

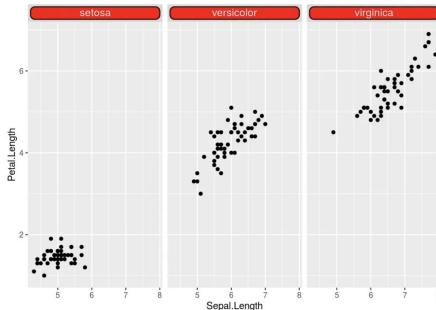


Using Pins to Ensure Reproducibility with Datasets

@javierluraschi
useR 2020 - RStudio PBC

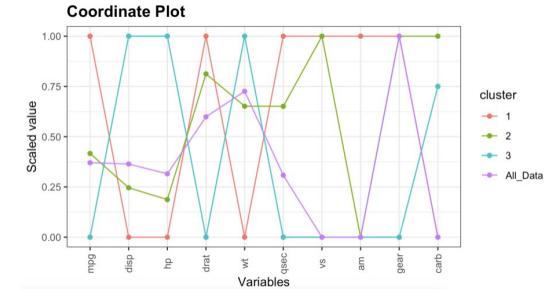
Data Workflows - Today

June 16th posts in R-bloggers use iris, mtcars, have missing data paths and require manually downloading datasets.



```
#####
# ## IMPORT RAW DATA
#
log_info("Loading data")
mydat = fread(data_path)
```

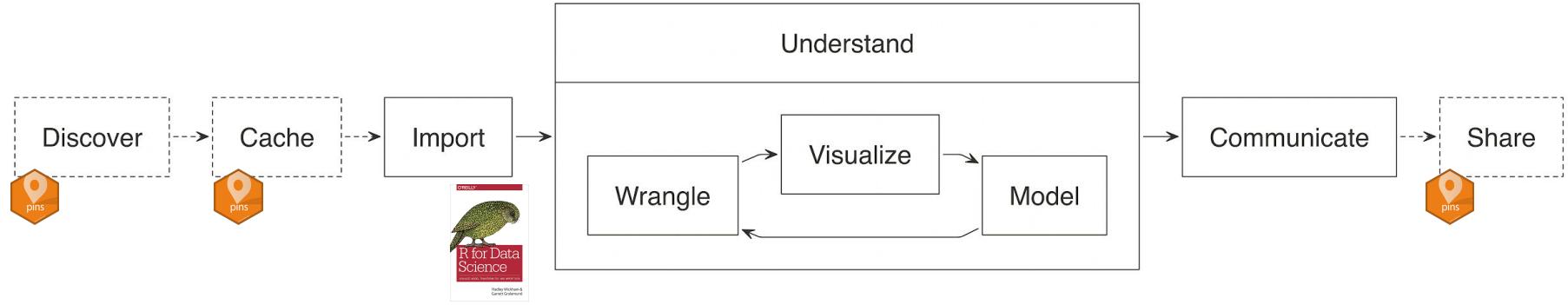
```
coord_plot(data=mtcars2, group_var="cluster",
group_func=median, print_table=TRUE)
```



```
# https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/
fichier_covid <- "donnees/covid.csv"
```

Data Workflows - In Data Science

We know from R for Data Science that in a typical project we usually need to import, tidy, understand and communicate knowledge.



However, it is often also required for us to discover which dataset to use, cache it locally, and share our datasets with colleagues.

Data Workflows - Reproducibility

But more importantly, properly caching and sharing our datasets allows us to make Data Science more reproducible in tools like R Markdown and Jupyter.

The collage includes:

- A Jupyter logo with a hexagonal pattern background.
- A screenshot of the Jupyter Notebook interface showing code cells and output.
- A screenshot of the Jupyter website (jupyter.org) featuring the Jupyter logo and navigation links.
- A screenshot of the R Markdown from R Studio interface, showing a plot of a volcano dataset.
- A large central image showing a plot of the Lorenz system with mathematical equations above it.
- A screenshot of the R Markdown website (rmarkdown.rstudio.com) with the tagline "Analyze. Share. Reproduce."
- A screenshot of the RStudio IDE showing a "Merge With Volcano" plot and the R console.

Language of choice
Jupyter supports over 40 programming languages, including Python, R, Julia, and Scala.

R Markdown documents are fully reproducible. Use a productive [notebook interface](#) to weave together narrative text and code to produce elegantly formatted output. Use [multiple languages](#) including R, Python, and SQL.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Your data tells a story. Tell it with R Markdown. Turn your analyses into high quality documents, reports, presentations and dashboards.

Data Workflows - Ideally



- Learn a **single tool** regardless of where data lives.
- Easy **switch storage** and share it anywhere.
- Easy to build **automated workflows**.
- Can share datasets of **any size**.
- Can easily use **interesting datasets!**
- Data loads **fast** and works **offline**.

What is the package?



- **Pin** remote resources locally with `pin()`, work offline and cache results.
- **Discover** new resources across different boards using `pin_find()`.
- **Share resources** in local folders, GitHub, Kaggle, and RStudio Connect by registering new boards with `board_register()`.

Links

Download from CRAN at
<https://cloud.r-project.org/package=pins>

Browse source code at
<https://github.com/rstudio/pins>

Report a bug at
<https://github.com/rstudio/pins/issues>

License

Apache License 2.0

Developers

Javier Luraschi
Author, maintainer

[All authors...](#)

Dev status

build passing

CRAN 0.4.1

codecov 77%

downloads 3954/month

lifecycle maturing

chat on gitter

stars 107

How can I use Pins?

```
library(pins)

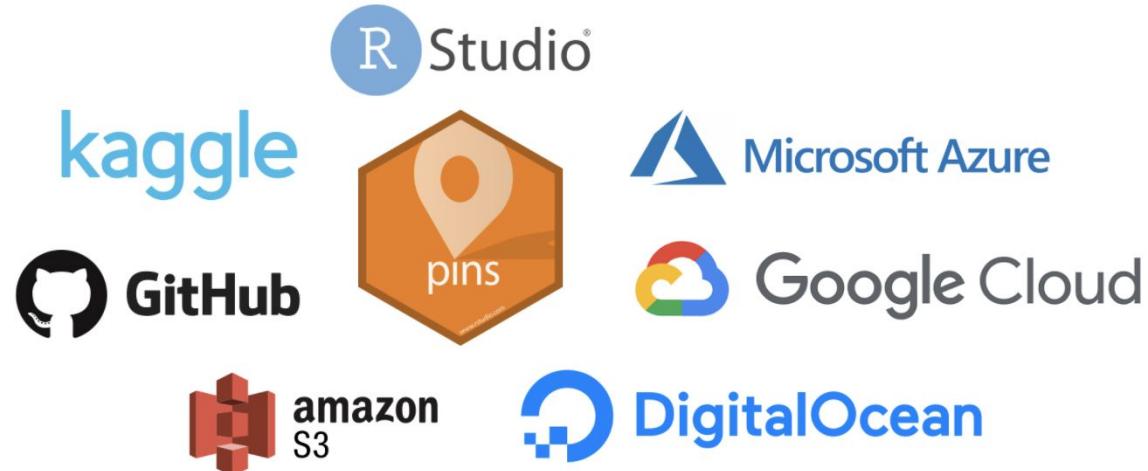
c(2, 3, 5, 7, 11, 13, 17, 19) %>% pin(name = "primes")
pin_get("primes")

pin("http://cs231n.stanford.edu/tiny-imagenet-200.zip ", name = "tinyimagenet")
pin_get("tinyimagenet")

pin_find()
```

What are Boards?

A board is a storage location, like your local hard drive, but there are also many additional places where you can store your datasets like RStudio Connect, Microsoft Azure, Google Cloud, Digital Ocean, Amazon S3, GitHub and Kaggle.



How can I use Boards?

```
library(pins)
board_register("_____")

c(2, 3, 5, 7, 11, 13, 17, 19) %>% pin(name = "primes", board = "_____")
pin_get("primes")

pin("http://cs231n.stanford.edu/tiny-imagenet-200.zip",
    name = "tinyimagenet", board = "_____")
pin_get("tinyimagenet")

pin_find()
```

Boards and Pins

Pins works in the same way across all storage services; however, each service provides a slightly different user interface, permissions, workflows, etc.

The image displays four separate screenshots illustrating the 'pins' interface across different platforms:

- Google Sheets:** Shows a table titled 'mtcars' with columns: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb. A sidebar on the left shows 'My Pins' and 'Recent'. A note at the top says 'This document is shared with "javierluraschi".' A 'Share' button is visible at the top right.
- AWS S3:** Shows the 'pinscontainer' bucket in the 'Amazon S3' console. It lists objects: 'iris' (size 2 KB), 'mtcars' (size 3.1 KB), and 'data.txt' (size 206.0 KB). The 'Actions' dropdown menu is open. The top navigation bar includes 'Services', 'Resource Groups', and 'Javier - Global - Support'.
- Databricks:** Shows a 'Datasets' page with three datasets: 'The iris data set' (public, size 2 KB), 'The motor trend cars data set' (private, size 2 KB), and 'YOUR DATASETS' (private, size 3.1 KB). A search bar and filter button are at the top. The bottom right corner says 'End of results'.
- Github:** Shows a repository page for 'javierluraschi/mtcars'. Key statistics are: 466 commits, 2 branches, 0 releases, and 1 contributor. A 'Branch: tests' dropdown and a 'New pull request' button are at the top. Below, a message says 'This branch is 464 commits ahead, 1 commit behind master.' A list of recent commits includes:
 - javierluraschi update mtcars (Latest commit 7274a18 12 seconds ago)
 - iris update iris (20 seconds ago)
 - mtcars update mtcars (13 seconds ago)
 - data.txt update mtcars (12 seconds ago)Buttons for 'Create new file', 'Upload files', 'Find File', and 'Clone or download' are at the bottom.

Website Boards

Pins uses datatxt.org to describe datasets, a simple YAML file to locate resources and optional metadata. You can then auto-generate data websites (like cellar.kasa.ai) with the datatxt package.

Cellar Datasets

Summary

Datasets

- [FIMA NFIP Redacted Claims Data Set](#)
- [French Third-Party Liability \(Claims\)](#)
- [French Third-Party Liability \(Policies\)](#)
- [Schedule P Data](#)
- [SOA Lapse Study Data](#)

© Kasa AI 2020

Name	Piped data
Number of rows	26639
Number of columns	2

Column type frequency:

numeric	2
---------	---

Group variables

None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
policy_id	0	1	2279863.83	1577201.81	139	1087642.50	2137413	3180162.00	6113971	
claim_amount	0	1	2278.54	29297.48	1	686.81	1172	1228.08	4075401	

Source

R package 'CASdatasets', <http://dutangc.free.fr/pub/RRepos/web/CASdatasets-index.html>

Versioning

You can optionally enable or disable tracking versions, some boards support this by default, in others you opt-in when the board is registered.

```
board_register("github", repo = "javierluraschi/datasets", branch = "datasets")

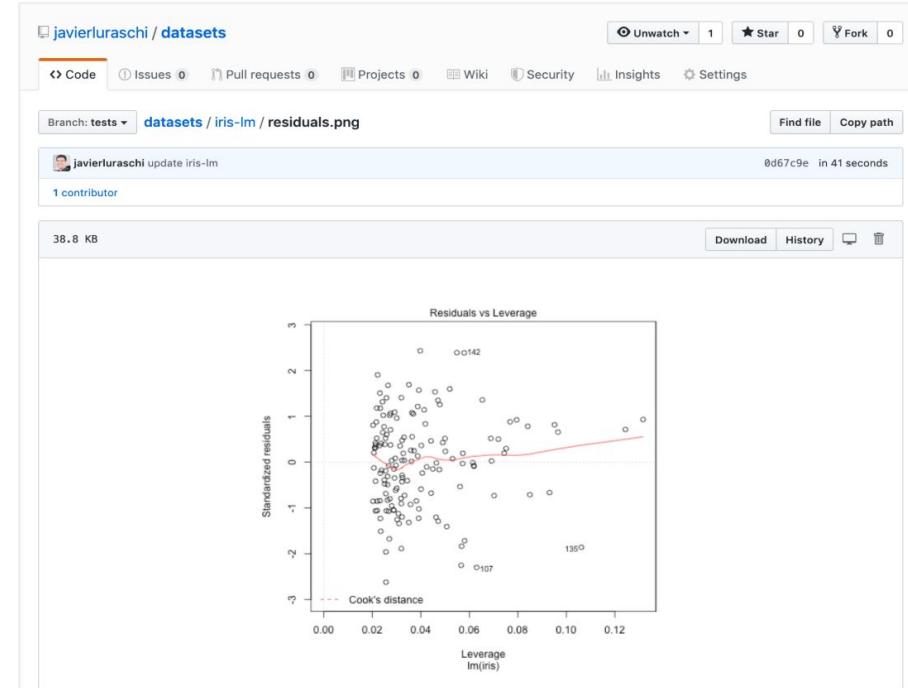
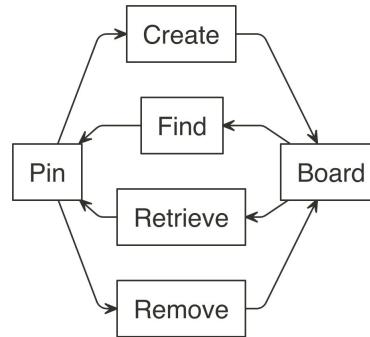
pin(iris, name = "versioned", board = "github", commit = "use iris...")
pin(mtcars, name = "versioned", board = "github", commit = "slight
preference...")

pin_versions("versioned", board = "github")
```

```
# A tibble: 2 x 4
  version created      author      message
  <chr>    <chr>        <chr>        <chr>
1 6e6c320 2020-04-02T21:28:07Z javierluraschi slight preference to mtcars
2 01f8ddf 2020-04-02T21:27:59Z javierluraschi use iris as the main dataset
```

Extending Pins and Boards

You can store R objects with specific file formats and also create new boards by implementing the pins API.



Thank You!

pins.rstudio.com

pins: Pin, Discover and Share Resources

Overview
You can use the `pins` package to:

- Pin remote resources locally with `pin()`, work offline and cache results.
- Discover new resources across different boards using `pin_find()`.
- Share resources in local folders, GHHub, Kaggle, and RStudio Connect by registering new boards with `board_register()`.

Installation
Install the released version from CRAN:
`install.packages("pins")`

To get a bug fix, or use a feature from the development version, you can install pins from GHHub.

```
# install.packages("remotes")
remotes::install_github("rstudio/pins")
```

Usage
`library(pins)`

Links
Downloaded from CRAN at <https://cloud.r-project.org/package=pins>
Browse source code at <https://github.com/rstudio/pins>
Report a bug at <https://github.com/rstudio/pins/issues>

License
Apache License 2.0

Developers
Javier Luraschi
Author, maintainer
All authors...

Dev status
 CHAN: X.4.1
 codecov: 77%
 downloads: 1054/month
 R package info
 stars: 107

blogs.rstudio.com/ai

RStudio AI Blog

Towards privacy: Encrypted deep learning with Syft and Keras
April 28, 2020

Deep learning need not be incompatible with privacy protection. Federated learning enables on-device distributed model training; encryption keeps model and gradients private; differential privacy protects the training data from leaking. As of today, private and secure deep learning is an emerging technology. In this post, we introduce Syft, a federated learning library that integrates with PyTorch as well as TensorFlow. In our example use case, we obtain private predictions from a Keras model.

sparklyr 1.2: ForEach, Spark 3.0 and Databricks Connect
April 21, 2020

A new sparklyr release is now available. This sparklyr 1.2 release features new functionalities such as support for Databricks Connect, a Spark backend for the `foreach` package, interop improvements for working with Spark 3.0 preview, and many other minor bug fixes and improvements addressing user-visible pain points.

youtube.com/c/mlverse

mlverse
495 subscribers

Home **VIDEOS** **PLAYLISTS** **CHANNELS** **DISCUSSION** **ABOUT**

Created playlists

TensorFlow	Spark	Datasets	MLflow	Visualization
Updated yesterday	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST
TensorFlow	Spark	Datasets	MLflow	Visualization
1	17	8	2	3

SUBSCRIPTIONS

TensorFlow	Spark	Datasets	MLflow	Visualization
Popular on YouTube	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST	VIEW FULL PLAYLIST
Music	1	1	1	1
Sports	1	1	1	1
Gaming	1	1	1	1

MORE FROM YOUTUBE

Meta, and all that	Platform	Loved Videos	Linux	TensorFlow Probability
VIEW FULL PLAYLIST				

pins.rstudio.com

blogs.rstudio.com/ai

youtube.com/c/mlverse



@zkajdan



@dfalbel



@javierluraschi



@yl790