When a Plant Disease Epidemiologist works on the data science

Sith Jaisong
Plant Disease Management Group, CESD, IRRI
Los Baños, Philippines
s.jaisong@irri.org

The data generally are raw and needed to pass processes in order to extract meaning from data or solve the problems. There are many similarities in data processes between plant disease epidemiology and data science. Epidemiologists often deal with more than one factors possibly causing the plant disease. Humidity, temperature, soil pH, soil type, plant varieties, crop density are some examples of variables considering the causes of plant disease development. When the number of variables were added in the study, the dataset will become massive, and require a powerful tool to manage. Data science is emerging to meet the challenge of processing very large data. It applies the techniques from statistics and computer science extracting the meaning from the data and product the data product (graphs or models, etc.) in order to answer the questions. The process [2] starts with collecting the data that are able to answer the questions or hypothesis. Joining, combining, or wrangling the data are the processes after getting them. In reality data we have may be needed to clean (outlier and missing value removal). Once we have this clean dataset, exploratory data analysis may be applied. Next, Designing the model by using some algorithm k-nearest neighbor (k-NN), linear regression etc. The model selected depend on the type of questions you try to answer. Then, we can interpret, visualize, report or communicate the result. The process of data science is in common with plant disease epidemiology. Epidemiologists often deal with more than one variable possibly causing the plant disease. Plant diseases develop from interaction between plant and its pathogen but also environment (weather) and human activities referring to the agricultural practices. The models are effective features capturing these relationships. They are developed for many uses or objectives. The most common ones are description, understating, prediction, comparison, and communication [1].The ideas and the process of tackling the problem are in the same concept. Therefore, data science process is the promising framework that epidemiologist should have in mind.

References

[1] Laurence V Madden, Gareth Hughes, and Frank Van den Bosch. *The study of plant disease epidemics*. American Phytopathological Society St Paul, MN, 2007.

[2] Rachel Schutt and Cathy O'Neil. *Doing Data Science: Straight Talk from the Frontline*. " O'Reilly Media, Inc.", 2013.

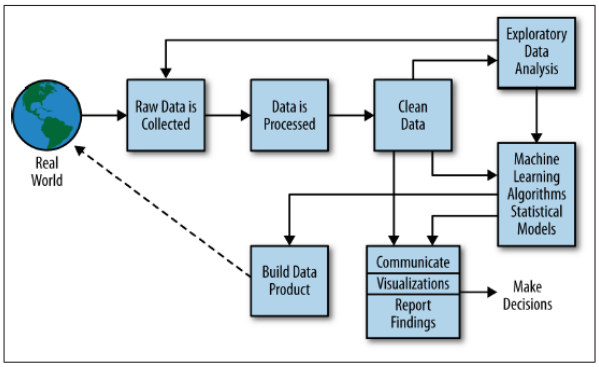

Figure 1: Data science process borrow from [2]