

AS03: Reproducible Open Source Environments for Advanced Analytics

Sean Lopp
March 2020

Reproducible

Open Source

Advanced Analytics

Reproducible

Open Source

Advanced Analytics

Personal Healthcare

Diagnostics

Early Stage Research

Clinical Trials

Genentech



Reproducible

Open Source

Advanced Analytics

Durable and
repeatable,
regardless of
economic means

Bigger pool of
innovation



Reproducible

Open Source

Advanced Analytics

Durable and repeatable, regardless of economic means

Inspectable

Validation and Qualification

Reproducible

Open Source

Advanced Analytics

Durable and repeatable, regardless of economic means

Inspectable

Validation and Qualification

*How do I **trust** the tools I use?*

<https://www.pharmar.org>

Reproducible

Open Source

Advanced Analytics

Durable and repeatable, regardless of economic means

Inspectable

Validation and **Qualification**

*How do I **consistently use** the same tools?*

Reproducible

Open Source

Advanced Analytics

Durable and repeatable, regardless of economic means

Inspectable

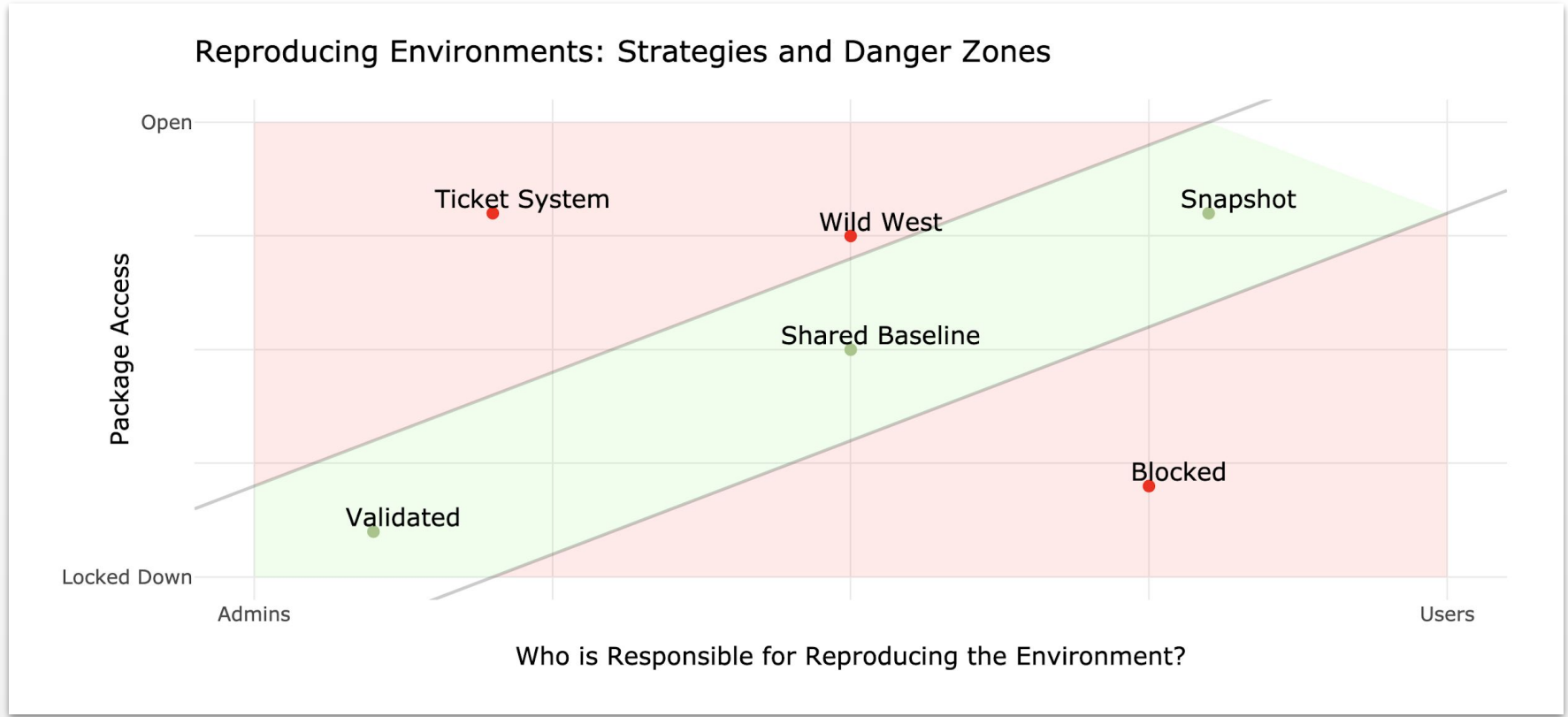
Validation and **Qualification**

*How do I **consistently use** the same tools?*

Reproducible Environments

- Code that used to run no longer runs, even though the code has not changed.
- You are afraid to upgrade or install a new package, because it might break your code or someone else's.
- Typing `install.packages` in your environment doesn't do anything, or doesn't do the *right* thing.

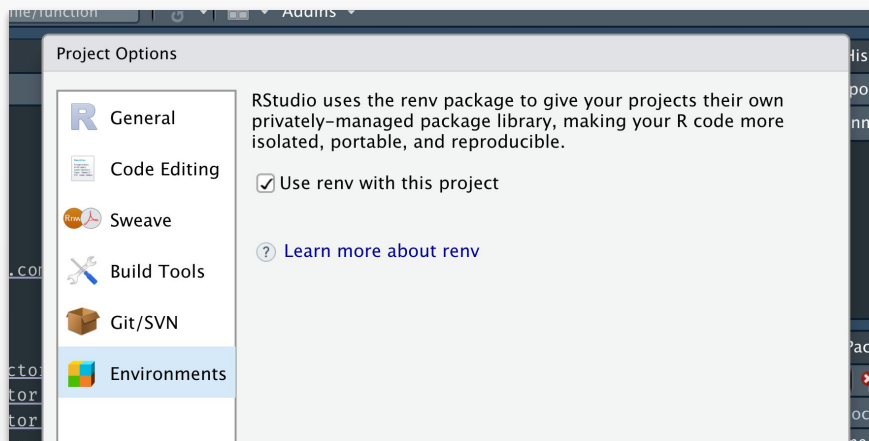
3 Strategies



Snapshot

- **Individual users** take a **snapshot** of a **project library**

```
renv::init()  
renv::snapshot()  
renv::restore()
```

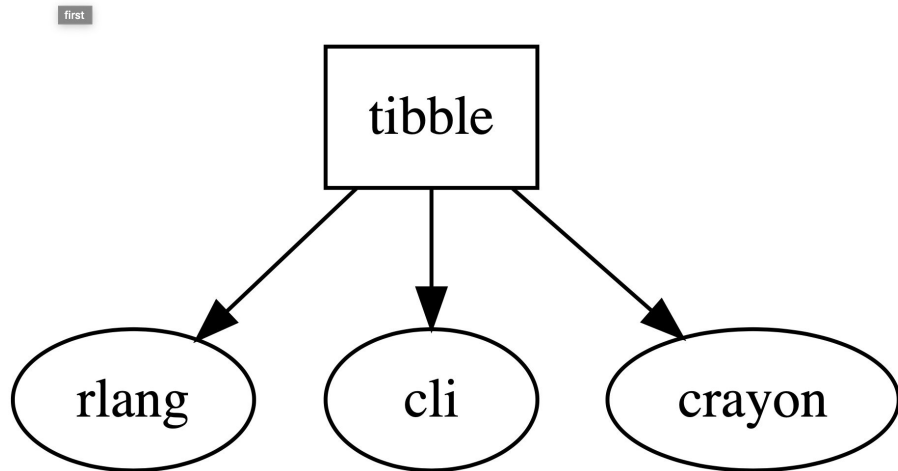


RStudio 1.3

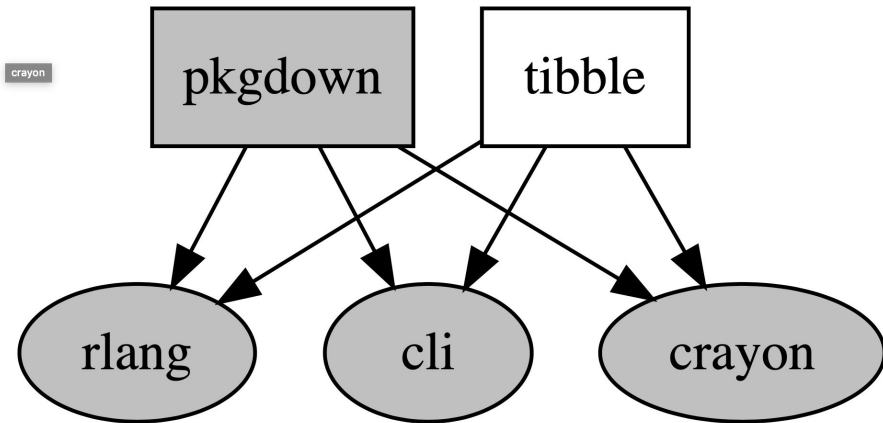
```
1 [renv]  
2 Version=0.3.0-19  
3  
4 [R]  
5 Version=3.5.3  
6 Repositories=  
7   CRAN=https://cran.rstudio.com/  
8  
9 [Bioconductor]  
10 Repositories=  
11   BioCsoft=https://bioconductor.org/packages/3.  
12   BioCann=https://bioconductor.org/packages/3.8  
13   BioCexp=https://bioconductor.org/packages/3.8  
14   BioCworkflows=https://bioconductor.org/packag  
15   CRAN=https://cran.rstudio.com/  
16  
17 [R/Package/MASS]  
18 Package=MASS  
19 Version=7.3-51.5  
20 Source=CRAN  
21 Hash=d23c0864ec6a1fc5afd320e1433a7c71  
22  
23 [R/Package/Matrix]  
24 Package=Matrix  
25 Version=1.2-18
```

Shared Baseline - Why

- **Everyone** uses packages **from the same point in time**



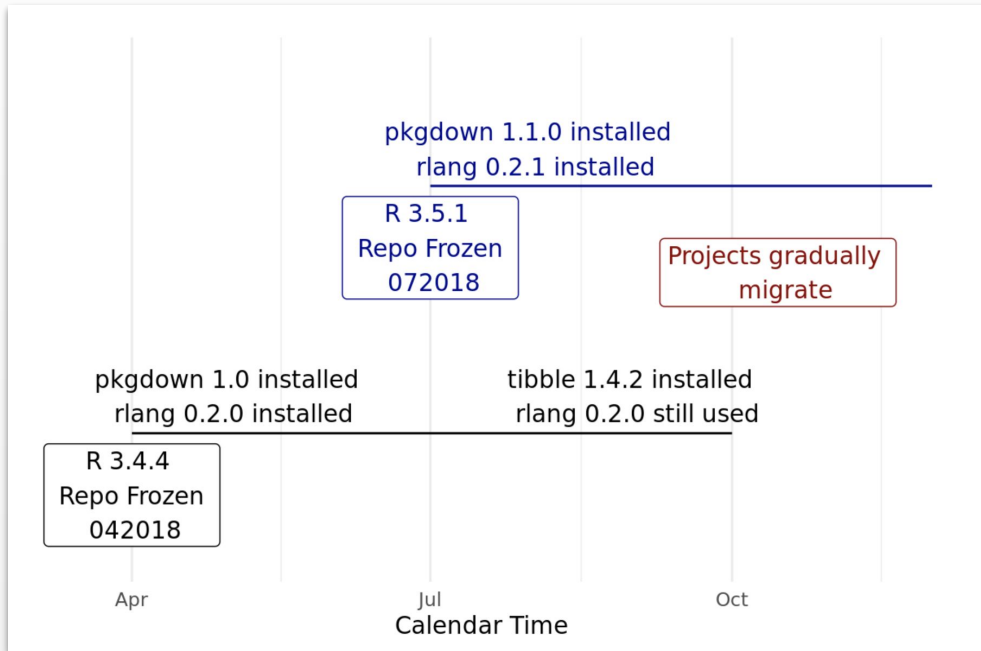
May 2019



December 2019

Shared Baseline - How

- **Everyone** uses packages **from the same point in time**



Repository URL

You can choose to use the latest packages, or freeze your set of packages to a particular checkpoint.

2019							December 2019							January 2020					
N	TUE	WED	THU	FRI	SAT		SUN	MON	TUE	WED	THU	FRI	SAT	SUN	MON	TUE	WED	THU	
				1	2		1	2	3	4	5	6	7					1	2
	5	6	7	8	9		8	9	10	11	12	13	14	5	6	7	8	9	
1	12	13	14	15	16		15	16	17	18	19	20	21	12	13	14	15	16	
3	19	20	21	22	23		22	23	24	25	26	27	28	19	20	21	22	23	
5	26	27	28	29	30		29	30	31					26	27	28	29	30	

You can also choose to use source or binary packages. ⓘ

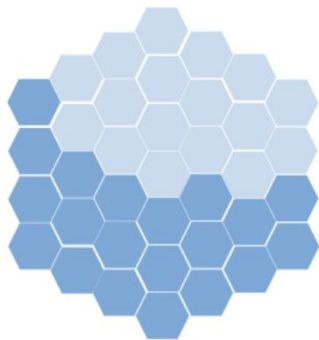
CentOS/RHEL 7 ▾

Use this URL to work with the binary packages available as of December 15, 2019 for CentOS/R

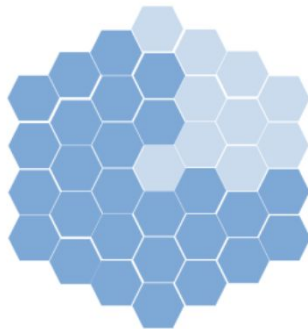
https://demo.rstudiopm.com/all/__linux__/centos7/2331

Validated

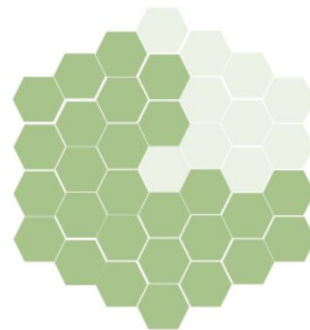
- **Risk Approved** packages **from the same point in time**



1. Admin selects approved packages



2. New packages can be *added* from the same point in CRAN's time



3. The set of packages can be *updated* to a new point in CRAN's time

Rest of the picture

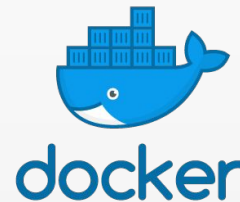
- Package Strategy + **Best Software Development Processes**



Code



System / Server



Resources

pharmar.org

environments.rstudio.com

rstudio.github.io/renv

demo.rstudiopm.com