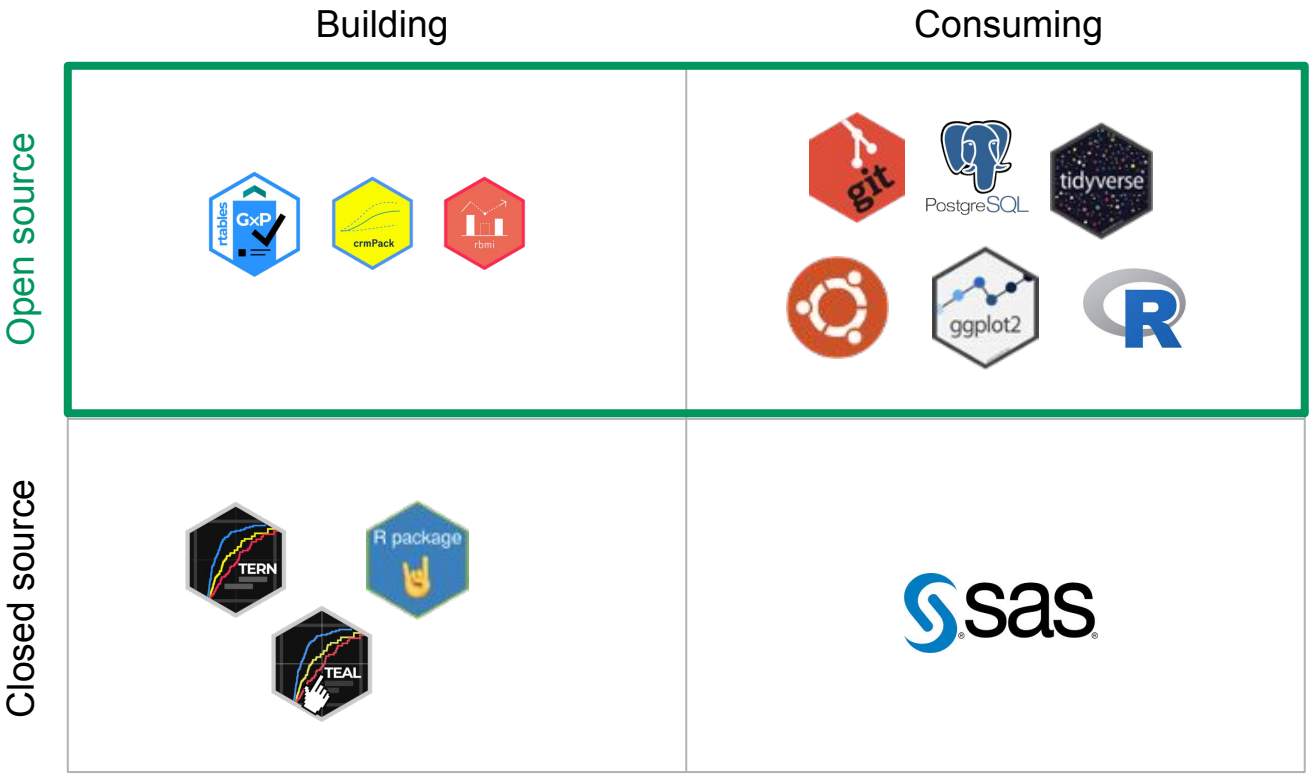
Smoothing out our path to open source pan-pharma code collaboration

James Black,
Associate Director, Insights Engineering

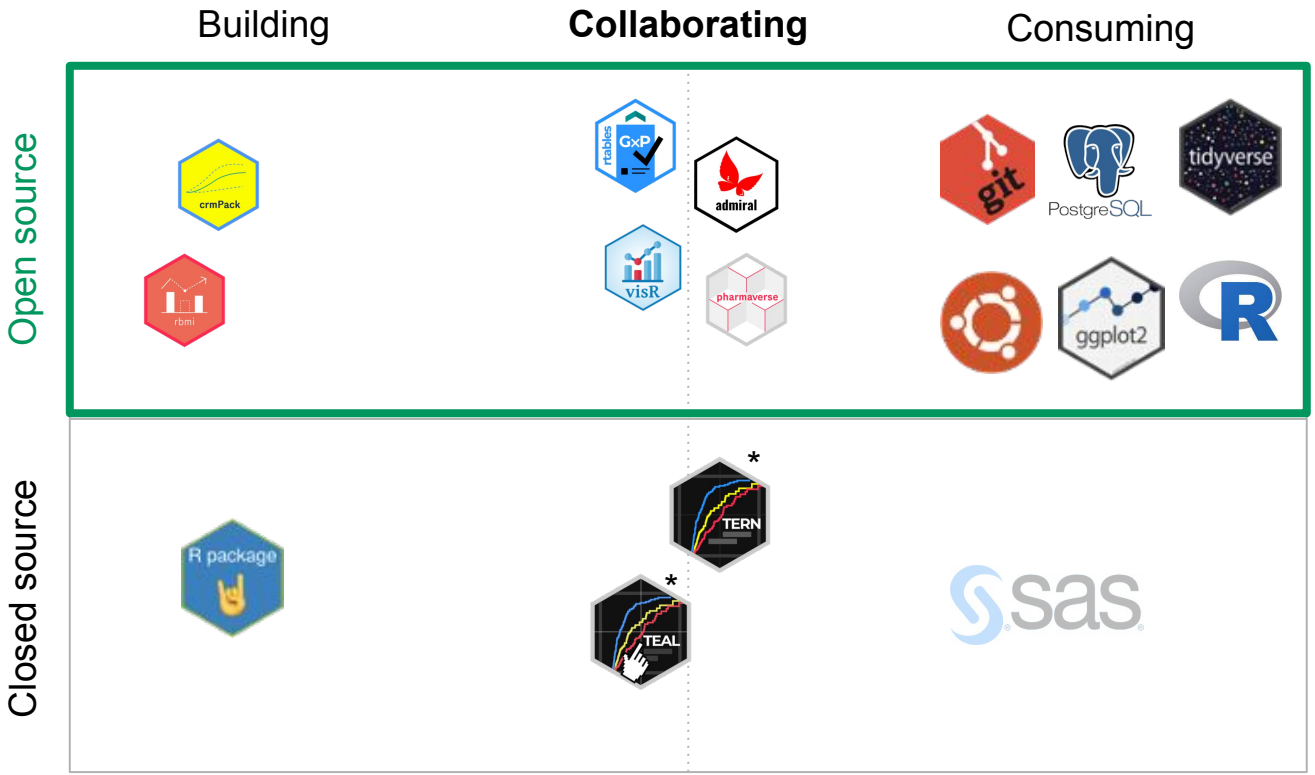
R/Pharma - 2021-11-04

Contact: epijim.uk/contact

Historically - open sourcing has been mostly about pushing out code, and consuming others code



Today - open sourcing has been *complicated* by collaboration between companies



*in the process of being open sourced

Complications in open sourcing big collaborations

From individuals contributing open source as a by-product of projects to departments assigning FTE on open source projects

What R code should we prioritise to collaborate on in an open setting?

Which licenses should we use, and what are the ramifications?

Our big open source projects (e.g. Nest, Admiral) are corporate investments - so how can we map open source governance models to business critical, collaboratively built, open source tools?

Where can we get immediate benefit focussing efforts as a collaborative shared codebase?

Pre-competitive

Generic statistical / ML / methods

Implementations of novel methods (e.g. scorer, graph dbase)

Competitive

Molecule effectiveness

Molecule discovery and design

Post-competitive

Templates and standards for TLGs and ADaMs

Orchestrate analysis ready data and TLGs

Which licenses should we use, and what are the ramifications?

Copy-left licenses are *reciprocal* in that anything that includes the code must be open sourced under the same license. This would prevent you including the code within a proprietary product, so rules out options like making dependency free executables. Most common example is the GPL family.

Permissive licenses are *permissive* and enable you to distribute derivative code under the license of your choosing. MIT and Apache 2.0 are common examples.

Which licenses should we use, and what are the ramifications?

We need contracts for everyone contributing code!

Apache 2.0 includes several important conditions that simplify contributions by **making sure that you can accept code contributions from third parties** (e.g. a pull request):

What if they decide to close source and monetise, or take the project in a direction we disagree with!

The code, up till the point the copyright holder changes the license, will always have this license. **If we decided to switch an Apache 2.0 licenced repository to private, you can simply keep your copy and continue to use and improve that fork.**

How we tried to streamline the process at Roche Pharma Data Sciences

- Created guidance on why we open source, what types of projects are appropriate, license recommendations, and descriptions of Roche Pharma Data Science products with different governance models (open source and by contract between companies).
  - This is a collaboration between Data Sciences and Legal.
  - It's a living document that we add to as questions arise

- Continuing to discuss with several Pharma's ideal governance models for 'master branch'
  - R packages like those in Nest have ~30 FTE invested and will be used in thousands of GxP reporting events. How to ensure a shared roadmap remains on track for each company and our investments get the return we need?

- We have created (and still adding) 'fit for purpose' contracts for early sharing of pre-open source code and governing 'master branch' of projects between a small group of companies.

Doing now what patients need next