Workshop proposal: Data management, visualization and analysis for entomological surveillance, research, and action

COURSE OBJECTIVE: Empower participants to fully engage with their data by learning the steps of the data lifecycle: data collection, storage, processing, visualization, analysis, and communication. By the end of the course, participants will feel comfortable using the R statistical environment to carry out research and epidemiological/entomological surveillance related tasks such as data cleaning, merging, mapping, and basic statistical hypothesis testing.

PREREQUISITES: Students need a laptop computer, basic proficiency in English, and a desire to learn. That's all.

METHODS: The course will be active and participatory. It will be challenging, but fun. Using real demographic, meteorological, geospatial, and epidemiological data, each participant will progressively build two "capstone" projects: an interactive data dashboard and a formal analysis write-up. The course can be one or two weeks in duration.

COST AND OPTIONS: Cost is a function of the options chosen:

- 1 or 2 two teachers
- 1 or 2 weeks
- Additional pre/post-course virtual modules (during the period 4 weeks before and after the course)

We recommend two teachers for a full two-week course, with the virtual pre/post-course modules. Cost can be reduced by doing only one week ("basic package") or by having only one teacher. But our experience suggests that students do best when two teachers are available (one for instruction and one for simultaneous hands-on help) and when they have enough time to dig deep into the lessons (ie two weeks). Even if cost-constrained, we strongly recommend the pre/post-course virtual modules, since the time in-person will be of much greater value if students have the guidance and structure to begin learning prior to the (in-person) course and continue learning after completion.

Course Outline

Week 1: basic package

The one week "basic package" can be a stand-alone course, or the first half of the "advanced package" (2 week course).

- Day 1
    - Introduction to R and RStudio
    - Basic calculations and object oriented programming
    - Best practices for workspace and file management
    - Creating and managing a cloud-based survey for data collection
- Day 2
    - Reading in data (flat files)
    - Introduction to libraries
    - R syntax, data structures, and variable types
    - Exploratory analysis through visualization

- Day 3
  - Interactive visualization
  - Maps and GIS
  - Introduction to "reproducible research" and Rmarkdown
  - Dashboard basics
- Day 4
  - Deep-dive into dashboards and applications in surveillance
  - Widgets, pop-ups, and user experience
  - Automating analysis "pipelines"
  - Principles of documentation
- Day 5
  - Basic hypothesis testing through statistics
  - Bibliographic management in the R environment
  - LaTeX, Rmarkdown, html generation, and other formats for data communication
  - How to help yourself - resources for further learning

Week 2: advanced package only

The two week "advanced" package starts with the one week "basic package" (above) before digging deeper into matters of interactivity, database access, collaboration and version control, advanced data "munging" (cleaning, re-shaping), using web APIs, and deploying knowledge products (dashboards and reports) to the web.

- Day 1
  - The shiny web framework: bringing data visualization to life
  - Accessing database-stored data through R: dplyr, DBI, RPostgreSQL, and RMySQL
- Day 2
  - Deep-dive into dplyr: the verbs of data processing
  - Advanced Rmarkdown: templates, formats, runtimes, etc.
- Day 3
  - Getting data from the web: APIs and web-scraping
  - Git and version control
  - Deeper-dive into interactive visualizations: leaflet, rCharts, highcharter, networkD3
- Day 4
  - "Literate programming" and "reproducible research": applying the principles to academia and public health
  - Deploying projects online
- Day 5
  - Writing functions and documentation.
  - Building an R package
  - Deploying to the web

INSTRUCTORS:

Joe Brew | Data scientist | joe@databrew.cc | databrew.cc/joe.pdf

Ben Brew | Statistician, Economist | ben@databrew.cc | databrew.cc/ben.pdf

TENTATIVE DATES:

Week 1: April 2-6, 2018

(Week 2: April 9-13, 2018)