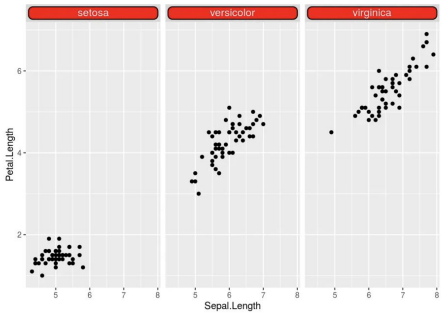
Using Pins to Ensure Reproducibility with Datasets

@javierluraschi

useR 2020 - RStudio PBC

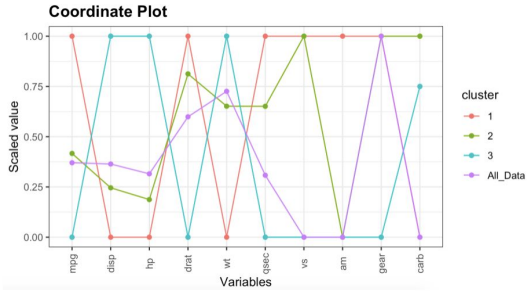
Data Workflows - Today

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




```
32.  ### IMPORT RAW DATA
33.
34.  log_info("Loading data")
35.
36.  mydat = fread(data_path)
37.
```

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

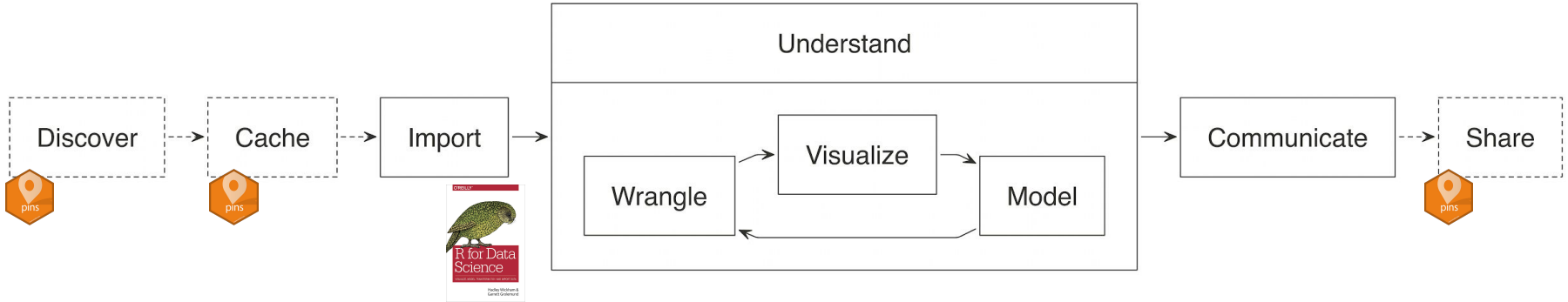



```
# https://www.data.gouv.fr/fr/datasets/donnees-
hospitalieres-relatives-a-lepidemie-de-covid-19/
fichier_covid <- "donnees/covid.csv"
```

R-bloggers
R news and tutorials contributed by hundreds of R bloggers

Data Workflows - In Data Science

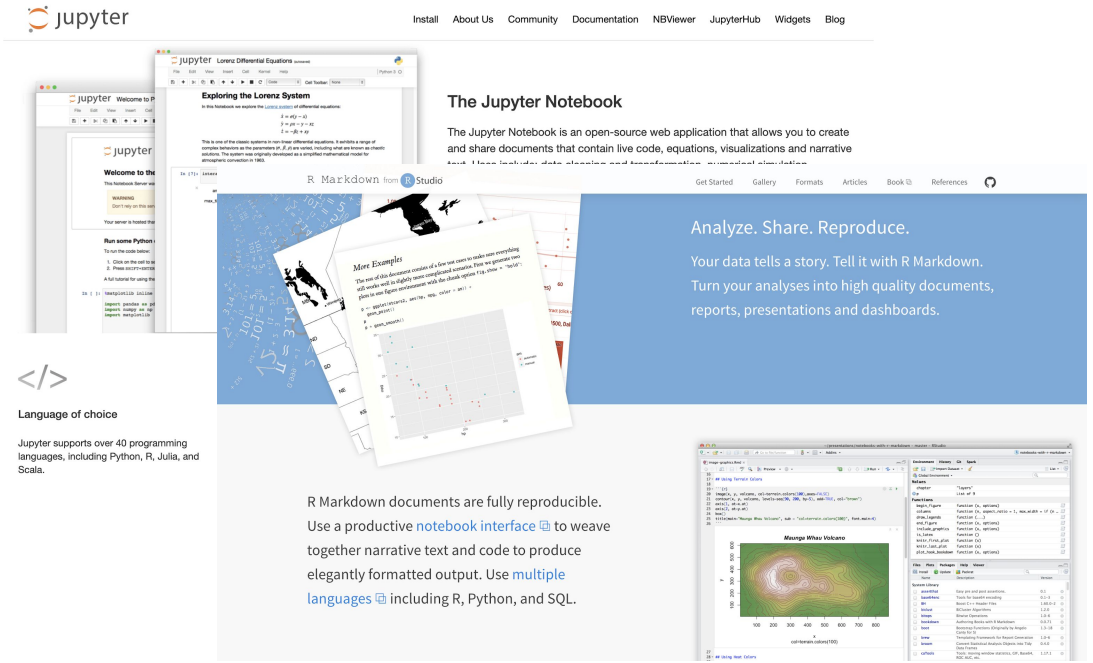
We know from R for Data Science that in a typical project we usually need to import, tidy, understand and communicate knowledge.



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.



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...

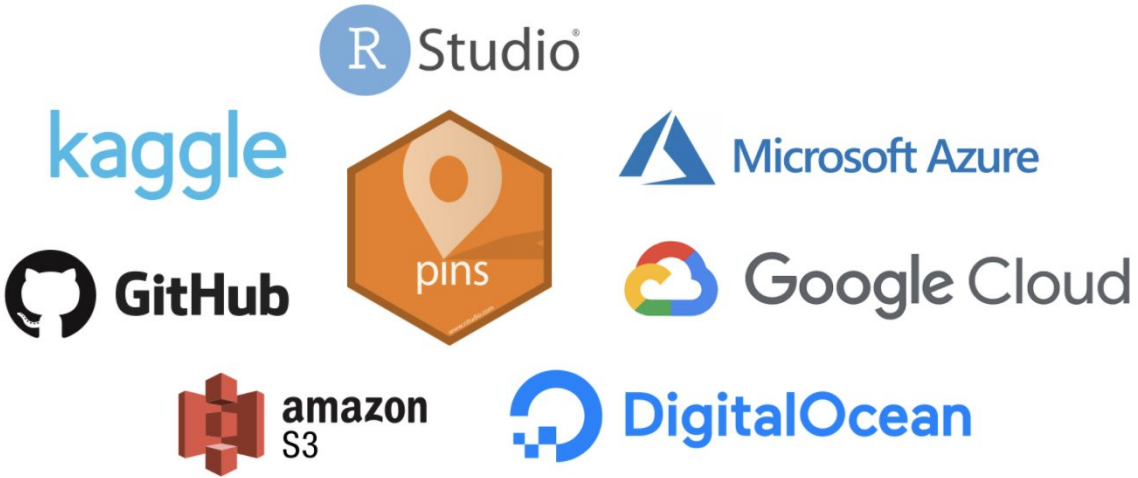
Dev status

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.

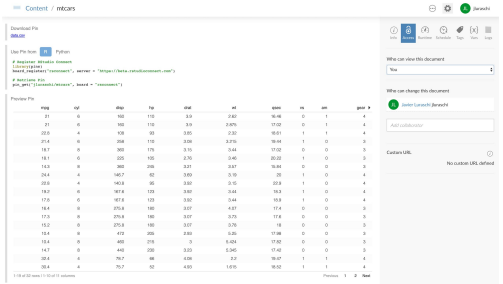


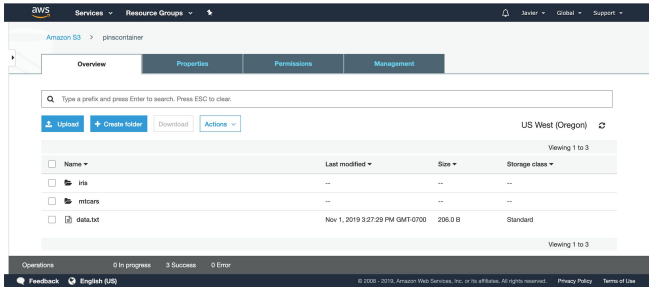
How can I use Boards?

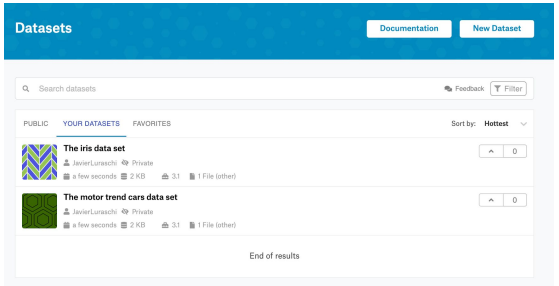
```
library(pins)
board_register("______")

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()
```

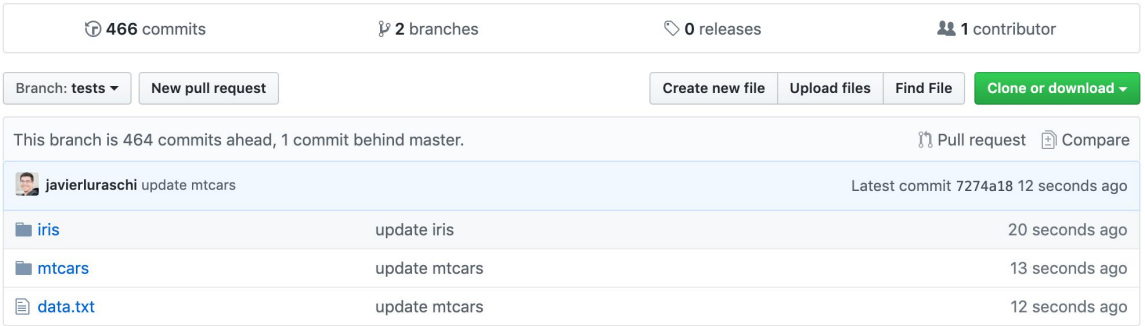
Boards and Pins

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



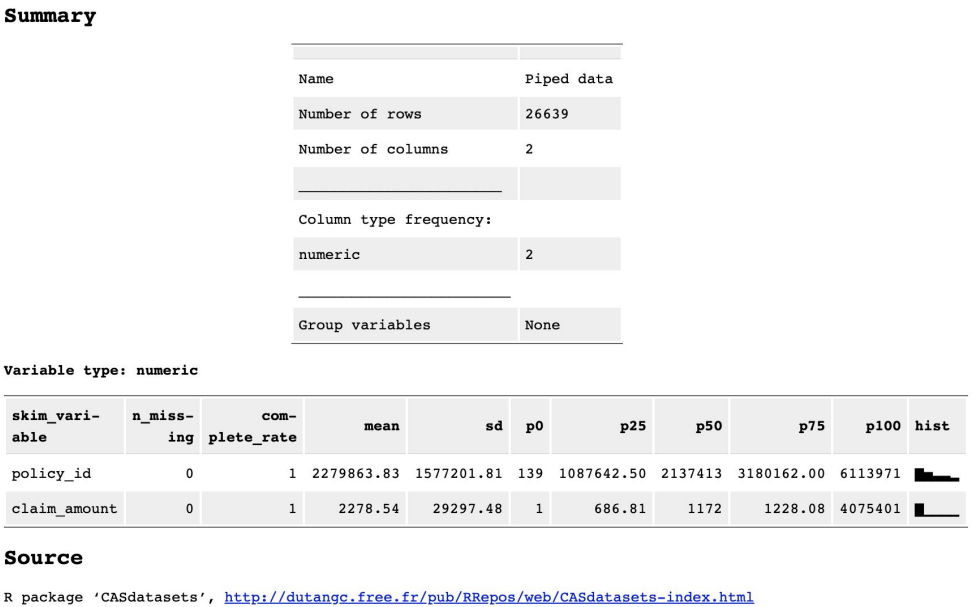






Website Boards

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.



Summary

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

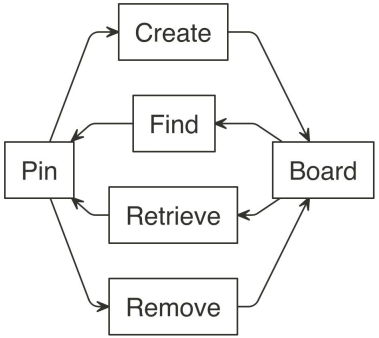
You can optionally enable or disable tracking versions, some boards support this by default, in others you opt-in when the board is registered.

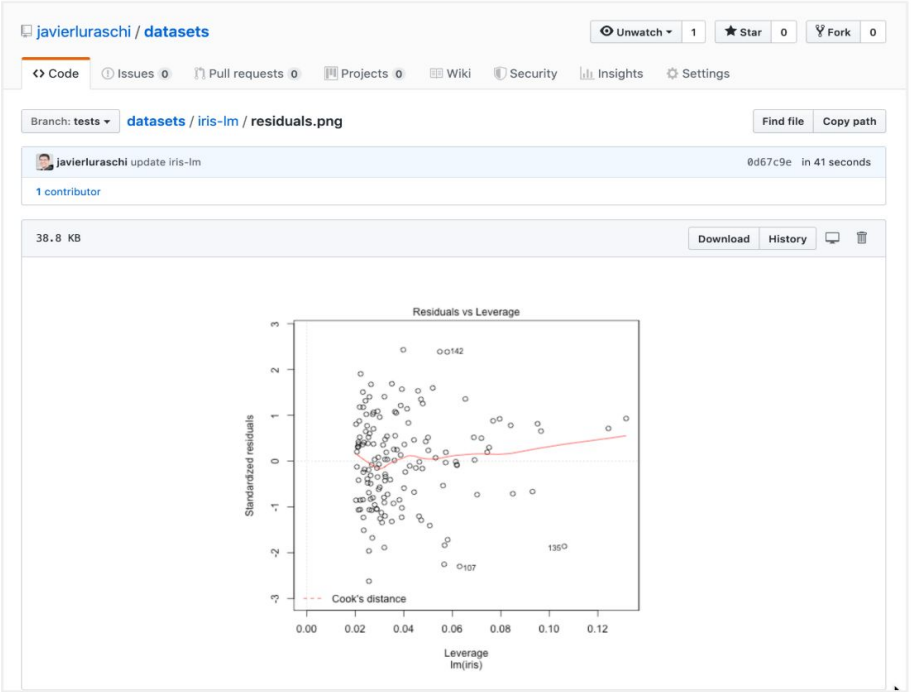
```
board_register("github", repo = "javierluraschi/datasets", branch = "datasets")

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
```

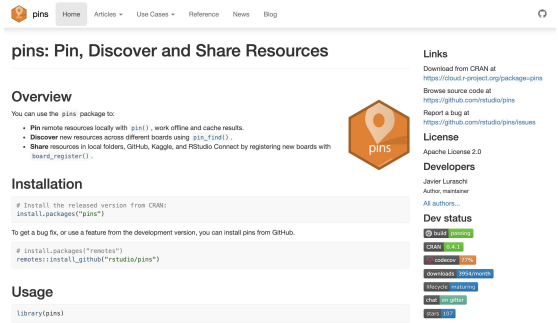
Extending Pins and Boards

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

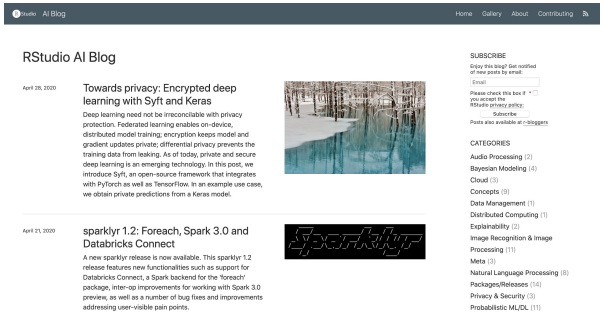




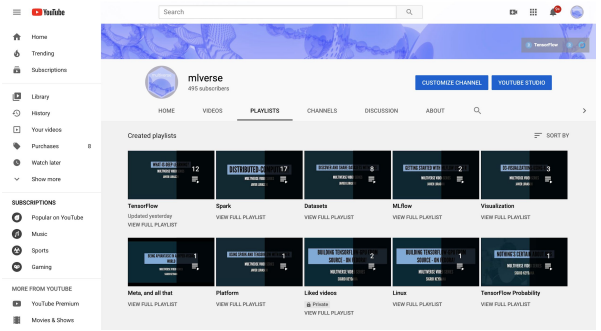
Thank You!



pins.rstudio.com



blogs.rstudio.com/ai



youtube.com/c/mlverse

@zkajdan

@javierluraschi

@dfalbel

@yl790