ATMOSPHERIC SCIENCES 6910

Programming for Environmental Sciences

Syllabus

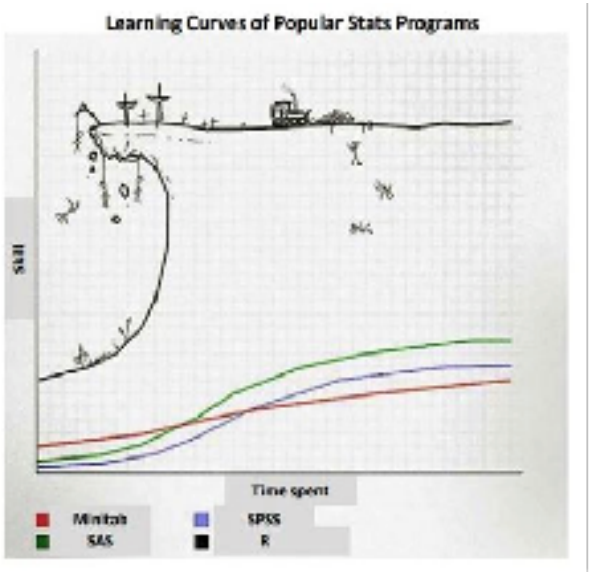
- MWF 9:40 – 10:30 in WBB 711
- Instructors:
  - Sally Benson
  - Chris Galli

- First 10 minuets of class reviewing readings and homework is applicable, followed by lecture.
- Homework will be available on class website by the start of class.
- To turn in homework email your answer / program to the instructor who assigned it.
- Class HTML page at:

  - https://github.com/cgalli/ATMOS_6910_2018

Languages

- Environmental sciences is a big field encompassing many different disciplines, employing the use of an array of different analytical practices.
- Languages: R, Fortran, Python, MatLab, Idl, C, ect…
- Data Structures: .csv, Netcdf, HDF5
- Analytical Methods: Statistical, visual, spatial, time-series
- This class will be language agnostic or neutral. We are assuming some basic understanding of coding.



Languages

- Each research group will employ different data structures, methods and languages to examine their particular field.
- Each language has its benefits and limitations. Some languages have strong communities built around them (R, Python). Open source languages tend to have better communities which can be a real advantage.
- Stack Overflow!
- Our Goal: to provide you with a background in computer science / programming to help you answer your research objectives in the most powerful and time efficient ways while avoiding common pitfalls.

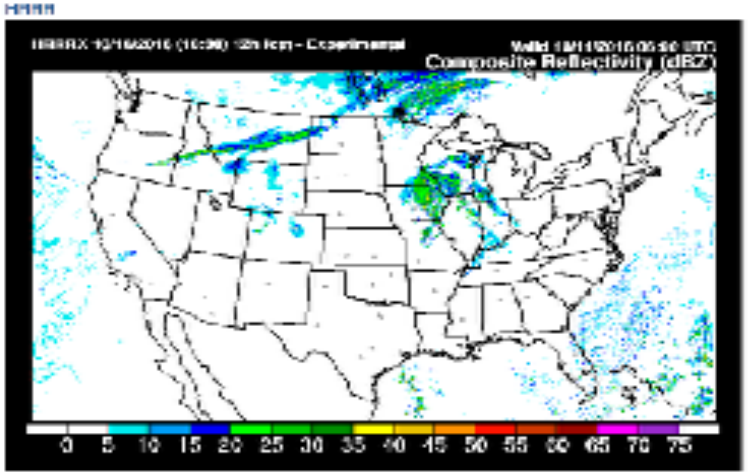
Workflow

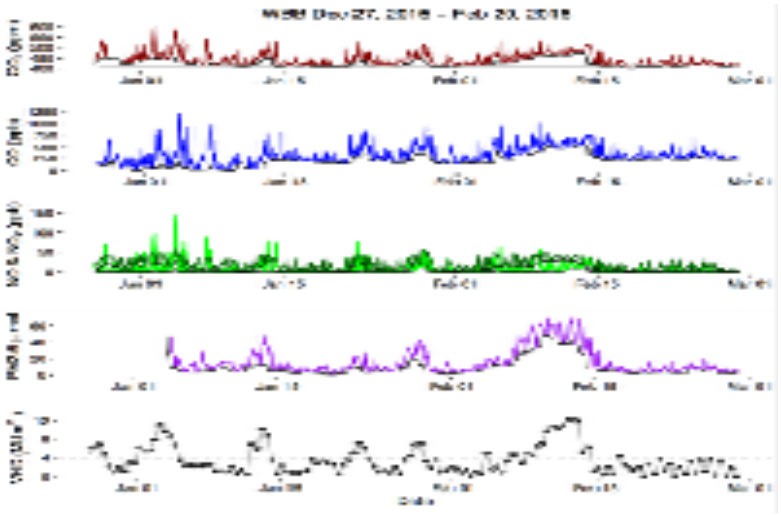
Regardless of which toolsets used, the general flow of each of your projects will follow a similar workflow.

1) Get data
2) Bring data and packages into environment
3) Frist look
4) QA/QC
5) Pair additional data sets
6) Data transformations
7) Sub-setting / isolating data
8) Analysis
9) Plotting / data visualization
10) Final results

Workflow: Data

- Step 1: Get your data
  - Data API's
  - Measurements
  - Laboratory Experiments
- Understand your data
  - Products vs. Measurements
  - Expectations (freq., sensitivity)
  - Data Quality
  - Accuracy and Precision
  - Data Type
- Where to store the data?
  - CHPC, personal computer, other server





Workflow: Data In

- Step 2: Get the data into an environment

```
11
12  ### Load in the data
13
14  setwd("~/UQA Winter:ine PM2.5 StudyVData/WBB")
15  dat = read.csv("WBB_Uni_Aggregated_160930.csv", stringsAsFactors = FALSE)
16  dat$Time <- as.POSIXct(dat$Time, tz='MST')
17  dat.inv = subset(dat, dat$WHO >= 4)
18  dat.clean = subset(dat, dat$WHO < 4)
19
20  setwd("~/UQA Winter:ine PM2.5 StudyVData/Backgrounds")
21  background = read.csv("background_160930.csv")
22  background$Time = as.POSIXct(background$Time, tz = 'MST')
23
24
25  setwd("~/UQA Winter:ine PM2.5 StudyVData")
26  sites = read.csv("Site_locations.csv", header = TRUE)
27  sites$Site.ID = as.character(sites$Site.ID)
28
29  setwd("~/UQA Winter:ine PM2.5 StudyVData/hawthorne")
30  haw = read.csv("haw_161003.csv", stringsAsFactors = FALSE)
31  haw$Time <- as.POSIXct(haw$Time, tz='MST')
32
33  setwd("~/UQA Winter:ine PM2.5 StudyVData/Mobile")
34  mobile = read.csv("mobile subset agg 161003.csv")
35
36
```

- Load necessary packages

```
1  library(ggmap)
2  library(ggplot2)
3  library(leaflet)
4  library(sp)
5  library(maptools)
6  library(rgdal)
7  library(maps)
8  require("rgdal") # requires sp, will use proj.4 if installed
9  require("maptools")
10 require("ggplot2")
11 require("plyr")
12
```

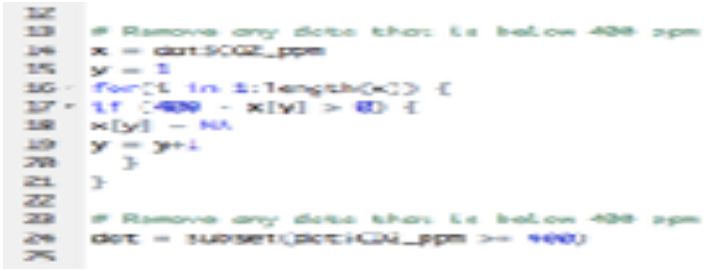
Workflow: First Look

- Step 3: First look
- Headers
- Data structure
- Variables
- Data types
- Data frequency

```
> str(mobile)
'data.frame':    89617 obs. of  13 variables:
 $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Time_UTC      : Factor w/ 72095 levels "2016-01-25 22:07:38",..: 1286 1287 1288 1289 1290 1
 $ lat           : num  40.7 40.7 40.7 40.7 40.7 ...
 $ lon           : num  -112 -112 -112 -112 -112 ...
 $ CO2d_ppm      : int  467 467 464 462 458 453 448 446 443 441 ...
 $ CH4d_ppm      : num  1.95 1.95 1.96 1.95 1.95 ...
 $ CO_ppm        : num  0.61 0.61 0.61 0.61 0.738 ...
 $ pm2.5         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ NOx           : num  NA NA NA NA NA 43 NA NA NA NA ...
 $ O3_ppbv       : num  NA 24.3 26.7 NA NA 28.5 28.9 NA NA 27.2 ...
 $ CO_ppb        : num  610 612 610 610 738 ...
 $ CO_CO2_slope: num  35.8 35.8 35.8 35.8 35.8 35.8 35.8 35.8 35.8 35.8 ...
 $ ID            : chr  "I15" "I15" "I15" "I15" ...
>
```

Workflow: QA/QC

- Step 4: QA/QC the data
- Visual / Interactive
  - Plot the data and visually identify problematic areas
  - Limited to smaller datasets and can be subjective
  - http://www.inscc.utah.edu/~u0079358/atmos6910/Class_1/qaqc.html
- Statistical / Mathematical
  - STDEV, > or <, y-x, smoothing, quintiles
  - Great for large data sets and automated scripts but can miss complex problems and can remove valid data
- Pre-define data levels
  - Raw, qa/qc, final

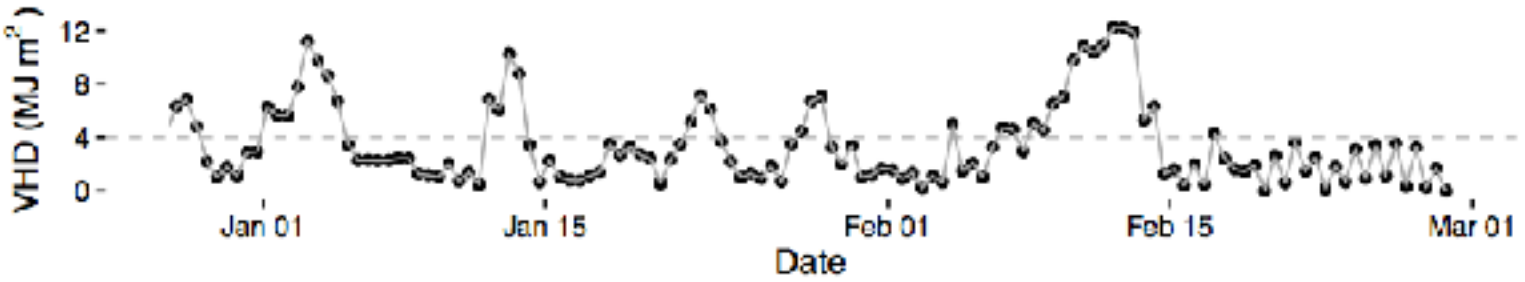


Workflow: Pairing Additional Data

- Step 5: Pair additional datasets
- Repeat steps 1:4 and make sure to QC/QC new data before pairing!!!
- Matching variable
  - Common time stamps / gridded location
- Matching data frequencies
- Double check that the matching variables line up….
  - Pay attention to time zones

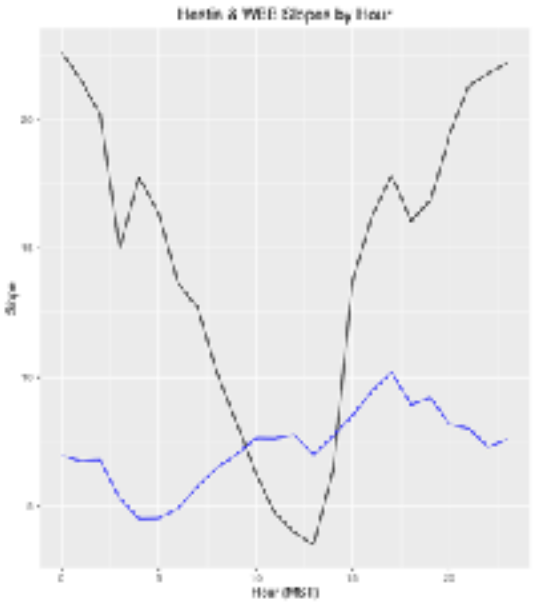
Workflow: Data Transformations

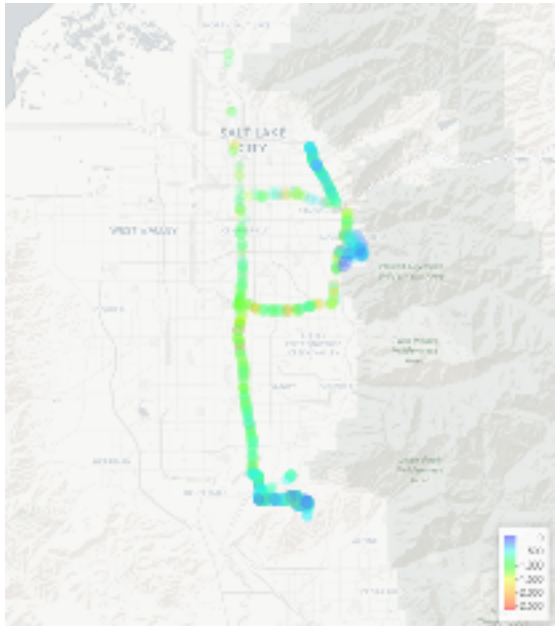
- Step 6: Data Transformations
  - Unit Conversions
  - Data Type Changes
    - as.numeric(), as.POSIXct(), as.boolean()
  - Calculations
  - Interpolations
  - Averages

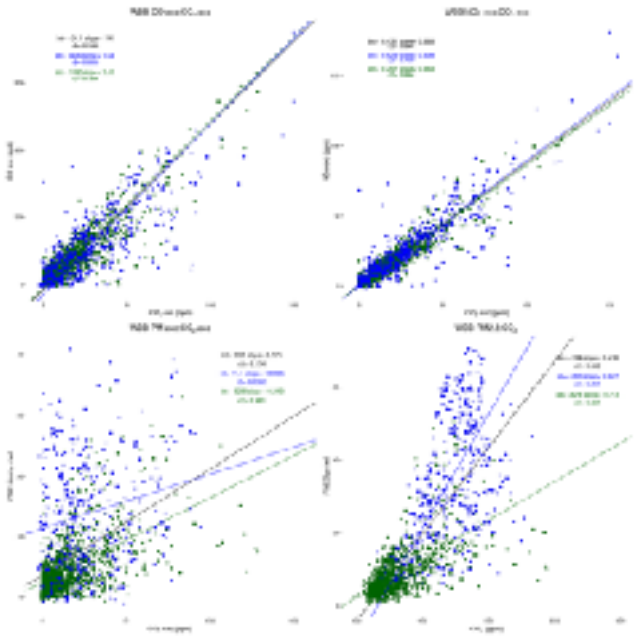


Workflow: Analysis

- Step 7: Analysis
- Statistical
- Spatial
- Validation

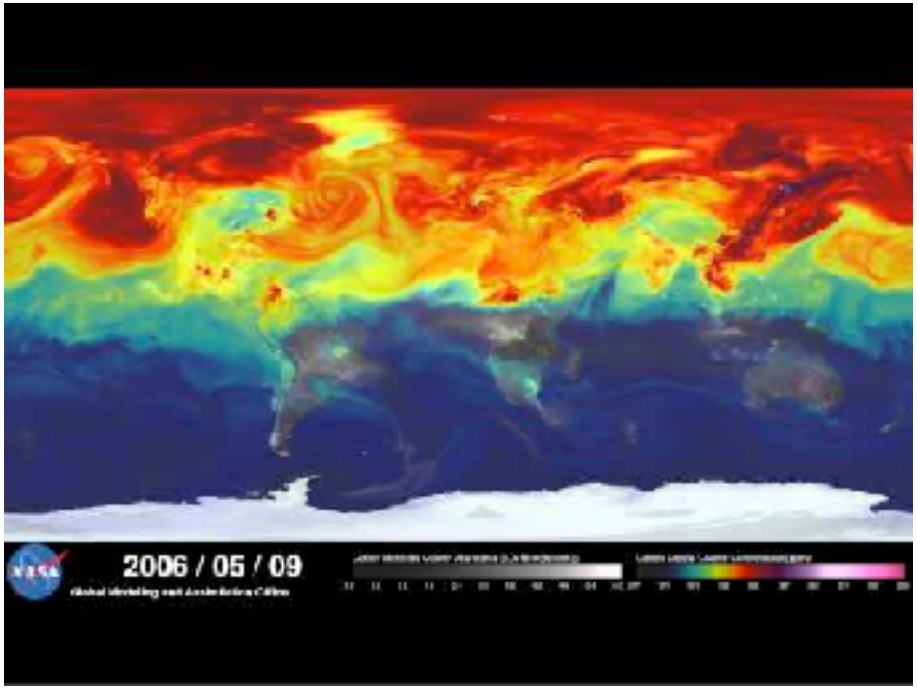


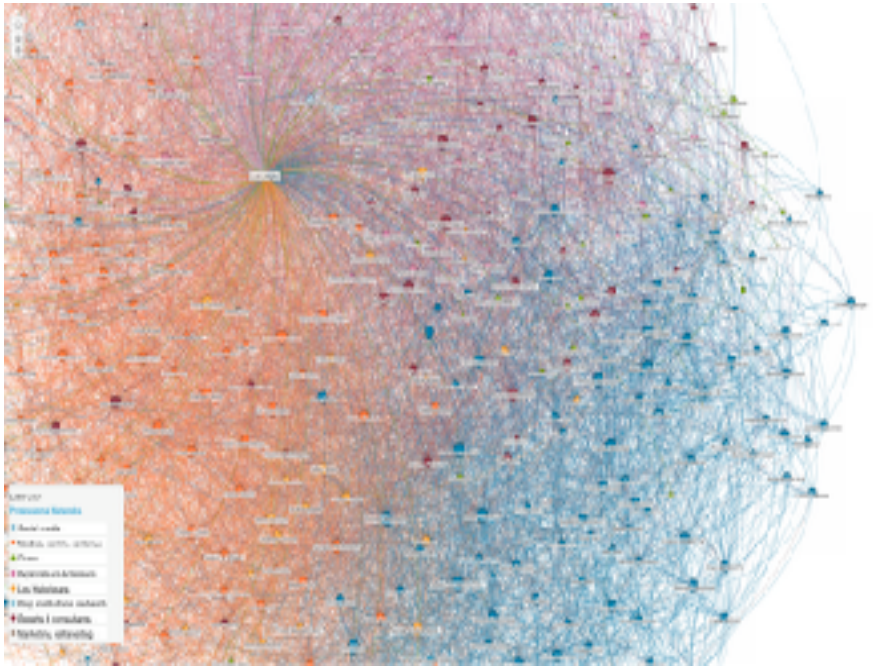




Workflow: Visualization

- Step 8: Data Visualization / Plotting





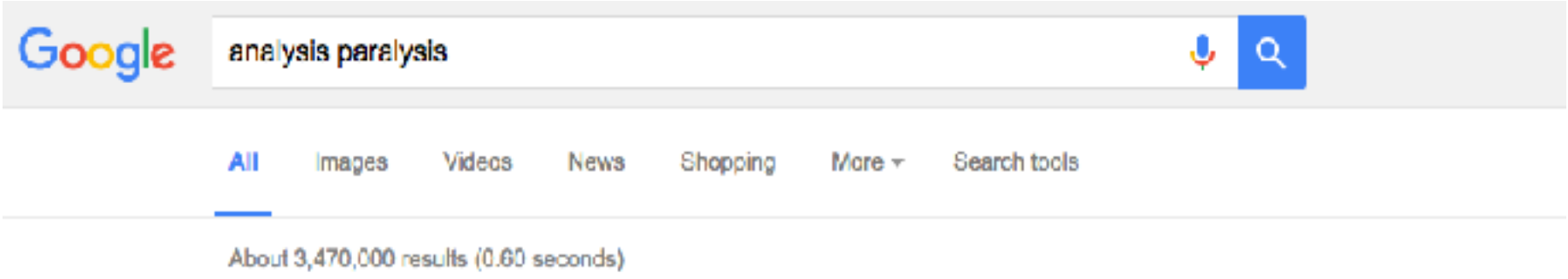
https://www.youtube.com/watch?v=x1SgmFa0r04

Workflow: Final Results

- Step 9: Final Results
- Science is a never ending redefining of the knowledge we currently have.
- A good analysis will identify more questions than it will answer.
- Repeat steps 1 – 8…
- Publish or perish!

Analysis Paralysis

- The prospect of taking on a multi-year research project can be overwhelming at times. There are so many decisions to be made and obstacles to overcome.
- It is easy to fall in to analysis paralysis, or a “state of over-analyzing (over-thinking) a situation so that a decision or action is never taken, in effect paralyzing the outcome”.



Overcoming Analysis Paralysis

- Differentiate between big and small decisions
  - Small
  - Mid-level
  - Big
- Identify your objectives. And remember to re-analyze your objectives as new knowledge is gained
- Perfection is not the goal.
  - “Perfection is the enemy of good enough” – Jim Ehleringer
- Eliminate the bad options
- Pick one and go. Don’t second guess
- Set a hard time limit! Deadlines are the key to preventing procrastination

Overcoming Analysis Paralysis

- Delegate the decision to someone else
  - Your advisors are there for a reasons… to advise you
- Get the opinion of someone you trust
  - Other gradstudents have been there before. Ask them!
- Make to do lists.
  - There is something incredibly powerful about writing down an objective then crossing it off. Take your larger objective and break it down into smaller, yet actionable events.
- Don't go down the rabbit hole. "Curb your curiosity".
- The moons will never be perfectly aligned.