Cascadia R Conference 2019

DRAKE-AGE:

Lessons Learned While Package-ing {drake}

Tiernan Martin

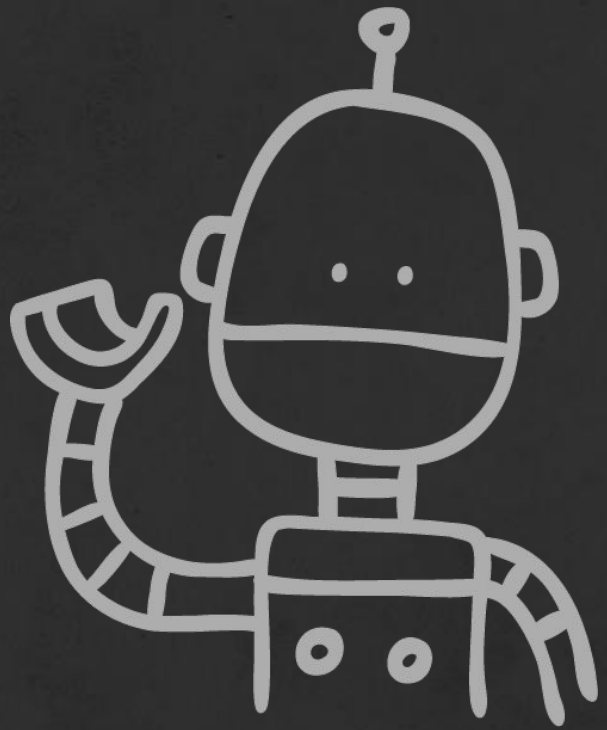
futurewise.org

RopenSci

an R-focused pipeline toolkit for reproducibility and high-performance computing

author: Will Landau

who has used
(or heard of)
drake before?



who has used
(or heard of)
drake before?

image: robot by ester barbato from the Noun Project

1. introduction to drake

1. introduction to drake

2. tips on making a drake-driven R package

1. introduction to drake

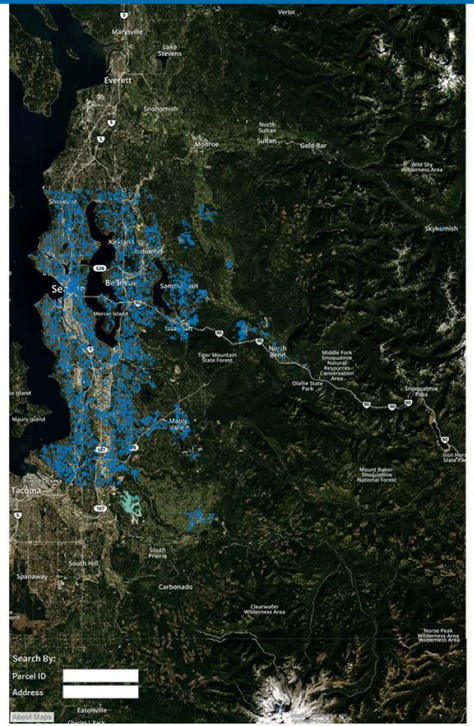
2. tips on making a drake-driven R package

3. links & resources

a quick story...

Home & Hope: Site Mapper

First time using the tool?
Click the question mark to check out the user guide.



Location

City
(All)

King County Council District
(All)

Legislative District
(All)

School District
(All)

Seattle City Council District
(All)

Proximity

Transit Stop
Any distance

Play Space
Any distance

Early Learning Facility
Any distance

Exclude Lots:

Within 1,000 Feet of Cannabis Business
Yes

Within 1/4 Mile of Income-Restricted Housing
No

Ownership

Owner Category
(Multiple values)

Public Owner Type
(All)

Funding Eligibility

Opportunity Zone
(All)

Qualified Census Tract
(All)

Difficult to Develop Area
(All)

New Market Tax Credit
(All)

Site Characteristics

Site Utilization Ratio
(All)

Zoning Description
(All)

Show Only Lots Presently Used as Surface Parking?
No

Lot Size (Sq Ft)
Min 0
Max 3,649,040

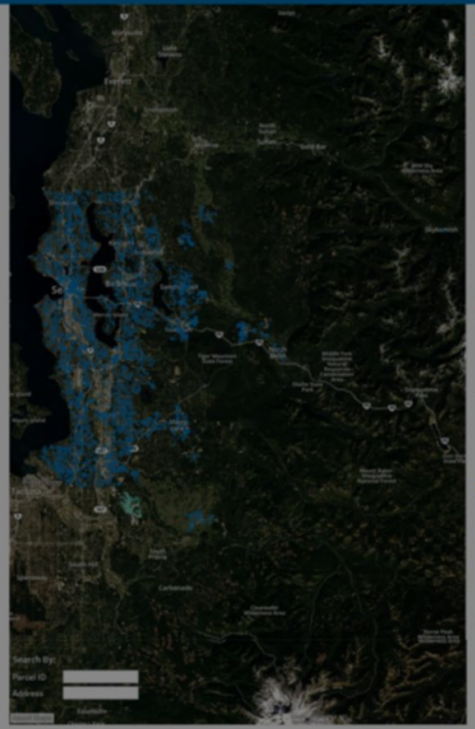
Exclude Lots:

With Known Contamination
Yes

With Environmental Restrictions
Yes

Home & Hope: Site Mapper

First time using the tool?
Click the question mark to check out the user guide.



Location

City
(All)

King County Council District
(All)

Legislative District
(All)

School District
(All)

Seattle City Council District
(All)

Proximity

Transit Stop
Any distance

Play Space
Any distance

Early Learning Facility
Any distance

Exclude Lots:

Within 1,000 Feet of Cannabis Business
Yes

Within 1/4 Mile of Income-Restricted Housing
No

Ownership

Owner Category
(Multiple values)

Public Owner Type
(All)

Funding Eligibility

Opportunity Zone
(All)

Qualified Census Tract
(All)

Difficult to Develop Area
(All)

New Market Tax Credit
(All)

Site Characteristics

Site Utilization Ratio
(All)

Zoning Description
(All)

Show Only Lots Presently Used as Surface Parking?
No

Lot Size (Sq Ft)
Min 0
Max 3,649,040

Exclude Lots:

With Known Contamination
Yes

With Environmental Restrictions
Yes

City
(All)

King County Council District
(All)

Legislative District
(All)

School District
(All)

Seattle City Council District
(All)

Transit Stop
Any distance

Play Space
Any distance

Early Learning Facility
Any distance

Exclude Lots:

Within 1,000 Feet of Cannabis Business
Yes

Within 1/4 Mile of Income-Restricted Housing
No

Owner Category
(Multiple values)

Public Owner Type
(All)

Opportunity Zone
(All)

Qualified Census Tract
(All)

Difficult to Develop Area
(All)

New Market Tax Credit
(All)

Site Utilization Ratio
(All)

Zoning Description
(All)

Show Only Lots Presently Used as Surface Parking?
No

Lot Size (Sq Ft)

Min	Max
0	3,649,040

Exclude Lots:

With Known Contamination
Yes

With Environmental Restrictions
Yes

City
(All)

King County Council District
(All)

Legislative District
(All)

School District
(All)

Seattle City Council District
(All)

Transit Stop
Any distance

Play Space
Any distance

Early Learning Facility
Any distance

Exclude Lots:

Within 1,000 Feet of Cannabis Business
Yes

Within 1/4 Mile of Income-Restricted Housing
No

Owner Category
(Multiple values)

Public Owner Type
(All)

Opportunity Zone
(All)

Qualified Census Tract
(All)

Difficult to Develop Area
(All)

New Market Tax Credit
(All)

Site Utilization Ratio
(All)

Zoning Description
(All)

Show Only Lots Presently Used as Surface Parking?
No

Lot Size (Sq Ft)

Min	Max
0	3,649,040

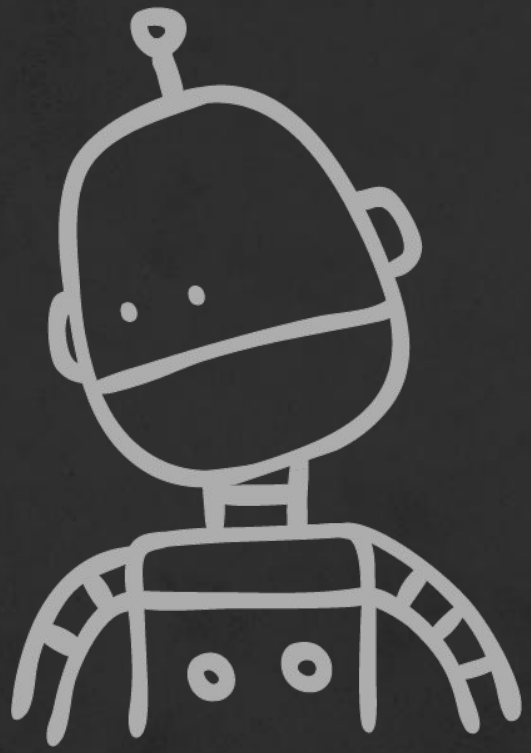
Exclude Lots:

With Known Contamination
Yes

With Environmental Restrictions
Yes


City
(All)
King County Council District
(All)
Legislative District
(All)
School District
(All)
Seattle City Council District
(All)
Transit Stop
Any distance
Play Space
Any distance
Early Learning Facility
Any distance
Exclude Lots:
Within 1,000 Feet of Cannabis Business
Yes
Within 1/4 Mile of Income-Restricted Housing
No
Owner Category
(Multiple values)
Public Owner Type
(All)
Opportunity Zone
(All)
Qualified Census Tract
(All)
Difficult to Develop Area
(All)
New Market Tax Credit
(All)
Site Utilization Ratio
(All)
Zoning Description
(All)
Show Only Lots Presently Used as Surface Parking?
No
Lot Size (Sq Ft)
Min
0
Max
3,649,040
Exclude Lots:
With Known Contamination
Yes
With Environmental Restrictions
Yes
CSV
x2
SHP
x5

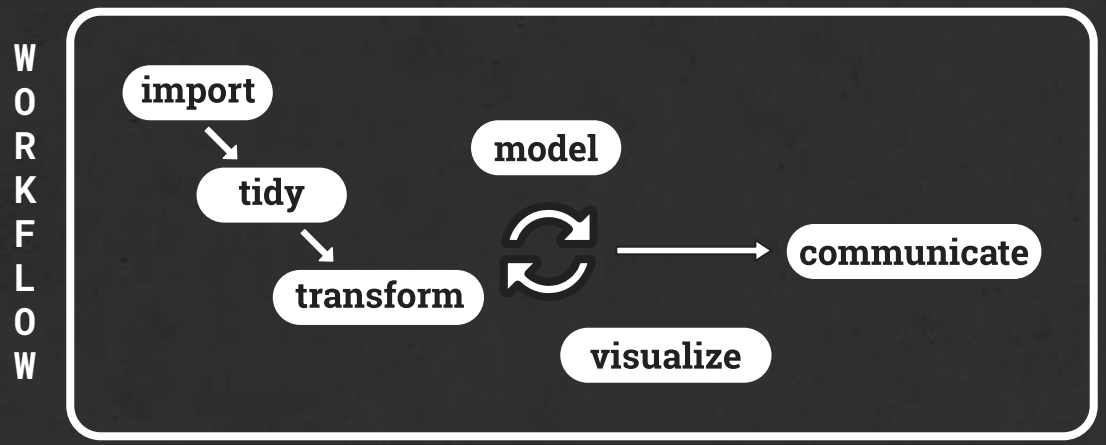

City
(All)
King County Council District
(All)
Legislative District
(All)
School District
(All)
Seattle City Council District
(All)
Transit Stop
Any distance
Play Space
Any distance
Early Learning Facility
Any distance
Exclude Lots:
Within 1,000 Feet of Cannabis Business
Yes
Within 1/4 Mile of Income-Restricted Housing
No
Owner Category
(Multiple values)
Public Owner Type
(All)
Opportunity Zone
(All)
Qualified Census Tract
(All)
Difficult to Develop Area
(All)
New Market Tax Credit
(All)
Site Utilization Ratio
(All)
Zoning Description
(All)
Show Only Lots Presently Used as Surface Parking?
No
Lot Size (Sq Ft)
Min
0
Max
3,649,040
Exclude Lots:
With Known Contamination
Yes
With Environmental Restrictions
Yes
CSV x2
SHP x5
CSV x2
SHP x5
XLS x3
CSV x2
SHP x5
XLS x3
CSV x2
SHP x5
CSV x2
TXT x4
CSV x2
SHP x5
XLS x6



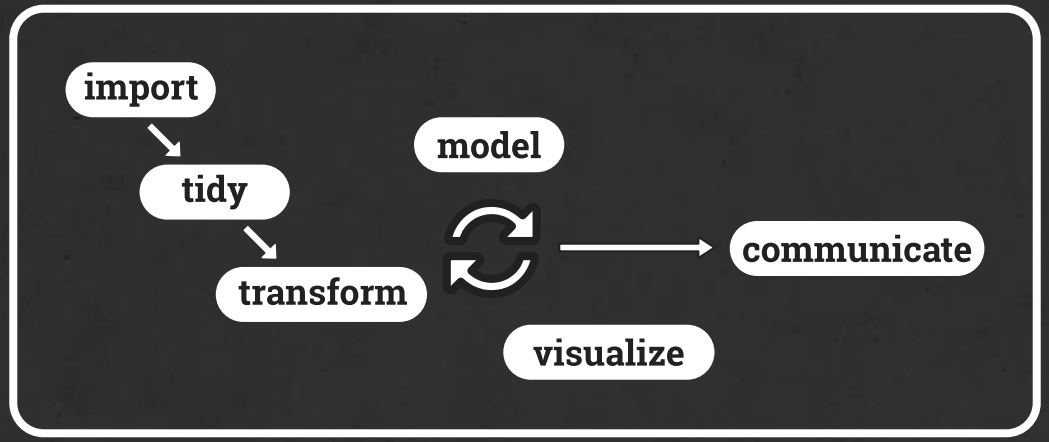
{drake}

an R-focused pipeline toolkit for reproducibility and high-performance computing

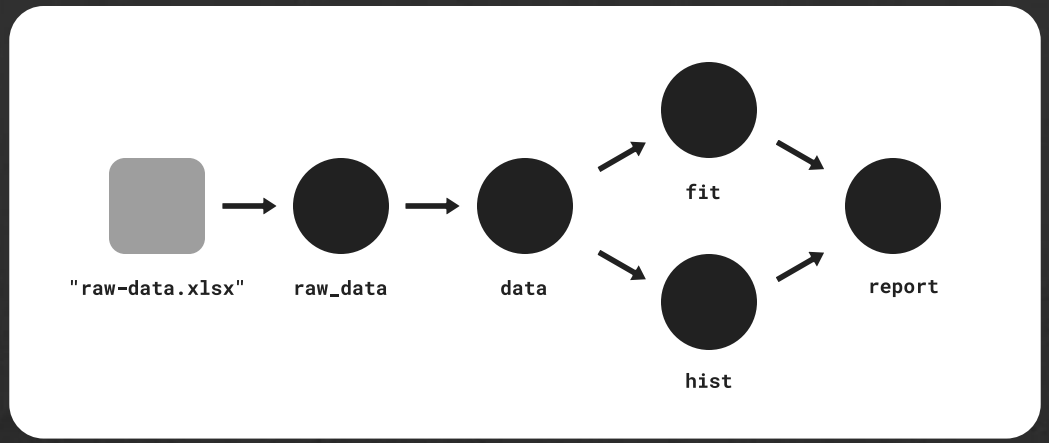
github.com/ropensci/drake/


WORKFLOW
import
tidy
transform
model
visualize
communicate

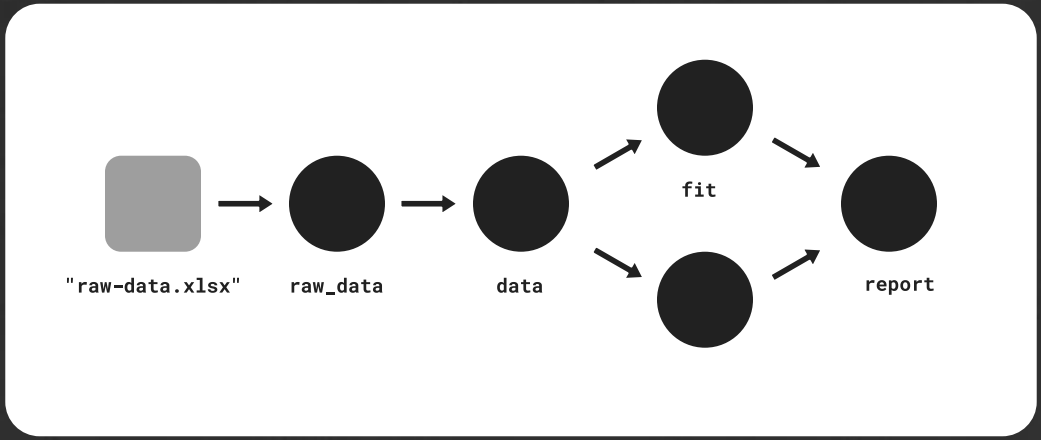
WORKFLOW



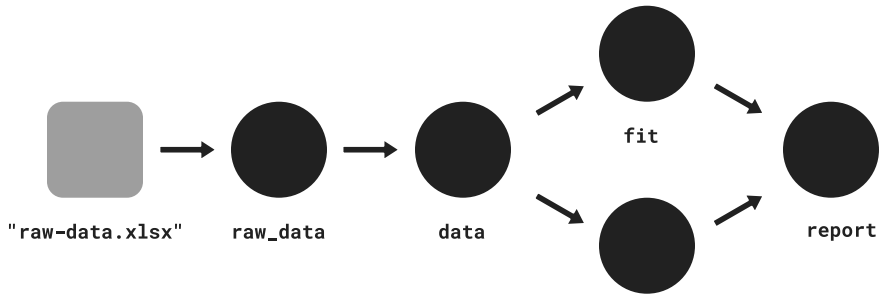
DEPENDENCIES



DEPENDENCIES



DEPENDENCIES



DRAKE PLAN

```
> plan

# A tibble: 5 x 2
  target   command
  <chr>    <expr>
1 raw_data readxl::read_excel(file_in("raw_data.xlsx"))                          ~
2 data     raw_data %>% mutate(Species = forcats::fct_inorder(Species))          ~
3 hist     create_plot(data)                                                     ~
4 fit      lm(Sepal.Width ~ Petal.Width + Species, data)                         ~
5 report   rmarkdown::render(knitr_in("report.Rmd"), output_file = file_out("report.~
```

DRAKE PLAN

```
> plan

# A tibble: 5 x 2
  target   command
  <chr>    <expr>
1 raw_data readxl::read_excel(file_in("raw_data.xlsx"))                          ~
2 data     raw_data %>% mutate(Species = forcats::fct_inorder(Species))          ~
3 hist     create_plot(data)                                                     ~
4 fit      lm(Sepal.Width ~ Petal.Width + Species, data)                         ~
5 report   rmarkdown::render(knitr_in("report.Rmd"), output_file = file_out("report.~
```

DRAKE PLAN

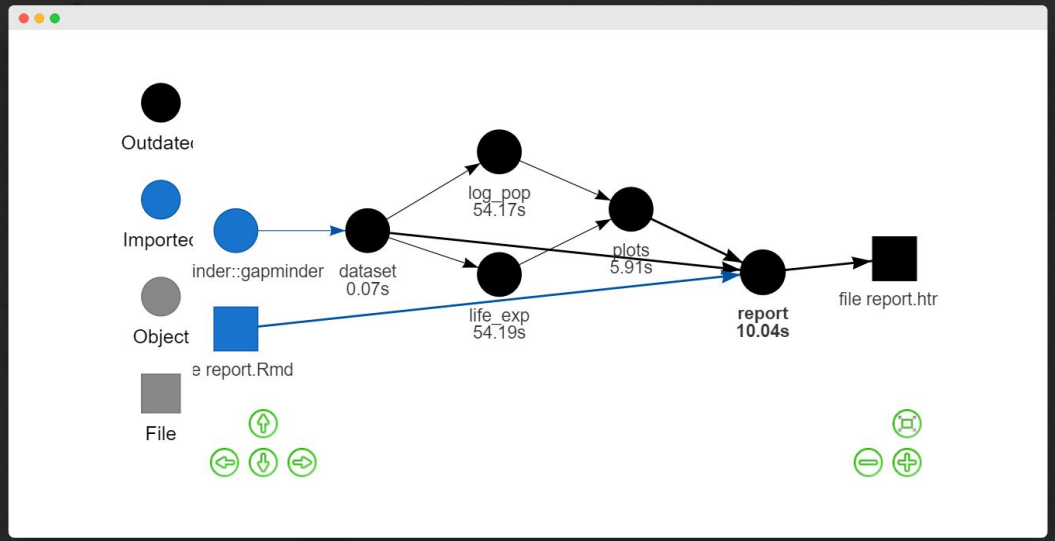
```
> plan

# A tibble: 5 x 2
  target   command
  <chr>    <expr>
1 raw_data readxl::read_excel(file_in("raw_data.xlsx"))                          ~
2 data     raw_data %>% mutate(Species = forcats::fct_inorder(Species))          ~
3 hist     create_plot(data)                                                     ~
4 fit      lm(Sepal.Width ~ Petal.Width + Species, data)                         ~
5 report   rmarkdown::render(knitr_in("report.Rmd"), output_file = file_out("report.~
```

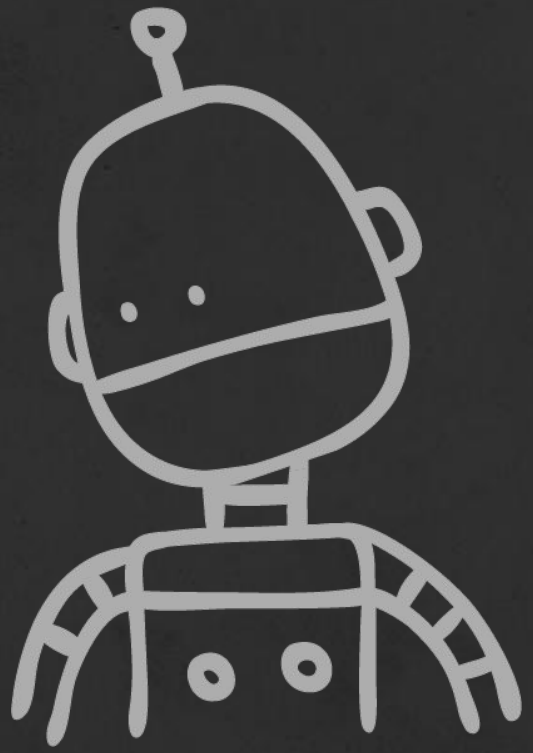
```
  target
1 raw_data
2 data
3 hist
4 fit
5 report
```

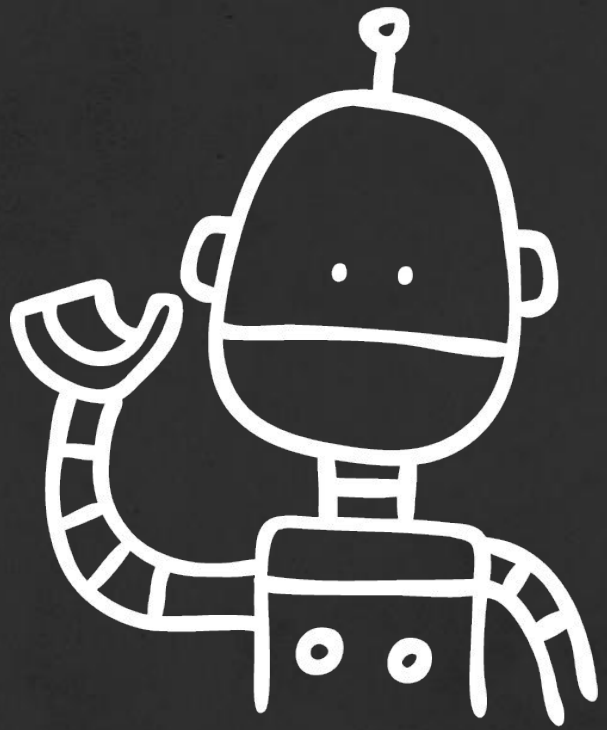
cache

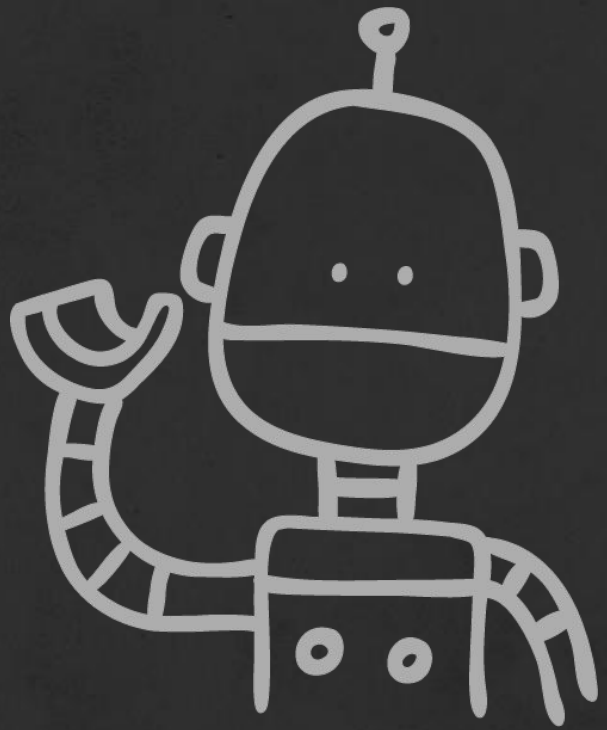
GAPMINDER PLAN



drake provides interactive visualizations of your workflow's dependency graph







with **drake**,
what gets done
stays done

PROJECT PLAN

```
# A tibble: 100 x 2
  Target              command
  <chr>               <expr>
1 filter_variable_01  make_filtervar01(file_in("raw_data_HUD.csv"))
2 filter_variable_02  make_filtervar02(file_in("raw_data_EPA.xlsx"))
3 filter_variable_03  make_filtervar03(file_in("raw_data_OH.shp"))
4 filter_variable_04  make_filtervar04(file_in("raw_data_CHAS.csv"))
5 filter_variable_05  make_filtervar05(file_in("raw_data_KC.csv"))
```

PROJECT PLAN

```
# A tibble: 100 x 2
  Target                command

1 filter_variable_01    make_filtervar01(file_in("raw_data_HUD.csv"))
2 filter_variable_02    make_filtervar02(file_in("raw_data_EPA.xlsx"))
3 filter_variable_03    make_filtervar03(file_in("raw_data_OH.shp"))
4 filter_variable_04    make_filtervar04(file_in("raw_data_CHAS.csv"))
5 filter_variable_05    make_filtervar05(file_in("raw_data_KC.csv"))
```

R SCRIPT

make_filtervar05.R

```
# MAKE_FILTERVAR05 ---------------------------
# data provider: King County Assessor's Office
# download url: ftp://kcassessordata/e0gw9
make_filtervar05 <- function(filepath){
  raw_dat <- readr::read_csv(filepath)
  raw_dat %>%
...
```

PROJECT PLAN

```
# A tibble: 100 x 2
  Target                command

1 filter_variable_01    make_filtervar01(file_in("raw_data_HUD.csv"))
2 filter_variable_02    make_filtervar02(file_in("raw_data_EPA.xlsx"))
3 filter_variable_03    make_filtervar03(file_in("raw_data_OH.shp"))
4 filter_variable_04    make_filtervar04(file_in("raw_data_CHAS.csv"))
5 filter_variable_05    make_filtervar05(file_in("raw_data_KC.csv"))
```

R SCRIPT

make_filtervar05.R

```
# MAKE_FILTERVAR05 ---------------------------
# data provider: King County Assessor's Office
# download url: ftp://kcassessordata/e0gw9
make_filtervar05 <- function(filepath){
  raw_dat <- readr::read_csv(filepath)
  raw_dat %>%
...
```

metadata in comments ... not ideal

⏳ restructuring ...

PROJECT PLAN

```
# A tibble: 100 x 2
  Target                command

1 filter_variable_01    make_filtervar01(file_in("raw_data_HUD.csv"))
2 filter_variable_02    make_filtervar02(file_in("raw_data_EPA.xlsx"))
3 filter_variable_03    make_filtervar03(file_in("raw_data_OH.shp"))
4 filter_variable_04    make_filtervar04(file_in("raw_data_CHAS.csv"))
5 filter_variable_05    make_filtervar05(file_in("raw_data_KC.csv"))
```

R SCRIPT

make_filtervar05.R

```
#' @title Make Filter Variable #5
#' @param filepath the data file's filepath
#' @notes data source ftp://kcassessordata/...
#' @export
make_filtervar05 <- function(filepath){
  raw_dat <- readr::read_csv(filepath)
  raw_dat %>%
```

PROJECT PLAN

```
# A tibble: 100 x 2
  Target                command

1 filter_variable_01    make_filtervar01(file_in("raw_data_HUD.csv"))
2 filter_variable_02    make_filtervar02(file_in("raw_data_EPA.xlsx"))
3 filter_variable_03    make_filtervar03(file_in("raw_data_OH.shp"))
4 filter_variable_04    make_filtervar04(file_in("raw_data_CHAS.csv"))
5 filter_variable_05    make_filtervar05(file_in("raw_data_KC.csv"))
```

R SCRIPT

make_filtervar05.R

```
#' @title Make Filter Variable #5
#' @param filepath the data file's filepath
#' @notes data source ftp://kcassessordata/...
#' @export
make_filtervar05 <- function(filepath){
  raw_dat <- readr::read_csv(filepath)
  raw_dat %>%
```

make_filtervar05 {projectpkg} R Documentation

Filter Variable #5

Description

This variable creates the XX filter.

It is part of the package drake plan.

Usage

make_filtervar05(filepath)

Arguments

filepath the data file's file path

Value

Returns a dataframe

Notes

data source: ftp://kcassessordata/…

Examples

make_filtervar05("raw_data_KC.csv")

[Package projectpkg version 0.0.0.9001 Index]

{drakepkg}

Github, Inc. [US] | https://github.com/tiernanmartin/drakepkg

README.md

repo status WIP

drakepkg

The goal of `drakepkg` is to demonstrate how a `drake` workflow can be organized as an R package.

Why do this? Because the package system in R provides a widely-adopted method of structuring, documenting, testing, and sharing R code. While most R packages are general purpose, this approach applies the same framework to a specific workflow (or set of workflows). It increases the reproducibility of a complex workflow without requiring users to recreate the workflow's environment with a container image (although that approach is compatible with `drakepkg` - see januz/drakepkg).

The `drakepkg` package is experimental in nature and currently requires some inconvenient steps (see the drake manual - 7.1.4 Workflows as R packages); please use caution when applying this approach to your own work.

Installation

You can install the released version of `drakepkg` from its Github repository with:

```
devtools::install_packages("tiernanmartin/drakepkg")
```

Usage

The following table shows how each feature of a `drake` workflow is made accessible within an R package:

`drake`	R Package
plans, commands	functions (`R/*.R`)
targets	stored in the cache (`.drake/`)
input files, output files	internal data (`inst/intdata/*`), external data (`inst/extdata/*`), images and documents (`inst/documents/*`)

The package comes with two example `drake` plans, both of which are loosely based on the `main` example included in the `drake` package:

1. An introductory plan: `drakepkg::get_example_plan_simple()`
2. A plan that involves downloading external data: `drakepkg::get_example_plan_external()`

The first plan looks like this:

{drakepkg}

- ✔ examples & vignettes,
- ✔ learning resources,
- ✔ and this slide deck

what’s the benefit?

1. uses a familiar structure

2. easy transfer `devtools::install_github()`

3. handy tools! unit testing, coverage, checks, etc.

(potential) stumbling blocks

1. package development learning curve

2. `drake::expose_imports()`

3. doesn’t guarantee reproducibility

(but pairs well with tools like packrat, docker, etc.)

take-home message

thanks

{drake}

- package website

  https://ropensci.github.io/drake/

- 6 minute video

  https://player.vimeo.com/video/288956463

- manual

  https://ropenscilabs.github.io/drake-manual/

research compendia

- Karthik Ram's talk

  https://resources.rstudio.com/rstudio-conf-2019/a-guide-to-modern-reproducible-data-science-with-r

- Marwick et al., 2018

  https://doi.org/10.1080/00031305.2017.1375986

- Open Science Framework

  https://osf.io

{drakepkg}

- Github repo

  https://github.com/tiernanmartin/drakepkg

- real-world example

  https://github.com/tiernanmartin/hhsitemapper

- drakepkg w/ docker

  https://github.com/januz/drakepkg