auk

extracting & processing eBird Data in R

Matt Strimas-Mackey

Cornell Lab of Ornithology • strimas.com • github.com/mstrimas

500 million observations

253

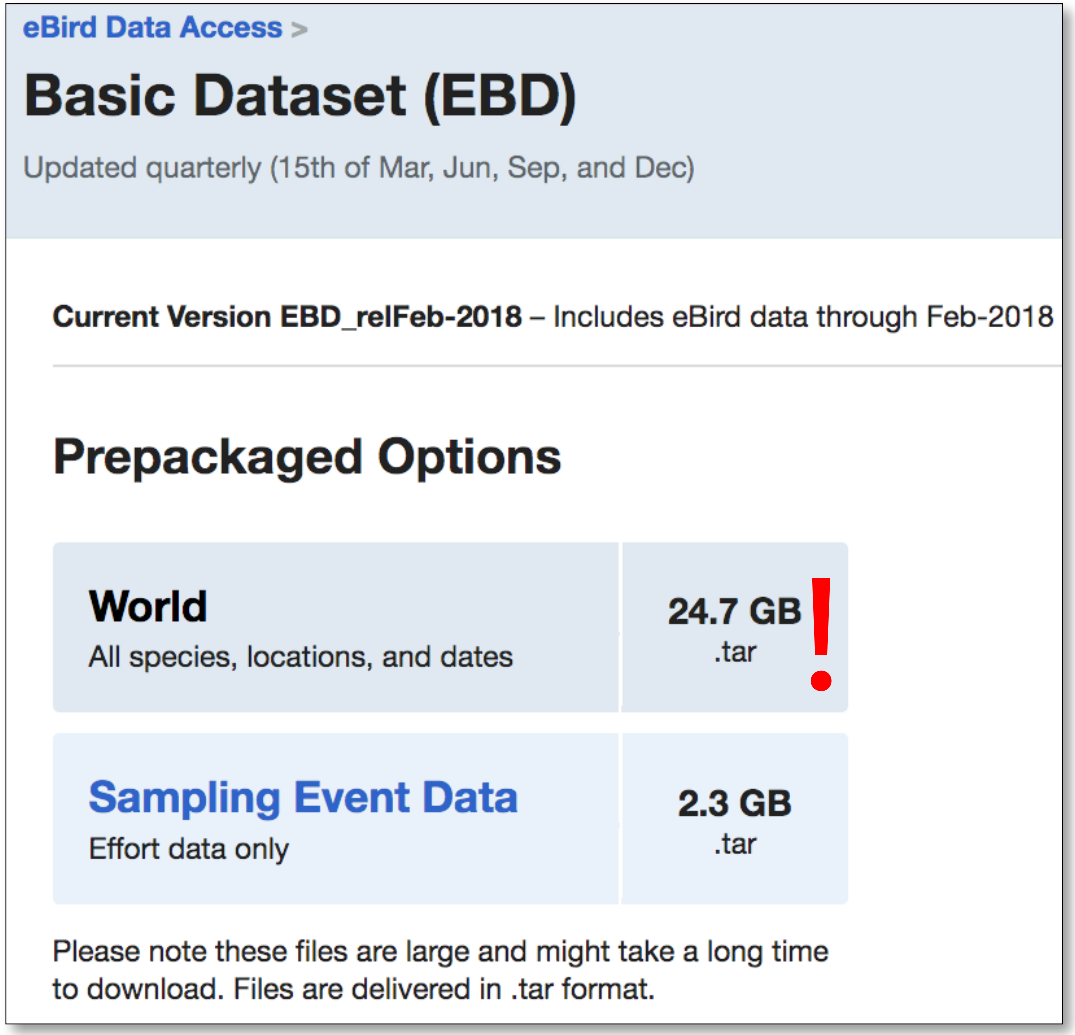
countries

10,374

species

SHOW ME THE DATA!!!

ebird.org/data/download/ebd



eBird Data Access >

Basic Dataset (EBD)

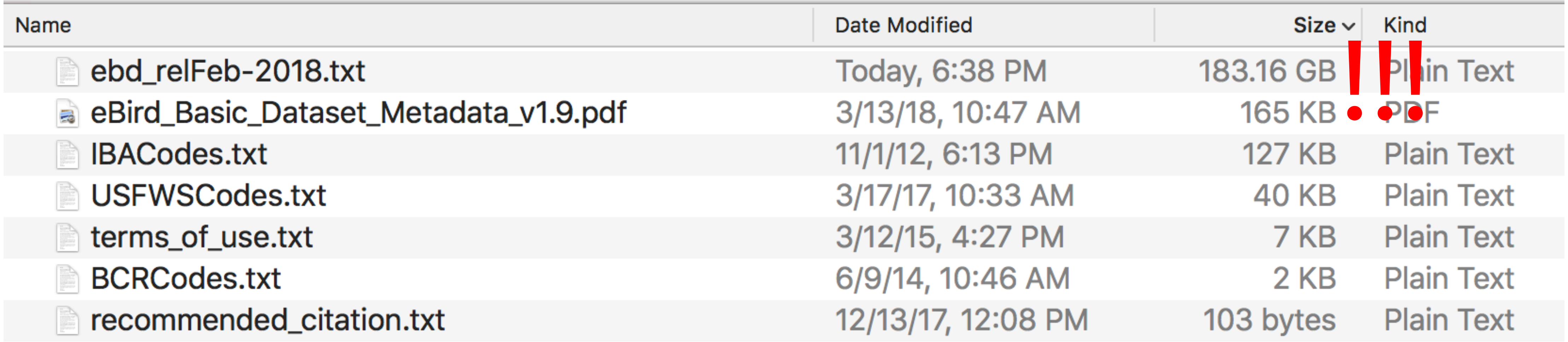
Updated quarterly (15th of Mar, Jun, Sep, and Dec)

Current Version EBD_relFeb-2018 – Includes eBird data through Feb-2018

Prepackaged Options

World All species, locations, and dates	**24.7 GB** .tar !
Sampling Event Data Effort data only	**2.3 GB** .tar

Please note these files are large and might take a long time to download. Files are delivered in .tar format.



Name	Date Modified	Size	Kind
ebd_relFeb-2018.txt	Today, 6:38 PM	183.16 GB	Plain Text
eBird_Basic_Dataset_Metadata_v1.9.pdf	3/13/18, 10:47 AM	165 KB	PDF
IBACodes.txt	11/1/12, 6:13 PM	127 KB	Plain Text
USFWSCodes.txt	3/17/17, 10:33 AM	40 KB	Plain Text
terms_of_use.txt	3/12/15, 4:27 PM	7 KB	Plain Text
BCRCodes.txt	6/9/14, 10:46 AM	2 KB	Plain Text
recommended_citation.txt	12/13/17, 12:08 PM	103 bytes	Plain Text

X

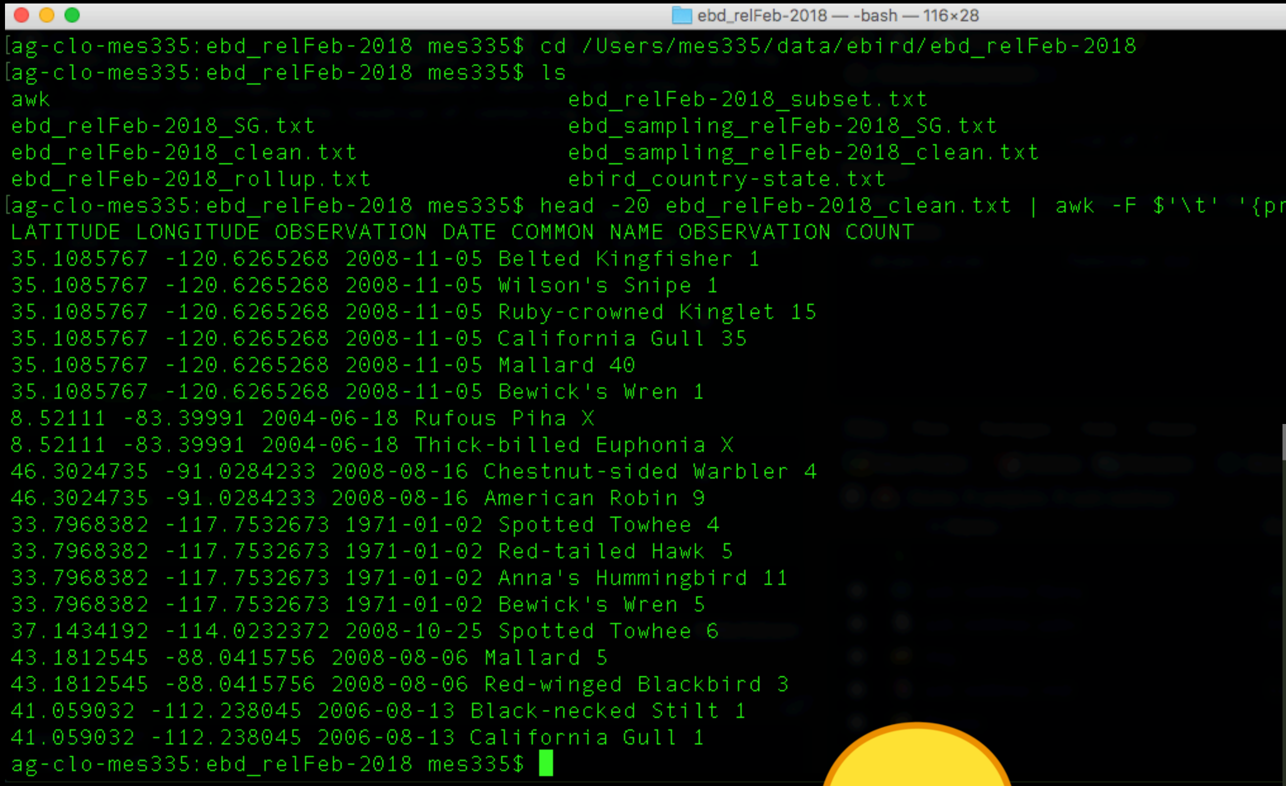
R

X

we need to make this

File

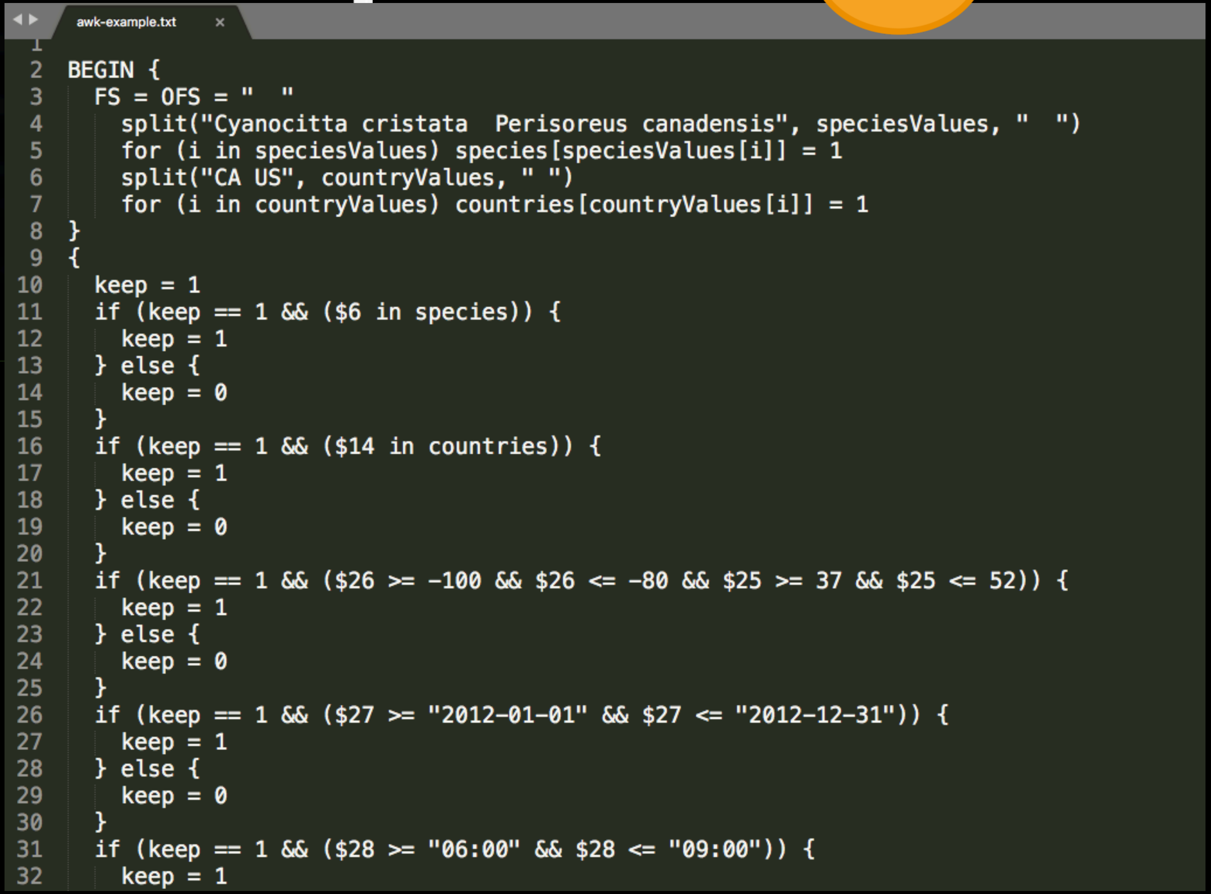
smaller



```
[ag-clo-mes335:ebd_relFeb-2018 mes335$ cd /Users/mes335/data/ebird/ebd_relFeb-2018
[ag-clo-mes335:ebd_relFeb-2018 mes335$ ls                                                                          ]
 awk                                        ebd_relFeb-2018_subset.txt
 ebd_relFeb-2018_SG.txt                     ebd_sampling_relFeb-2018_SG.txt
 ebd_relFeb-2018_clean.txt                  ebd_sampling_relFeb-2018_clean.txt
 ebd_relFeb-2018_rollup.txt                 ebird_country-state.txt
[ag-clo-mes335:ebd_relFeb-2018 mes335$ head -20 ebd_relFeb-2018_clean.txt | awk -F $'\t' '{print $25,$26,$27,$5,$9'} ]
 LATITUDE LONGITUDE OBSERVATION DATE COMMON NAME OBSERVATION COUNT
 35.1085767 -120.6265268 2008-11-05 Belted Kingfisher 1
 35.1085767 -120.6265268 2008-11-05 Wilson's Snipe 1
 35.1085767 -120.6265268 2008-11-05 Ruby-crowned Kinglet 15
 35.1085767 -120.6265268 2008-11-05 California Gull 35
 35.1085767 -120.6265268 2008-11-05 Mallard 40
 35.1085767 -120.6265268 2008-11-05 Bewick's Wren 1
 8.52111 -83.39991 2004-06-18 Rufous Piha X
 8.52111 -83.39991 2004-06-18 Thick-billed Euphonia X
 46.3024735 -91.0284233 2008-08-16 Chestnut-sided Warbler 4
 46.3024735 -91.0284233 2008-08-16 American Robin 9
 33.7968382 -117.7532673 1971-01-02 Spotted Towhee 4
 33.7968382 -117.7532673 1971-01-02 Red-tailed Hawk 5
 33.7968382 -117.7532673 1971-01-02 Anna's Hummingbird 11
 33.7968382 -117.7532673 1971-01-02 Bewick's Wren 5
 37.1434192 -114.0232372 2008-10-25 Spotted Towhee 6
 43.1812545 -88.0415756 2008-08-06 Mallard 5
 43.1812545 -88.0415756 2008-08-06 Red-winged Blackbird 3
 41.059032 -112.238045 2006-08-13 Black-necked Stilt 1
 41.059032 -112.238045 2006-08-13 California Gull 1
 ag-clo-mes335:ebd_relFeb-2018 mes335$ 
```

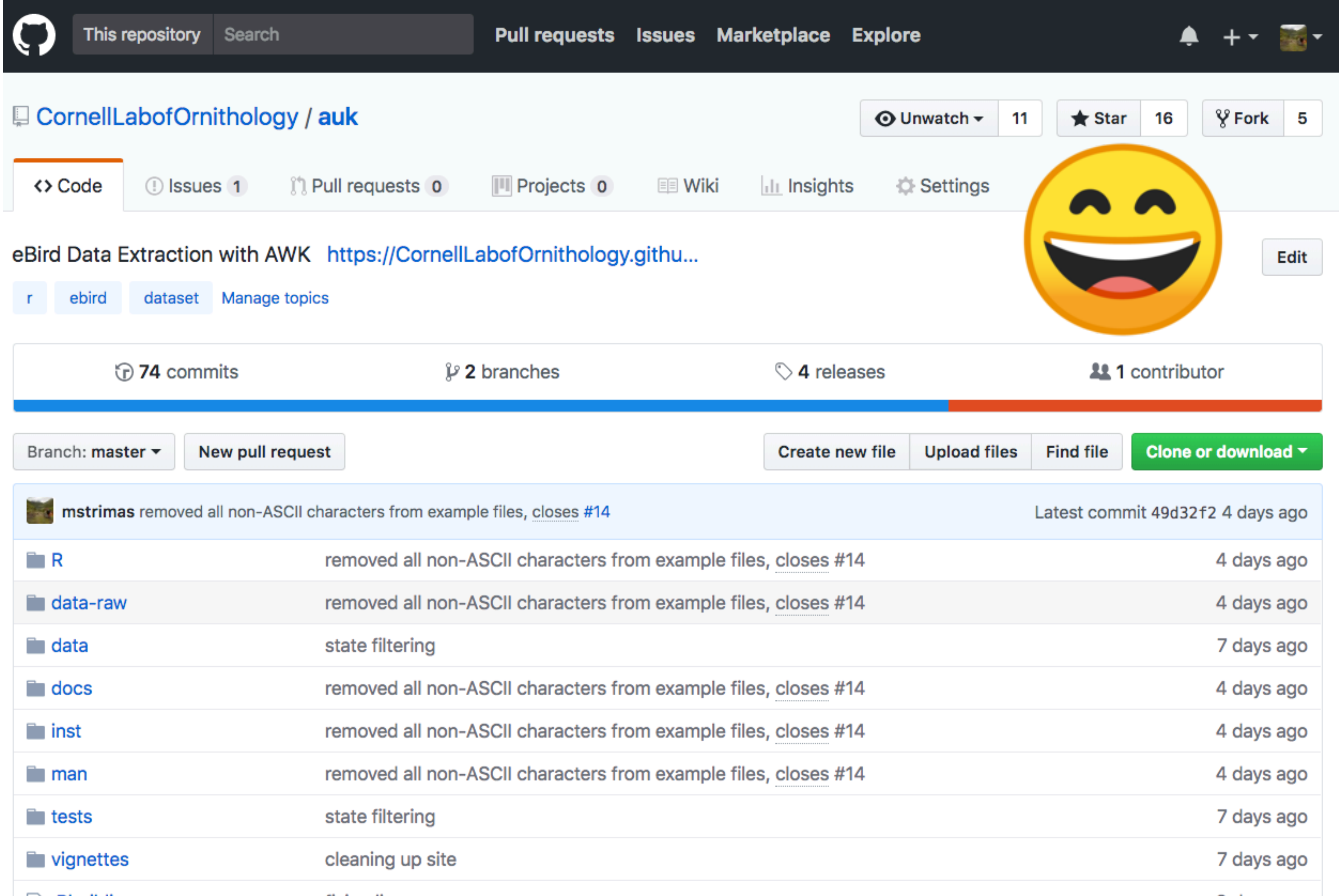
Terminal 😑

AWK 😕



```

BEGIN {
  FS = OFS = "  "
    split("Cyanocitta cristata  Perisoreus canadensis", speciesValues, "  ")
    for (i in speciesValues) species[speciesValues[i]] = 1
    split("CA US", countryValues, " ")
    for (i in countryValues) countries[countryValues[i]] = 1
}
{
  keep = 1
  if (keep == 1 && ($6 in species)) {
    keep = 1
  } else {
    keep = 0
  }
  if (keep == 1 && ($14 in countries)) {
    keep = 1
  } else {
    keep = 0
  }
  if (keep == 1 && ($26 >= -100 && $26 <= -80 && $25 >= 37 && $25 <= 52)) {
    keep = 1
  } else {
    keep = 0
  }
  if (keep == 1 && ($27 >= "2012-01-01" && $27 <= "2012-12-31")) {
    keep = 1
  } else {
    keep = 0
  }
  if (keep == 1 && ($28 >= "06:00" && $28 <= "09:00")) {
    keep = 1
```



CornellLabofOrnithology / auk

eBird Data Extraction with AWK https://CornellLabofOrnithology.githu...

r ebird dataset

74 commits 2 branches 4 releases 1 contributor

mstrimas removed all non-ASCII characters from example files, closes #14 Latest commit 49d32f2 4 days ago

R	removed all non-ASCII characters from example files, closes #14	4 days ago
data-raw	removed all non-ASCII characters from example files, closes #14	4 days ago
data	state filtering	7 days ago
docs	removed all non-ASCII characters from example files, closes #14	4 days ago
inst	removed all non-ASCII characters from example files, closes #14	4 days ago
man	removed all non-ASCII characters from example files, closes #14	4 days ago
tests	state filtering	7 days ago
vignettes	cleaning up site	7 days ago

auk workflow

1. clean
2. filter
3. import
4. pre-process
5. zero-fill

1 clean

remove problematic rows
drop free-text columns

2

filter

extract just what you need
reduce file to a manageable size

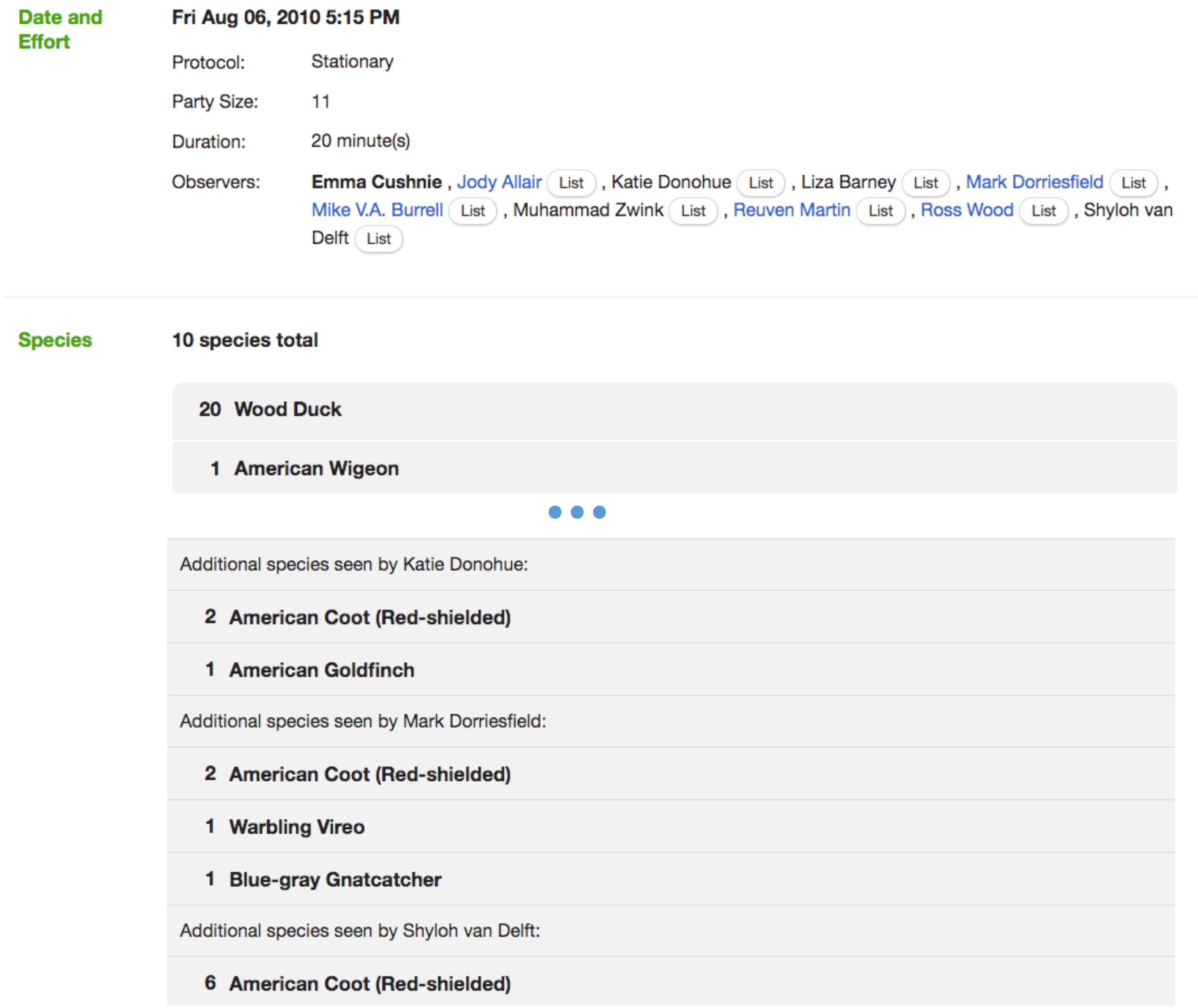
3

import

import to a data frame
choose clean names and sensible data types

4 pre-process

collapse group checklists



Date and Effort

Fri Aug 06, 2010 5:15 PM

Protocol: Stationary

Party Size: 11

Duration: 20 minute(s)

Observers: **Emma Cushnie** , Jody Allair List , Katie Donohue List , Liza Barney List , Mark Dorriesfield List , Mike V.A. Burrell List , Muhammad Zwink List , Reuven Martin List , Ross Wood List , Shyloh van Delft List

Species

10 species total

20 Wood Duck

1 American Wigeon

● ● ●

Additional species seen by Katie Donohue:

2 American Coot (Red-shielded)

1 American Goldfinch

Additional species seen by Mark Dorriesfield:

2 American Coot (Red-shielded)

1 Warbling Vireo

1 Blue-gray Gnatcatcher

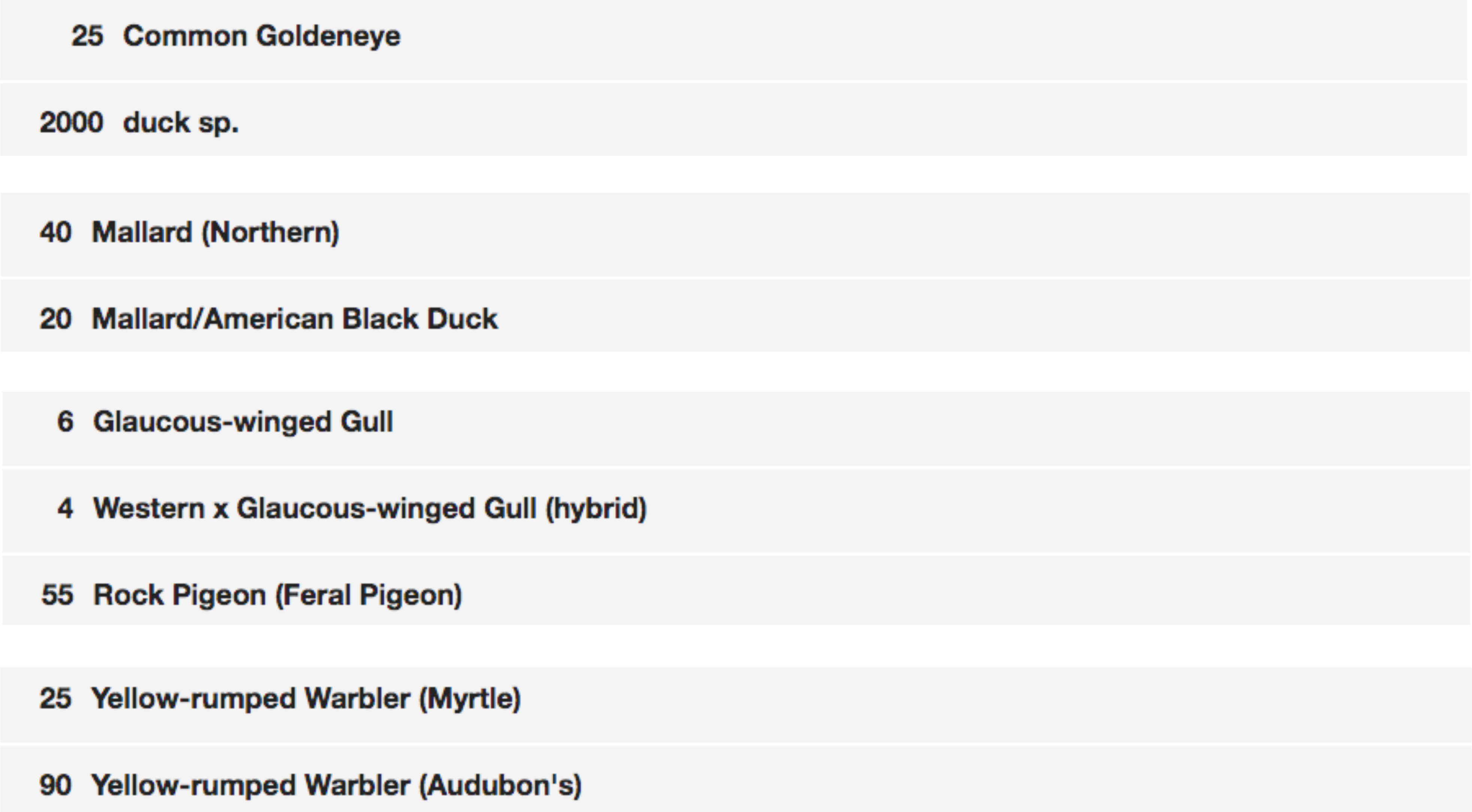
Additional species seen by Shyloh van Delft:

6 American Coot (Red-shielded)

4

pre-process

taxonomic roll-up

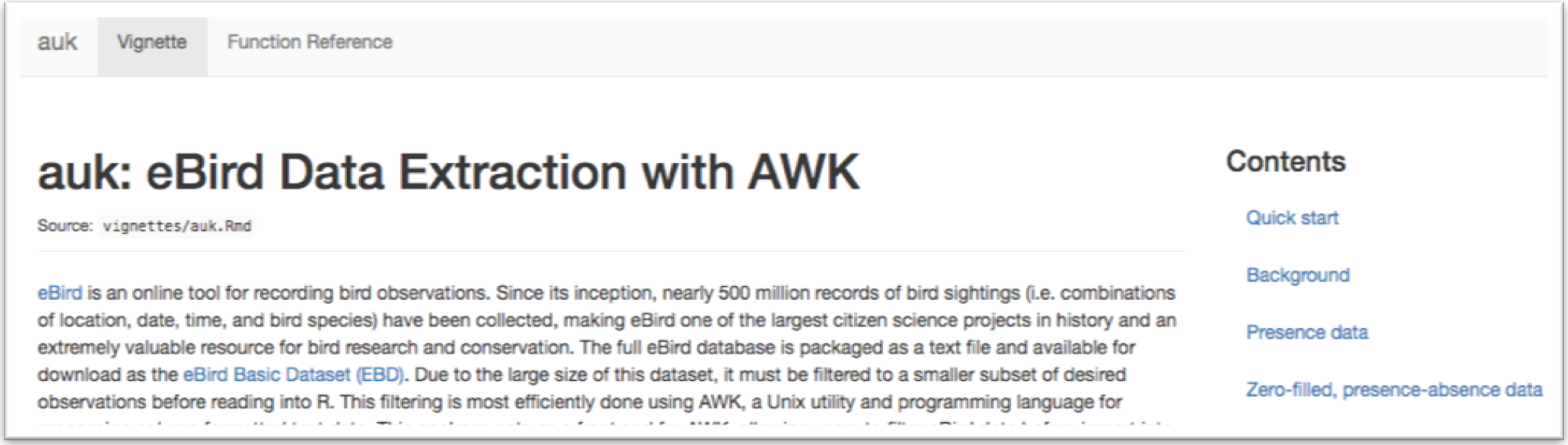


25	Common Goldeneye
2000	duck sp.
40	Mallard (Northern)
20	Mallard/American Black Duck
6	Glaucous-winged Gull
4	Western x Glaucous-winged Gull (hybrid)
55	Rock Pigeon (Feral Pigeon)
25	Yellow-rumped Warbler (Myrtle)
90	Yellow-rumped Warbler (Audubon's)

5 zero-fill

complete checklists give implicit zeros
use sampling event data to make zeros explicit

Read the vignette: CornellLabofOrnithology.github.io/auk



auk: eBird Data Extraction with AWK

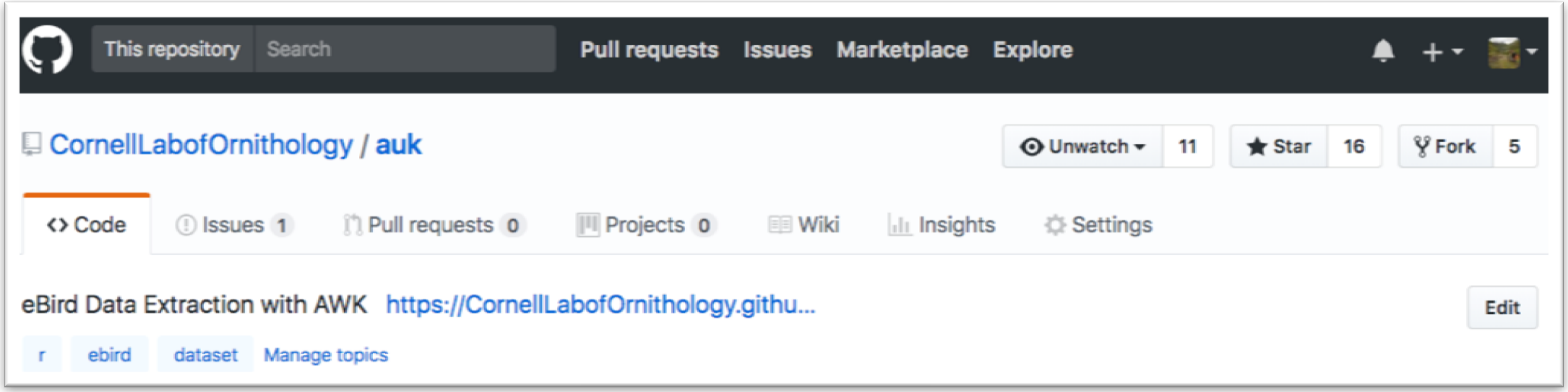
Source: `vignettes/auk.Rmd`

eBird is an online tool for recording bird observations. Since its inception, nearly 500 million records of bird sightings (i.e. combinations of location, date, time, and bird species) have been collected, making eBird one of the largest citizen science projects in history and an extremely valuable resource for bird research and conservation. The full eBird database is packaged as a text file and available for download as the eBird Basic Dataset (EBD). Due to the large size of this dataset, it must be filtered to a smaller subset of desired observations before reading into R. This filtering is most efficiently done using AWK, a Unix utility and programming language for

Contents

- Quick start
- Background
- Presence data
- Zero-filled, presence-absence data

File an issue: github.com/CornellLabofOrnithology/auk



eBird Data Extraction with AWK https://CornellLabofOrnithology.githu...

Send me an email: mes335@cornell.edu