

# Kaggle in the Classroom:

Using R and GitHub to run predictive modeling competitions

Colin Rundel  
Duke University

rstudio::conf 2018 - San Diego



# Context

## Statistical Computing (Sta 323)

---

- 2nd / 3rd year undergrads
- Elective
- ~40 students
- Offered each Spring

## Statistical Programming (Sta 523)

---

- 1st year master's
- Required
- ~40 students
- Offered each Fall

# Context

## Statistical Computing (Sta 323)

- 2nd / 3rd year undergrads
- Elective
- ~40 students
- Offered each Spring

## Statistical Programming (Sta 523)

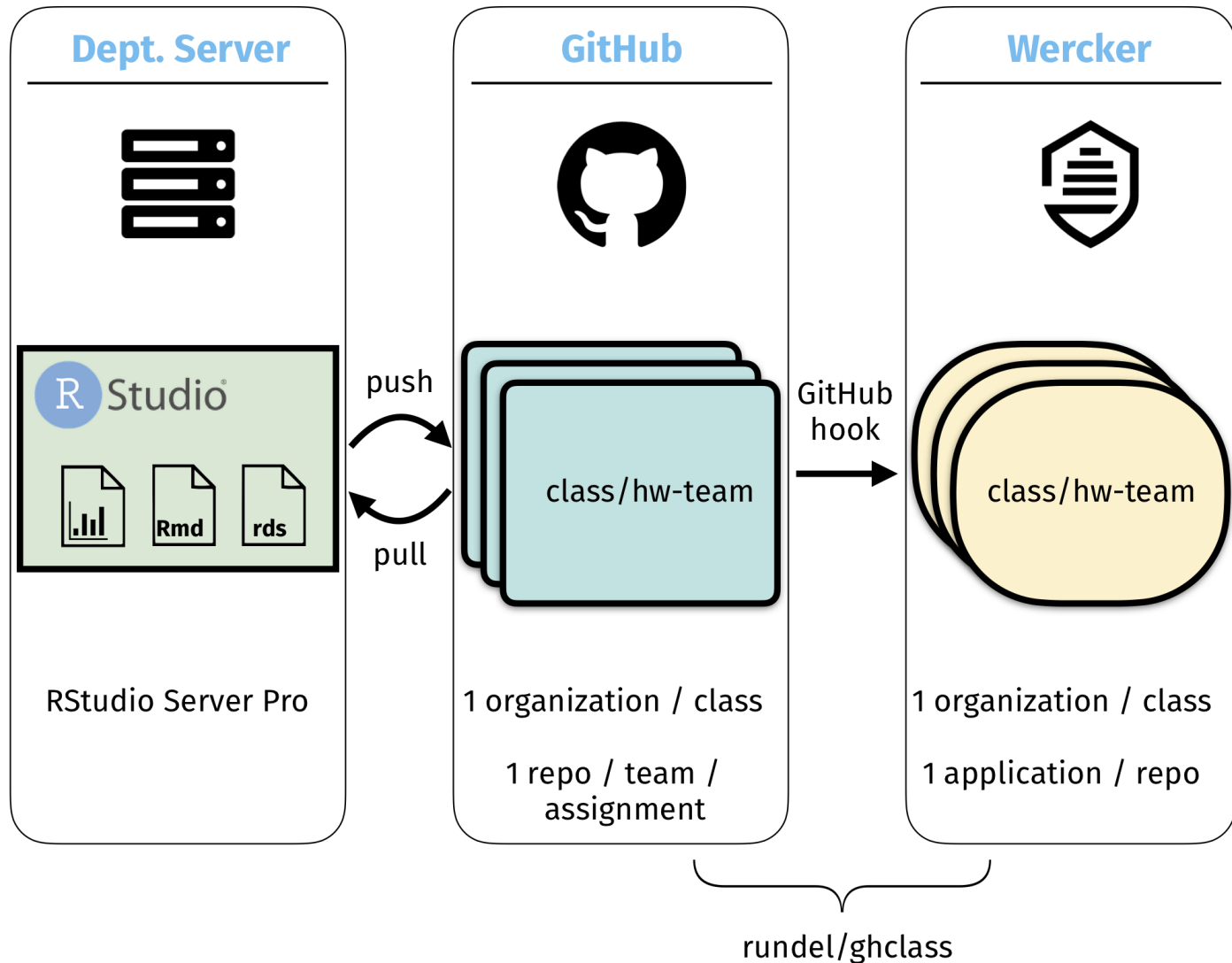
- 1st year master's
- Required
- ~40 students
- Offered each Fall

*Programming course with statistics*

vs.

*Statistics course with programming*

# Workflow



# Automatic testing ...

- for reproducibility
- for file organization
- for style (styler, lintr)
- ...
- for correctness

# wercker.yml

```
box: rocker/tidyverse
build:
  steps:
    - script:
        name: Check for allowed files
        code: |
          Rscript -e "source(paste0('https://raw.githubusercontent.com/',
                                   'Sta323-Sp18/hw1/master/hw1_whitelist.R'))"
    - script:
        name: Render R markdown
        code: |
          Rscript -e "library(rmarkdown);render('hw1.Rmd')"
```

ORACLE + wercker Pipelines

Sta323-Sp18 / hw1

Runs

Workflows

Access

Environment

Options

✗ Update hw1.Rmd

#### Steps

✓ get code

✓ setup environment

✓ wercker-init

✓ Check for allowed files

✗ Render R markdown

*Command cancelled due to error*

```
export WERCKER_STEP_ROOT="/pipeline/script-4ef0c4fc-7df2-4c86-b439-67e
export WERCKER_STEP_ID="script-4ef0c4fc-7df2-4c86-b439-67e3ba5208fc"
export WERCKER_STEP_OWNER="wercker"
export WERCKER_STEP_NAME="script"
export WERCKER_REPORT_NUMBERS_FILE="/report/script-4ef0c4fc-7df2-4c86-
export WERCKER_REPORT_MESSAGE_FILE="/report/script-4ef0c4fc-7df2-4c86-
export WERCKER_REPORT_ARTIFACTS_DIR="/report/script-4ef0c4fc-7df2-4c86-
source "/pipeline/script-4ef0c4fc-7df2-4c86-b439-67e3ba5208fc/run.sh"
```

```
processing file: hw1.Rmd
|.....|
ordinary text without R code
```

```
|.....|
label: unnamed-chunk-1
Quitting from lines 15-21 (hw1.Rmd)
Error in setwd("/dont/set_me/on_fire") : cannot change working directo
Calls: render ... withCallingHandlers -> withVisible -> eval -> eval -
```

ORACLE + wercker Pipelines

Sta323-Sp18 / hw1

Runs

Workflows

Access

Environment

Options

✓ Update README.md

#### Steps

✓ get code

✓ setup environment

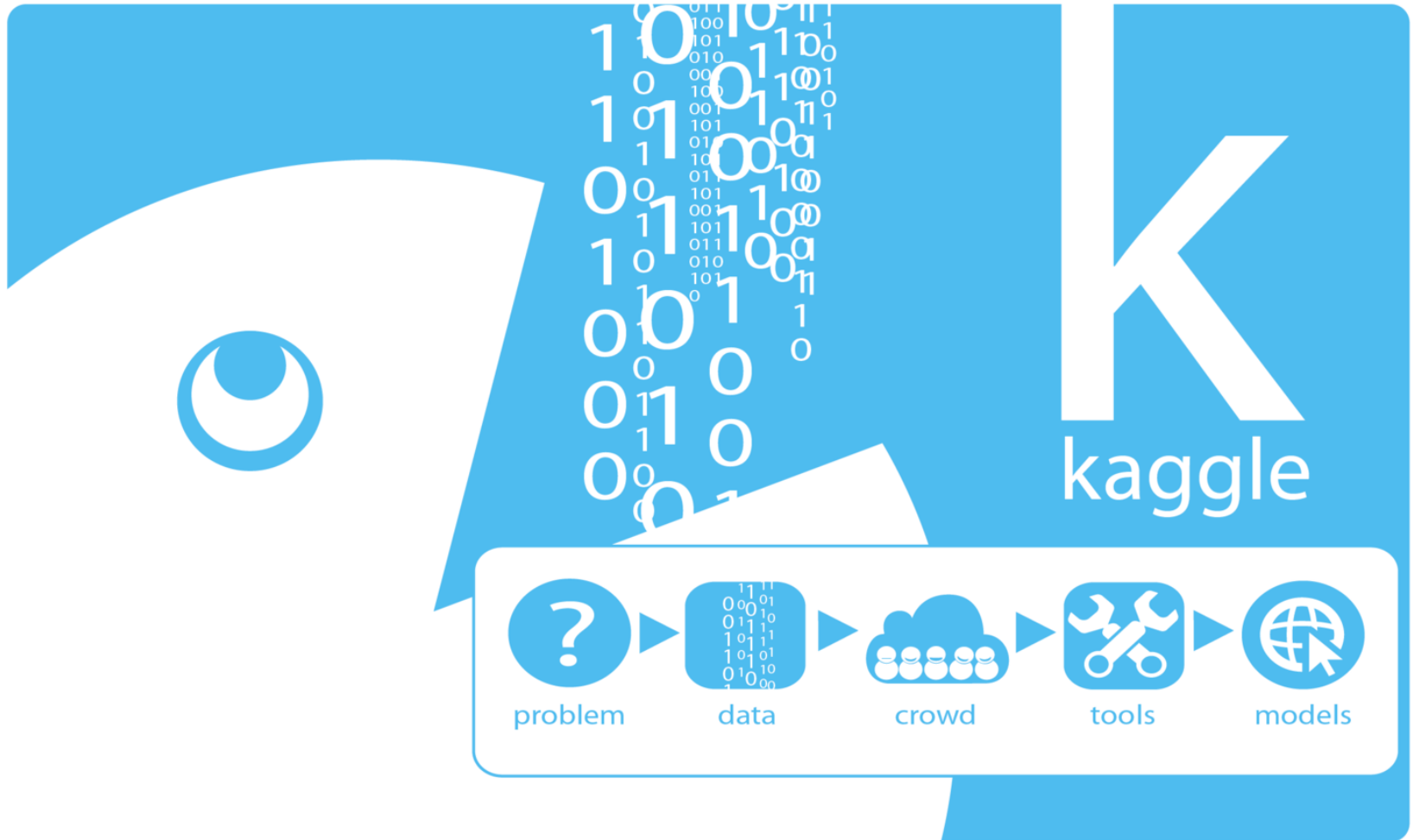
✓ wercker-init

✓ Check for allowed files

✓ Render R markdown

✓ store

# Kaggle™?





# Features

- Automatic scoring
- Feedback / diagnostics
- Leaderboard
- History

# Tickets, taxes, and precincts ...



## Parking Violations

- 9.1M tickets w/ 43 variables
- 1.7G csv
- *Human generated*

[source](#)

## Property Taxes

- MapPLUTO
- Shapefiles of property boundaries
- Includes addresses

[source](#)

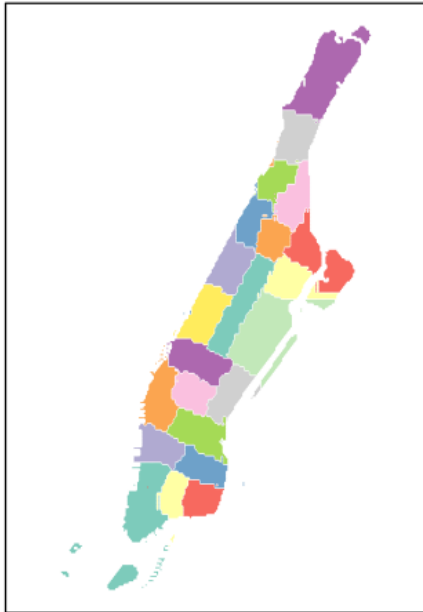
## Police Precincts

- Shapefiles of boundaries
- 74 precincts
- 24 in Manhattan

[source](#)

# Scoring

**Prediction**



**Precincts**



**Score**



# Approach #1 - Wercker

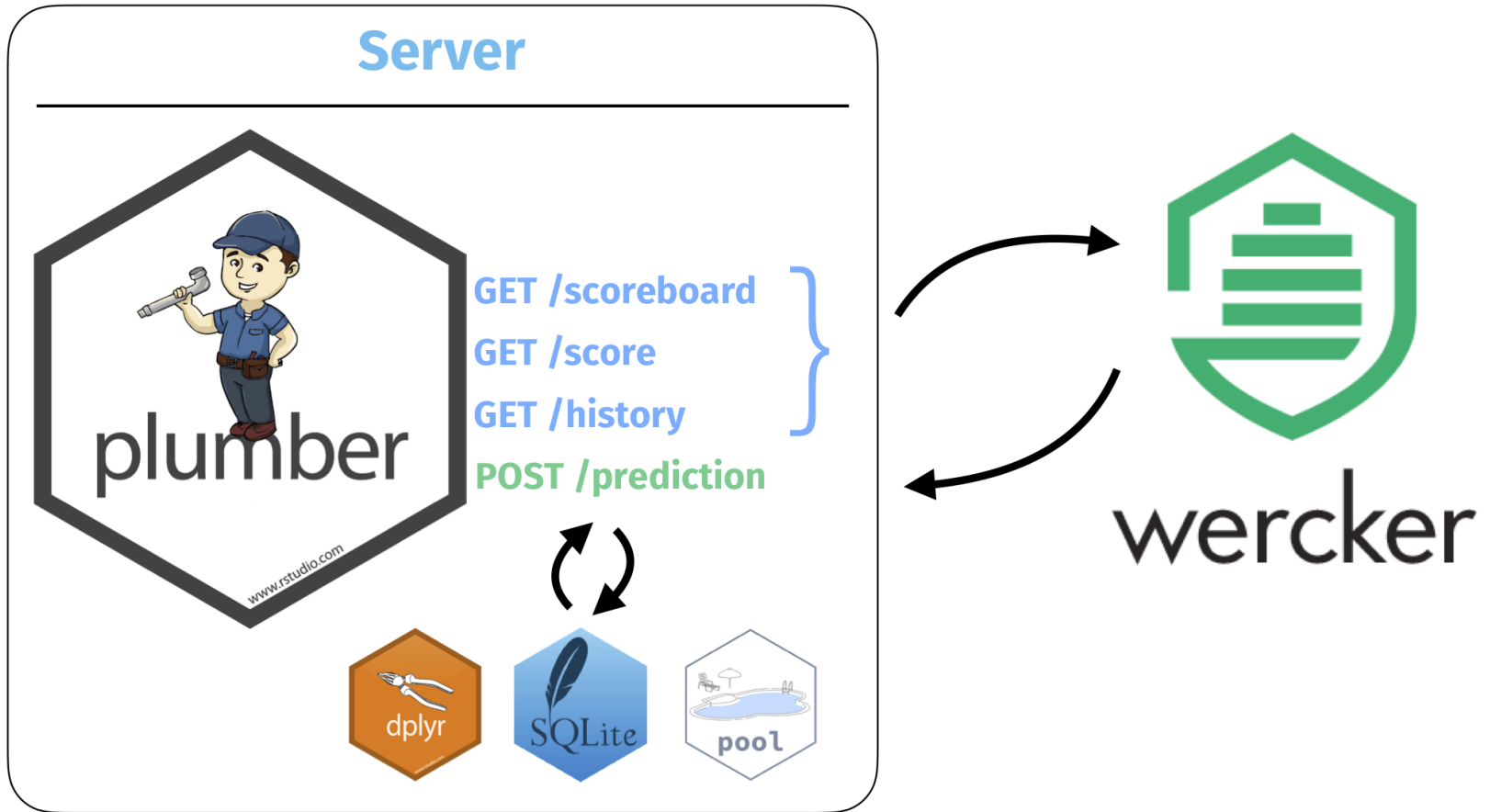
The screenshot displays the Wercker CI interface for a build run. At the top, a red navigation bar contains tabs for 'Runs', 'Workflows', 'Access', 'Environment', and 'Options'. The 'Runs' tab is active. Below the navigation bar, a green status bar indicates the build is successful with a green checkmark and the text 'cleaned up RmD'. To the right of this bar, it shows the branch 'master' and the build name 'build', followed by a blue 'Actions' button with a magnifying glass icon. The main section is titled 'Steps' and lists the following steps, each with a green checkmark indicating success:

Step Name	Duration
get code	1 second
setup environment	51 seconds
wercker-init	0 seconds
Update scores	5 seconds
Show score	2 seconds
Show leaderboard	2 seconds
Show history	2 seconds
Check for allowed files	2 seconds

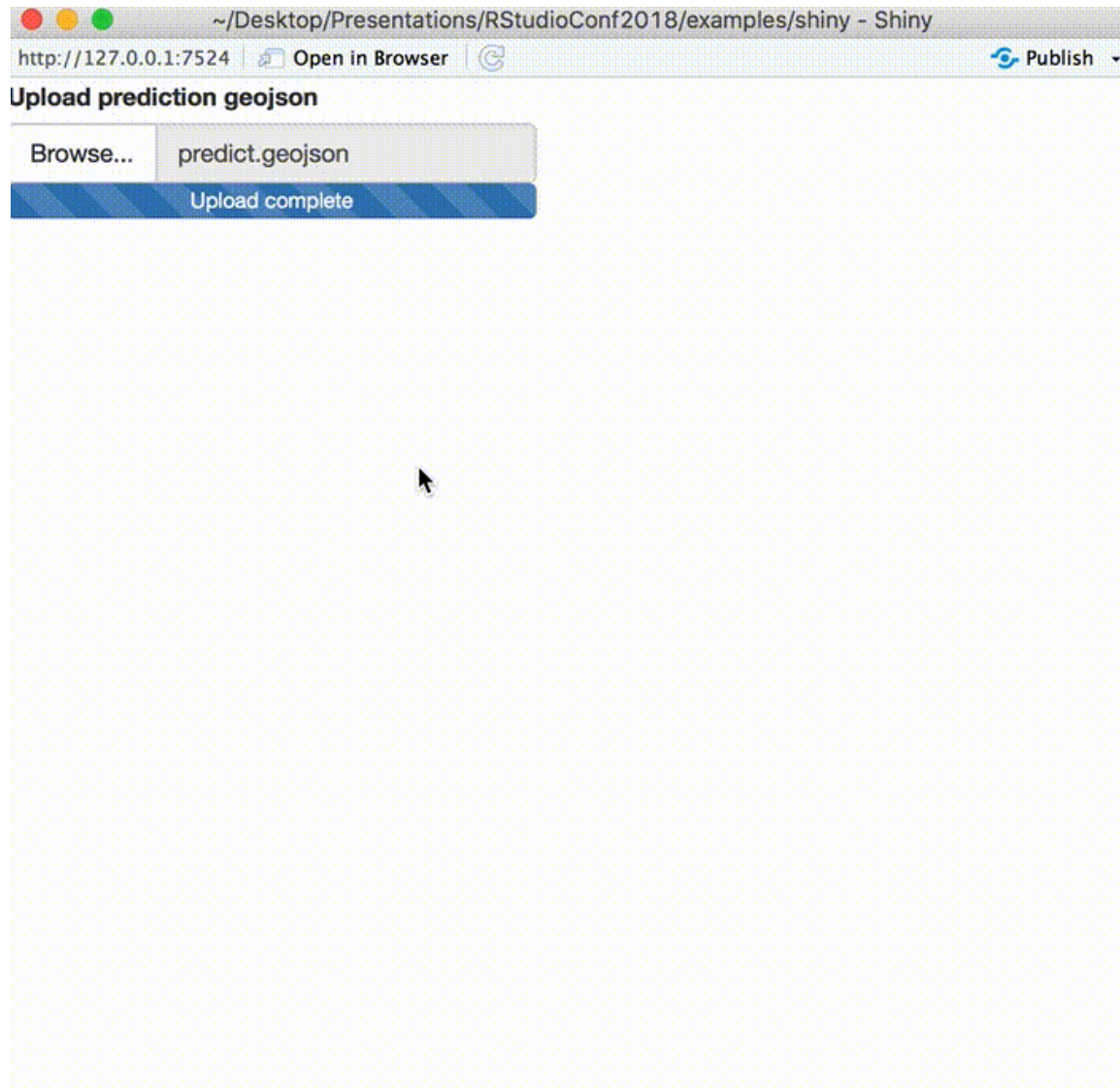
# wercker.yml

```
box: rocker/tidyverse
build:
  steps:
    - script:
      name: Update scores
      code: |
        Rscript -e "httr::stop_for_status(httr::POST('http://saxon.stat.duke.edu:7887,
    - script:
      name: Show score
      code: |
        wget --quiet -O - http://saxon.stat.duke.edu:7887/score?t=$team
    - script:
      name: Show leaderboard
      code: |
        wget --quiet -O - http://saxon.stat.duke.edu:7887/scoreboard
    - script:
      name: Show history
      code: |
        wget --quiet -O - http://saxon.stat.duke.edu:7887/history?t=$team
    - script:
      name: Check for allowed files
      code: |
        Rscript -e "source('https://raw.githubusercontent.com/Sta523-Fa17/hw6/master/1
```











# Details



# Approach #2 - Shiny



# Overall

	Wercker	Shiny
Automatic scoring		
Feedback / Diagnostics		
Leaderboard		
History		
Simplicity		



# Overall

	Wercker	Shiny	Shiny + DB
Automatic scoring			
Feedback / Diagnostics			
Leaderboard			
History			
Simplicity			

# Thank you!



[rundel@gmail.com](mailto:rundel@gmail.com)



[github.com/rundel](https://github.com/rundel)



[@rundel](https://twitter.com/rundel)



[bit.ly/rundel\\_rstudioconf2018](https://bit.ly/rundel_rstudioconf2018)



[stat.duke.edu/~cr173](https://stat.duke.edu/~cr173)