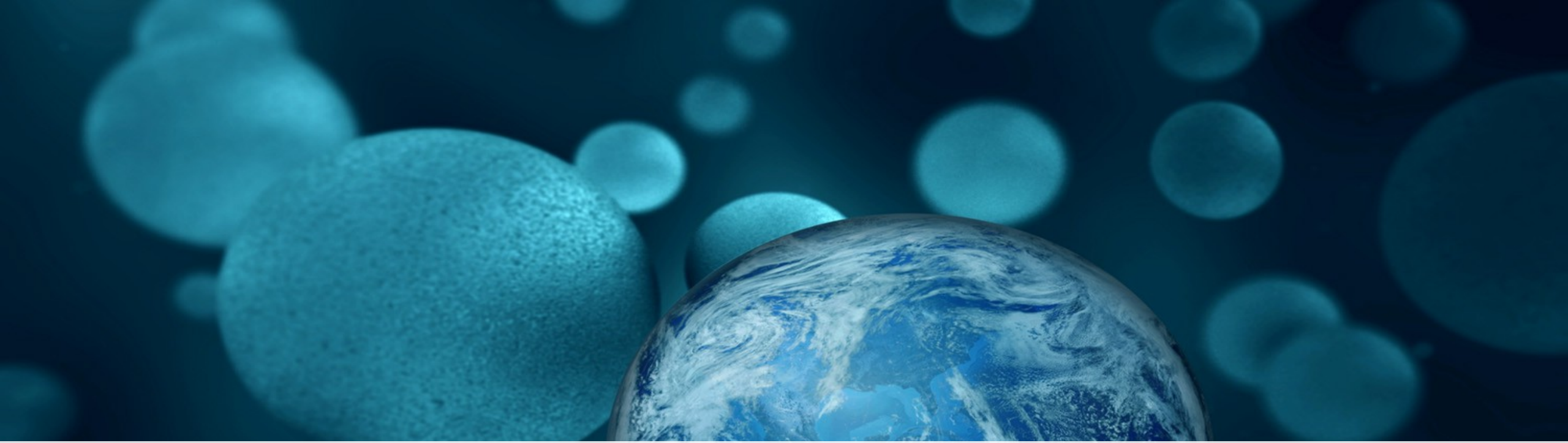

ThermoFisher SCIENTIFIC

Creating R Packages at For-Profit Companies

Jae Brodsky
06-03-2017

SNPolisher: R Package from Thermo Fisher Scientific (Affymetrix)

SNPolisher is an R package with 11 functions for use with analyzing post-genotyped data

- Human and ag/bio genetics
- QC data checks
- Visualization
- Algorithms and Data Analysis group uses SNPolisher as an R&D playground that can be quickly updated and sent to customers
  - Updating the full software products can take months longer

Two major differences with R packages at for-profit companies: CRAN and the intended users

Difference #1: CRAN

Comprehensive R Archive Network (https://cran.r-project.org)

- Vast majority of R packages are available through CRAN
- CRAN repository policies must be followed (https://cran.r-project.org/web/packages/policies.html)
- License policy is generally problematic for packages from for-profit companies
  - Must have a license from the list (https://svn.r-project.org/R/trunk/share/licenses/license.db)
  - Most companies have their own licenses approved by their legal department
  - CRAN policy with distribution and licenses can be unacceptable for a for-profit company

"The package's license must give the right for CRAN to distribute the package in perpetuity."

My company's legal department will not let us put a license that lets CRAN distribute our package in perpetuity on the SNPolisher package.

Difference #2: Intended Users

Who is the intended user for most R packages? Statisticians, other scientists, and people who are expected to figure out things on their own when they need to.

Who is the intended user for a package from a for-profit company? The company's customers.

- Probably not statisticians
  - Our customers are geneticists, some of whom have a strong statistical background
- May not have ever used R before
  - Command-line programming may be new to some customers
- The user guide for the package must be very comprehensive and detailed
  - Customers may not pay money for the actual package but they have paid money for your product
  - User guide should include how to install R, some basics of R, give links to R introductions, explain how to call man pages, etc.
  - Company's support team must also learn the package and how to answer common questions
  - SNPolisher: 11 functions, user guide is ~250 pages
  - Include useful examples/vignettes!

My Rules for SNPolisher

Just a list of what I do at work for our R package:

- Make it CRAN-compliant (except the license); pretend that you will upload it to CRAN/Bioconductor
- Make sure that every customer can install and run it
  - User guide has screen shots of every step of installation, code examples for every function, and lists of common errors
  - Goal: try to reduce the number of problems that get sent to the support team
- Get customer feedback and requests from the support team
  - They deal with the customers directly and I don't, so they have a much better idea of how the package is going over with the customers

DOCUMENT EVERYTHING

- Of course all of the code is commented – this rule is for everything. I made a guide to making R packages on our system that goes through every step, explains the difference between man files and the user guide and vignettes, how we name our functions, how to contact the legal department to get a license update, how to get the package put up on our website for distribution, what extra files should go in our package folder besides the actual package when we distribute it, and a few other things that I've forgotten about.
- Assume that whoever takes over the package creation from you won't have ever used R before and document your process to that level.
  - My predecessor had never used R and taught herself how to make packages just using the CRAN guide.

Acknowledgements

- Teresa Webster
- Ali Pirani
- Jeremy Gollub
- Dorothy Oliver
- Mingham Wu
- Marcos Woehrmann

Statistics Without Borders

https://www.facebook.com/statisticswithoutborders
http://community.amstat.org/statisticswithoutborders/