Scientific Data Wrangling

Colin Robertson

2021-11-16

Contents

Chapter 1

prerequisites

Moving from data acquired by a sensor or in the field to a model or visualization that can provide insights to a question often requires an extensive amount of work. It is estimated that 'data wrangling' - cleaning, loading, processing, integrating their data comprises at least half of a data scientist's time, and that may be even higher in the context of environemental data science.

1.1 textbook

Wickham H, Grolemund G. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media. Chicago, available online at http://r4ds.had.co.nz/

1.2 additional readings

Broman KW, Woo KH. 2018. Data Organization in Spreadsheets. The American Statistician 72(1): 2-10. https://doi.org/10.1080/00031305.2017.1375989

Bryan J, et al. 2018. Happy Git and GitHub for the useR. http://happygitwithr.com/

Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, Jones MB, Lenhardt WC, MacDonald A, Michener WK, Mudge J, Pourmokhtarian A, Schildhauer MP, Woo KH, Zimmerman N. 2015. The Tao of open science for ecology. Ecosphere 6(7):120. http://dx.doi.org/10.1890/ES14-00402.1

Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, et al. 2016. Ten Simple Rules for Digital Data Storage. PLoS Comput Biol12(10): e1005097. https://doi.org/10.1371/journal.pcbi.1005097