE REPRODUCIBLE CODE**

Organizing Data Analysis

Thursday, August 11, 2016 1:25 PM

- Data Analysis Files:
    - Data
        - Raw data
        - Processed data
    - Figures
        - Exploratory
        - Final
    - R code
        - Raw/unused scripts
        - Final scripts
        - R Markdwon files
    - Text
        - README files
        - Text of analysis/report
- Raw data
    - In analysis folder
    - log file describing source
- Processed Data
    - Document how raw -> processed
- Exploratory figures
    - Made during course of analysis
    - Don't need to be pretty
    - Not necessarily part of final report
- Final figures
    - Labeled/annotated well
    - Maximally utilize colors and legends
    - Only a subset of original plots
- Raw scripts
    - Less commented
    - May be multiple versions
    - Sometimes have discarded analyses
- Final scripts
    - Clearly commented
        - Small comments liberally
        - Bigger comments for whole sections
    - Only analyses in final write-up
- R-markdown files
    - Reproducible results
    - Text + code in one spot (look @ Literal Stastical Programming inConcepts and Ideas)
    - Sometimes alleviates point of README
- Text of doc
    - Title, intro, methods, results, and conclusions
    - **Tell a story**
    - Not all analyses, enough that are relevant to understand process

knitr

Thursday, August 11, 2016 4:51 PM

- Big tool for making Literal Statistical Programming
- knitr supports variety of documentation languages
- Creating reproducibility
  - Decide to do so early
  - Keep track of things
  - Don't save output
  - Don't use proprietary formats
- Literate Programming
  - Pros
    - Text + code in one place
    - Code is live when building doc
  - Cons
    - Text + code in one place… long and extensive (both of them)
    - Can slow processing of data
- knitr: R Markdown, LaTeX, HTML
- Good for:
  - Manuals
  - Technical docs
  - Tutorials
  - Reports
  - Data preprocessing docs/summaries
- Bad for:
  - Long research articles
  - Complex time-consuming computations
  - Docs that need precise formatting

Communicating Results

Friday, August 12, 2016 11:11 AM

- Often useful to break down results of analysis into different levels of granularity/detail
- Hierarchy of Information: Research Paper
  - Title/Author List
  - Abstract
  - Body/Results
  - Supplementary Materials/the gory details
  - Code/Data/really gory details
- Hierarchy of Email Presentation
  - Subject line/Sender info
  - Email body
  - Attachments
    - Reports
  - Links to supp materials
    - Github!
- Choose what levels of hierarchy based on audience and requirements

Reproducible Research Checklist

Friday, August 12, 2016 11:16 AM

- DO: Start With Good Science
  - Coherent, focused question simplifies many problems
  - Working with good collaborators and reinforces good practices
  - Be interested in it
- DON'T: DO Things by hand
  - Editing spreadsheets and data
    - Removing outliers
    - Validating certain measurements
  - Downloading data via links **(oh shit we've been doing this)**
  - **NO "We're just doing this once"**
- DON'T: Use GUIs and "Point & Click" Softwares
  - Very difficult to reproduce process later
- DO: Teach a Computer
  - Ie make your computer able to repeatedly do it via scripts, programs, etc.
  - Makes it repeatable
  - Easily modifiable
  - Modularized specific steps
- DO: Version code
  - Slow things down
  - See process and thought process of deicisions
- DO: Keep Track of Software Environment:
  - Computer Architecture
  - OS
  - Software toolchain
    - Compilers
    - Interpreters
    - Command shell
    - Language
  - Supporting software (libraries, packages, etc.)
  - External dependencies
  - Version numbers
- DON'T: Save Output
  - Avoid saving data analysis output
  - Just save and know the process (ie original data and processing code)
- DO: Set your seed
  - Make sure you can exactly reproduce the results you got
- DO: Think About the Entire Pipeline
  - Data analysis is a lengthy process
  - Raw -> processed -> analysis -> report

Evidence-based Data Analysis

Saturday, August 13, 2016 1:26 PM

- Replication and Reproducibility
    - Replication
        - Standard for proving scientific claims
    - Reproducibility
        - Increases validity, ability to replicate, etc.
- Background and Underlying Trends
    - Databases can be merged
    - Some studies cannot be replicated due to resources
    - Computational application exists all fields
- Result
    - Difficult to replicate
- Reproducibility
    - Transparency
    - Data Availability
    - Improved Transfer of Knowledge
    - Don't get
        - Validity (NOT A GIVEN)
    - Problems:
        - Addresses
- Who reproduces research?
    - Scientists, people trying to prove you wrong
    - Not reproducible can make proving you're right very difficult
- Evidence-based Data Analysis
    - Most data analyses involve stringing together different tools and methods
    - Aim to have a mostly standardized set of data analyses, specified for different situations
    - Create analytic pipelines from evidence-based components
        - Analysis with a "transparent box"
        - Reduce "researcher degrees of freedom"
    - Write as a deterministic state machine- distinct states that are moved through
- Case Study: Estimating Acute Effects of Air Pollution
    - DSM Modules for Time Series Studies of Air Pollution and Health
        - Check for outliers, overdispersion, etc
        - Fill in missing data? NO (Doesn't usually work out well!)
        - Model selection: estimate degrees of freedom to adjust for unmeasured confounders

Caching Computations

Saturday, August 13, 2016 1:58 PM

- cacher package
  - Evaluates code
  - Stores results in a key-value database
- Using cacher as an Author
  - Parse R file
  - Cycle through expressions
  - If cache result exists for specific parts retrieve results, otherwise if new run computation
  - Knowing the id for results at specific points can help when authoring a paper, provide the id for different points