

# The Rockerverse: Packages and Applications for Containerisation with R

by Daniel Nüst, Dirk Eddelbuettel, Dom Bennett, Robrecht Cannoodt, Dav Clark, Gergely Daroczi, Mark Edmondson, Colin Fay, Ellis Hughes, Lars Kjeldgaard, Sean Lopp, Ben Marwick, Heather Nolis, Jacqueline Nolis, Hong Ooi, Karthik Ram, Noam Ross, Lori Shepherd, Péter Sóllymos, Tyson Lee Swetnam, Nitesh Turaga, Charlotte Van Petegem, Jason Williams, Craig Willis, Nan Xiao

**Abstract** The R Project provides widely used Docker images for R across different application scenarios. This article surveys downstream projects that build upon the R Project images and presents the current state of R packages for managing Docker images and controlling containers. These use cases cover diverse topics such as package development, reproducible research, collaborative work, cloud-based data processing, and production deployment of services. The variety of applications demonstrates the power of the R Project specifically and containerisation in general. Across the diverse ways to use containers, we identified common themes: reproducible environments, scalability and efficiency, and portability across clouds. We conclude that the current growth and diversification of use cases is likely to continue its positive impact, but see the need for consolidating the Rockerverse ecosystem of packages, developing common practices for applications, and exploring alternative containerisation software.

## Introduction

The R community continues to grow. This can be seen in the number of new packages on CRAN, which is still on growing exponentially (Hornik et al., 2019), but also in the numbers of conferences, open educational resources, meetups, unconferences, and companies that are adopting R, as exemplified by the useR! conference series<sup>1</sup>, the global growth of the R and R-Ladies user groups<sup>2</sup>, or the foundation and impact of the R Consortium<sup>3</sup>. These trends cement the role of R as the *lingua franca* of statistics, data visualisation, and computational research. During the last few years, and coinciding with the rise of R, has been the advent of Docker as a general tool for distributing and deploying of server applications—in fact, Docker can be called the *lingua franca* of describing computing environments and packaging software. Combining both these topics, the R Project (<https://www.r-project.org/>) provides Docker images with R (see the next section for more details). The considerable uptake and continued evolution of the R Project has led to numerous projects that extend or build upon R Project images, ranging from reproducible<sup>4</sup> research to production deployments. As such, this article presents what we may call the *Rockerverse* of projects across all development stages: early demonstrations, working prototypes, and mature products. We also introduce related activities that connect the R language and environment with other containerisation solutions. Our main contributions is a coherent picture of the current status of using containers in, with, and for R.

The article continues with a brief introduction of containerisation basics and the R Project, followed by use cases and applications, starting with the R packages specifically for interacting with Docker, next the second-level packages that use containers indirectly or only for specific features, and finally some complex use cases that leverage containers. We conclude by reflecting on the landscape of packages and applications and point out future directions of development.

## Containerisation and R

Docker, an application and service provided by the eponymous company, has, in just a few short years, risen to prominence for developing, testing, deploying and distributing computer software (cf. Datadog, 2018; Muñoz, 2019). While related approaches exist, such as LXC<sup>5</sup> or Singularity (Kurtzer et al., 2017), Docker has become synonymous with “containerisation”—the method of taking software artefacts and bundling them in such a way that use becomes standardized and portable across operating systems. In doing so, Docker had recognised and validated the importance of one very important thread that had been emerging, namely virtualisation. By allowing (one or possibly) multiple

<sup>1</sup><https://www.r-project.org/conferences/>

<sup>2</sup><https://www.r-consortium.org/blog/2019/09/09/r-community-explorer-r-user-groups>, <https://www.r-consortium.org/blog/2019/08/12/r-community-explorer>

<sup>3</sup><https://www.r-consortium.org/news/announcements>, <https://www.r-consortium.org/blog/2019/11/14/data-driven-tracking-and-discovery-of-r-consortium-activities>

<sup>4</sup>“Reproducible” in the sense of the Claerbout/Donoho/Peng terminology (Barba, 2018).

<sup>5</sup><https://en.wikipedia.org/wiki/LXC>

applications or services to run concurrently on one host machine without any fear of interference between them, Docker provides an important scalability opportunity. Beyond this though, Docker has improved this compartmentalisation by accessing the host system—generally Linux—through a much thinner and smaller shim than a full operating system emulation or virtualisation. This containerisation, also called operating-system-level virtualisation (Wikipedia contributors, 2020b), makes efficient use of operating system resources (Felter et al., 2015) and allows another order of magnitude in terms of scalability of deployment (cf. Datadog, 2018), because virtualisation may emulate a whole operating system, a container typically runs only one process. The single process together with sharing the host’s kernel results in a reduced footprint and faster start times. While Docker makes use of Linux kernel features, it has become important enough that some required aspects of running Docker have been added to other operating systems so that those systems can more efficiently support Docker (Microsoft, 2019b). The success of Docker has even paved the way for industry collaboration and standardisation (OCI, 2019).

The key accomplishment of Docker as an “application” is to make a “bundled” aggregation of software, the so-called “image”, available to any system equipped to run Docker, without requiring much else from the host besides the actual Docker application installation. This is a rather attractive proposition, and Docker’s very easy to operate user interface has led to widespread adoption and use of Docker in a variety of domains, e.g., cloud computing infrastructure (e.g., Bernstein, 2014), data science (e.g., Boettiger, 2015), and edge computing (e.g., Alam et al., 2018). It has also proven to be a natural match for “cloud deployment” which runs, or at least appears to run, “seamlessly” without much explicit reference to the underlying machine, architecture or operating system: Containers are portable and can be deployed with very little dependencies on the host system—only the container runtime is required. These Docker images are normally built from plain text documents called Dockerfiles; a Dockerfile has a specific set of instructions to create and document a well-defined environment, i.e., install specific software and expose specific ports.

For statistical computing and analysis centered around R, the **Rocker Project** has provided a variety of Docker containers since it began in 2014 (Boettiger and Eddelbuettel, 2017). The Rocker Project provides several lines of containers spanning from building blocks with R-release or R-devel, via containers with RStudio Server and Shiny Server, to domain-specific containers such as rocker/geospatial (Boettiger et al., 2019). These containers form *image stacks*, building on top of each other for easier maintainability (i.e., smaller Dockerfiles), better composability, and to reduce build time. Also of note is a series of “versioned” containers which match the R release they contain with the *then-current* set of packages via the MRAN Snapshot views of CRAN (Microsoft, 2019a). The Rocker Project’s impact and importance was acknowledged by the Chan Zuckerberg Initiative’s *Essential Open Source Software for Science*, which provides funding for the projects’s sustainable maintenance, community growth, and targeting new hardware platforms including GPUs (Chan Zuckerberg Initiative et al., 2019).

Docker is not the only containerisation software. An alternative stemming from the domain of high-performance computing is **Singularity** (Kurtzer et al., 2017). Singularity can run Docker images. Rocker images work out of the box if the main process is R, e.g., in rocker/r-base, but Singularity does not succeed in running images where there is an init script, e.g., in containers that by default run RStudio Server. In the latter case, a Singularity file, a recipe akin to a Dockerfile, needs to be used to make necessary adjustments. To date, no comparable image stack to the Rocker Project’s images exists on Singularity Hub. A further tool for running containers is **podman**, which can also build Dockerfiles and run Docker images. Proof of concepts exists for using podman to build and run Rocker containers<sup>6</sup>, but the prevalence of Docker, especially in the broader user community beyond experts or niche systems and the vast amount of blog posts and courses for Docker currently cap specific development efforts for both Singularity and podman in the R community. This might quickly change if the usability and spread of Singularity or podman increase, or if security features such as rootless/unprivileged containers, which both these tools support out of the box, become more sought after.

## Interfaces for Docker in R

Users interact with the Docker daemon typically through the **Docker Command Line Interface** (Docker CLI). However, moving back and forth between an R console and the command line can create friction in workflows and reduce reproducibility because of manual steps. A number of first-order R packages provide an interface to the Docker CLI, allowing for the interaction with the Docker CLI from an R console. Table 1 gives an overview of packages with client functionality, each of which provides functions for interacting with the Docker daemon. The packages focus on different aspects and support different stages of a container’s life cycle. As such, the choice of which package is most useful depends on the use case at hand as well as on the user’s level of expertise.

<sup>6</sup>See <https://github.com/nuest/rodman> and <https://github.com/rocker-org/rocker-versioned/issues/187>

Functionality	AzureContainers	babelwhale	dockermachine	dockyard	harbor	stevedore
Generate a Dockerfile				✓		
Build an image	✓			✓		
Execute a container locally or remotely	✓	✓	✓	✓	✓	✓
Deploy or manage instances in the cloud	✓		✓		✓	✓
Interact with an instance (e.g., file transfer)		✓	✓			✓
Manage storage of images					✓	✓
Supports Docker and Singularity		✓				
Direct access to Docker API instead of using the CLI						✓
Installing Docker software			✓			

**Table 1:** R packages with Docker client functionality.

**harbor** (<https://github.com/wch/harbor>) is no longer actively maintained, but it should be honorably mentioned as the first R package for managing Docker images and containers. It uses the **sys** package (Ooms, 2019) to run system commands against the Docker CLI, both locally and through an SSH connection, and it has convenience functions, e.g., for listing and removing containers/images and for accessing logs. The outputs of container executions are converted to appropriate R types. The Docker CLI's basic functionality, although it evolves quickly and with little concern for avoiding breaking changes, has remained unchanged in core functions, meaning that a core function such as `harbor::docker_run(image = "hello-world")` still works despite its stopped development.

**stevedore** is currently the most powerful Docker client in R (FitzJohn, 2020). It interfaces with the Docker daemon over the Docker HTTP API<sup>7</sup> via a Unix socket on Linux or MacOS, over a named pipe on Windows, or over an HTTP/TCP connection. The package is the only one not using system calls to the docker CLI tool for managing images and containers. The package thereby enables connections to remote Docker instances without direct configuration of the local Docker daemon. Furthermore using the API gives access to information in a structured way, is system independent, and is likely more reliable than parsing command line output. **stevedore**'s own interface is automatically generated based on the OpenAPI specification of the Docker daemon, but it is still similar to the Docker CLI. The interface is similar to R6 objects, in that an object of class "stevedore\_object" has a number of functions attached to it that can be called, and multiple specific versions of the Docker API can be supported thanks to the automatic generation<sup>8</sup>.

**AzureContainers** is an interface to a number of container-related services in Microsoft's **Azure Cloud** (Ooi, 2019). While it is mainly intended for working with Azure, as a convenience feature it includes lightweight, cross-platform shells to Docker and Kubernetes (tools `kubect1` and `helm`). These can be used to create and manage arbitrary Docker images and containers, as well as Kubernetes clusters on any platform or cloud service.

**babelwhale** provides a unified interface to interact with Docker and Singularity containers (Cannoodt and Saelens, 2019). Users can, for example, execute a command inside a container, mount a volume, or copy a file with the same R commands for both container runtimes.

**dockyard** (<https://github.com/thebioengineer/dockyard>) has the goal of lowering the barrier to creating Dockerfiles, building Docker images, and deploying Docker containers. The package follows the increasingly used piping paradigm of the Tidyverse-style (Wickham et al., 2019) of programming for chaining R functions representing the instructions in a Dockerfile. An existing Dockerfile can be used as a template. **dockyard** also includes wrappers for common steps, such as installing an R package or copying files, and provides built-in functions for building an image and running a container, which make Docker more approachable within a single R-based user interface.

<sup>7</sup><https://docs.docker.com/engine/api/latest/>

<sup>8</sup>See <https://github.com/richfitz/stevedore/blob/master/development.md>.

**dockermachine** (<https://github.com/cboettig/dockermachine>) is an R package to provide a convenient interface to **Docker Machine** from R. The CLI tool `docker-machine` allows users to create and manage a virtual host on local computers, local data centers, or at cloud providers. A local Docker installation can be configured to transparently forward all commands issued on the local Docker CLI to a selected (remote) virtual host. Docker Machine was especially crucial for local use in the early days of Docker, when no native support was available for Mac or Windows computers, but it remains relevant for provisioning on remote systems. The package has not received any updates for two years, but it is functional with a current version of `docker-machine` (0.16.2). It potentially lowers the barriers for R users to run containers on various hosts if they perceive that using the Docker Machine CLI directly as a barrier and it enables scripted workflows with remote processing.

## Use cases and applications

### Image stacks for communities of practice

**Bioconductor** (<https://bioconductor.org/>) is an open source, open development project for the analysis and comprehension of genomic data (Gentleman et al., 2004). As of October 30th 2019, the project consists of 1823 R software packages, as well as packages containing annotation or experiment data. *Bioconductor* has a semi-annual release cycle, where each release is associated with a particular version of R, and Docker images are provided for current and past versions of *Bioconductor* for convenience and reproducibility. All images, which are described on the *Bioconductor* web site (see <https://bioconductor.org/help/docker/>), are created with Dockerfiles maintained on GitHub and distributed through Docker Hub<sup>9</sup>. *Bioconductor*'s "base" Docker images are built on top of the `rocker/rstudio` image. *Bioconductor* installs packages based on the R version in combination with the *Bioconductor* version and, therefore, uses *Bioconductor* version tagging `devel` and `RELEASE_X_Y`, e.g., `RELEASE_3_10`. Past and current combinations of R and *Bioconductor* will therefore be accessible via a specific image tags.

The *Bioconductor* Dockerfile selects the desired version of R from *Rocker*, adds required system dependencies, and uses the **BiocManager** package for installing appropriate versions of *Bioconductor* packages (Morgan, 2019). A strength of this approach is that the responsibility for complex software configuration (including customized development) is shifted from the user to the experienced *Bioconductor* core team. However, a recent audit of the *Bioconductor* image stack Dockerfile led to the deprecation of several community-maintained images, because the numerous specific images became too hard to understand, complex to maintain, and cumbersome to customise. As part of the simplification, a recent innovation is to produce a `bioconductor_docker:devel` image to emulate the *Bioconductor* nightly Linux build machine as closely as possible. This image contains the build system environment variables and the *system dependencies* needed to install and check almost all (1813 out of 1823) *Bioconductor* software packages, and it saves users and package developers from managing these themselves. Furthermore, the image is configured so that `.libPaths()` has `'/usr/local/lib/R/host-site-library'` as the first location. Users mounting a location on the host file system to this location can persistently manage installed packages across Docker containers or image updates. Many R users pursue flexible work flows tailored to particular analysis needs rather than standardized work flows. The new `bioconductor_docker` image is well suited for this preference, while `bioconductor_docker:devel` provides developers with a test environment close to *Bioconductor*'s build system.

**Data science** is a widely discussed topic in all academic disciplines (e.g., Donoho, 2017). These discussions have shed a light on the tools and craftsmanship behind the analysis of data with computational methods. The practice of data science often involves combining tools and software stacks and requires a cross-cutting skillset. This complexity and an inherent concern for openness and reproducibility in the data science community has led to Docker being used widely. The remainder of this section presents example Docker images and image stacks featuring R intended for data science.

- The *Jupyter Docker Stacks* project is a set of ready-to-run Docker images containing Jupyter applications and interactive computing tools (Jupyter, 2018). The `jupyter/r-notebook` image includes R and "popular packages", and naturally also the *IRKernel* (<https://irkernel.github.io/>), an R kernel for Jupyter, so that Jupyter Notebooks can contain R code cells. R is also included in the catchall `jupyter/datascience-notebook` image<sup>10</sup>. For example, these images allow users to quickly start a Jupyter Notebook server locally or build their own specialised images on top of stable toolsets. R is installed using the Conda package manager<sup>11</sup>, which can

<sup>9</sup>See [https://github.com/Bioconductor/bioconductor\\_docker](https://github.com/Bioconductor/bioconductor_docker) and <https://hub.docker.com/u/bioconductor> respectively.

<sup>10</sup><https://jupyter-docker-stacks.readthedocs.io/en/latest/using/selecting.html>

<sup>11</sup><https://conda.io/>



manage environments for various programming languages, pinning both the R version and the versions of R packages<sup>12</sup>.

- *Kaggle* provides the [gcr.io/kaggle-images/rstats](https://gcr.io/kaggle-images/rstats) image (previously `kaggle/rstats`) and corresponding `Dockerfile` for usage in their Machine Learning competitions and easy access to the associated datasets. It includes machine learning libraries such as Tensorflow and Keras (see also image `rocker/ml` in Section [Common or public work environments](#)), and it also configures the `reticulate` package (Ushey et al., 2019). The image uses a base image with *all packages from CRAN*, `gcr.io/kaggle-images/rcran`, which requires a Google Cloud Build because Docker Hub would time out<sup>13</sup>. The final extracted image size is over 25 GB, which calls into question whether having everything available is actually convenient.
- The *Radiant* project provides several images, e.g., `vnijs/rsm-msba-spark`, for their browser-based business analytics interface based on *Shiny* (Chang et al., 2019), and for use in education as part of an MSc course<sup>14</sup>. As data science often applies a multitude of tools, this image favours inclusion over selection and features Python, Postgres, JupyterLab and Visual Studio Code besides R and RStudio, bringing the image size up to 9 GB.
- *Gigantum* (<http://gigantum.com/>) is a platform for open and decentralized data science with a focus on using automation and user-friendly tools for easy sharing of reproducible computational workflows. Gigantum builds on the *Gigantum Client* (running either locally or on a remote server) for development and execution of data-focused *Projects*, which can be stored and shared via the *Gigantum Hub* or via a zipfile export. The Client is a user-friendly interface to a backend using Docker containers to package, build, and run Gigantum projects. It is configured to use a default set of Docker base images (<https://github.com/gigantum/base-images>), and users are able to define and configure their own custom images. The available images include two with R based on Ubuntu Linux and these have the `c2d4u CRAN PPA` pre-configured for installation of binary R packages<sup>15</sup>. The R images vary in the included authoring environment, i.e., Jupyter in `r-tidyverse` or both Jupyter & RStudio in `rstudio-server`. The independent image stack can be traced back to the Gigantum environment and its features. The R images are based on Gigantum's `python3-minimal` image, originally to keep the existing front-end configuration, but also to provide a consistent Python-to-R interoperability. The Dockerfiles also use build args to specify bases, for example for different versions of NVIDIA CUDA for GPU processing<sup>16</sup>, so that appropriate GPU drivers can be enabled automatically when supported. Furthermore, Gigantum's focus lies on environment management via GUI and ensuring a smooth user interaction, e.g., with reliable and easy conflict detection and resolution. For this reason, project repositories store authoritative package information in a separate file per package, allowing Git to directly detect conflicts and changes. A Dockerfile is generated from this description that inherits from the specified base image, and additional custom Docker instructions may be appended by users, though Gigantum's default base images do not currently include the `littler` tool, which is used by `Rocker` to install packages within Dockerfiles. Because of these specifics, instructions from `rocker/r-ubuntu` could *not* be readily re-used in this image stack (see Section [Conclusions](#)). Both approaches enable the `apt` package manager (Wikipedia contributors, 2020a) as an installation method, and this is exposed via the GUI-based environment management<sup>17</sup>. The image build and publication process is scripted with Python and JSON template configuration files, unlike `Rocker` images which rely on plain Dockerfiles. A further reason in the creation of an independent image stack were project constraints requiring a `Rocker`-incompatible licensing of the Dockerfiles, i.e., the MIT License.

## Capture and create environments

Community-maintained images provide a solid basis so users can meet their own individual requirements. Several second-order R packages attempt to streamline the process of creating Docker images and using containers for specific tasks, such as running tests or rendering reproducible reports. While authoring and managing an environment with Docker by hand is possible and feasible for experts<sup>18</sup>,

<sup>12</sup>See `jupyter/datascience-notebook`'s `Dockerfile` at <https://github.com/jupyter/docker-stacks/blob/master/datascience-notebook/Dockerfile#L47>.

<sup>13</sup>Originally, a stacked collection of over 20 images with automated builds on Docker Hub was used, see <https://web.archive.org/web/20190606043353/http://blog.kaggle.com/2016/02/05/how-to-get-started-with-data-science-in-containers/> and <https://hub.docker.com/r/kaggle/rcran/dockerfile>

<sup>14</sup>'Dockerfile' available on GitHub: <https://github.com/radiant-rstats/docker>.

<sup>15</sup><https://docs.gigantum.com/docs/using-r>

<sup>16</sup>See [https://github.com/gigantum/base-images/blob/master/\\_templates/python3-minimal-template/Dockerfile](https://github.com/gigantum/base-images/blob/master/_templates/python3-minimal-template/Dockerfile) for the Dockerfile of `python3-minimal`.

<sup>17</sup>See <https://docs.gigantum.com/docs/environment-management>

<sup>18</sup>See, e.g., this tutorial by RStudio on how to manage environments and package versions and to ensure deterministic image builds with Docker: <https://environments.rstudio.com/docker>.

the following examples show that when environments become too cumbersome to create manually, automation is a powerful tool. In particular, the practice of *version pinning*, with system package managers for different operating systems and with packages **remotes** and **versions** or by using MRAN for R, can greatly increase the reproducibility of built images and are common approaches.

**dockerfiler** is an R package designed for building Dockerfiles straight from R (Fay, 2019). A scripted creation of a Dockerfile enables iteration and automation, for example for packaging applications for deployment (see [Deployment and continuous delivery](#)). Developers can retrieve system requirements and package dependencies to write a Dockerfile, for example, by leveraging the tools available in R to parse a DESCRIPTION file.

**containerit** (<https://github.com/o2r-project/containerit/>) attempts to take this one step further and includes these tools to automatically create a Dockerfile that can execute a given workflow (Nüst and Hinz, 2019). **containerit** accepts an R object of classes "sessionInfo" or "session\_info" as input and provides helper functions to derive these from workflows, e.g., an R script or R Markdown document, by analysing the session state at the end of the workflow. It relies on the **sysreqs** (<https://github.com/r-hub/sysreqs/>) package and its mapping of package system dependencies to platform specific installation package names<sup>19</sup>. **containerit** uses **stevedore** to streamline the user interaction and improve the created Dockerfiles, e.g., by running a container for the desired base image to extract the already available R packages.

**dockr** is a similar package focusing on the generation of Docker images for R packages, in which the package itself and all of the R dependencies, including local non-CRAN packages, are available (Kjeldgaard, 2019a,b). **dockr** facilitates the organisation of code in the R package structure and the resulting Docker image mirrors the package versions of the current R session. Users can manually add statements for non-R dependencies to the Dockerfile.

**liftr** aims to solve the problem of persistent reproducible reporting in statistical computing based on the R Markdown format (Xie et al., 2018) for dynamic documents (<https://nanx.me/liftr/>, Xiao, 2019). The irreproducibility of authoring environments can become an issue for collaborative documents and large-scale platforms for processing documents. **liftr** makes the document the main and sole workflow control file and the only file that needs to be shared between collaborators for consistent environments, e.g. demonstrated in the DockFlow project (<https://dockflow.org>). It introduces new fields to the R Markdown document header, allowing users to manually declare the dependencies, including versions, for rendering the document. The package then generates a Dockerfile from this metadata and provides a utility function to render the document inside a Docker container, i.e., `render_docker("foo.Rmd")`. An RStudio addin even allows compilation of documents with the single push of a button.

System dependencies are the domain of Docker, but for a full description of the computing environment one must also manage the R version and the R packages. R versions are available via the versioned Rocker image stack (Boettiger and Eddelbuettel, 2017). **r-online** leverages these images and provides an app for helping users to detect breaking changes between different R versions and for historic exploration of R. With a standalone NodeJS app or **r-online**, the user can compare a piece of code run in two separate versions of R. Internally, **r-online** opens one or two Docker instances with the given version of R based on Rocker images, executes a given piece of code, and returns the result to the user. Regarding R package management, this can be achieved with MRAN, or with packages such as **checkpoint** (Ooi et al., 2020) and **renv** (Ushey, 2020), which can naturally be applied within images and containers. For example, **renv** helps users to manage the state of the R library in a reproducible way, further providing isolation and portability. While **renv** does not cover system dependencies, the **renv**-based environment can be transferred into a container either by restoring the environment based on the main configuration file `renv.lock` or by storing the **renv**-cache on the host and not in the container (Ushey, 2019). With both the system dependencies and R packages consciously managed in a Docker image, users can start using containers as the *only* environment for their workflows, which allows them to work independently of physical computers<sup>20</sup> and to assert a specific degree of confidence in the stability of a developed software (cf. README.Rmd in Marwick, 2017).

## Development, debugging, and testing

Containers can also serve as useful playgrounds to create environments ad hoc or to provide very specific environments that are not needed or not easily available in day-to-day development for the purposes of developing R packages. These environments may have specific versions of R, of R extension packages, and of system libraries used by R extension packages, and all of the above in a specific combination.

<sup>19</sup>See <https://sysreqs.r-hub.io/>.

<sup>20</sup>Allowing them to be digital "nomads", cf. J. Bryan's <https://github.com/jennybc/docker-why>.

First, such containers can greatly facilitate **fixing bugs and quick evaluation**, because developers and users can readily start and later discard a container to investigate a bug report or try out a piece of software (cf. Ooms, 2017). The container does not affect their regular system. Using the Rocker images with RStudio, these disposable environments lack no development comfort (cf. Section [Packaging research reproducibly](#)). Ooms (2017) describes how `docker exec` can be used to get a root shell in a container for customisation during software evaluation. Eddelbuettel and Koenker (2019) describes an example of how a Docker container was used to debug an issue with a package only occurring with a particular version of Fortran, and using tools which are not readily available on all platforms (e.g., not on macOS).

Second, the strong integration of **system libraries in core packages** in the [R-spatial community](#) makes containers essential for stable and proactive development of common classes for geospatial data modelling and analysis. For example, GDAL (GDAL/OGR contributors, 2019) is a crucial library in the geospatial domain. GDAL is a system dependency allowing R packages such as `sf`, which provides the core data model for geospatial vector data, or `rgdal`, to accommodate users to be able to read and write hundreds of different spatial raster and vector formats (Pebesma, 2018; Bivand et al., 2019). `sf` and `rgdal` have hundreds of indirect reverse imports and dependencies and, therefore, the maintainers spend a lot of effort trying not to break them. Purpose-built Docker images are used to prepare for upcoming releases of system libraries, individual bug reports, and for the lowest supported versions of system libraries<sup>21</sup>.

Third, special-purpose images exist for identifying problems beyond the mere R code, such as **debugging R memory problems**. These images significantly reduce the barriers to following complex steps for fixing memory allocation bugs (cf. Section 4.3 in [R Core Team, 1999](#)). These problems are hard to debug and critical, because when they do occur they lead to fatal crashing processes. `rocker/r-devel-san` and `rocker/r-devel-ubsan-clang` are Docker images that have a particularly configured version of R to trace such problems with gcc and clang compilers, respectively (cf. [sanitizers](#) for examples, Eddelbuettel, 2014). `wch/r-debug` is a purpose-built Docker image with multiple instrumented builds of R, each with a different diagnostic utility activated.

Fourth, containers are useful for **testing R code** during development. To submit a package to CRAN, an R package must work with the development version of R, which must be compiled locally; this can be a challenge for some users. The **R-hub** project provides “a collection of services to help R package development”, with the package builder as the most prominent one (R-hub project, 2019). R-hub makes it easy to ensure that no errors occur, but fixing errors still often warrants a local setup, e.g., using the image `rocker/r-devel`, as is testing packages with native code, which can make the process more complex (cf. Eckert, 2018). The R-hub Docker images can also be used to debug problems locally using various combinations of Linux platforms, R versions, and compilers<sup>22</sup>. The images go beyond the configurations, or *flavours*, used by CRAN for checking packages<sup>23</sup>, e.g., with CentOS-based images, but they lack a container for checking on Windows or OS X. The images greatly support package developers to provide support on operating systems with which they are not familiar. The package **dockertest** (<https://github.com/traitecoevo/dockertest/>) is a proof of concept for automatically generating Docker files and building images specifically to run tests<sup>24</sup>. These images are accompanied by a special launch script so the tested source code is not stored in the image; instead, the currently checked in version from a local Git repository is cloned into the container at runtime. This approach clearly separates test environment, test code, and current working copy of the code. Another use case where a container can help to standardise tests across operating systems is detailed the vignettes of the package **RSelenium** (Harrison, 2019). The package recommends Docker for running the **Selenium** Server application needed to execute test suites on browser-based user interfaces and webpages, but it requires users to manually manage the containers.

Fifth, Docker images can be used on **continuous integration (CI) platforms** to streamline the testing of packages. Ye (2019) describes how they speed up the process of testing by running tasks on **Travis CI** within a container using `docker exec`, e.g., the package check or rendering of documentation. Cardozo (2018) also saved time with Travis CI by re-using the testing image as the basis for an image intended for publication on Docker Hub. `r-ci` is, in turn, used with **GitLab CI**, which itself is built on top of Docker images: the user specifies a base Docker image and control code, and the whole set of tests is run inside a container. The `r-ci` image stack combines rocker versioning and a series of tools specifically designed for testing in a fixed environment with a customized list of preinstalled packages. Especially for long-running tests or complex system dependencies, these approaches to separate installation of build dependencies with code testing streamline the development process. Containers

<sup>21</sup>Cf. <https://github.com/r-spatial/sf/tree/master/inst/docker>, [https://github.com/Nowosad/rsatial\\_proj6](https://github.com/Nowosad/rsatial_proj6), and <https://github.com/r-spatial/sf/issues/1231>

<sup>22</sup>See <https://r-hub.github.io/rhub/articles/local-debugging.html> and <https://blog.r-hub.io/2019/04/25/r-devel-linux-x86-64-debian-clang/>

<sup>23</sup>[https://cran.r-project.org/web/checks/check\\_flavors.html](https://cran.r-project.org/web/checks/check_flavors.html)

<sup>24</sup>**dockertest** is not actively maintained, but mentioned still because of its interesting approach.

can also simplify the integration of R software into larger, multi-language CI pipelines. Furthermore, with each change, even this manuscript is rendered into a PDF and deployed to a GitHub-hosted website (see `.travis.yml` and `Dockerfile` in the manuscript repository), not because of concern about time, but to control the environment used on a CI server. This gives, on the one hand, easy access after every update of the R Markdown source code and, on the other hand a second controlled environment to make sure that the article renders successfully and correctly.

## Processing

The portability of containerised environments becomes particularly useful for improving expensive processing of data or shipping complex processing pipelines. First, it is possible to **offload complex processing to a server** or clouds and also to execute processes in parallel to speed up or to serve many users. **batchtools** provides a parallel implementation of the Map function for various schedulers (Lang et al., 2017). For example, the package can **schedule jobs with Docker Swarm**. **googleComputeEngineR** has the function `gce_vm_cluster()` to create clusters of 2 or more virtual machines, running multi-CPU architectures (Edmondson, 2019). Instead of running a local R script with the local CPU and RAM restrictions, the same code can be processed on all CPU threads of the cluster of machines in the cloud, all running a Docker container with the same R environments. **googleComputeEngineR** integrates with the R parallelisation package **future** (Bengtsson, 2020a) to enable this with only a few lines of R code<sup>25</sup>. **Google Cloud Run** is a CaaS (Containers as a Service) platform. Users can launch containers using any Docker image without worrying about underlying infrastructure in a so-called serverless configuration. The service takes care of network ingress, scaling machines up and down, authentication, and authorisation—all features which are non-trivial for a developer to build and maintain on their own. This can be used to scale up R code to millions of instances if need be with little or no changes to existing code, as demonstrated by the proof of concept `cloudRunR`<sup>26</sup>, which uses Cloud Run to create a scalable R-based API using **plumber** (Trestle Technology, LLC, 2018). **Google Cloud Build** and the Google Container Registry are a continuous integration service and an image registry, respectively, that offload building of images to the cloud, while serving the needs of commercial environments such as private Docker images or image stacks. As Google Cloud Build itself can run any container, the package **googleCloudRunner** demonstrates how R can be used as the control language for one-time or batch processing jobs and scheduling of jobs<sup>27</sup>. **drake** is a workflow manager for data science projects (Landau, 2018). It features implicit parallel computing and automated detection of the parts of the work that actually needs to be re-executed. **drake** has been demonstrated to run inside containers for high reproducibility<sup>28</sup>. Furthermore, **drake** workflows have been shown to use **future** package's function `makeClusterPSOCK()` for sending parts of the workflow to a Docker image for execution<sup>29</sup> (see package's function documentation; Bengtsson, 2020b). In the latter case, the container control code must be written by the user, and the **future** package ensures that the host and worker can connect for communicating over socket connections. **RStudio Server Pro** includes a functionality called **Launcher** (since version 1.2, released in 2019). It gives users the ability to spawn R sessions and background/batch jobs in a scalable way on external clusters, e.g., **Kubernetes based on Docker images** or **Slurm** clusters, and optionally, with Singularity containers. A benefit of the proprietary Launcher software is the ability for R and Python users to leverage containerisation's advantages in RStudio without writing specific deployment scripts or learning about Docker or managing clusters at all.

Second, containers are perfectly suited for **packaging and executing software pipelines** and required data. Containers allow for building complex processing pipelines that are independent of the host programming language. Due to its original use case (see **Introduction**), Docker has no standard mechanisms for chaining containers together; it lacks definitions and protocols for how to use environment variables, volume mounts, and/or ports that could enable transfer of input (parameters and data) and output (results) to and from containers. Some packages, e.g., **containerit**, provide Docker images that can be used very similar to a CLI, but this usage is cumbersome<sup>30</sup>. **outsider** (<https://docs.ropensci.org/outsider/>) tackles the problem of integrating external programs into an R workflow without the need for users to directly interact with containers (Bennett et al., 2020). Installation and usage of external programs can be difficult, convoluted and even impossible if the platform is incompatible. Therefore, **outsider** uses the platform-independent Docker images to encapsulate processes in **outsider modules**. Each outsider module has a `Dockerfile` and an R package

<sup>25</sup><https://cloudyr.github.io/googleComputeEngineR/articles/massive-parallel.html>

<sup>26</sup><https://github.com/MarkEdmondson1234/cloudRunR>

<sup>27</sup><https://code.markedmondson.me/googleCloudRunner/articles/cloudbuild.html>

<sup>28</sup>See for example <https://github.com/joelnitta/pleurosoripsis> or <https://gitlab.com/ecohealthalliance/drake-gitlab-docker-example>, the latter even running in a continuous integration platform (cf. **Development, debugging, and testing**).

<sup>29</sup><https://docs.ropensci.org/drake/index.html?q=docker#with-docker>

<sup>30</sup><https://o2r.info/containerit/articles/container.html>



with functions for interacting with the encapsulated tool. Using only R functions, an end user can install a module with the **outsider** package and then call module code to seamlessly integrate a tool into their own R-based workflow. The **outsider** package and module manage the containers and handle the transmission of arguments and the transfer of files to and from a container. These functionalities also allow a user to launch module code on a remote machine via SSH, expanding the potential computational scale. Outsider modules can be hosted code-sharing services, e.g., on GitHub, and **outsider** contains discovery functions for them.

## Deployment and continuous delivery

The cloud is the natural environment for containers, and, therefore, containers are the go-to mechanism for deploying R server applications. More and more continuous integration (CI) and continuous delivery (CD) services also use containers, opening up new options for use. The controlled nature of containers, i.e., the possibility to abstract internal software environment from a minimal dependency outside of the container is crucial, for example to match test or build environments with production environments or transfer runnable entities to as-a-service infrastructures.

First, different packages use containers for the **deployment of R and Shiny apps**. **Shiny** is a popular package for creating interactive online dashboards with R, and it enables users with very diverse backgrounds to create stable and user-friendly web applications (Chang et al., 2019). **ShinyProxy** (<https://www.shinyproxy.io/>) is an open source tool to deploy Shiny apps in an enterprise context, where it features single sign-on, but it can also be used in scientific use cases (e.g., Savini et al., 2019; Glouzon et al., 2017). ShinyProxy uses Docker containers to isolate user sessions and to achieve scalability for multi-user scenarios with multiple apps. ShinyProxy itself is written in Java to accommodate corporate requirements and may itself run in a container for stability and availability. The tool is built on ContainerProxy (<https://www.containerproxy.io/>), which provides similar features for executing long-running R jobs or interactive R sessions. The started containers can run on a regular Docker host but also in clusters. Continuous integration and deployment (CI/CD) for Shiny applications using Shinyproxy can be achieved, e.g., via **GitLab pipelines**. A pipeline can include building and checking R packages and Shiny apps. After the code has passed the checks, Docker images are built and pushed to GitLab's container registry. The pipeline finishes with triggering a webhook on the server, where the deployment script is executed. The script updates configurations, pulls the new Docker images, and restarts the services.

Another example is the package **golem**, which makes an heavy use of **dockerfiler** when it comes to creating the Dockerfile for building and deploying production-grade Shiny applications (Guyader et al., 2019). **googleComputeEngineR** enables quick deployments of key R services, such as RStudio and Shiny, onto cloud virtual machines (VMs) with Google Cloud Compute Engine (Edmondson, 2019). The package utilises Dockerfiles to move the labour of setting up those services from the user to a premade Docker image, which is configured and run in the cloud VM. For example, by specifying the template `template="rstudio"` in functions `gce_vm_template()` and `gce_vm()` an up-to-date RStudio Server image is launched for development work, whereas specifying `template="rstudio-gpu"` will launch an RStudio Server image with a GPU attached, etc.

Second, containers can be used to create **platform installation packages** in a DevOps setting. The **OpenCPU** system provides an HTTP API for data analysis based on R. Ooms (2017) describes how various platform-specific installation files for OpenCPU are created using Docker Hub. The automated builds install the software stack from the source code on different operating systems; afterwards a script file downloads the images and extracts the OpenCPU binaries.

Third, containers can greatly facilitate the **deployment to existing infrastructures**. **Kubernetes** (<https://kubernetes.io/>) is a container-orchestration system for managing container-based application deployment and scaling. A *cluster* of containers, orchestrated as a single deployment, e.g., with Kubernetes, can mitigate limitations on request volumes or a container occupied with computationally intensive task. A cluster features load-balancing, autoscaling of containers across numerous servers (in the cloud or on premise), and restarting failed ones. Many organisations already use a Kubernetes cluster for other applications, or a managed cluster can be acquired from service providers. Docker containers are used within Kubernetes clusters to hold native code, for which Kubernetes creates a framework around network connections and scaling of resources up and down. Kubernetes can thereby host R applications, big parallel tasks, or scheduled batch jobs in a scalable way, and the deployment can even be triggered by changes to code repositories (i.e., CD, see Edmondson, 2018). The package **googleKubernetesR** (<https://github.com/RhysJackson/googleKubernetesR>) is a proof of concept for wrapping the Google Kubernetes Engine API, Google's hosted Kubernetes solution, in an easy-to-use R package. The package **analogsea** provides a way to programmatically create and destroy cloud VMs on the **Digital Ocean** platform (Chamberlain et al., 2019). It also includes R wrapper functions to install Docker in such a VM, manage images, and control containers straight

from R functions. These functions are translated to Docker CLI commands and transferred transparently to the respective remote machine using SSH. **AzureContainers** is an umbrella package that provides interfaces for three commercial services of Microsoft's Azure Cloud, namely **Container Instances** for running individual containers, **Container Registry** for private image distribution, and **Kubernetes Service** for orchestrated deployments. While a package like **plumber** provides the infrastructure for turning an R workflow into a web service, for production purposes it is usually necessary to take into account scalability, reliability and ease of management. AzureContainers provides an R-based interface to these features and, thereby, simplifies complex infrastructure management to a number of R function calls, given an Azure account with sufficient credit<sup>31</sup>. **Heroku** is another cloud platform as a service provider, and it supports container-based applications. **heroku-docker-r** (<https://github.com/virtualstaticvoid/heroku-docker-r>) is an independent project providing a template for deploying R applications based on Heroku's image stack, including multiple examples for interfacing R with other programming languages. Yet the approach requires a manual management of the computing environment.

Independent integrations of R for different cloud providers lead to repeated efforts and code fragmentation. To mitigate these problems and to avoid vendor lock-in motivated the **OpenFaaS project**. OpenFaaS facilitates the deployment of functions and microservices to Kubernetes or Docker Swarm. It is language agnostic and provides auto-scaling, metrics, and an API gateway. Reduced boilerplate code is achieved via templates. Template for R<sup>32</sup> are provided based on Rocker's Debian and R-hub's **r-minimal** Alpine images. The templates use **multi-stage Docker builds** to combine R base images with the OpenFaaS 'watchdog', a tiny Golang web server. The watchdog marshals an HTTP request and invokes the actual application. The R session uses **plumber** or similar packages for the API endpoint with packages and data preloaded, thus minimizing response times.

The prevalence of Docker in industry naturally leads to the use of containers with R in such settings as well, as companies already manage platforms in Docker containers. These products often entail a large amount of open source software in combination with proprietary layers adding the relevant commercialisation features. One such example is RStudio's data science platform **RStudio Team**. It allows teams of data scientists and their respective IT/DevOps groups to develop and deploy code in R and Python around the RStudio Open Source Server inside of Docker images, without requiring users to learn new tools or directly interact with containers. The best practices for **running RStudio with Docker containers** as well as **Docker images** for RStudio's commercial products are publicly available.

## Using R to power enterprise software in production environments

R has been historically viewed as a tool for analysis and scientific research, but not for creating software that corporations can rely on for production services. However, thanks to advancements in R running as a web service, along with the ability to deploy R in Docker containers, modern enterprises are now capable of having real-time machine learning powered by R. A number of packages and projects have enabled R to respond to client requests over TCP/IP and local socket servers, such as **Rserve** (Urbanek, 2019), **svSocket** (Grosjean, 2019), **rApache** and more recently **plumber** (<https://www.rplumber.io/>) and **RestRserve** (<http://restrserve.org>), which even processes incoming requests in parallel with forked processes using **Rserve**. The latter two also provide documentation for deployment with Docker or ready-to-use images with automated builds<sup>33</sup>. These software allow other (remote) processes and programming languages to interact with R and to expose R-based function in a service architecture with HTTP APIs. APIs based on these packages can be deployed with scalability and high availability using containers. This pattern of deploying code matches those used by software engineering services created in more established languages in the enterprise domain, such as Java or Python, and R can be used alongside those languages as a first-class member of a software engineering technical stack.

CARD.com implemented a web application for the optimisation of the acquisition flow and the real-time analysis of debit card transactions. The software used **Rserve** and **rApache** and was deployed in Docker containers. The R session behind **Rserve** acted as a read-only in-memory database, which was extremely fast and scalable, for the many concurrent **rApache** processes responding to the live-scoring requests of various divisions of the company. Similarly deodorised R scripts were responsible for the ETL processes and even the client-facing email, text message and push notification alerts sent in real-time based on card transactions. The related Docker images were made available at <https://github.com/cardcorp/card-rocker>. The images extended rocker/r-base and additionally entailed an SSH client and a workaround for being able to mount SSH keys from the host, Pandoc,

<sup>31</sup>See "Deploying a prediction service with Plumber" vignette for details: [https://cran.r-project.org/web/packages/AzureContainers/vignettes/vig01\\_plumber\\_deploy.html](https://cran.r-project.org/web/packages/AzureContainers/vignettes/vig01_plumber_deploy.html).

<sup>32</sup>See OpenFaaS R templates at <https://github.com/analythium/openfaas-rstats-templates>.

<sup>33</sup>See <https://www.rplumber.io/docs/hosting.html#docker>, <https://hub.docker.com/r/trestletech/plumber/> and <https://hub.docker.com/r/rexyai/restrserve/>.

the Amazon Web Services (AWS) SDK, and Java, which is required by the AWS SDK. The AWS SDK allowed for running R consumers reading from real-time data processing streams of [AWS Kinesis](#)<sup>34</sup>. The applications were deployed on Amazon Elastic Container Service (ECS). The main takeaways from using R in Docker were not only that pinning the R package versions via MRAN is important, but also that moving away from Debian testing to a distribution with long-term support can be necessary. For the use case at hand, this switch allowed for more control over upstream updates and for minimising the risk of breaking the automated builds of the Docker images and production jobs.

The AI @ T-Mobile team created a set of neural network machine learning natural language processing models to help customer care agents manage text-based messages for customers ([T-Mobile et al., 2018](#)). For example, one model quickly identifies whether a message is from a customer (see [Shiny-based demo](#) further described by [Nolis and Werdell, 2019](#)), and others tell which customers are likely to make a repeat purchase. If a data scientist creates a machine learning model and exposes it through a **plumber** API, then someone else on the marketing team could write software that sends different emails depending on that real-time prediction. The models are convolutional neural networks that use the **keras** package ([Allaire and Chollet, 2019](#)) and run in a Rocker container. The corresponding Dockerfiles are published [on GitHub](#). Since the models power tools for agents and customers, they need to have extremely high uptime and reliability. The AI @ T-Mobile team found that the models performed well, and today these models power real-time services that are called over a million times a day.

### Common or public work environments

The fact that Docker images are portable and well defined make them useful when more than one person needs access to the same computing environment. This is even more useful when some of the users do not have the expertise to create such an environment themselves, and when these environments can be run in public or using shared infrastructure. For example, *RCloud* (<https://rcloud.social>) is a cloud-based platform for data analysis, visualisation and collaboration using R. It provides a rocker/drdr base image for easy evaluation of the platform<sup>35</sup>.

The **Binder** project, maintained by the team behind Jupyter, makes it possible for users to **create and share computing environments** with others ([Jupyter et al., 2018](#)). A *BinderHub* allows anyone with access to a web browser and an internet connection to launch a temporary instance of these custom environments and execute any workflows contained within. From a reproducibility standpoint, Binder makes it exceedingly easy to compile a paper, visualize data, and run small examples from papers or tutorials without the need for any local installation. To set up Binder for a project, a user typically starts at an instance of a BinderHub and passes the location of a repository with a workspace, e.g., a hosted Git repository, or a data repository like Zenodo. Binder's core internal tool is *repo2docker*. It deterministically builds a Docker image by parsing the contents of a repository, e.g., project dependency configurations or simple configuration files<sup>36</sup>. In the most powerful case, *repo2docker* builds a given Dockerfile. While this approach works well for most run-of-the-mill Python projects, it is not so seamless for R projects. This is partly because *repo2docker* does not support arbitrary base images due to the complex auto-generation of the Dockerfile instructions.

Two approaches make using Binder easier for R users. First, **holepunch** (<https://github.com/karthik/holepunch>) is an R package that was designed to make sharing work environments accessible to novice R users based on Binder. For any R projects that use the Tidyverse suite ([Wickham et al., 2019](#)), the time and resources required to build all dependencies from source can often time out before completion, making it frustrating for the average R user. **holepunch** removes some of these limitations by leveraging Rocker images that contain the Tidyverse along special Jupyter dependencies, and only installs additional packages from CRAN and Bioconductor that are not already part of these images. It short circuits the configuration file parsing in *repo2docker* and starts with the Binder/Tidyverse base images, which eliminates a large part of the build time and, in most cases, results in a Binder instance launching within a minute. **holepunch** also creates a DESCRIPTION file for essential metadata and dependency specification, and thereby turns any project into a research compendium (see [Packaging research reproducibly](#)). The Dockerfile included with the project can also be used to launch an RStudio Server instance locally, i.e., independent of Binder, which is especially useful when more or special computational resources can be provided there. The local image usage reduces the number of separately managed environments and, thereby, reduces work and increases portability and reproducibility.

Second, the **Whole Tale** project (<https://wholetale.org>) combines the strengths of the Rocker Project's curated Docker images with *repo2docker*. Whole Tale is a National Science Foundation (NSF)

<sup>34</sup>See useR!2017 talk "Stream processing with R in AWS".

<sup>35</sup><https://github.com/att/rcloud/tree/master/docker>

<sup>36</sup>See supported file types at [https://repo2docker.readthedocs.io/en/latest/config\\_files.html](https://repo2docker.readthedocs.io/en/latest/config_files.html). For R, the

funded project developing a scalable, open source, multi-user platform for reproducible research (Brinckman et al., 2019; Chard et al., 2019b). A central goal of the platform is to enable researchers to easily create and publish executable research objects<sup>37</sup> associated with published research (Chard et al., 2019a). Using Whole Tale, researchers can create and publish Rocker-based reproducible research objects to a growing number of repositories including DataONE member nodes, Zenodo and soon Dataverse. Additionally, Whole Tale supports automatic data citation and is working on capabilities for image preservation and provenance capture to improve the transparency of published computational research artefacts (Mecum et al., 2018; McPhillips et al., 2019). For R users, Whole Tale extends the Jupyter Project's repo2docker tool to simplify the customisation of R-based environments for researchers with limited experience with either Docker or Git. Multiple options have been discussed to allow users to change the Ubuntu LTS (long-term support, currently *Bionic Beaver*) base image, buildpack-deps:bionic, used in repo2docker. Whole Tale implemented a custom RockerBuildPack<sup>38</sup>. The build pack combines a rocker/geospatial image with repo2docker's composability<sup>39</sup>. This works because both Rocker images and the repo2docker base image use distributions with APT (Wikipedia contributors, 2020a) so that the instructions created by the latter work because of the compatible shell and package manager.

In **high-performance computing**, one use for containers is to run workflows on shared local hardware where teams manage their own high-performance servers. This can follow one of several design patterns: Users may deploy containers to hardware as a work environment for a specific project, containers may provide per-user persistent environments, or a single container can act as a common multi-user environment for a server. In all cases, though, the containerised approach provides several advantages: First, users may use the same image and thus work environment on desktop and laptop computers. The first two patterns provide modularity, while the last approach is most similar to a simple shared server. Second, software updates can be achieved by updating and redeploying the container rather than by tracking local installs on each server. Third, the containerised environment can be quickly deployed to other hardware, cloud or local, if more resources are necessary or in case of server destruction or failure. In any of these cases, users need a method to interact with the containers, be it an IDE exposed over an HTTP port or command-line access via tools such as SSH. A suitable method must be added to the container recipes. The Rocker Project provides containers pre-installed with the RStudio IDE. In cases where users store nontrivial amounts of data for their projects, the data needs to persist beyond the life of the container. This may be in shared disks, attached network volumes, or in separate storage where it is uploaded between sessions. In the case of shared disks or network-attached volumes, care must be taken to match user permissions, and of course backups are still necessary.

CyVerse is an open source, NSF-funded cyberinfrastructure platform for the life sciences providing easy access to computing and storage resources (Merchant et al., 2016). CyVerse has a browser-based 'data science workbench' called the **Discovery Environment** (DE). The DE uses a combination of HTCondor and Kubernetes for orchestrating container-based analysis and integrates with external HPC, i.e., NSF-XSEDE, through TAPIS (TACC-API's). CyVerse hosts a multi-petabyte Data Store based on iRODS with shared access by its users. The DE runs Docker containers on demand, with users able to integrate bespoke containers from DockerHub or other registries (Devisetty et al., 2016). Rocker image integration in the DE is designed to provide researchers with scalable, compute-intensive, R analysis capabilities for large and complex datasets (e.g., genomics/multi-omics, GWAS, phenotypic data, geospatial data, etc.). These capabilities give users flexibility similar to Binder, but allow containers to be run on larger computational resources (RAM, CPU, Disk, GPU), and for longer periods of time (days to weeks). The Rocker Project's RStudio and Shiny are integrated into the DE by deriving new images from Rocker images<sup>40</sup>. These new images include a reverse proxy using nginx to handle communication with CyVerse's authentication system (RStudio Support, 2020); CyVerse also allows owners to invite other registered users to securely access the same instance. The CyVerse Rocker images further include tools for connecting to its Data Store, such as the CLI utility `icommands` for iRODS. CyVerse accounts are free (with some limitations for non-US users), and the **CyVerse Learning Center** provides community members with information about the platform, including training and education opportunities.

Using GPUs (graphical processing units) as a specialised hardware from containerised common work environments is also possible and useful (Haydel et al., 2015). GPUs are increasingly popular for

<sup>37</sup>In Whole Tale a *tale* is a research object that contains metadata, data (by copy or reference), code, narrative, documentation, provenance, and information about the computational environment to support computational reproducibility.

<sup>38</sup>See [https://github.com/whole-tale/repo2docker\\_wholetale](https://github.com/whole-tale/repo2docker_wholetale).

<sup>39</sup>Composability refers to the ability to combine multiple package managers and their configuration files, such as R, 'pip', and 'conda'; see Section **Common or public work environments** for details.

<sup>40</sup>See <https://github.com/cyverse-vice/> for Dockerfiles and configuration scripts; images are auto-built on DockerHub at <https://hub.docker.com/u/cyversevice>.



compute-intensive machine learning (ML) tasks, e.g., deep artificial neural networks (Schmidhuber, 2015). Although in this case containers are not completely portable between hardware environments, but the software stack for ML with GPUs is so complex to set up that a ready-to-use container is helpful. Containers running GPU software require drivers and libraries specific to GPU models and versions, and containers require a specialized runtime to connect to the underlying GPU hardware. For NVIDIA GPUs, the [NVIDIA Container Toolkit](#) includes a specialized runtime plugin for Docker and a set of base images with appropriate drivers and libraries. The Rocker Project [has a repository](#) with (beta) images based on these that include GPU-enabled versions of machine-learning R packages, e.g., `rocker/ml` and `rocker/tensorflow-gpu`.

**Teaching** is a further example of where shared computing environments and sandboxing can greatly improve the process. First, *prepared environments for teaching* are especially helpful for (a) introductory courses, where students often struggle with the first step of installation and configuration (Çetinkaya Rundel and Rundel, 2018), and (b) courses that require access to a relatively complex setup of software tools, e.g., database systems. Çetinkaya Rundel and Rundel (2018) describe how a Docker-based deployment of RStudio (i) avoided problems with troubleshooting individual students' computers and greatly increased engagement through very quickly showing tangible outcomes, e.g., a visualisation, and (ii) reduced demand on teaching and IT staff. Each student received access to a personal RStudio instance running in a container after authentication with the university login, which gives the benefits of sandboxing and the possibility of limiting resources. Çetinkaya Rundel and Rundel (2018) found that for the courses at hand, actual usage of the UI is intermittent so a single cloud-based VM with four cores and 28 GB RAM sufficed for over 100 containers. An example for mitigating *complex setups* is teaching databases. R is very useful tool for interfacing with databases, because almost every open source and proprietary database system has an R package that allows users to connect and interact with it. This flexibility is even broadened by [DBI \(R Special Interest Group on Databases \(R-SIG-DB\) et al., 2019\)](#), which allows for creating a common API for interfacing these databases, or the [dbplyr](#) package (Wickham and Ruiz, 2019), which runs [dplyr](#) (Wickham et al., 2020) code straight against the database as queries. But learning and teaching these tools comes with the cost of deploying or having access to an environment with the software and drivers installed. For people teaching R, it can become a barrier if they need to install local versions of database drivers or connect to remote instances which might or might not be made available by IT services. Giving access to a sandbox for the most common environments for teaching databases is the idea behind `r-db`, a Docker image that contains everything needed to connect to a database from R. Notably, with `r-db`, users do not have to install complex drivers or configure their machine in a specific way. The `rocker/tidyverse` base image ensures that users can also readily use packages for analysis, display, and reporting.

Second, the idea of a common environment and partitioning allows for using *containers in teaching for secure execution and automated testing* of submissions by students. [Dodona](#) is a web platform developed at Ghent University that is used to teach students basic programming skills, and it uses Docker containers to test submissions by students. This means that both the code testing the students' submissions and the submission itself are executed in a predictable environment, avoiding compatibility issues between the wide variety of configurations used by students. The containerisation is also used to shield the Dodona servers from bad or even malicious code: memory, time and I/O limits are used to make sure students cannot overload the system. The web application managing the containers communicates with them by sending configuration information as a JSON document over standard input. Every Dodona Docker image shares a `main.sh` file that passes through this information to the actual testing framework, while setting up some error handling. The testing process in the Docker containers sends back the test results by writing a JSON document to its standard output channel. In June 2019, R support was added to Dodona using an image derived from the `rocker/r-base` image that sets up the runner user and a `main.sh` file expected by Dodona<sup>41</sup>. It also installs the packages required for the testing framework and the exercises so that this does not have to happen every time a student's submission is evaluated. The actual testing of R exercises is done using a custom framework loosely based on [testthat](#) (Wickham, 2011). During the development of the testing framework, it was found that the [testthat](#) framework did not provide enough information to its reporter system to send back all the fields required by Dodona to render its feedback. Right now, multiple statistics courses are developing exercises to automate the feedback for their lab classes.

## Packaging research reproducibly

Containers provide a high degree of isolation that is often desirable when attempting to capture a specific computational environment so that others can reproduce and extend a research result. Many computationally intensive research projects depend on specific versions of original and third-party

<sup>41</sup><https://github.com/dodona-edu/docker-images/blob/master/dodona-r.dockerfile>

software packages in diverse languages, joined together to form a pipeline through which data flows. New releases of even just a single piece of software in this pipeline can break the entire workflow, making it difficult to find the error and difficult for others to reuse existing pipelines. These breakages can make the original results irreproducible and, and the chance of a substantial disruption like this is high in a multi-year research project where key pieces of third-party software may have several major updates over the duration of the project. The classical “paper” article is insufficient to adequately communicate the knowledge behind such research projects (cf. Donoho, 2010; Marwick, 2015).

Gentleman and Lang (2007) coined the term **Research Compendium** for a dynamic document together with supporting data and code. They used the R package system (R Core Team, 1999) for the functional prototype all the way to structuring, validating, and distributing research compendia. This concept has been taken up and extended<sup>42</sup>, not in the least by applying containerisation and other methods for managing computing environments—see Section **Capture and create environments**. Containers give the researcher an isolated environment to assemble these research pipelines with specific versions of software to minimize problems with breaking changes and make workflows easier to share (cf. Boettiger, 2015; Marwick et al., 2018). Research workflows in containers are safe from contamination from other activities that occur on the researcher’s computer, for example the installation of newest version of packages for teaching demonstrations or specific versions for evaluation of others’ works. Given the users in this scenario, i.e., often academics with limited formal software development training, templates and assistance with containers around research compendia is essential. In many fields we see that a typical unit of research for a container is a research report or journal article, where the container holds the compendium, or self-contained set of data (or connections to data elsewhere) and code files needed to fully reproduce the article (Marwick et al., 2018). The package **rrtools** (<https://github.com/benmarwick/rrtools>) provides a template and convenience functions to apply good practices for research compendia, including a starter Dockerfile. Images of compendium containers can be hosted on services such as Docker Hub for convenient sharing among collaborators and others. Similarly, packages such as **containerit** and **dockerfiler** can be used to manage the Dockerfile to be archived with a compendium on a data repository (e.g. Zenodo, Dataverse, Figshare, OSF). A typical compendium’s Dockerfile will pull a rocker image fixed to a specific version of R, and install R packages from the MRAN repository to ensure the package versions are tied to a specific date, rather than the most recent version. A more extreme case is the *dynverse* project (<https://dynverse.org/>), which packages over 50 computational methods with different environments (R, Python, C++, etc.) in Docker images, which can be executed from R. *dynverse* uses a CI platform (see **Development, debugging, and testing**) to build Rocker-derived images, test them, and, if the tests succeed, publish them on Docker Hub.

Future researchers can download the compendium from the repository and run the included Dockerfile to build a new image that recreates the computational environment used to produce the original research results. If building the image fails, the human-readable instructions in a Dockerfile are the starting point for rebuilding the environment. When combined with CI (see **Development, debugging, and testing**), a research compendium set-up can enable *continuous analysis* with easier verification of reproducibility and audits trails (Beaulieu-Jones and Greene, 2017).

Further safeguarding practices are currently under development or not part of common practice yet, such as preserving images (Emsley and De Roure, 2018), storing both images and Dockerfiles (cf. Nüst et al., 2017), or pinning system libraries beyond the tagged base images, which may be seen as stable or dynamic depending on the applied time scale (see discussion on `debian:testing` base image in Boettiger and Eddelbuettel, 2017). A recommendation of the recent National Academies’ report on *Reproducibility and Replicability in Science* is that journals “consider ways to ensure computational reproducibility for publications that make claims based on computations” (Committee on Reproducibility and Replicability in Science, 2019). In fields such as political science and economics, journals are increasingly adopting policies that require authors to publish the code and data required to reproduce computational findings reported in published manuscripts, subject to independent verification (Jacoby et al., 2017; Vilhuber, 2019; Alvarez et al., 2018; Christian et al., 2018; Eubank, 2016; King, 1995). Problems with the computational environment, installation and availability of software dependencies are common. R is gaining popularity in these communities, such as for creating a research compendium. In a sample of 105 replication packages published by the *American Journal of Political Science* (AJPS), over 65% use R. The NSF-funded Whole Tale project, which was mentioned above, uses the Rocker Project community images with the goal of improving the reproducibility of published research artefacts and simplifying the publication and verification process for both authors and reviewers by reducing errors and time spent specifying the environment.

<sup>42</sup>See full literature list at <https://research-compendium.science/>.

## Conclusions

This article is a snapshot of the R corner in a universe of applications built with a many-faced piece of software, Docker. Dockerfiles and Docker images are the go-to methods for collaboration between roles in an organisation, such as development and IT operations teams, and between parties in the communication of knowledge, such as research workflows or education. Docker has become synonymous with applying the concept of containerisation to solve challenges of reproducible environments, e.g., in research and in development & production, and of scalable deployments as it can easily move processing between machines (e.g., locally, a cloud provider's VM, another cloud provider's Container-as-a-Service). Reproducible environments, scalability and efficiency, and portability across platforms/infrastructures are the common themes behind R packages, use cases, and applications in this work.

The R packages and use cases presented above show the growing number of users, developers, and real-world applications in the community and the resulting innovations. But the applications also point to the challenges of keeping up with a continuously evolving landscape. The use cases contributed by co-authors also have a degree of overlap, which can be expected as a common language and understanding of good practices is still taking shape. Also, the ease with which one can create a complex software systems with Docker, such as an independent Docker image stack, to serve one's specific needs leads to parallel developments. This ease-of-DIY in combination with the difficulty of reusing parts from or composing multiple Dockerfiles is a further reason for **fragmentation**. Instructions can be outsourced into distributable scripts and then copied into the image during build, but that makes Dockerfiles harder to read. Scripts added to a Dockerfile also add a layer of complexity and increase the risk of incomplete recipes. Despite the different image stacks presented here, the pervasiveness of Docker images can be traced back to its maintainers and the user community valuing collaboration and shared starting points over impulses to create individual solutions. Aside from that, fragmentation may not be a bad sign but may instead be a reflection of a growing market that is able to sustain multiple related efforts. With the maturing of core building blocks, such as the Docker Project's suite of images, more systems will be built successfully, but they may also be built behind the curtains. Docker alone, as a flexible core technology, is not a feasible level of collaboration and abstraction. Instead, the use cases and applications observed in this work provide a more useful division.

Nonetheless, at least on the level of R packages some **consolidation** seems in order, e.g., to reduce the number of packages creating Dockerfiles from R code or controlling the Docker daemon with R code. It remains to be seen which approach to control Docker, via the Docker API as **stevedore** or via system calls as **dockyard**/**docker**/**dockr**, is more sustainable, or whether the question will be answered by the endurance of maintainers and sufficient funding. Similarly, capturing environments and their serialisation in form of a Dockerfile currently is happening at different levels of abstraction, and re-use of functionality seems reasonable, e.g., **liftr** could generate the environment with **containerit**, which in turn may use **dockfiler** for low-level R objects representing a Dockerfile and its instructions. In this consolidation of R packages, the Docker Project could play the role of a coordinating entity. Nonetheless, for the moment, it seems that the Docker Project will focus on maintaining and extending its image stacks, e.g., images for GPU-based computing and artificial intelligence. Even with coding being more and more accepted as a required and achievable skill, an easier access, for example by exposing containerisation benefits via simple user interfaces in the users' IDE, could be an important next step, since currently containerisation happens more in the background for UI-based development (e.g., a **docker/rstudio** image in the cloud). Furthermore, the maturing of the Dockerverse packages for managing containers may lead to them being adopted in situations where manual coding is currently required, e.g. in the case of **RSelenium** or **drake** (see Sections [Development, debugging, and testing](#) and [Processing](#) respectively). In some cases, e.g., for **analogsea**, the interaction with the Docker daemon may remain too specific to re-use first-order packages to control Docker.

New features which make complex workflows accessible and reproducible and the variety in packages connected with containerisation, even when they have overlapping features, are a signal and a support for a growing user base. This growth is possibly the most important goal for the foreseeable future in the Dockerverse, and, just like the Docker images have matured over years of use and millions of runs, the new ideas and prototypes will have to prove themselves. It should be noted that the dominant position is that Docker is a blessing and a curse for these goals. It might be wise to start experimenting with non-Docker containerisation tools now, e.g., R packages interfacing with other container engines, such as **podman**/**buildah**, or an R package for creating Singularity files. Such efforts might help to avoid lock-in and to design sustainable workflows based on concepts of *containerisation*, not on their implementation in Docker. If adoption of containerisation and R continue to grow, the missing pieces for a success predominantly lie in (a) coordination and documentation of activities to reduce repeated work in favour of open collaboration, (b) the sharing of lessons learned

from use cases to build common knowledge and language, and (c) a sustainable continuation and funding for development, community support, and education. A first concrete effort to work towards these missing pieces should be sustain the structure and captured status quo from this work in form of a *CRAN Task View on containerisation*.

## Author contributions

The ordering of authors following DN and DE is alphabetical. DN conceived the article idea, [initialised the formation of the writing team](#), wrote sections not mentioned below, and revised all sections. DE wrote the introduction and the section about containerisation and the Rocker Project, and reviewed all sections. DB wrote the section on **outsider**. GD contributed the CARD.com use case. RC & EH contributed to the section on interfaces for Docker in R (*dynverse* and *dynwrap*). DC contributed content on Gigantum. ME contributed to the section on processing and deployment to cloud services. CF wrote paragraphs about **r-online**, **dockerfiler**, **r-ci** and **r-db**. LK contributed content on **dockr**. SL contributed content on RStudio's usage of Docker. BM wrote the section on research compendia and made the project Binder ready. HN & JN co-wrote the section on the T-Mobile use case. KR wrote the section about **holepunch**. NR wrote paragraphs about shared work environments and GPUs. LS & NT wrote the section on Bioconductor. PS wrote the paragraphs about GitLab CI/CD pipelines with Shinyproxy and OpenFaaS templates. TS & JW wrote the section on CyVerse. CvP wrote the section on the usage of Docker containers in Dodona. CW wrote the sections on Whole Tale and contributed content about publication reproducibility audits. NX contributed content on **liftr**. All authors approved the final version. This articles was collaboratively written at <https://github.com/nuest/rockerverse-paper/>. The [contributors page](#), [commit history](#), and [discussion issues](#) provide a detailed view on the respective contributions.

## Acknowledgements

DN is supported by the project Opening Reproducible Research (**o2r**) funded by the German Research Foundation (DFG) under project number [PE 1632/17-1](#). The funders had no role in data collection and analysis, decision to publish, or preparation of the manuscript. KR was supported in part by a grant from The Leona M. and Harry B. Helmsley Charitable Trust, award number 2016PG-BRI004. LS and NT are supported by US NIH / NHGRI awards U41HG00405 and U24HG010263. CW is supported by the Whole Tale project (<https://wholetale.org>) funded by the US National Science Foundation (NSF) under award [OAC-1541450](#). We would like to thank Celeste R. Brennecka from the Science Writing Support Service of the University of Münster for her editorial support.

## Bibliography

- M. Alam, J. Rufino, J. Ferreira, S. H. Ahmed, N. Shah, and Y. Chen. Orchestration of Microservices for IoT Using Docker and Edge Computing. *IEEE Communications Magazine*, 56(9):118–123, Sept. 2018. ISSN 0163-6804, 1558-1896. doi: 10.1109/MCOM.2018.1701233. URL <https://doi.org/10.1109/MCOM.2018.1701233>. [p2]
- J. Allaire and F. Chollet. *keras: R Interface to 'Keras'*, 2019. URL <https://CRAN.R-project.org/package=keras>. R package version 2.2.5.0. [p11]
- R. M. Alvarez, E. M. Key, and L. Núñez. Research replication: Practical considerations. *PS: Political Science & Politics*, 51(2):422–426, Apr 2018. ISSN 1049-0965, 1537-5935. doi: 10.1017/S1049096517002566. URL <https://doi.org/10.1017/S1049096517002566>. [p14]
- L. A. Barba. Terminologies for Reproducible Research. *arXiv:1802.03311 [cs]*, Feb. 2018. URL <http://arxiv.org/abs/1802.03311>. arXiv: 1802.03311. [p1]
- B. K. Beaulieu-Jones and C. S. Greene. Reproducibility of computational workflows is automated using continuous analysis. *Nature Biotechnology*, advance online publication, Mar. 2017. ISSN 1087-0156. doi: 10.1038/nbt.3780. URL <https://doi.org/10.1038/nbt.3780>. [p14]
- H. Bengtsson. *future: Unified Parallel and Distributed Processing in R for Everyone*, 2020a. URL <https://CRAN.R-project.org/package=future>. R package version 1.16.0. [p8]
- H. Bengtsson. *future: Unified Parallel and Distributed Processing in R for Everyone*, 2020b. URL <https://CRAN.R-project.org/package=future>. R package version 1.16.0. [p8]



- D. Bennett, H. Hettling, D. Silvestro, R. Vos, and A. Antonelli. outsider: Install and run programs, outside of r, inside of r (under review). *Journal of Open Source Software*, 5(45):2038, 2020. doi: 10.21105/joss.02038. URL <https://doi.org/10.21105/joss.02038>. [p8]
- D. Bernstein. Containers and cloud: From LXC to docker to kubernetes. *IEEE Cloud Computing*, 1(3): 81–84, Sept. 2014. doi: 10.1109/mcc.2014.51. URL <https://doi.org/10.1109/mcc.2014.51>. [p2]
- R. Bivand, T. Keitt, and B. Rowlingson. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*, 2019. URL <https://CRAN.R-project.org/package=rgdal>. R package version 1.4-8. [p7]
- C. Boettiger. An introduction to Docker for reproducible research, with examples from the R environment. *ACM SIGOPS Operating Systems Review*, 49(1):71–79, Jan. 2015. ISSN 01635980. doi: 10.1145/2723872.2723882. URL <https://doi.org/10.1145/2723872.2723882>. [p2, 14]
- C. Boettiger and D. Eddelbuettel. An Introduction to Rocker: Docker Containers for R. *The R Journal*, 9(2):527–536, 2017. doi: 10.32614/RJ-2017-065. URL <https://doi.org/10.32614/RJ-2017-065>. [p2, 6, 14]
- C. Boettiger, R. Lovelace, M. Howe, and J. Lamb. rocker-org/geospatial, Dec. 2019. URL <https://doi.org/10.5281/zenodo.1013940>. [p2]
- A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, S. Kulasekaran, B. Ludäscher, B. D. Mecum, J. Nabrzyski, et al. Computing environments for reproducibility: Capturing the “Whole Tale”. *Future Generation Computer Systems*, 94:854–867, 2019. doi: 10.1016/j.future.2017.12.029. URL <https://doi.org/10.1016/j.future.2017.12.029>. [p12]
- R. Cannoodt and W. Saelens. *babelwhale: Talking to 'Docker' and 'Singularity' Containers*, 2019. URL <https://CRAN.R-project.org/package=babelwhale>. R package version 1.0.1. [p3]
- L. Cardozo. Faster Docker builds in Travis CI for R packages, 2018. URL <https://lecardozo.github.io/2018/03/01/automated-docker-build.html>. [p7]
- S. Chamberlain, H. Wickham, and W. Chang. *analogsea: Interface to 'Digital Ocean'*, 2019. URL <https://CRAN.R-project.org/package=analogsea>. R package version 0.7.2. [p9]
- Chan Zuckerberg Initiative, C. Boettiger, N. Ross, and D. Eddelbuettel. Maintaining Rocker: Sustainability for Containerized Reproducible Analyses, 2019. URL <https://chanzuckerberg.com/eoss/proposals/maintaining-rocker-sustainability-for-containerized-reproducible-analyses/>. [p2]
- W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2019. URL <https://CRAN.R-project.org/package=shiny>. R package version 1.4.0. [p5, 9]
- K. Chard, N. Gaffney, M. B. Jones, K. Kowalik, B. Ludäscher, T. McPhillips, J. Nabrzyski, V. Stodden, I. Taylor, T. Thelen, M. J. Turk, and C. Willis. Application of BagIt-Serialized Research Object Bundles for Packaging and Re-execution of Computational Analyses. 2019a. doi: 10.5281/zenodo.3381754. URL <https://doi.org/10.5281/zenodo.3381754>. To appear in 2019 IEEE 15th International Conference on e-Science (e-Science). [p12]
- K. Chard, N. Gaffney, M. B. Jones, K. Kowalik, B. Ludäscher, J. Nabrzyski, V. Stodden, I. Taylor, M. J. Turk, and C. Willis. Implementing computational reproducibility in the whole tale environment. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS '19, pages 17–22, 2019b. doi: 10.1145/3322790.3330594. URL <https://doi.org/10.1145/3322790.3330594>. [p12]
- T.-M. Christian, W. G. Jacoby, S. Lafferty-Hess, and T. Carsey. Operationalizing the replication standard. *International Journal of Digital Curation*, 13(1), 2018. doi: 10.2218/ijdc.v13i1.555. URL <https://doi.org/10.2218/ijdc.v13i1.555>. [p14]
- Committee on Reproducibility and Replicability in Science. *Reproducibility and Replicability in Science*. National Academies Press, 2019. ISBN 978-0-309-48616-3. doi: 10.17226/25303. URL <https://doi.org/10.17226/25303>. [p14]
- Datadog. 8 surprising facts about real Docker adoption, June 2018. URL <https://www.datadoghq.com/docker-adoption/>. [p1, 2]
- U. K. Devisetty, K. Kennedy, P. Sarando, N. Merchant, and E. Lyons. Bringing your tools to CyVerse Discovery Environment using Docker. *F1000Research*, 5:1442, Dec. 2016. ISSN 2046-1402. doi: 10.12688/f1000research.8935.3. URL <https://doi.org/10.12688/f1000research.8935.3>. [p12]

- D. Donoho. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, Oct. 2017. ISSN 1061-8600. doi: 10.1080/10618600.2017.1384734. URL <https://doi.org/10.1080/10618600.2017.1384734>. [p4]
- D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, July 2010. ISSN 1465-4644. doi: 10.1093/biostatistics/kxq028. URL <https://doi.org/10.1093/biostatistics/kxq028>. [p14]
- A. Eckert. Building and testing R packages with latest R-Devel, Feb. 2018. URL <https://alexandereckert.com/post/testing-r-packages-with-latest-r-devel/>. [p7]
- D. Eddelbuettel. *sanitizers: C/C++ source code to trigger Address and Undefined Behaviour Sanitizers*, 2014. URL <https://CRAN.R-project.org/package=sanitizers>. R package version 0.1.0. [p7]
- D. Eddelbuettel and R. Koenker. Debugging with Docker and Rocker – A Concrete Example helping on macOS, Aug. 2019. URL <http://dirk.eddelbuettel.com/blog/2019/08/05/>. [p7]
- M. Edmondson. R on Kubernetes - serverless Shiny, R APIs and scheduled scripts, May 2018. URL <https://code.markedmondson.me/r-on-kubernetes-serverless-shiny-r-apis-and-scheduled-scripts/>. [p9]
- M. Edmondson. *googleComputeEngineR: R Interface with Google Compute Engine*, 2019. URL <https://CRAN.R-project.org/package=googleComputeEngineR>. R package version 0.3.0. [p8, 9]
- I. Emsley and D. De Roure. A Framework for the Preservation of a Docker Container | International Journal of Digital Curation, 12(2), Apr. 2018. doi: 10.2218/ijdc.v12i2.509. URL <https://doi.org/10.2218/ijdc.v12i2.509>. [p14]
- N. Eubank. Lessons from a decade of replications at the Quarterly Journal of Political Science. *PS: Political Science & Politics*, 49(2):273–276, Apr 2016. ISSN 1049-0965, 1537-5935. doi: 10.1017/S1049096516000196. URL <https://doi.org/10.1017/S1049096516000196>. [p14]
- C. Fay. *dockerfiler: Easy Dockerfile Creation from R*, 2019. URL <https://CRAN.R-project.org/package=dockerfiler>. R package version 0.1.3. [p6]
- W. Felter, A. Ferreira, R. Rajamony, and J. Rubio. An updated performance comparison of virtual machines and Linux containers. In *2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 171–172, Mar. 2015. doi: 10.1109/ISPASS.2015.7095802. URL <https://doi.org/10.1109/ISPASS.2015.7095802>. [p2]
- R. FitzJohn. *stevedore: Docker Client*, 2020. URL <https://CRAN.R-project.org/package=stevedore>. R package version 0.9.3. [p3]
- GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2019. URL <https://gdal.org>. [p7]
- R. Gentleman and D. T. Lang. Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, 16(1):1–23, Mar. 2007. ISSN 1061-8600. doi: 10.1198/106186007X178663. URL <https://doi.org/10.1198/106186007X178663>. [p14]
- R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, Sept. 2004. ISSN 1474-760X. doi: 10.1186/gb-2004-5-10-r80. URL <https://doi.org/10.1186/gb-2004-5-10-r80>. [p4]
- J.-P. S. Glouzon, J.-P. Perreault, and S. Wang. Structurexplor: a platform for the exploration of structural features of RNA secondary structures. *Bioinformatics*, 33(19):3117–3120, Oct. 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx323. URL <https://doi.org/10.1093/bioinformatics/btx323>. [p9]
- P. Grosjean. *SciViews-R: A GUI API for R*. UMONS, MONS, Belgium, 2019. URL <http://www.sciviews.org/SciViews-R>. [p10]
- V. Guyader, C. Fay, S. Rochette, and C. Girard. *golem: A Framework for Robust Shiny Applications*, 2019. URL <https://CRAN.R-project.org/package=golem>. R package version 0.1. [p9]
- J. Harrison. *RSelenium: R Bindings for 'Selenium WebDriver'*, 2019. URL <https://CRAN.R-project.org/package=RSelenium>. R package version 1.7.5. [p7]

- N. Haydel, G. Madey, S. Gesing, A. Dakkak, S. G. de Gonzalo, I. Taylor, and W.-m. W. Hwu. Enhancing the Usability and Utilization of Accelerated Architectures via Docker. In *Proceedings of the 8th International Conference on Utility and Cloud Computing, UCC '15*, pages 361–367. IEEE Press, 2015. ISBN 978-0-7695-5697-0. URL <http://dl.acm.org/citation.cfm?id=3233397.3233456>. [p12]
- K. Hornik, U. Ligges, and A. Zeileis. Changes on cran. *The R Journal*, 11(1):438–441, June 2019. URL <http://journal.r-project.org/archive/2019-1/cran.pdf>. [p1]
- W. G. Jacoby, S. Lafferty-Hess, and T.-M. Christian. Should journals be responsible for reproducibility? *Inside Higher Ed*, Jul 2017. URL <https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility>. [p14]
- P. Jupyter. Jupyter Docker Stacks — docker-stacks latest documentation, 2018. URL <https://jupyter-docker-stacks.readthedocs.io/en/latest/>. [p4]
- P. Jupyter, M. Bussonnier, J. Forde, J. Freeman, B. Granger, T. Head, C. Holdgraf, K. Kelley, G. Nalvarte, A. Osheroff, M. Pacer, Y. Panda, F. Perez, B. Ragan-Kelley, and C. Willing. Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. *Proceedings of the 17th Python in Science Conference*, pages 113–120, 2018. doi: 10.25080/Majora-4af1f417-011. URL <https://doi.org/10.25080/Majora-4af1f417-011>. [p11]
- G. King. Replication, replication. *PS: Political Science & Politics*, 28(3):444–452, Sep 1995. doi: 10.2307/420301. URL <https://doi.org/10.2307/420301>. [p14]
- L. Kjeldgaard. *dockr: Creation of Lightweight Docker Images for Your Packages*, 2019a. URL <https://CRAN.R-project.org/package=dockr>. R package version 0.8.6. [p6]
- L. Kjeldgaard. ‘dockr’: easy containerization for R - pRopaganda by smaakagen, Dec. 2019b. URL <http://smaakage85.netlify.com/2019/12/21/dockr-easy-containerization-for-r/>. [p6]
- G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, May 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177459. URL <https://doi.org/10.1371/journal.pone.0177459>. [p1, 2]
- W. M. Landau. The drake r package: a pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, 3(21), 2018. URL <https://doi.org/10.21105/joss.00550>. [p8]
- M. Lang, B. Bischl, and D. Surmann. batchtools: Tools for r to work on batch systems. *The Journal of Open Source Software*, 2(10):135, 2 2017. ISSN 2475-9066. doi: 10.21105/joss.00135. URL <https://doi.org/10.21105/joss.00135>. [p8]
- B. Marwick. How computers broke science – and what we can do to fix it, Nov. 2015. URL <http://theconversation.com/how-computers-broke-science-and-what-we-can-do-to-fix-it-49938>. [p14]
- B. Marwick. Research compendium for the 1989 excavations at Madjedbebe rockshelter, NT, Australia, July 2017. URL <https://doi.org/10.6084/m9.figshare.1297059.v4>. [p6]
- B. Marwick, C. Boettiger, and L. Mullen. Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician*, 72(1):80–88, Jan. 2018. ISSN 0003-1305. doi: 10.1080/00031305.2017.1375986. URL <https://doi.org/10.1080/00031305.2017.1375986>. [p14]
- T. McPhillips, C. Willis, M. Gryk, S. Nunez-Corrales, and B. Ludäscher. Reproducibility by Other Means: Transparent Research Objects. 2019. doi: 10.5281/zenodo.3382423. URL <https://doi.org/10.5281/zenodo.3381754>. To appear in 2019 IEEE 15th International Conference on e-Science (e-Science). [p12]
- B. Mecum, M. B. Jones, D. Vieglaiss, and C. Willis. Preserving reproducibility: Provenance and executable containers in dataone data packages. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 45–49. IEEE, 2018. doi: 10.1109/eScience.2018.00019. URL <https://doi.org/10.1109/eScience.2018.00019>. [p12]
- N. Merchant, E. Lyons, S. Goff, M. Vaughn, D. Ware, D. Micklos, and P. Antin. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLOS Biology*, 14(1): e1002342, Jan. 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002342. URL <https://doi.org/10.1371/journal.pbio.1002342>. [p12]
- Microsoft. CRAN Time Machine - MRAN, 2019a. URL <https://mran.microsoft.com/timemachine>. [p2]

- Microsoft. Linux Containers on Windows, Sept. 2019b. URL <https://docs.microsoft.com/en-us/virtualization/windowscontainers/deploy-containers/linux-containers>. [p2]
- M. Morgan. *BiocManager: Access the Bioconductor Project Package Repository*, 2019. URL <https://CRAN.R-project.org/package=BiocManager>. R package version 1.30.10. [p4]
- S. Muñoz. The history of Docker's climb in the container management market, June 2019. URL <https://searchservervirtualization.techtarget.com/feature/The-history-of-Dockers-climb-in-the-container-management-market>. [p1]
- J. Nolis and J. Werdell. Small data, big value, Dec. 2019. URL <https://medium.com/tmobile-tech/small-data-big-value-f783ceca4fdb>. [p11]
- D. Nüst and M. Hinz. containerit: Generating Dockerfiles for reproducible research with R. *Journal of Open Source Software*, 4(40):1603, Aug. 2019. ISSN 2475-9066. doi: 10.21105/joss.01603. URL <https://joss.theoj.org/papers/10.21105/joss.01603>. [p6]
- D. Nüst, M. Konkol, E. Pebesma, C. Kray, M. Schutzzeichel, H. Przibytzin, and J. Lorenz. Opening the Publication Process with Executable Research Compendia. *D-Lib Magazine*, 23(1/2), Jan. 2017. ISSN 1082-9873. doi: 10.1045/january2017-nuest. URL <https://doi.org/10.1045/january2017-nuest>. [p14]
- OCI. Open Containers Initiative - About, 2019. URL <https://www.opencontainers.org/about>. [p2]
- H. Ooi. *AzureContainers: Interface to 'Container Instances', 'Docker Registry' and 'Kubernetes' in 'Azure'*, 2019. URL <https://CRAN.R-project.org/package=AzureContainers>. R package version 1.2.0. [p3]
- H. Ooi, A. de Vries, and Microsoft. *checkpoint: Install Packages from Snapshots on the Checkpoint Server for Reproducibility*, 2020. URL <https://CRAN.R-project.org/package=checkpoint>. R package version 0.4.9. [p6]
- J. Ooms. OpenCPU - Why Use Docker with R? A DevOps Perspective, Oct. 2017. URL <https://www.opencpu.org/posts/opencpu-with-docker/>. [p7, 9]
- J. Ooms. *sys: Powerful and Reliable Tools for Running System Commands in R*, 2019. URL <https://CRAN.R-project.org/package=sys>. R package version 3.3. [p3]
- E. Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1): 439–446, 2018. doi: 10.32614/RJ-2018-009. URL <https://doi.org/10.32614/RJ-2018-009>. [p7]
- R Core Team. *Writing R extensions*, 1999. URL <https://cran.r-project.org/doc/manuals/r-devel/R-exts.html>. [p7, 14]
- R-hub project. R-hub Docs, 2019. URL <https://docs.r-hub.io/>. [p7]
- R Special Interest Group on Databases (R-SIG-DB), H. Wickham, and K. Müller. *DBI: R Database Interface*, 2019. URL <https://CRAN.R-project.org/package=DBI>. R package version 1.1.0. [p13]
- RStudio Support. Running rstudio server with a proxy, Jan. 2020. URL <https://support.rstudio.com/hc/en-us/articles/200552326-Running-RStudio-Server-with-a-Proxy>. [p12]
- L. Savini, L. Candeloro, S. Perticara, and A. Conte. EpiExploreR: A Shiny Web Application for the Analysis of Animal Disease Data. *Microorganisms*, 7(12):680, Dec. 2019. doi: 10.3390/microorganisms7120680. URL <https://doi.org/10.3390/microorganisms7120680>. [p9]
- J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan. 2015. ISSN 0893-6080. doi: 10.1016/j.neunet.2014.09.003. URL <https://doi.org/10.1016/j.neunet.2014.09.003>. [p13]
- T-Mobile, J. Nolis, and H. Nolis. Enterprise Web Services with Neural Networks Using R and TensorFlow, Nov. 2018. URL <https://opensource.t-mobile.com/blog/posts/r-tensorflow-api/>. [p11]
- Trestle Technology, LLC. *plumber: An API Generator for R*, 2018. URL <https://CRAN.R-project.org/package=plumber>. R package version 0.4.6. [p8]
- S. Urbanek. *Rserve: Binary R server*, 2019. URL <https://CRAN.R-project.org/package=Rserve>. R package version 1.7-3.1. [p10]
- K. Ushey. Using renv with Docker, 2019. URL <https://rstudio.github.io/renv/articles/docker.html>. [p6]




- K. Ushey. *renv: Project Environments*, 2020. URL <https://CRAN.R-project.org/package=renv>. R package version 0.9.3. [p6]
- K. Ushey, J. Allaire, and Y. Tang. *reticulate: Interface to 'Python'*, 2019. URL <https://CRAN.R-project.org/package=reticulate>. R package version 1.14. [p5]
- L. Vilhuber. Report by the AEA Data Editor. *AEA Papers and Proceedings*, 109:718–729, May 2019. ISSN 2574-0768. doi: 10.1257/pandp.109.718. [p14]
- H. Wickham. testthat: Get started with testing. *The R Journal*, 3:5–10, 2011. URL [https://journal.r-project.org/archive/2011-1/RJournal\\_2011-1\\_Wickham.pdf](https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf). [p13]
- H. Wickham and E. Ruiz. *dbplyr: A 'dplyr' Back End for Databases*, 2019. URL <https://CRAN.R-project.org/package=dbplyr>. R package version 1.4.2. [p13]
- H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Golemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686, Nov. 2019. ISSN 2475-9066. doi: 10.21105/joss.01686. URL <https://doi.org/10.21105/joss.01686>. [p3, 11]
- H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2020. URL <https://CRAN.R-project.org/package=dplyr>. R package version 0.8.4. [p13]
- Wikipedia contributors. APT (software), Feb. 2020a. URL [https://en.wikipedia.org/w/index.php?title=APT\\_\(software\)&oldid=939802209](https://en.wikipedia.org/w/index.php?title=APT_(software)&oldid=939802209). Page Version ID: 939802209. [p5, 12]
- Wikipedia contributors. OS-level virtualization, Jan. 2020b. URL [https://en.wikipedia.org/w/index.php?title=OS-level\\_virtualization&oldid=935110975](https://en.wikipedia.org/w/index.php?title=OS-level_virtualization&oldid=935110975). Page Version ID: 935110975. [p2]
- N. Xiao. *liftr: Containerize R Markdown Documents for Continuous Reproducibility*, 2019. URL <https://CRAN.R-project.org/package=liftr>. R package version 0.9.2. [p6]
- Y. Xie, J. J. Allaire, and G. Golemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, 2018. [p6]
- H. Ye. Docker Setup for R package Development, 2019. URL <https://haoye.us/post/2019-10-10-docker-for-r-package-development/>. [p7]
- M. Çetinkaya Rundel and C. Rundel. Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum. *The American Statistician*, 72(1):58–65, Jan. 2018. ISSN 0003-1305. doi: 10.1080/00031305.2017.1397549. URL <https://doi.org/10.1080/00031305.2017.1397549>. [p13]

Daniel Nüst  
University of Münster  
Institute for Geoinformatics  
Heisenbergstr. 2  
48149 Münster, Germany  
0000-0002-0024-5046  
[daniel.nuest@uni-muenster.de](mailto:daniel.nuest@uni-muenster.de)

Dirk Eddelbuettel  
University of Illinois at Urbana-Champaign  
Department of Statistics  
Illini Hall, 725 S Wright St  
Champaign, IL 61820, USA  
0000-0001-6419-907X  
[dirk@eddelbuettel.com](mailto:dirk@eddelbuettel.com)

Dom Bennett  
Gothenburg Global Biodiversity Centre, Sweden  
Carl Skottsbergs gata 22B  
413 19 Göteborg, Sweden  
0000-0003-2722-1359  
[dominic.john.bennett@gmail.com](mailto:dominic.john.bennett@gmail.com)

Robrecht Cannoodt  
Ghent University  
Data Mining and Modelling for Biomedicine group  
VIB Center for Inflammation Research  
Technologiepark 71  
9052 Ghent, Belgium  
 0000-0003-3641-729X  
[robrecht@cannoodt.dev](mailto:robrecht@cannoodt.dev)

Dav Clark  
Gigantum, Inc.  
1140 3rd Street NE  
Washington, D.C. 20002, USA  
 0000-0002-3982-4416  
[dav@gigantum.com](mailto:dav@gigantum.com)

Gergely Daroczi  
 0000-0003-3149-8537  
[daroczig@rapporter.net](mailto:daroczig@rapporter.net)


Mark Edmondson  
IIH Nordic A/S, Google Developer Expert for GCP  
[mark@markedmondson.me](mailto:mark@markedmondson.me)

Colin Fay  
ThinkR  
50 rue Arthur Rimbaud  
93300 Aubervilliers, France  
 0000-0001-7343-1846  
[contact@colinfay.me](mailto:contact@colinfay.me)

Ellis Hughes  
Fred Hutchinson Cancer Research Center  
Vaccine and Infectious Disease  
1100 Fairview Ave. N., P.O. Box 19024  
Seattle, WA 98109-1024, USA  
[ehhughes@fredhutch.org](mailto:ehhughes@fredhutch.org)

Lars Kjeldgaard  
Danish Tax Authorities  
Oestbanegade 123  
2100, København Oe  
[lars\\_kjeldgaard@hotmail.com](mailto:lars_kjeldgaard@hotmail.com)

Sean Lopp  
RStudio, Inc  
250 Northern Ave  
Boston, MA 02210, USA  
[sean@rstudio.com](mailto:sean@rstudio.com)

Ben Marwick  
University of Washington  
Department of Anthropology  
Denny Hall 230, Spokane Ln  
Seattle, WA 98105, USA  
 0000-0001-7879-4531  
[bmarwick@uw.edu](mailto:bmarwick@uw.edu)

Heather Nolis  
T-Mobile

12920 Se 38th St.  
Bellevue, WA, 98006, USA  
[heather.wensler1@t-mobile.com](mailto:heather.wensler1@t-mobile.com)

Jacqueline Nolis  
Nolis, LLC  
Seattle, WA, USA  
 0000-0001-9354-6501  
[jacqueline@nolisllc.com](mailto:jacqueline@nolisllc.com)

Hong Ooi  
Microsoft  
Level 5, 4 Freshwater Place  
Southbank, VIC 3006, Australia  
[hongooi@microsoft.com](mailto:hongooi@microsoft.com)

Karthik Ram  
Berkeley Institute for Data Science  
University of California  
Berkeley, CA 94720, USA  
 0000-0002-0233-1757  
[karthik.ram@berkeley.edu](mailto:karthik.ram@berkeley.edu)

Noam Ross  
EcoHealth Alliance  
460 W 34th St., Ste. 1701  
New York, NY 10001, USA  
 0000-0002-0233-1757  
[ross@ecohealthalliance.org](mailto:ross@ecohealthalliance.org)

Lori Shepherd  
Roswell Park Comprehensive Cancer Center  
Elm & Carlton Streets  
Buffalo, NY, 14263, USA  
 0000-0002-5910-4010  
[lori.shepherd@roswellpark.org](mailto:lori.shepherd@roswellpark.org)


Péter Sólymos  
Analythium Solutions  
#258 150 Chippewa Road  
Sherwood Park, AB, T8A 6A2, Canada  
 0000-0001-7337-1740  
[peter@analythium.io](mailto:peter@analythium.io)

Tyson Lee Swetnam  
University of Arizona  
1657 E Helen St.  
Tucson, AZ, 85721, USA  
 0000-0002-6639-7181  
[tswetnam@arizona.edu](mailto:tswetnam@arizona.edu)

Nitesh Turaga  
Roswell Park Comprehensive Cancer Center  
Elm & Carlton Streets  
Buffalo, NY, 14263, USA  
 0000-0002-0224-9817  
[nitesh.turaga@roswellpark.org](mailto:nitesh.turaga@roswellpark.org)

Charlotte Van Petegem  
Ghent University  
Department WE02  
Krijgslaan 281, S9

9000 Gent, Belgium

 0000-0003-0779-4897

[charlotte.vanpetegem@ugent.be](mailto:charlotte.vanpetegem@ugent.be)

Jason Williams

Cold Spring Harbor Laboratory

1 Bungtown Rd.

Cold Spring Harbor, NY, 11724, USA

 0000-0003-3049-2010


[williams@cshl.edu](mailto:williams@cshl.edu)

Craig Willis

University of Illinois at Urbana-Champaign

501 E. Daniel St.

Champaign, IL 61820, USA

 0000-0002-6148-7196

[willis8@illinois.edu](mailto:willis8@illinois.edu)

Nan Xiao

Seven Bridges Genomics

529 Main St, Suite 6610

Charlestown, MA 02129, USA

 0000-0002-0250-5673

[me@nanx.me](mailto:me@nanx.me)