Lecture 13

20 February 2017

Gaurav Kandlikar

Preliminaries

- Create a **directory** named classwork-21-Feb in your class work directory and commit this to your git repo.
- Within this directory, create a jupyter notebook for today's work
- Hacky hours tonight (Tuesday), **5.30-7.30 pm**
- Part 1 of HW due this morning; second part due Wed.
- Update on grades
- Find a partner for today's exercise

Course overview

Overall workflow of scientific computing:

- collect or retrieve data ;
- use a combination of tools (UNIX commands, python) to clean and reformat data
- use a combination of tools (python R) to summarize, analyse, model, and visualize the data
- use scientific typesetting tools (LaTeX, markdown) to write up the results for submission
- all the while, use version control (git) to manage the project

Review of Cooney et al and eBird datasets

- Downloaded csv files using `wget`
- Looked at what each row of data in each file contained
- In Cooney et al. ("nature") file, saw that first few rows were meta-data, so used `tail` to subset the file to exclude rows 1-3
- In eBird data, saw that there were commas included within some of the "columns"; used `sed` to replace certain commas with spaces

Review of Cooney et al and eBird datasets

- Each row in Coony et al. file represented a bird species and some measurements of its beak morphology
- Each row in the eBird dataset represented a single bird species and its taxonomy.
- Both files had idiosyncratic formats which made processing tricky- *this is super common!*

Guided in-class exercise

Today we will download data, understand how it is structured, clean it, and ask interesting biological questions.

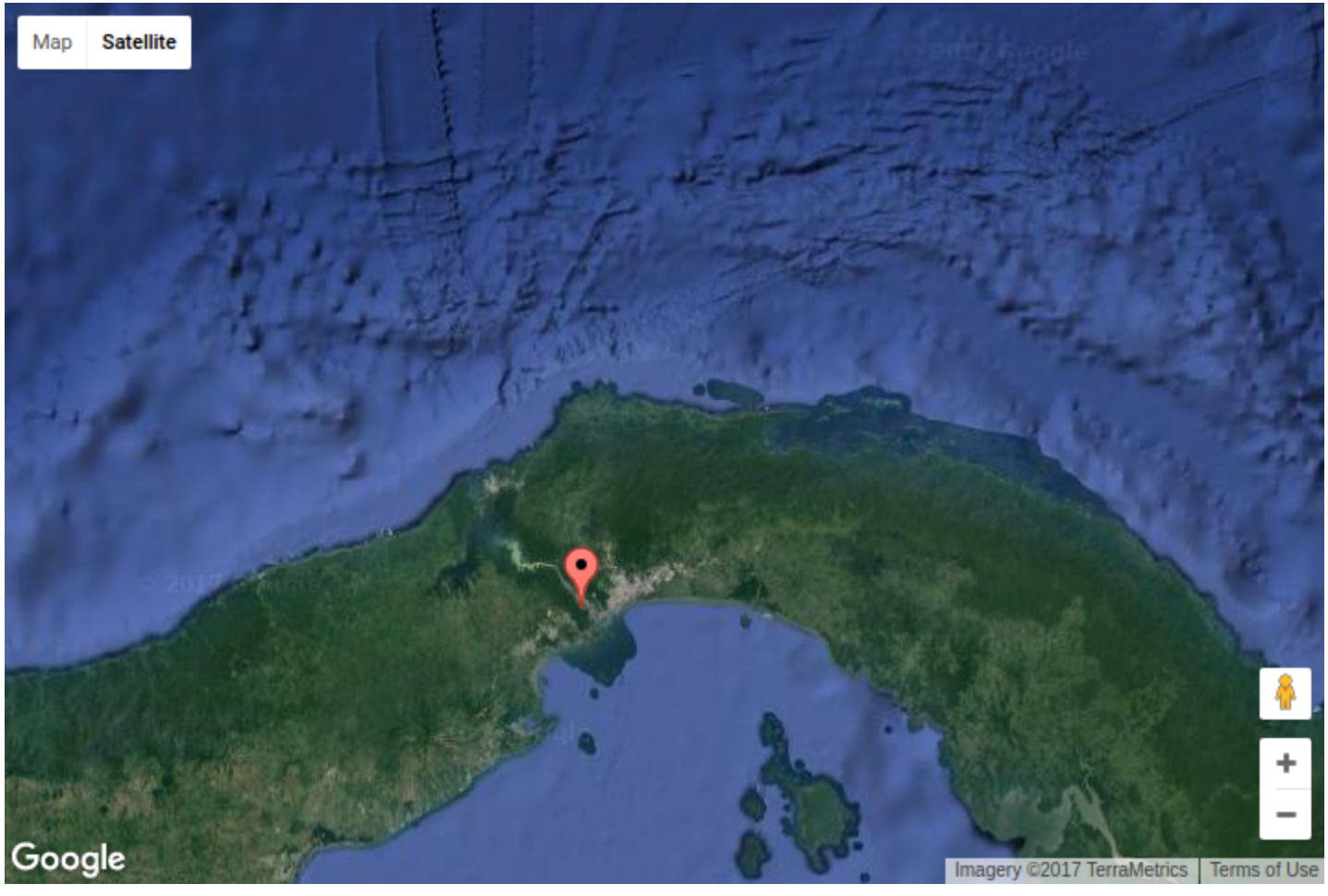
We will work with data from the Center for Tropical Forest Science from a Tropical Dry Forest in Panama.

Intro to CTFS datasets

Scientists from the Smithsonian have established plots in forests in which all individual stems above a certain size threshold are mapped, identified, and measured.
A few years later, scientists go back to the plot and re-measure each individual (or record it as being dead).

Note- this effort was initially spearheaded by Robin Foster of the Field Museum in Chicago and Steve Hubbell, now in our EEB department! Now, there are >60 such plots worldwide.

Cocoli forest



Cocoli data exercise

The data is available in a zip file: http://ctfs.si.edu/webatlas/datasets/cocoli/cocoli.zip.

- Downlad this file into your `classwork-21-Feb` directory directly from your terminal. (Record the command you used to complete this, and all other commands for today, in your python notebook).
- Unzip the downloaded file using the `unzip` command

Cocoli data exercise

- For the next few minutes, explore all of the files you have downloaded. In your python notebook, create a markdown block that describes what kind of information is present in each file in the unzipped folder.

Guided exploration of the data

Answer the following in a md block in the python notebook

- What does each row in `cocoli.txt` represent?
- Consider the following record from `cocoli.txt`:

```
tag      spcode  x       y       dbh1    dbh2    dbh3    re
000011   CAL2CA  4.0     13.2    69      69      65      A
```

- What does the entry in the column `multi2` refer to?
- What does the entry in the column `dbh2` refer to?
- How can we find the genus and species name of this record?
- What is the genus and species name of this record?

Guided exploration of the data

Consider the following records from `cocoli.txt`:

tag	spcode	x	y	dbh1	dbh2	dbh3	re
000001	PROTTE	3.0	0.9	171	267	277	A
000029	MICOAR	9.0	5.1	15	-1	-1	A

- Biologically, what happened to individual `000001` between 1994 and 1998?
- What about individual `000029`?

Guided exploration of the data

Given your knowledge of the dataset, brainstorm some interesting questions that you might be able to answer with these data.

Guided exploration of the data

- Which genus is the most species-rich in this community?
- hint: the following command may come in handy to reformat a file:
  `sed 's/\s/,/g'` (replace all whitespace with a single comma)
- How many genera have just one species represented in this community?
- This community has only one species in the family Bignoniaceae.
- i. What is the name of the species?
- ii. How many individuals of the species are in the community?

Tabebuia guayacan; image from SI

Guided exploration of the data

- One important way to measure tree performance is the **relative growth rate** (RGR), which is estimated as

$$RGR = \frac{dbh_{t+1} - dbh_t}{dbh_t}$$

- Use python to calculate the **yearly** RGR for each individual between 1994-1997; 1997-1998; 1994-1998

Guided exploration of the data

- Which individual had the highest yearly RGR between 1994-1998?
- *Note*: Remember that in python, dictionaries are **unordered collections**. How can you sort this?
- To which species does this individual belong?

Guided exploration of the data

- Write a Python function that will calculate the RGR from any pair of dictionaries (i.e. the function should take as input two dictionaries of sizes and the interval between the two measurements, and return a dictionary of the RGR between the two years)