Continuous Integration and Teaching Statistical Computing with R

Colin Rundel

UseR! 2016 - Stanford

Duke University
Department of Statistical Science

Sta (323|523) - Statistical (Computing|Programming)

Course details:

- Foundational computing course
  - 2nd/3rd year elective for BSS
  - Core course for MSS,
- Approximately 40 Students divided into teams of 4
- Biweekly team programming assignments
- Individual takehome midterms, team final project

Learning Objectives

1. R programming and ecosystem

   (R + Hadleyverse)

2. Reproducible Research

   (rmarkdown + knitr + make)

3. Software Engineering / Collaboration

   (shell + git + github)

Infrastructure

Dedicated departmental server

- RStudio Server Pro
- Individual departmental accounts
- System wide install of default packages

Github Organization

- 1 Organization / class
- 1 private repo / team / assignment
- Shared public repos (e.g. examples)
- CI / Testing via Wercker

Course Sketch

HW1 - FizzBuzz (Workflow basics)

HW2 - Graph Data Structures (Base R, testing)

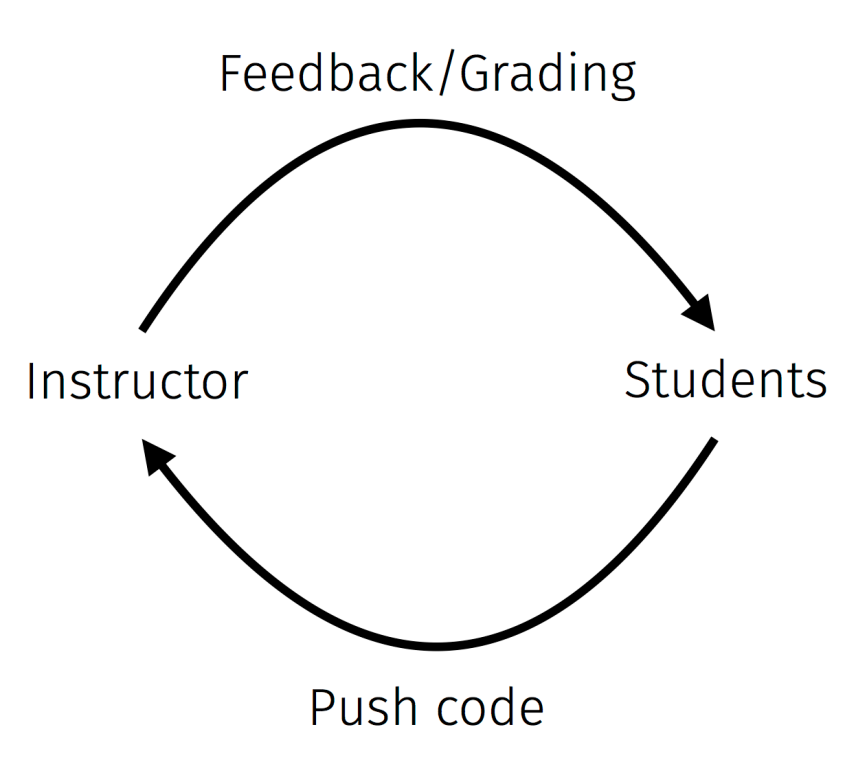
HW3 - La Quinta is Spanish for next to Denny's
(Web APIs, scraping, make)

HW4 - Karl Broman's Socks (Shiny, profiling, parallelization)

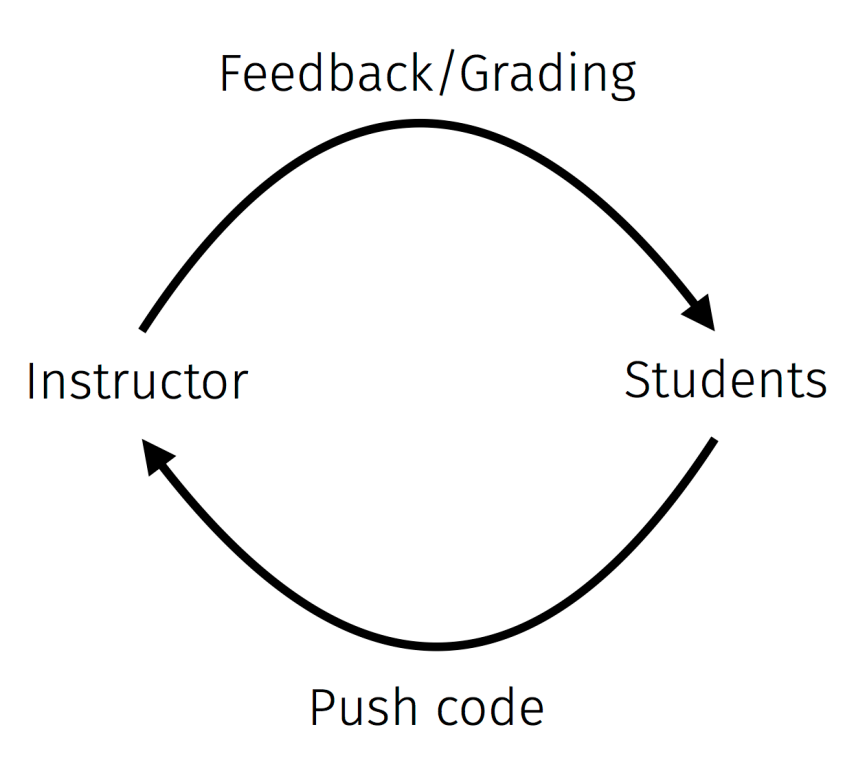
HW5 - Parking Wars: Manhattan (Data munging, prediction)

HW6 - How big is your data? (Hadoop, Spark)

Process Cartoon



Process Cartoon



Github is fantastic for this but doesn't address the fact that *the instructor / TAs are the rate limiting step* (we don't scale well).

A painfully common conversation

Student: We've submitted HW3!

+1 Day

Me: Your Rmd file doesn't knit, you used `setwd` *with an absolute path.*

+1 Day

Student: Ok we fixed that, does it work now?

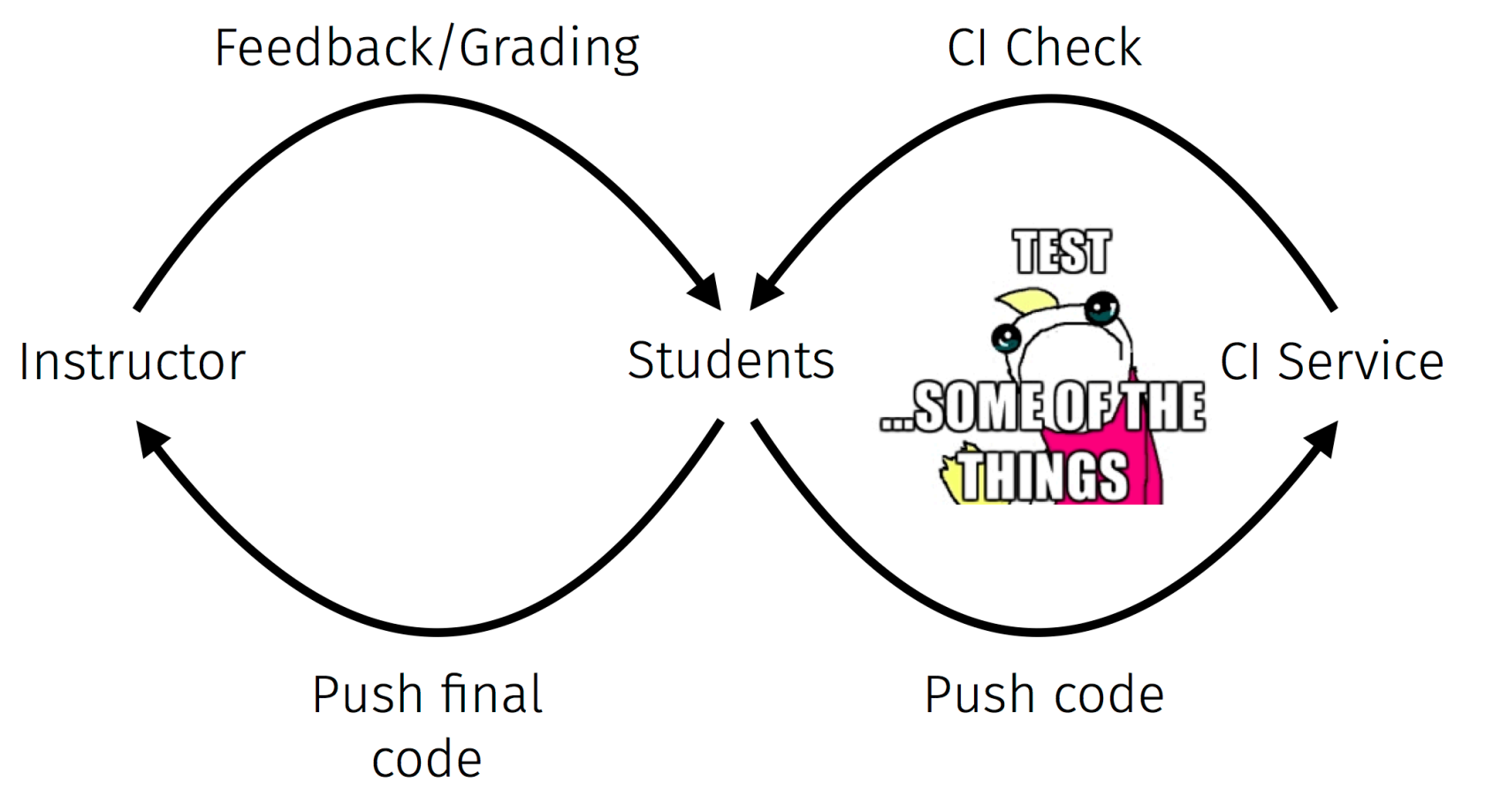
+1 Day

Me: Nope, you used `lme4` *without checking if it was installed.*

+1 Day

⋮

Course Process Cartoon - Improved



Our goal is not to test for correctness - test for process / reproducibility.

Why not TravisCI?

TravisCI

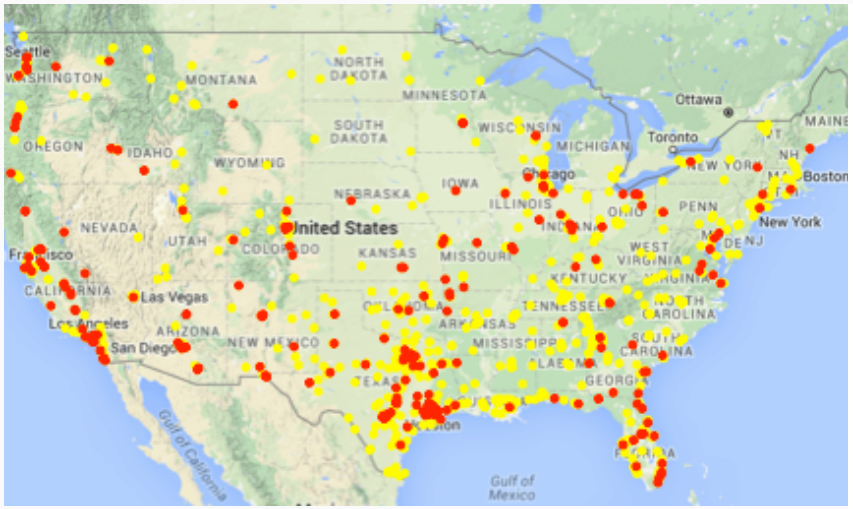
- Package focused (R CMD check)
- Explicit package installation
- Private repos cost $$$
- Mature API

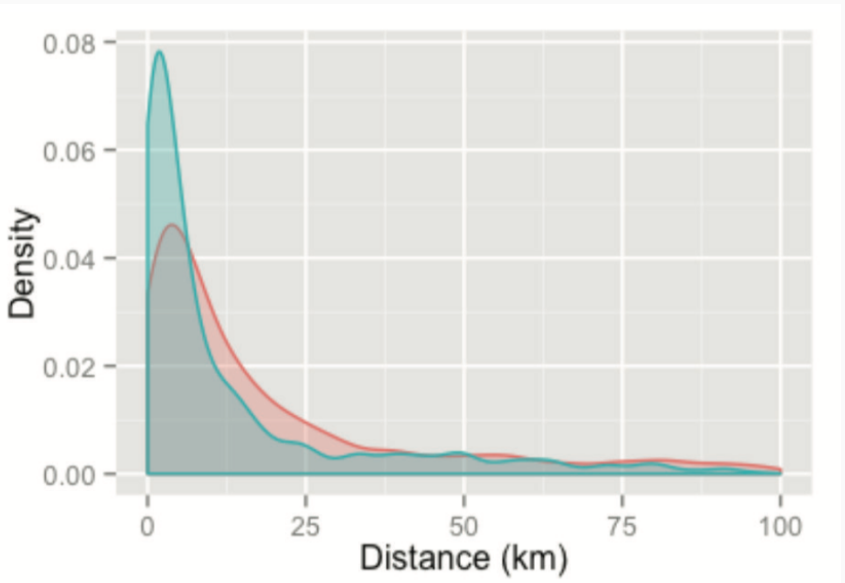
Wercker

- Steps
- Docker based (rocker/hadleyverse)
- Free* for public & private repos
- Manual configuration

La Quinta is Spanish for next to Denny's

La Quinta is Spanish for next to Denny's





Github Repo

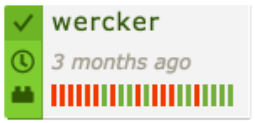
61 commits | 1 branch | 0 releases | 4 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

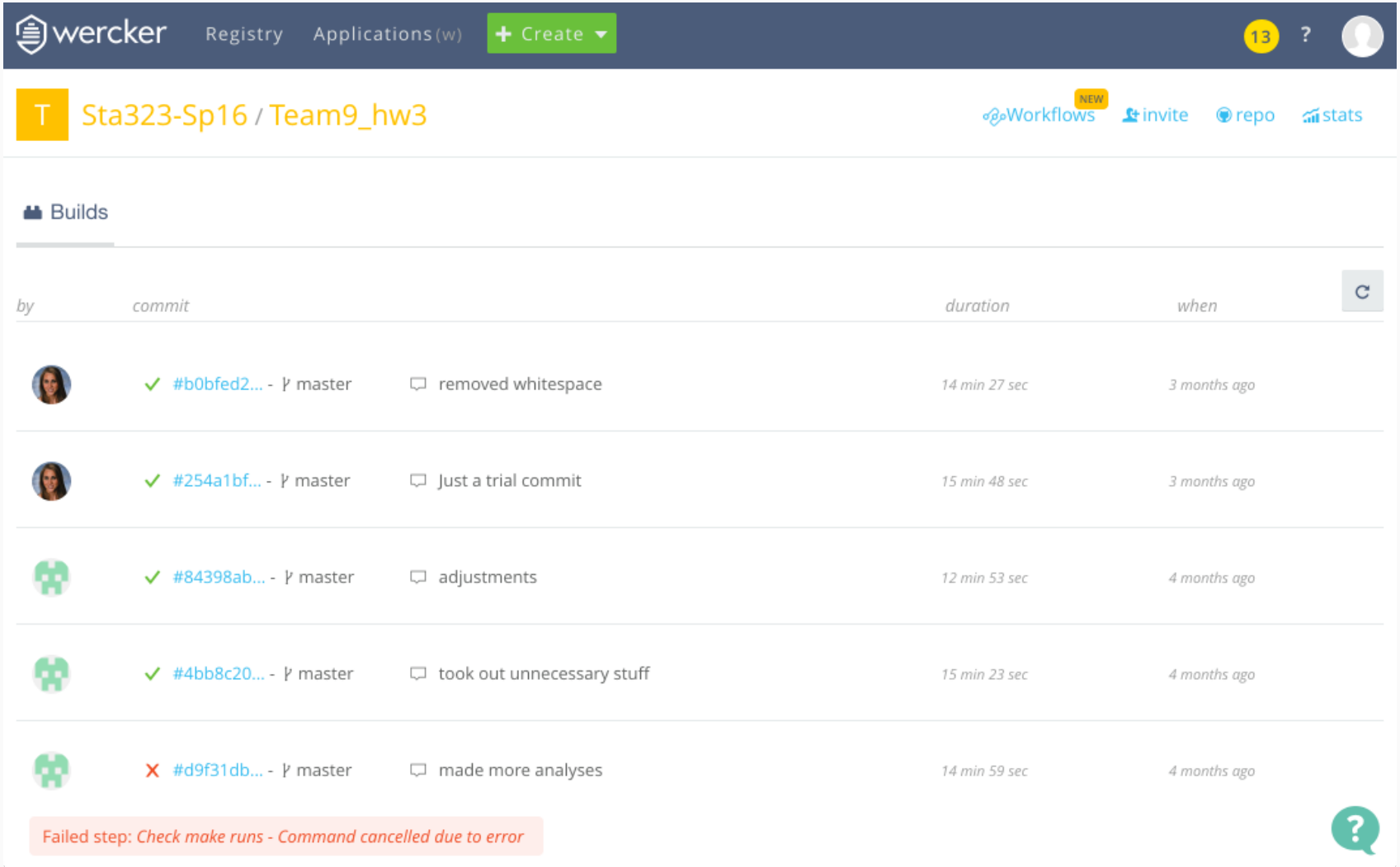
removed whitespace — Latest commit b0bfed2 on Mar 21

.gitignore	attempt to fix wercker error	4 months ago
Makefile	Create Makefile	4 months ago
README.md	Just a trial commit	3 months ago
get_dennys.R	Amanda cleaned up	4 months ago
get_lq.R	removed whitespace	3 months ago
hw3.Rmd	adjustments	4 months ago
parse_dennys.R	Amanda cleaned up	4 months ago
parse_lq.R	took out unnecessary stuff	4 months ago
wercker.yml	Add wercker build support	4 months ago

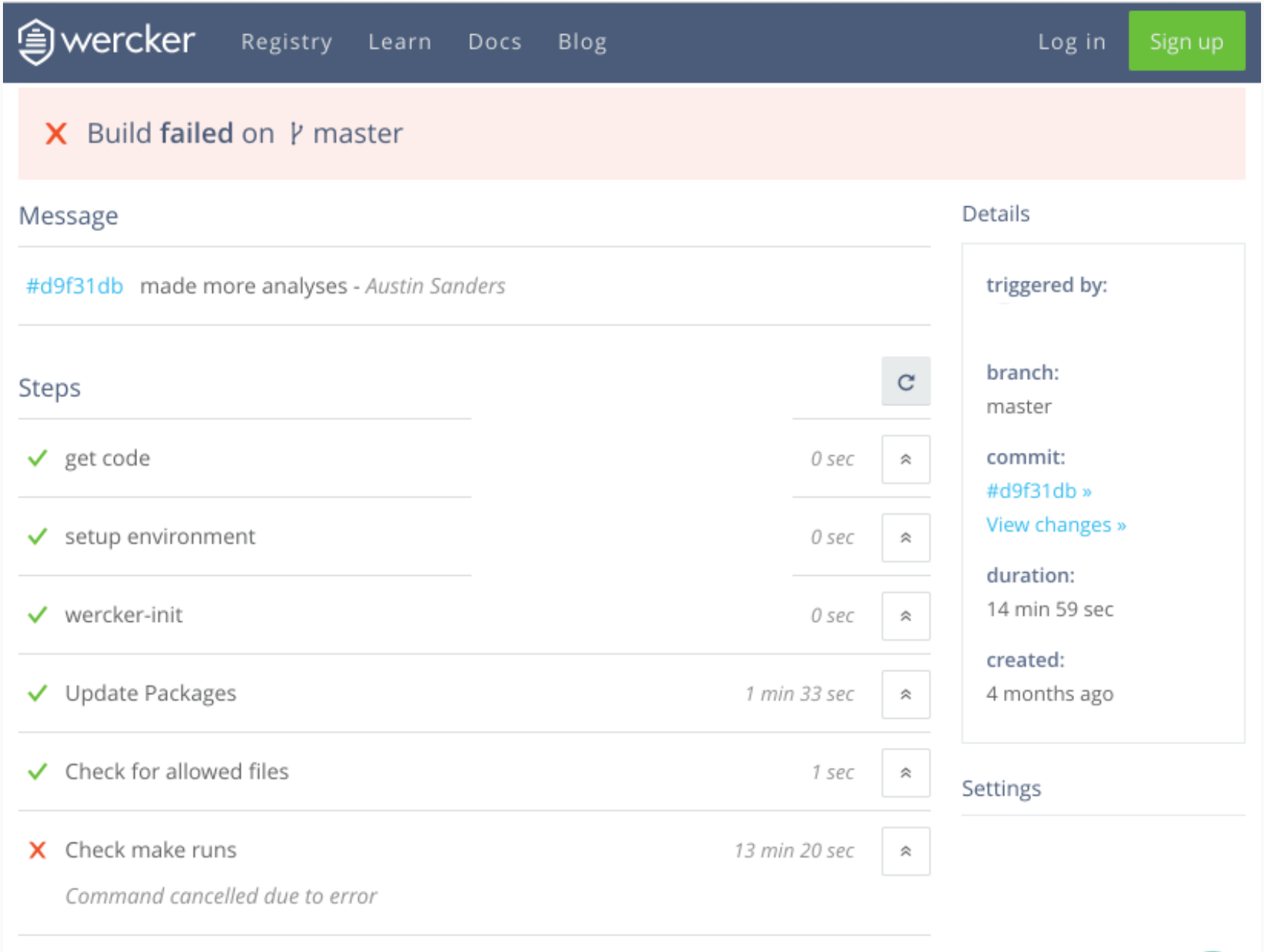
README.md



Wercker


wercker
Registry
Applications (w)
+ Create
13
?
T
Sta323-Sp16 / Team9_hw3
NEW
Workflows
invite
repo
stats
Builds
by
commit
duration
when
#b0bfed2... - master
removed whitespace
14 min 27 sec
3 months ago
#254a1bf... - master
Just a trial commit
15 min 48 sec
3 months ago
#84398ab... - master
adjustments
12 min 53 sec
4 months ago
#4bb8c20... - master
took out unnecessary stuff
15 min 23 sec
4 months ago
#d9f31db... - master
made more analyses
14 min 59 sec
4 months ago
Failed step: Check make runs - Command cancelled due to error

Wercker Steps



Wercker Error

✗ Check make runs — 13 min 20 sec

Command cancelled due to error

```
export WERCKER_STEP_ROOT="/pipeline/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9"
export WERCKER_STEP_ID="script-e51a4a54-1439-44ec-bec2-3e03e47d72f9"
export WERCKER_STEP_OWNER="wercker"
export WERCKER_STEP_NAME="script"
export WERCKER_REPORT_NUMBERS_FILE="/report/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/numbers.ini"
export WERCKER_REPORT_MESSAGE_FILE="/report/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/message.txt"
export WERCKER_REPORT_ARTIFACTS_DIR="/report/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/artifacts"
source "/pipeline/script-e51a4a54-1439-44ec-bec2-3e03e47d72f9/run.sh" < /dev/null
```

⋮

```
label: unnamed-chunk-3
  |.........................................                    |  69%
  ordinary text without R code

  |.............................................                |  77%
label: unnamed-chunk-4
Quitting from lines 94-99 (hw3.Rmd)
Error in data.frame(distance = LaQuinta$distToDennys, state = df_lq$state,  :
  arguments imply differing number of rows: 890, 0
Calls: render ... withCallingHandlers -> withVisible -> eval -> eval -> data.frame

Execution halted
Makefile:4: recipe for target 'hw3.html' failed
make: *** [hw3.html] Error 1
```

wercker.yml

```
box: rocker/hadleyverse

build:
  steps:
    - script:
        name: Update Packages
        code: |
          Rscript -e "update.packages(ask = FALSE)"
    - script:
        name: Check for allowed files
        code: |
          Rscript -e "source('https://raw.githubusercontent.com/Sta323-Sp16/Homework/master/hw3/hw3_whitelist.R')"
    - script:
        name: Check make runs
        code: |
          make
          Rscript -e "stopifnot(file.exists('hw3.html'))"
    - script:
        name: Check make clean runs
        code: |
          make clean
          Rscript -e "source('https://raw.githubusercontent.com/Sta323-Sp16/Homework/master/hw3/hw3_whitelist.R')"
```

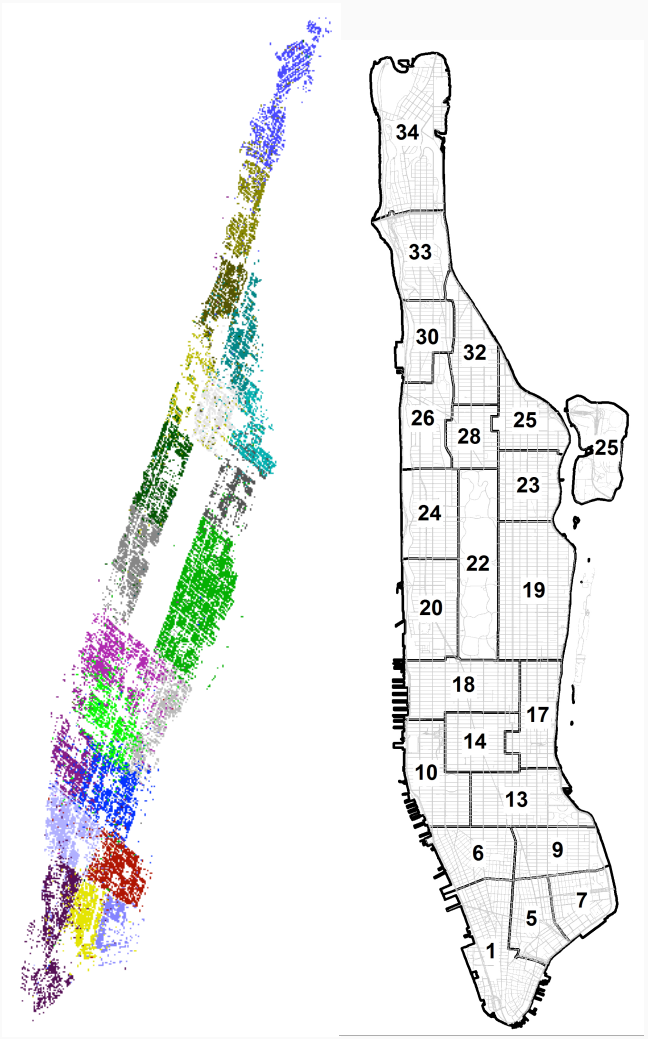
Parking Wars: Manhattan

NYC OpenData

Starting with

- Parking violations FY2014
  9.1M tickets
- MapPLUTO (Digital Tax Map)
  43K boundaries

find the geographic boundaries of all 22 police precincts in Manhattan.



wercker.yml

```
box: rocker/hadleyverse

build:
  steps:
    - script:
      name: Install libraries
      code: |
        printf "deb http://httpredir.debian.org/debian testing main\ndeb http://httpredir.debian.org/debian testing-upda
        apt-get update
        apt-get install -y --no-install-recommends curl libgdal-dev libgeos-dev libproj-dev
    - script:
      name: Install packages
      code: |
        Rscript -e "install.packages(c('jsonlite','rgdal','rgeos'), repos='https://cran.rstudio.com')"
    - script:
      name: Get scores
      code: |
        curl -s "$PP" > pp.Rdata
        curl -s "https://api.orchestrate.io/v0/hw5/Team0"  -u "$ORCH:" > Team0.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team1"  -u "$ORCH:" > Team1.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team2"  -u "$ORCH:" > Team2.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team3"  -u "$ORCH:" > Team3.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team4"  -u "$ORCH:" > Team4.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team5"  -u "$ORCH:" > Team5.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team6"  -u "$ORCH:" > Team6.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team7"  -u "$ORCH:" > Team7.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team8"  -u "$ORCH:" > Team8.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team9"  -u "$ORCH:" > Team9.json
        curl -s "https://api.orchestrate.io/v0/hw5/Team10" -u "$ORCH:" > Team10.json
    - script:
      name: Update scores
      code: |
          Rscript -e "source('https://raw.githubusercontent.com/Sta323-Sp16/Homework/master/hw5/update_score.R')" $TEAM
          curl -s "https://api.orchestrate.io/v0/hw5/$TEAM" \
            -XPUT \
            -H "Content-Type: application/json" \
            -u "$ORCH:" \
            -d "@$TEAM.json"
    - script:
      name: Show Leaderboard
      code: |
        Rscript -e "source('https://raw.githubusercontent.com/Sta323-Sp16/Homework/master/hw5/leaderboard.R')"
```

Lessons Learned

- Use github* for everything
  - Organizations + Teams are immensely useful in the classroom
  - Leverage the ecosystem
- Small investments in scripting / automation pay off
  - use the API (github, rgithub, httr)
- Think about ordering
  - More (explicit) testing -> less testing
  - Consider limitations (data size, infrastructure)

Questions, Comments?

rundel@gmail.com

github.com/rundel/

github.com/rundel/Presentations/

bit.ly/Sta523_2014

bit.ly/Sta523_2015

bit.ly/Sta323_2016