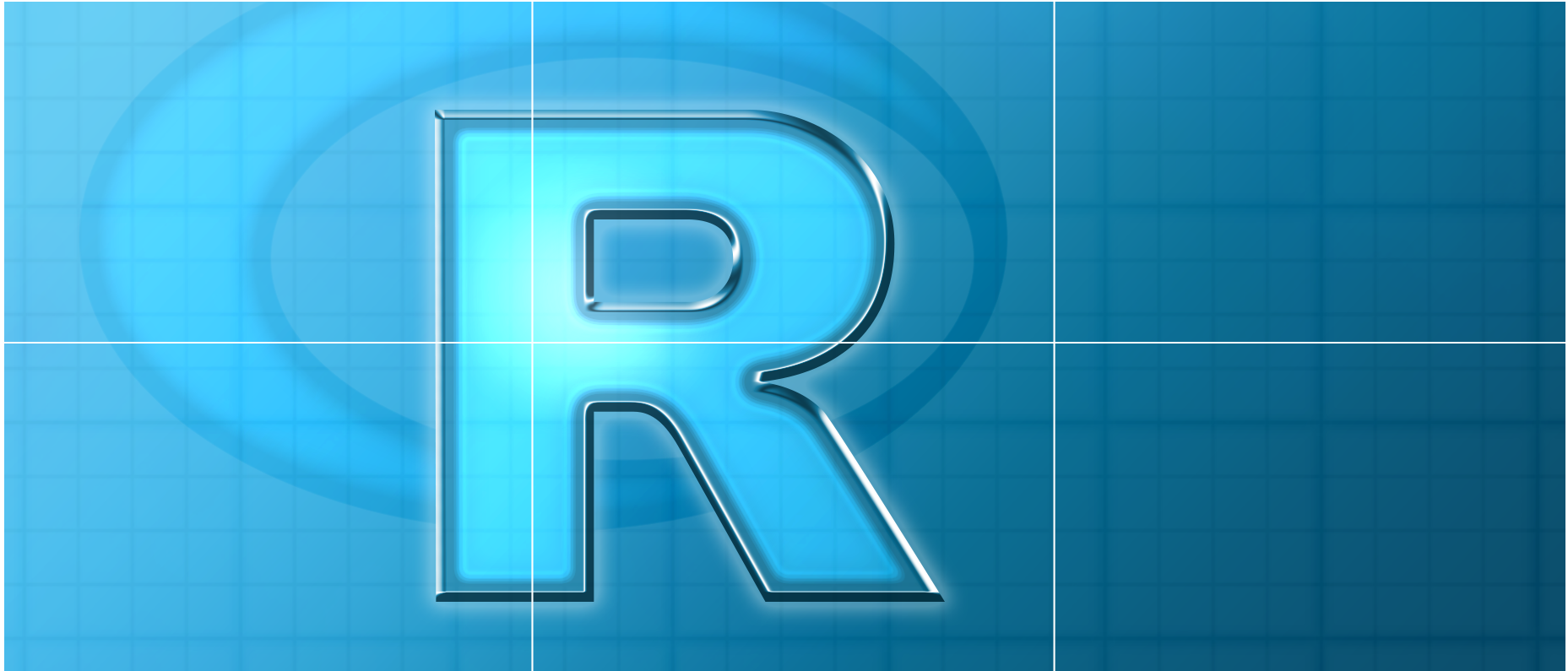


Developing an R Validation Framework

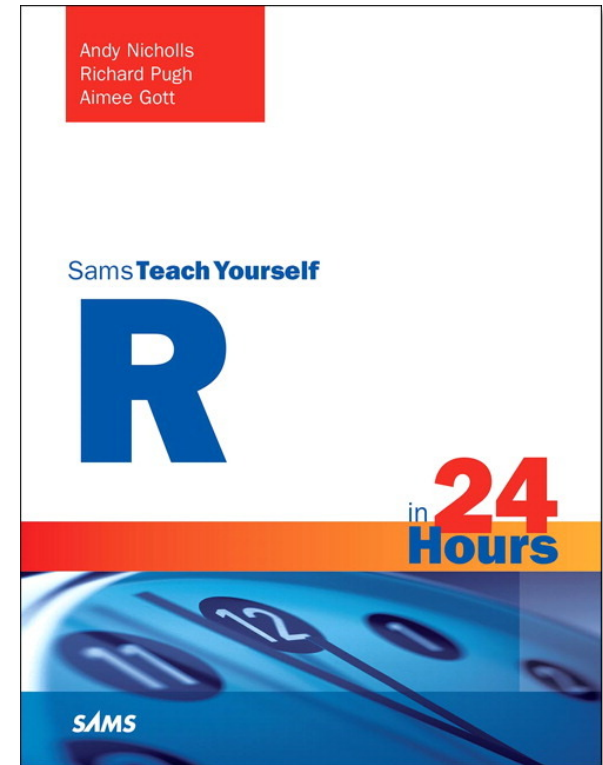


Andy Nicholls
Head of Consultancy

About Me

@AndyOfSmeg

- Previously a Statistician at GSK
- Programmer in R, SAS, S-PLUS, Python
- Now Head of Consultancy at Mango Solutions



About Mango

@MangoTheCat, @MangoPharma



- Private company founded in 2002
- Employee ~50 people
- Providers of analytic software and services
- Premier provider of services for the R language
- Development and services in line with ISO 9001 and FDA guidelines

Agenda

- What is validation?
- Validating R
- ValidR process
- ValidR Enterprise
- VRE as a product
- Future

Establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes

Glossary of Computerized System and Software Development Terminology, FDA
1995

*Establishing **documented evidence** which provides a **high degree of assurance** that a specific process will **consistently** produce a product meeting its **predetermined specifications** and quality attributes*

Glossary of Computerized System and Software Development Terminology, FDA
1995



Why?

Console ~/

R version 3.3.0 (2016-05-03) -- "Supposedly Educational"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software; you are welcome to redistribute it under certain conditions.
You are welcome to distribute the R binaries under the GPL provided that all the R binaries in the distribution are GPLed by the R Foundation.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' for how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

< > ↻ <https://cloud.r-project.org>



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)
[About R](#)

Contributed Packages

Available Packages

Currently, the CRAN package repository features 9297 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

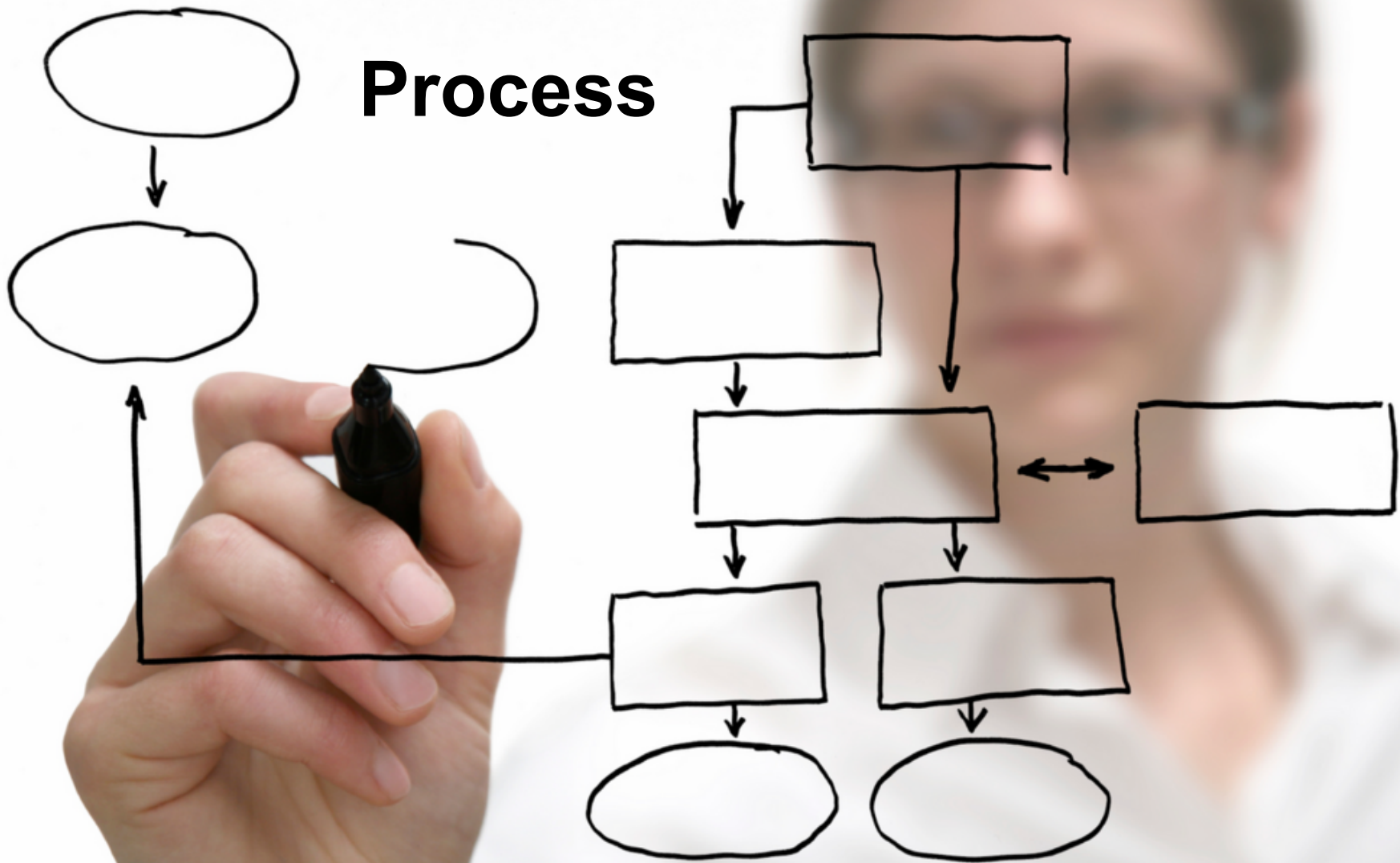
Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages. The manual [R Installation and Administration](#) (also contained in the R base sources)

Why?

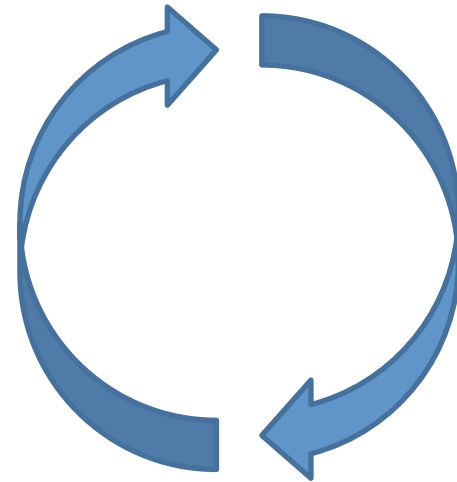
- R itself consists of 30 packages
- >9,000 additional packages available on CRAN
- >1,000 on BioConductor
- < 17% of packages on CRAN have a formal test framework

Process



A Validation Process

- Define requirements
- Design software
- Write software
- Write tests, run tests
- Document the results



Requirements

“predetermined specifications”

- Start with requirements
- R packages have already been written!
- Defined in:
 - DESCRIPTION files
 - Vignettes
 - Websites
- Very manual!

Example DESCRIPTION: lme4

Fit linear and generalized linear mixed-effects models. The models and their components are represented using S4 classes and methods. The core computational algorithms are implemented using the 'Eigen' C++ library for numerical linear algebra and 'RcppEigen' "glue".

Risk Assessment

“high degree of assurance”



Metrics

- Developed *packageMetrics* to automate the collection of metrics
- Metrics collected from RStudio CRAN logs

Package Metrics

Automatically collected package metrics. Note that in the investigation stage the table can be edited to fix incorrect values, e.g. some vignettes are sometimes not detected.

	Description	Risk	Result
ARR	Number of times = is used for assignment	Low	3
ATC	Author Test Coverage	Low	59.48
CCP	Cyclomatic Complexity	Low	1.91
DEP	Num of Dependencies	Medium	18
DPD	Number of Reverse-Dependencies	Low	277
DWL	Number of Downloads	Low	1412359
FLE	Average number of code lines per function	Low	7.16
FRE	Date of First Release	Low	2014-01-16
LIB	Number of library and require calls	High	17
LLE	Number of code lines longer than 80 characters	Medium	121
LNC	Number of lines of compiled code	Low	4515
LNR	Number of lines of R code	Medium	9016
LRE	Date of Last Release	Low	2015-09-01
NAT	Number of attach and detach calls	Low	0
NTF	Number of times T/F is used instead of TRUE/FALSE	Low	0
NUP	Updates During the Last 6 Months	Medium	0
OGH	Whether the package is on GitHub	Low	1
SAP	Number of supply calls	Medium	7
SEM	Number of trailing semicolons in the code	Low	2
SEQ	1:length(vec) is avoided	Low	0
SWD	Number of setwd calls	Low	0
VIG	Number of vignettes	Low	8

Save Metrics Table

Validation Application

VRE 1.0.0-ab09b29

Validation

Summary Tables

Package

dplyr

Status: passed

dplyr — 0.4.3

Title

A Grammar of Data Manipulation

Maintainer

Hadley Wickham <hadley@rstudio.com>

Author

Hadley Wickham [aut, cre], Romain Francois [aut], RStudio [cph]

Description

A fast, consistent tool for working with data frame like objects, both in memory and out of memory.

Source code

<https://github.com/cran/dplyr/tree/0.4.3>

Investigation

Investigation Review

Remediation

Remediation Review

Final Review

Approval

Report

Investigator Comments

Package Conclusion

```
1 | dplyr` is an incredibly popular package for data manipulation in R. It was first
2 | released in 2014, and, due to its popularity and usefulness within the R communi
3 | has had minor releases every 3-4 months.
4 |
5 | Authored by Hadley Wickham - a well-respected member of the R community - the co
6 | de in this package is of a very high standard. Further, the package is available on
7 | CRAN, and is actively maintained by a wide range of contributors. This implies
8 | both that the risks of bugs existing is small, and that any identified bugs are
9 | fixed and improvements are made frequently.
10 |
11 | `dplyr` has 8 vignettes: *Introduction to dplyr*, *Data frame performance*,
12 | *Databases*, *Two-table verbs*, *Adding a new SQL backend*, *Hybrid evaluation*,
13 | *Non-standard evaluation* and *Window functions and grouped mutate/filter*,
14 | which contain extensive information on how to use the main functions within the
15 | package.
16 |
17 | The fact that the package is available on CRAN is an indication that it meets
18 | the R CMD check, a form of quality criteria, and as such the risk associated
19 | with the package is expected to be reduced.
20 |
21 | Overall, due to the excellent reputation of the package authors,
22 | the wide exposure of the package and a lack of maturity,
23 | we would deem this to be of medium risk.
```

dplyr is an incredibly popular package for data manipulation in R. It was first released in 2014, and, due to its popularity and usefulness within the R community has had minor releases every 3-4 months.

Authored by Hadley Wickham - a well-respected member of the R community - the code in this package is of a very high standard. Further, the package is available on CRAN, and is actively maintained by a wide range of contributors. This implies both that the risks of bugs existing is small, and that any identified bugs are fixed and improvements are made frequently.

dplyr has 8 vignettes: *Introduction to dplyr*, *Data frame performance*, *Databases*, *Two-table verbs*, *Adding a new SQL backend*, *Hybrid evaluation*, *Non-standard evaluation* and *Window functions and grouped mutate/filter*, which contain extensive information on how to use the main functions within the package.

The fact that the package is available on CRAN is an indication that it meets the R CMD check, a form of quality criteria, and as such the risk associated with the package is expected to be reduced.

Overall, due to the excellent reputation of the package authors, the wide exposure of the package and a lack of maturity, we would deem this to be of medium risk.

Package Functionality

```
1 | ## Functionality Summary ##
```

Validation Application

- Built using Shiny
- Used to manage validation process
 - Define requirements
 - Assess risk
 - Determine level of remediation
 - Generate package validation reports
- Information written to database

Reports

Document: codeltools 0.2-8 Validation Report	Project/Product: ValidR	Job No: 2150	Document Revision: 1.0	Document Date: May 7, 2014	MANGOSOLUTIONS	
9 Validation Certificate						
9.1 Tests						
All tests identified in the report above were written, with no deviations.						
9.2 Test Execution Environment						
The tests were executed in the following environment:						
Attribute	Value					
Computer	validr-ci.mango.local					
Operating System	Linux 2.6.32-220.el6.x86_64					
Build Number	R version 3.0.2 (2013-09-15) , svn revision: 63987					
9.3 Test Outcomes						
All tests were executed successfully in the test execution environment. One test was written by the package author, and this test failed in the test execution environment. However, this failure does not affect the validity of the package, as this is a low priority test.						
9.4 Validation Conclusion						
The package has been validated. However, in order for the validR system to be deemed valid, all critical tests must pass on the client deployment environment.						



Testing

```
> library(testthat)
> myPi <- 3.14159
>
> # Passing test
> expect_equal(myPi, pi, tolerance = 0.00001)
>
> # Failing test
> expect_equal(myPi, pi)
Error: `myPi` not equal to lazyLoadDBfetch(c(483539L,
39L), datafile, compressed, envhook).
1/1 mismatches
[1] 3.14 - 3.14 == -2.65e-06
```

Testing

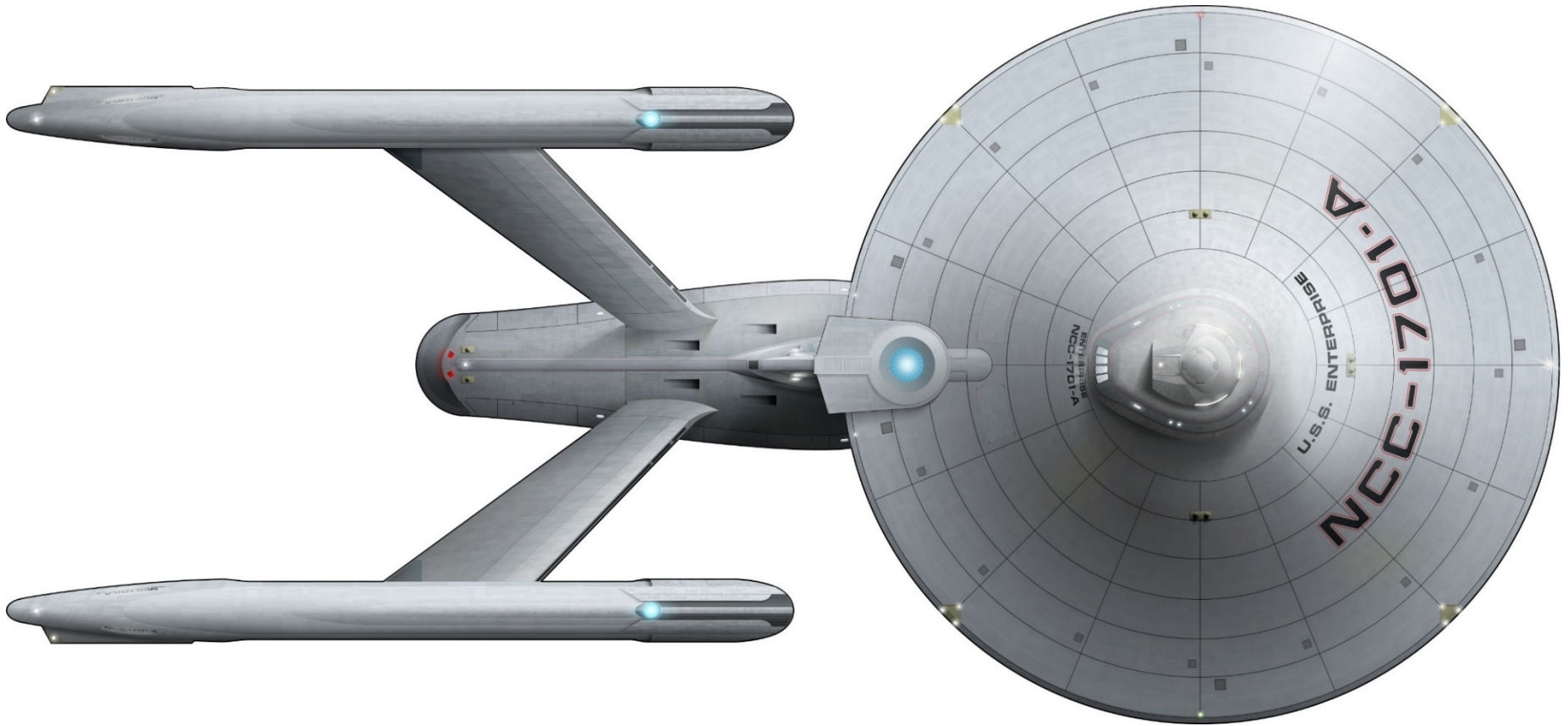
```
context('[critical]VR-480 digest: Must create hash function  
digests of R objects')
```

```
tmp <- digest(list(LETTERS,  
                  data.frame(a=letters[1:5],  
                             b=matrix(1:10,ncol=2)))  
)
```

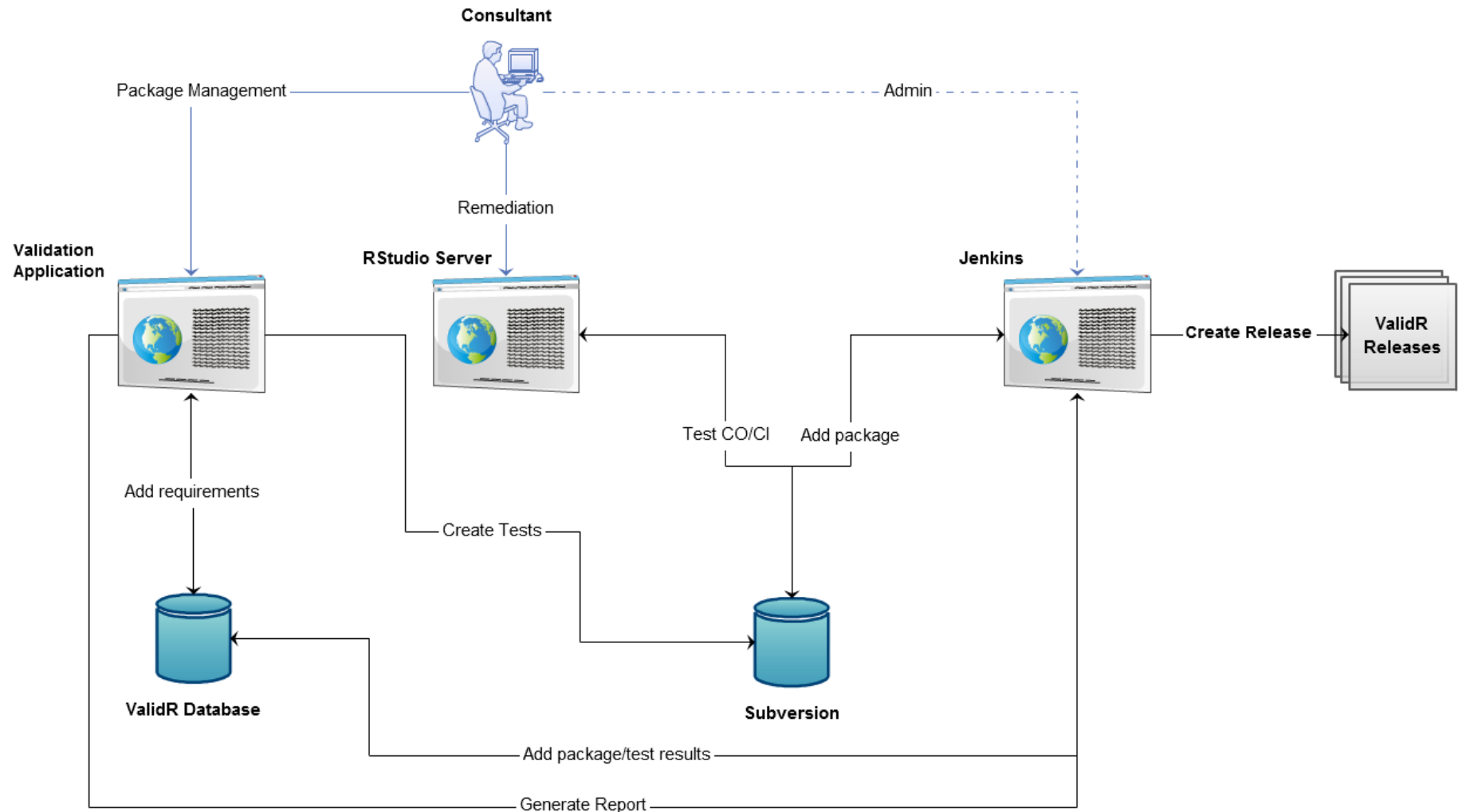
```
test_that('Can create hash for list object', {  
  expect_identical(tmp,"8b23d6c96da88f525c0a4a4909cce01f")  
})
```

...

ValidR Enterprise



Architecture



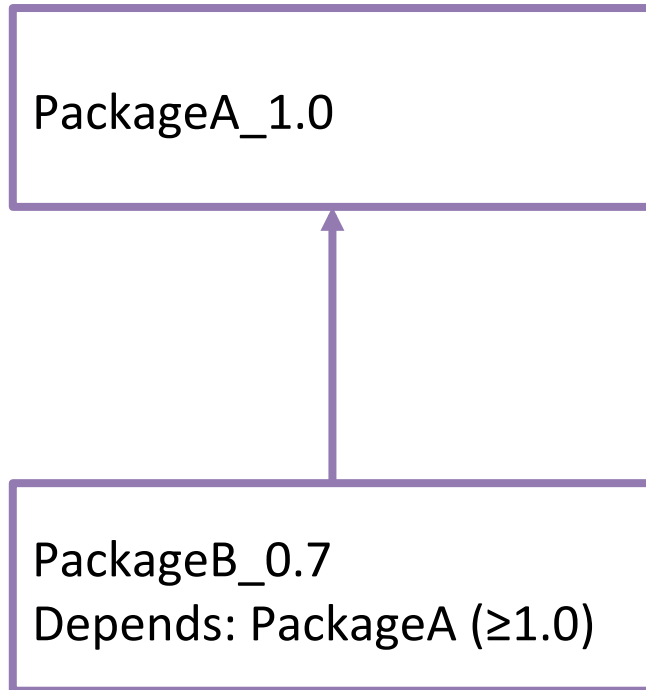


Jenkins

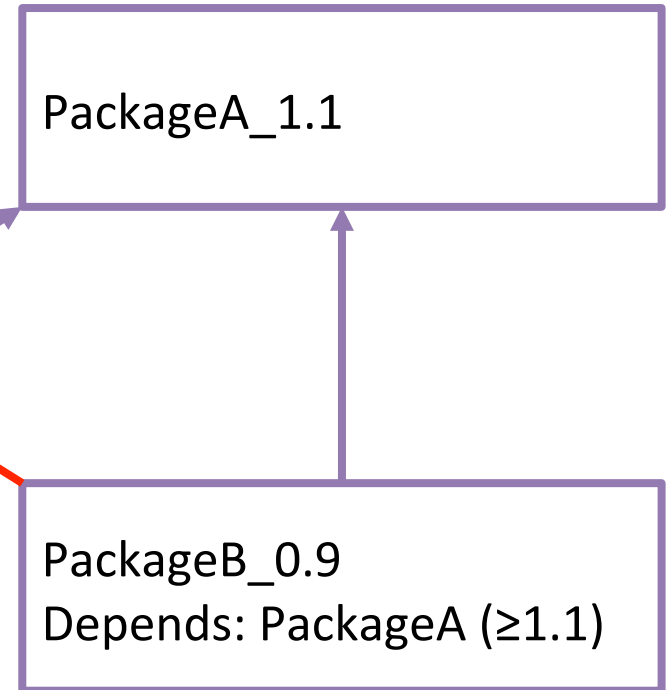
- Continuous Integration tool
- Adds packages
 - Retrieve the package from the appropriate location
 - Create a directory in SVN to store source
 - Run *packageMetrics* and write results to database
 - Expose package to consultants via RStudio Server
- Deploy and test ValidR builds

Resolving Dependencies

Repository Today



Repository Next Week



Resolving Dependencies

- Microsoft's MRAN tool provides CRAN snapshots

Now viewing CRAN snapshot from: 2016-05-04



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

Contributed Packages

Available Packages

Currently, the CRAN package repository features 8331 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The

Other Components

- **Database** stores package and report information
- **RStudio Server** enables consultants to develop tests
- **CI** - Jenkins continuously executes package tests as they are committed to SVN
 - Complete builds deployed nightly

Delivery



Delivery

- Orchestrated via RPMs
- Each package and its tests stored within a self-contained RPM
- Master RPM depends on each of the package RPMs
- Further RPMs for IQ, OQ and PQ

IQ, OQ, PQ

“documented evidence”

- Triggered by master RPM
- Checks everything installs to the correct locations
- Executes tests
- Re-executes test for consistency
- Automated report including overall PASS/FAIL for each stage

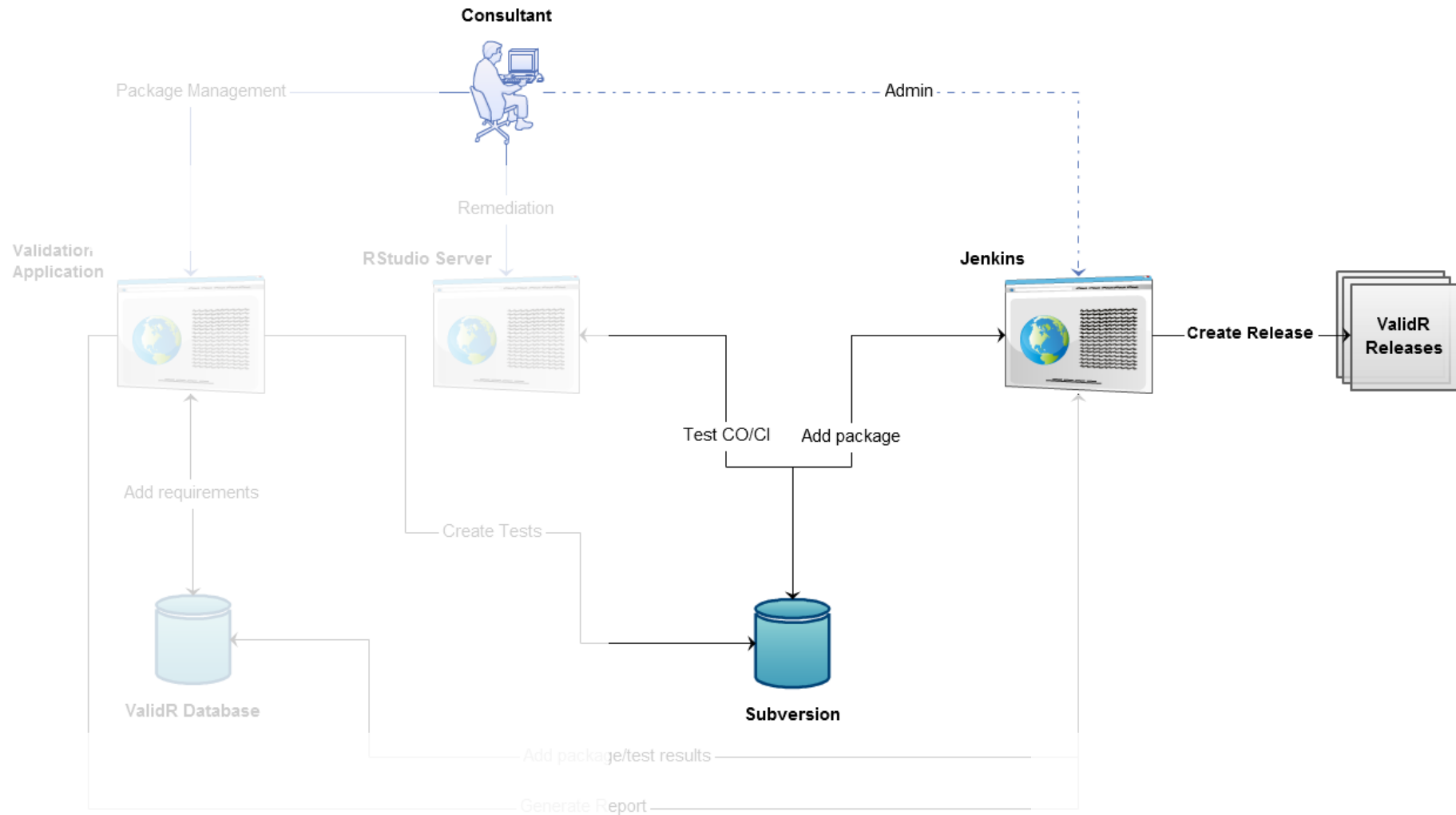
ValidR Enterprise as a Product



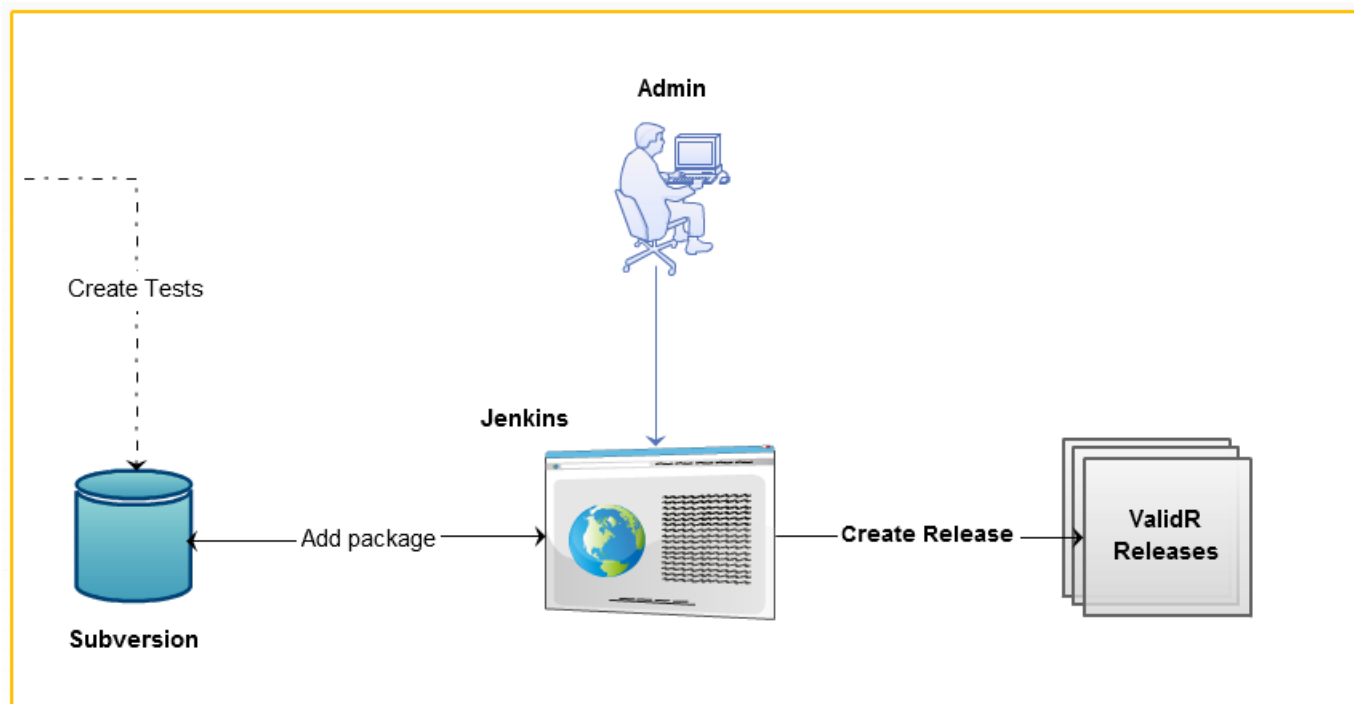
Key Drivers

Enable companies to ...

- Develop own internal R packages
- Add additional R packages
- Create custom R builds for different departments



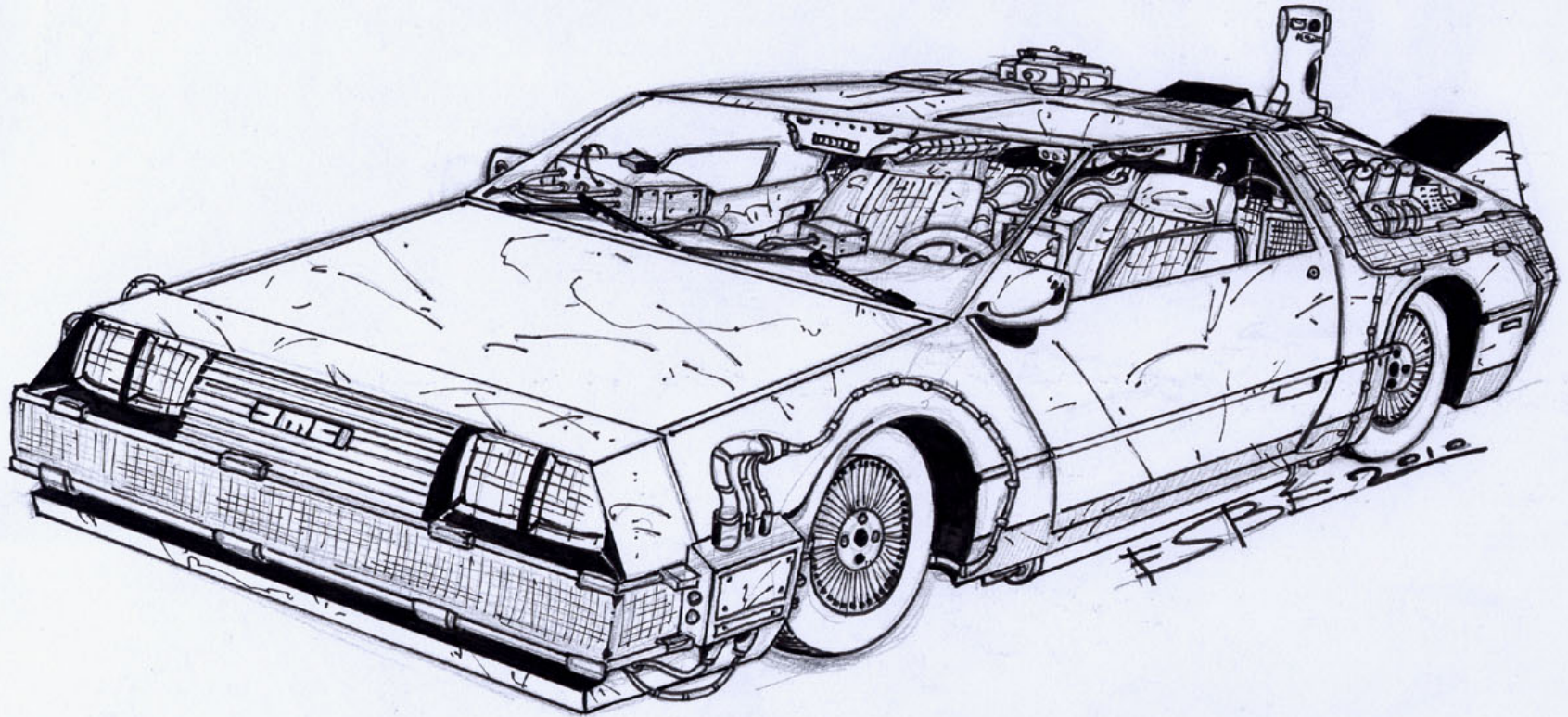
ValidR Enterprise Product Architecture



Valid Session?

- ValidR can be accessed:
 - Via command line / batch
 - Interactively via RStudio Server
- Session Tracking Application records whether loaded packages are valid or not
- The application also feeds into the product roadmap

The Future



The Future

- Session Tracker Enhancements for IT
- Release of Validation Application
- Develop Admin Application

Summary

- Validating R is not easy!
- Requires
 - Time
 - Skill
 - Many tools!

