Data Science | Lancaster University

Navigating open-source software: Can you rely on what you use?

Rebecca Killick(r.killick@lancs.ac.uk)
NHS R-conf, Nov 20

Key Principles of Science

Data Science | Lancaster University

- Reproducible
- Replicable
- Trustworthy

> *Among the very basic principles that guide scientists, as well as many other scholars, are those expressed as respect for the integrity of knowledge, collegiality, honesty, objectivity, and openness.*

Responsible Science: Ensuring the Integrity of the Research Process: Volume I. 1992

Trustworthy

- 16,549 packages available on CRAN
- <280,000 repositories using R on GitHub (most aren't packages though)

How do you find the "right" one(s) to use, the ones that you can trust?

> *I am a beginner R user. Since I've started to learn, I wonder who is it proved whether these packages work right? Who is checking for functions, particularly complicated ones whether they are written right? What is the best way of understanding a package is good or bad?*

Bahar Patlar's comment on Revolution Analytics 10,000 CRAN packages blog post.

Sources of Trust

Data Science | Lancaster University

- Recommendations from people you trust
- Reputation of authors
- Download / citation statistics
- Dependencies
- History of development
- Associated publications
- Listed on a CRAN Task View
- . . .

Formalizing Trust

Data Science | Lancaster University

- Validate CRAN packages
  - Unrealistic too large / expensive (money and time)
  - Can pay a company to validate for mission-critical packages
- Compare two independently developed packages in simple cases
- Create benchmarks for statistical software development
  - Not all “good” packages will adhere
  - What should these be?
  - How can these be both generic and specific?
- Create a risk strategy for R packages
  - What should the risk profile look like?
  - What attributes are less/more risky?

rOpenSci

rOpenSci fosters a culture that values open and reproducible research using shared data and reusable software

- Package review - focused on “helper” packages
- Community calls
- Events
- Forum
- Statistical Software Peer Review - new project

Statistical Software Peer Review

Sponsored by 

- 6 board members (including me) and around 1.5 full-time staff
- Goal: develop a set of agreed-upon standards for statistical package implementation and testing
- Launch a new peer-review process

Statistical Software Peer Review

It is really difficult!

- Statistics is diverse . . .
- . . . topic categories are diverse
- Thinking about general standards is tough

Categories considered

Data Science | Lancaster University

- Bayesian & Monte Carlo
- Dimensionality & Feature Reduction
- Machine Learning
- Regression, Splines, & Interpolation
- Statistical Indices and Scores
- Visualisation
- Exploratory Data Analysis (EDA)
- Probability Distributions
- Wrapper Packages
- Categorical Variables
- Networks
- Survival Analysis
- Workflow Software
- Summary Statistics
- Spatial Analysis
- Educational Software

So far . . .

Data Science | Lancaster University

- Currently looking at design and testing NOT correctness
- Have draft standards for 6 areas so far
  - Bayesian and Monte Carlo
  - Regression and supervised learning
  - Dimensionality Reduction, Clustering, and Unsupervised Learning
  - Exploratory Data Analysis
  - Time Series Software
  - Machine Learning Software
- First packages are undergoing review
- Seeking community review of the draft standards and welcome suggestions:

https://ropenscilabs.github.io/statistical-software-review-book/standards.html

R Validation Hub

Data Science | Lancaster University

R Validation Hub is a cross-industry initiative whose mission is to enable the use of R by the Bio-Pharmaceutical Industry in a regulatory setting, where the output may be used in submissions to regulatory agencies.

- Advocating risk assessment of packages
- Developing `riskmetric` R package
- Opportunity to contribute metrics as issues

https://github.com/pharmaR/riskmetric

Conclusion

- Selecting packages to use is often based on trust
- Trust in packages is hard to assess
- rOpenSci and R Validation Hub are trying to provide tools for users and developers

With prominent news stories about nationally important inference based on research code it is imperative that we address open-source software standards immediately.