```
$ cat title | cowsay -W 25

 _________________________
/ Obtaining, Scrubbing, \
| and Exploring Data at |
| the Command Line      |
|                       |
| Jeroen Janssens       |
\ @jeroenhjanssens      /
 -------------------------
        \   ^__^
         \  (oo)\_______
            (__)\       )\/\
                ||----w |
                ||     ||
```

Overview

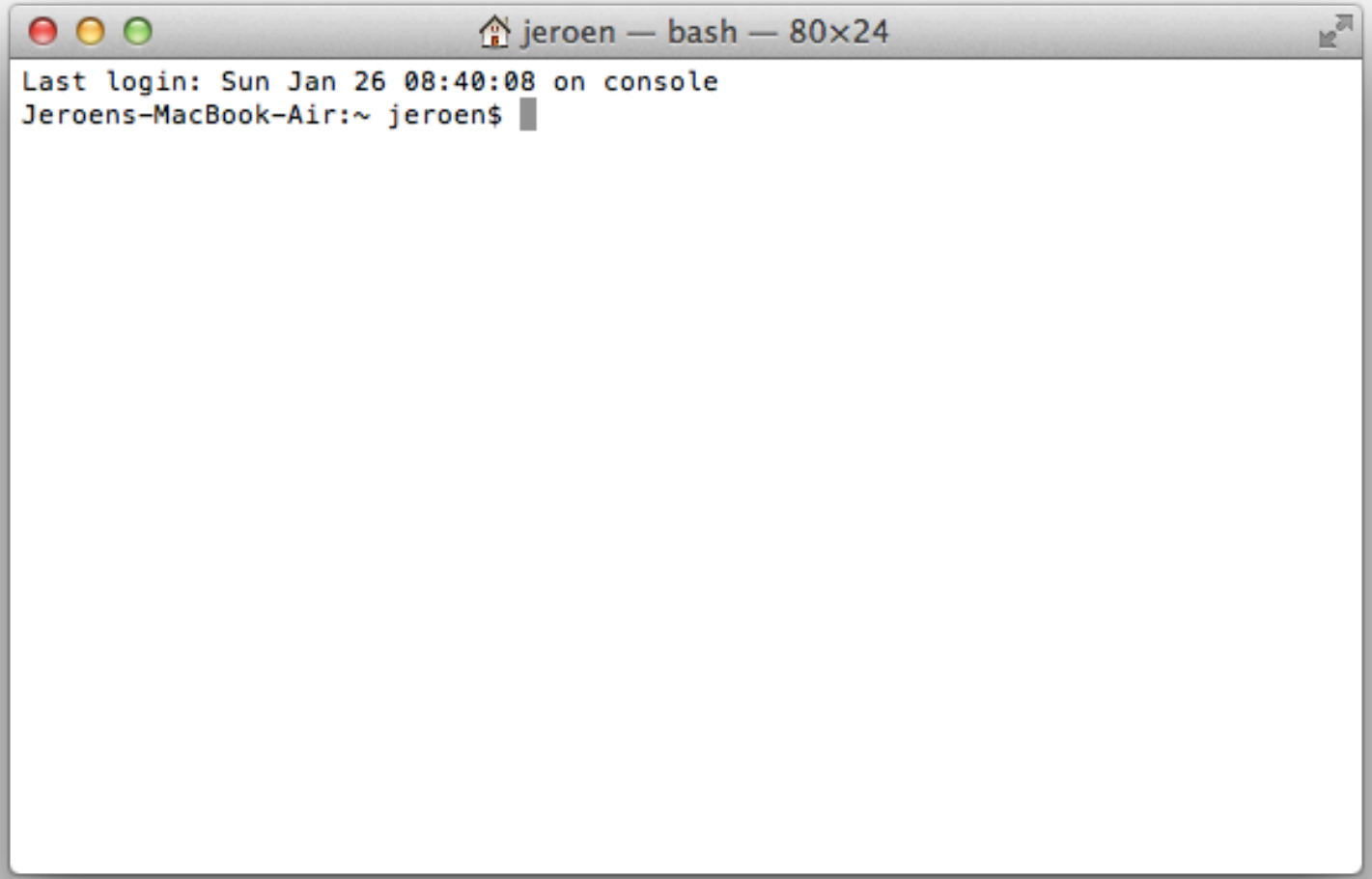
- Motivation
- Essential tools and concepts
- Web scraping
- Exploration
- Data science toolbox
- Parallelization
- Workflow management

Motivation

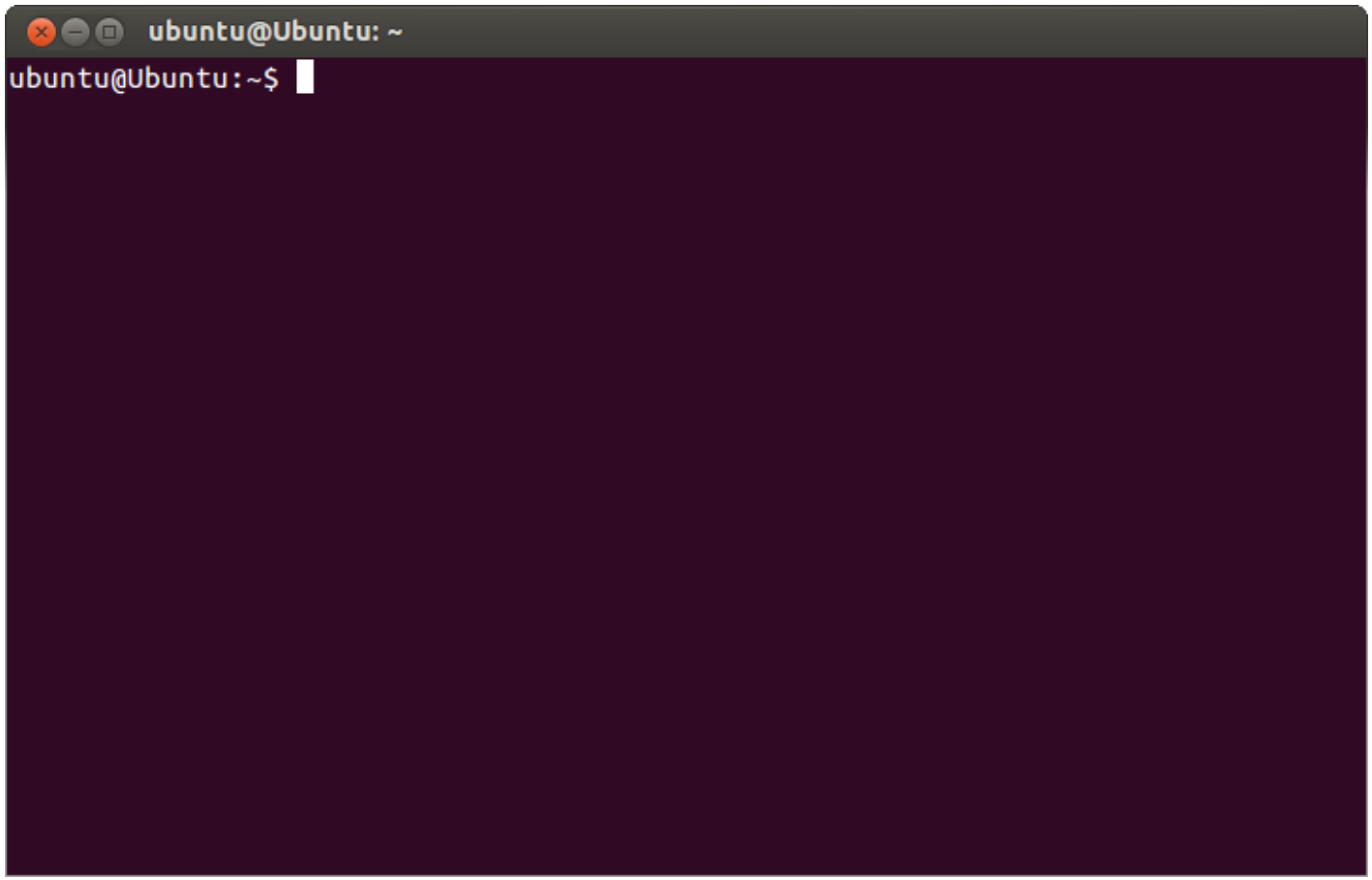
Data science is OSEMN

- **O**btaining data
- **S**crubbing data
- **E**xploring data
- **M**odeling data
- i**N**terpreting data

Command line on Mac OS X



Command line on Ubuntu



The command line is awesome

- Play with your data
- Combine tools
- Many tools available
- Automatable
- Many servers run Linux
- One overarching environment

Essential Tools and Concepts

Command-line tool is an umbrella term

- Executable
- Script
- One-liner
- Shell command
- Shell function
- Alias

Unix philosophy

Write command-line tools that:

- Do one thing and do it well
- Work together
- Handle text streams

Tips dataset

```
$ cat tips.csv
bill,tip,sex,smoker,day,time,size
16.99,1.01,Female,No,Sun,Dinner,2
10.34,1.66,Male,No,Sun,Dinner,3
21.01,3.5,Male,No,Sun,Dinner,3
23.68,3.31,Male,No,Sun,Dinner,2
24.59,3.61,Female,No,Sun,Dinner,4
25.29,4.71,Male,No,Sun,Dinner,4
8.77,2.0,Male,No,Sun,Dinner,2
26.88,3.12,Male,No,Sun,Dinner,4
15.04,1.96,Male,No,Sun,Dinner,2
14.78,3.23,Male,No,Sun,Dinner,2
10.27,1.71,Male,No,Sun,Dinner,2
35.26,5.0,Female,No,Sun,Dinner,4
```

Reference manual

```
$ man cat
CAT(1)                    User Commands                    CAT(1)

NAME
       cat  - concatenate files and print on the standard
       output
SYNOPSIS
       cat [OPTION]... [FILE]...
DESCRIPTION
       Concatenate FILE(s), or standard input,  to  stand
       ard output.

       -A, --show-all
              equivalent to -vET
```

Looking at files

```
$ cat tips.csv | csvlook
|-------+------+--------+--------+------+--------+-------|
|  bill | tip  | sex    | smoker | day  | time   | size  |
|-------+------+--------+--------+------+--------+-------|
|  16.99 | 1.01 | Female | No     | Sun  | Dinner | 2     |
|  10.34 | 1.66 | Male   | No     | Sun  | Dinner | 3     |
|  21.01 | 3.5  | Male   | No     | Sun  | Dinner | 3     |
|  23.68 | 3.31 | Male   | No     | Sun  | Dinner | 2     |
|  24.59 | 3.61 | Female | No     | Sun  | Dinner | 4     |
|  25.29 | 4.71 | Male   | No     | Sun  | Dinner | 4     |
|  8.77  | 2.0  | Male   | No     | Sun  | Dinner | 2     |
|  26.88 | 3.12 | Male   | No     | Sun  | Dinner | 4     |
|  15.04 | 1.96 | Male   | No     | Sun  | Dinner | 2     |
|  14.78 | 3.23 | Male   | No     | Sun  | Dinner | 2     |
```

Looking at files

```
$ cat tips.csv | less
$ cat tips.csv | head -n 3 | csvlook
|--------+------+--------+--------+-----+--------+--------|
|  bill  | tip  | sex    | smoker | day | time   | size   |
|--------+------+--------+--------+-----+--------+--------|
|  16.99 | 1.01 | Female | No     | Sun | Dinner | 2      |
|  10.34 | 1.66 | Male   | No     | Sun | Dinner | 3      |
|--------+------+--------+--------+-----+--------+--------|
$ < tips.csv tail -n 3 | csvlook -H
|-------+------+--------+-----+------+--------+----|
|  22.67 | 2.0  | Male   | Yes | Sat  | Dinner | 2  |
|  17.82 | 1.75 | Male   | No  | Sat  | Dinner | 2  |
|  18.78 | 3.0  | Female | No  | Thur | Dinner | 2  |
|-------+------+--------+-----+------+--------+----|
```

Filtering lines

```
$ grep 'Lunch' tips.csv | csvlook -H
|--------+------+--------+-----+------+--------+------|
|  27.2  | 4.0  | Male   | No  | Thur | Lunch  | 4    |
|  22.76 | 3.0  | Male   | No  | Thur | Lunch  | 2    |
|  17.29 | 2.71 | Male   | No  | Thur | Lunch  | 2    |
|  19.44 | 3.0  | Male   | Yes | Thur | Lunch  | 2    |
|  16.66 | 3.4  | Male   | No  | Thur | Lunch  | 2    |
|  10.07 | 1.83 | Female | No  | Thur | Lunch  | 1    |
|  32.68 | 5.0  | Male   | Yes | Thur | Lunch  | 2    |
|  15.98 | 2.03 | Male   | No  | Thur | Lunch  | 2    |
|  34.83 | 5.17 | Female | No  | Thur | Lunch  | 4    |
|  13.03 | 2.0  | Male   | No  | Thur | Lunch  | 2    |
|  18.28 | 4.0  | Male   | No  | Thur | Lunch  | 2    |
|  24.71 | 5.85 | Male   | No  | Thur | Lunch  | 2    |
```

Filtering lines

```
$ cat tips.csv | awk -F, '$7 !~ /[1-4]/' | csvlook
|-------+------+--------+--------+-----+--------+-------|
|  bill | tip  | sex    | smoker | day | time   | size  |
|-------+------+--------+--------+-----+--------+-------|
|  29.8 | 4.2  | Female | No     | Thur | Lunch  | 6     |
|  34.3 | 6.7  | Male   | No     | Thur | Lunch  | 6     |
|  41.19 | 5.0 | Male   | No     | Thur | Lunch  | 5     |
|  27.05 | 5.0 | Female | No     | Thur | Lunch  | 6     |
|  29.85 | 5.14 | Female | No     | Sun  | Dinner | 5     |
|  48.17 | 5.0 | Male   | No     | Sun  | Dinner | 6     |
|  20.69 | 5.0 | Male   | No     | Sun  | Dinner | 5     |
|  30.46 | 2.0 | Male   | Yes    | Sun  | Dinner | 5     |
|  28.15 | 3.0 | Male   | Yes    | Sat  | Dinner | 5     |
|-------+------+--------+--------+-----+--------+-------|
```

Filtering lines

```
$ csvgrep -c size -r "[1-4]" -i tips.csv | csvlook
|--------+------+--------+--------+------+--------+--------|
|  bill  | tip  | sex    | smoker | day  | time   | size   |
|--------+------+--------+--------+------+--------+--------|
|  29.8  | 4.2  | Female | No     | Thur | Lunch  | 6      |
|  34.3  | 6.7  | Male   | No     | Thur | Lunch  | 6      |
|  41.19 | 5.0  | Male   | No     | Thur | Lunch  | 5      |
|  27.05 | 5.0  | Female | No     | Thur | Lunch  | 6      |
|  29.85 | 5.14 | Female | No     | Sun  | Dinner | 5      |
|  48.17 | 5.0  | Male   | No     | Sun  | Dinner | 6      |
|  20.69 | 5.0  | Male   | No     | Sun  | Dinner | 5      |
|  30.46 | 2.0  | Male   | Yes    | Sun  | Dinner | 5      |
|  28.15 | 3.0  | Male   | Yes    | Sat  | Dinner | 5      |
|--------+------+--------+--------+------+--------+--------|
```

Extracting columns

```
$ csvgrep -c size -r "[1-4]" -i tips.csv > size56.csv
$ cut size56.csv -d, -f1,2
bill,tip
29.8,4.2
34.3,6.7
41.19,5.0
27.05,5.0
29.85,5.14
48.17,5.0
20.69,5.0
30.46,2.0
28.15,3.0
```

Extracting columns

```
$ awk -F, '{print $1","$2}' size56.csv
bill,tip
29.8,4.2
34.3,6.7
41.19,5.0
27.05,5.0
29.85,5.14
48.17,5.0
20.69,5.0
30.46,2.0
28.15,3.0
```

Extracting columns

```
$ csvcut size56.csv -c bill,tip
bill,tip
29.8,4.2
34.3,6.7
41.19,5.0
27.05,5.0
29.85,5.14
48.17,5.0
20.69,5.0
30.46,2.0
28.15,3.0
```

Extracting words

```
$ curl -s 'http://www.gutenberg.org/cache/epub/76/pg76.txt'|
> tee finn | grep -oE '\w+' | tee words
The
Project
Gutenberg
EBook
of
Adventures
of
Huckleberry
Finn
Complete
by
Mark
```

Sorting and counting

```
$ wc finn
 12361 114266 610157 finn

$ < words grep '^a' | grep 'e$' | sort | uniq -c | sort -rn
     77 are
     21 alone
     20 ashore
     19 above
     13 alive
      9 awhile
      9 apiece
      7 axe
      7 agree
      5 anywhere
```

Replacing data

```
$ < finn tr '[a-z]' '[A-Z]' > /dev/null
$ < finn tr '[:lower:]' '[:upper:]' | head -n 14

THE PROJECT GUTENBERG EBOOK OF ADVENTURES OF HUCKLEBERRY FINN,
BY MARK TWAIN (SAMUEL CLEMENS)

THIS EBOOK IS FOR THE USE OF ANYONE ANYWHERE AT NO COST AND WI
NO RESTRICTIONS WHATSOEVER. YOU MAY COPY IT, GIVE IT AWAY OR RE
IT UNDER THE TERMS OF THE PROJECT GUTENBERG LICENSE INCLUDED W
EBOOK OR ONLINE AT WWW.GUTENBERG.NET

TITLE: ADVENTURES OF HUCKLEBERRY FINN, COMPLETE

AUTHOR: MARK TWAIN (SAMUEL CLEMENS)
```

Replacing data

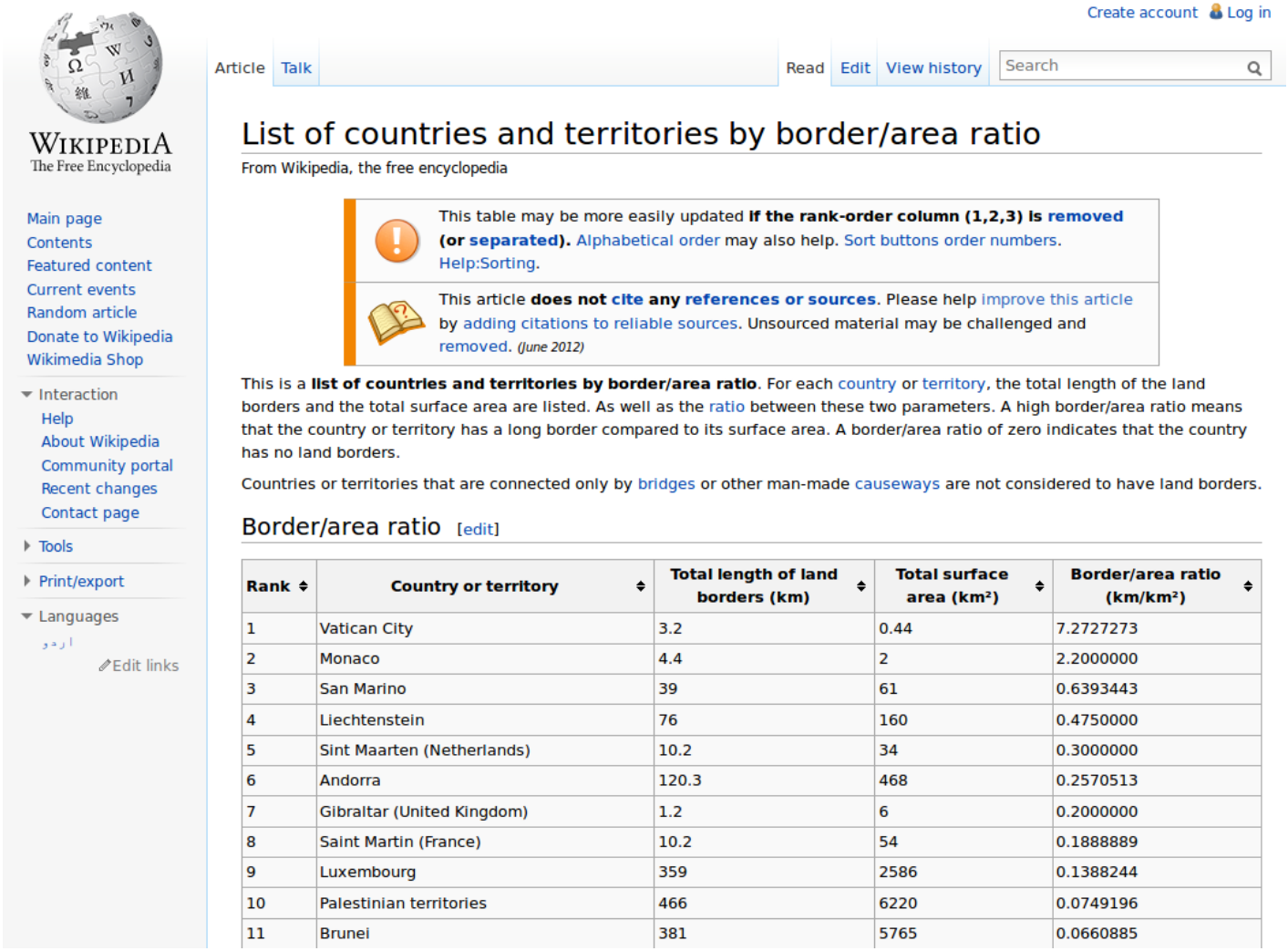
```
$ < finn sed 's/ /_/g' | head -n 14

The_Project_Gutenberg_EBook_of_Adventures_of_Huckleberry_Finn,_
by_Mark_Twain_(Samuel_Clemens)

This_eBook_is_for_the_use_of_anyone_anywhere_at_no_cost_and_wi
no_restrictions_whatsoever._You_may_copy_it,_give_it_away_or_re
it_under_the_terms_of_the_Project_Gutenberg_License_included_w
eBook_or_online_at_www.gutenberg.net

Title:_Adventures_of_Huckleberry_Finn,_Complete

Author:_Mark_Twain_(Samuel_Clemens)
```

Summing values

```
$ < tips.csv | tail -n +2 | cut -d, -f1 | paste -s -d+
16.99+10.34+21.01+23.68+24.59+25.29+8.77+26.88+15.04+14.78+
10.27+35.26+15.42+18.43+14.83+21.58+10.33+16.29+16.97+20.65
+17.92+20.29+15.77+39.42+19.82+17.81+13.37+12.69+21.7+19.65
+9.55+18.35+15.06+20.69+17.78+24.06+16.31+16.93+18.69+ ...

$ < tips.csv | tail -n +2 | cut -d, -f1 | paste -s -d+ | bc
4827.77

$ < tips.csv awk -F, '{ sum+=$1} END {print sum}'
4827.77

$ < tips.csv Rio -e 'sum(df$bill)'
[1] 4827.77
```

Web Scraping

Extracting data from HTML



WIKIPEDIA
The Free Encyclopedia

Create account Log in

Article Talk Read Edit View history Search

List of countries and territories by border/area ratio

From Wikipedia, the free encyclopedia

This table may be more easily updated **If the rank-order column (1,2,3) is removed (or separated).** Alphabetical order may also help. Sort buttons order numbers. Help:Sorting.

This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(June 2012)*

This is a **list of countries and territories by border/area ratio**. For each country or territory, the total length of the land borders and the total surface area are listed. As well as the ratio between these two parameters. A high border/area ratio means that the country or territory has a long border compared to its surface area. A border/area ratio of zero indicates that the country has no land borders.

Countries or territories that are connected only by bridges or other man-made causeways are not considered to have land borders.

Border/area ratio [edit]

Rank	Country or territory	Total length of land borders (km)	Total surface area (km²)	Border/area ratio (km/km²)
1	Vatican City	3.2	0.44	7.2727273
2	Monaco	4.4	2	2.2000000
3	San Marino	39	61	0.6393443
4	Liechtenstein	76	160	0.4750000
5	Sint Maarten (Netherlands)	10.2	34	0.3000000
6	Andorra	120.3	468	0.2570513
7	Gibraltar (United Kingdom)	1.2	6	0.2000000
8	Saint Martin (France)	10.2	54	0.1888889
9	Luxembourg	359	2586	0.1388244
10	Palestinian territories	466	6220	0.0749196
11	Brunei	381	5765	0.0660885

Download HTML using curl

```
$ curl -s 'http://en.wikipedia.org/wiki/List_of_countries_ar
<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-nojs">
<head>
<meta charset="UTF-8" /><title>List of countries and territo
<meta name="generator" content="MediaWiki 1.23wmf10" />
<link rel="alternate" type="application/x-wiki" title="Edit
<link rel="edit" title="Edit this page" href="/w/index.php?t
<link rel="apple-touch-icon" href="//bits.wikimedia.org/appl
<link rel="shortcut icon" href="//bits.wikimedia.org/favicon
<link rel="search" type="application/opensearchdescription+x
<link rel="EditURI" type="application/rsd+xml" href="//en.wi
<link rel="copyright" href="//creativecommons.org/licenses/b
```

Scrape element with CSS selectors

```
$ < wiki.html scrape -b -e 'table.wikitable > \
> tr:not(:first-child)'
<!DOCTYPE html>
<html>
<body>
<tr>
<td>1</td>
<td>Vatican City</td>
<td>3.2</td>
<td>0.44</td>
<td>7.2727273</td>
</tr>
```

Convert to JSON using xml2json

```
$ < table.html xml2json | jq '.'
{
  "html": {
    "body": {
      "tr": [
        {
          "td": [
            {
              "$t": "1"
            },
            {
              "$t": "Vatican City"
            },
```

Transform JSON using jq

```
$ < table.json jq -c '.html.body.tr[] | {country: .td[1][],
> border: .td[2][], surface: .td[3][], ratio: .td[4][]}'
{"ratio":"7.2727273","surface":"0.44","border":"3.2","countr
{"ratio":"2.2000000","surface":"2","border":"4.4","country"
{"ratio":"0.6393443","surface":"61","border":"39","country"
{"ratio":"0.4750000","surface":"160","border":"76","country
{"ratio":"0.3000000","surface":"34","border":"10.2","country
{"ratio":"0.2570513","surface":"468","border":"120.3","coun
{"ratio":"0.2000000","surface":"6","border":"1.2","country"
{"ratio":"0.1888889","surface":"54","border":"10.2","country
{"ratio":"0.1388244","surface":"2586","border":"359","countr
{"ratio":"0.0749196","surface":"6220","border":"466","countr
```

Convert to CSV with json2csv

```
$ < countries.json json2csv -p -k border,surface | csvlook
|----------+-----------|
|  border  | surface   |
|----------+-----------|
|  3.2     | 0.44      |
|  4.4     | 2         |
|  39      | 61        |
|  76      | 160       |
|  10.2    | 34        |
|  120.3   | 468       |
|  1.2     | 6         |
|  10.2    | 54        |
|  359     | 2586      |
|  466     | 6220      |
```

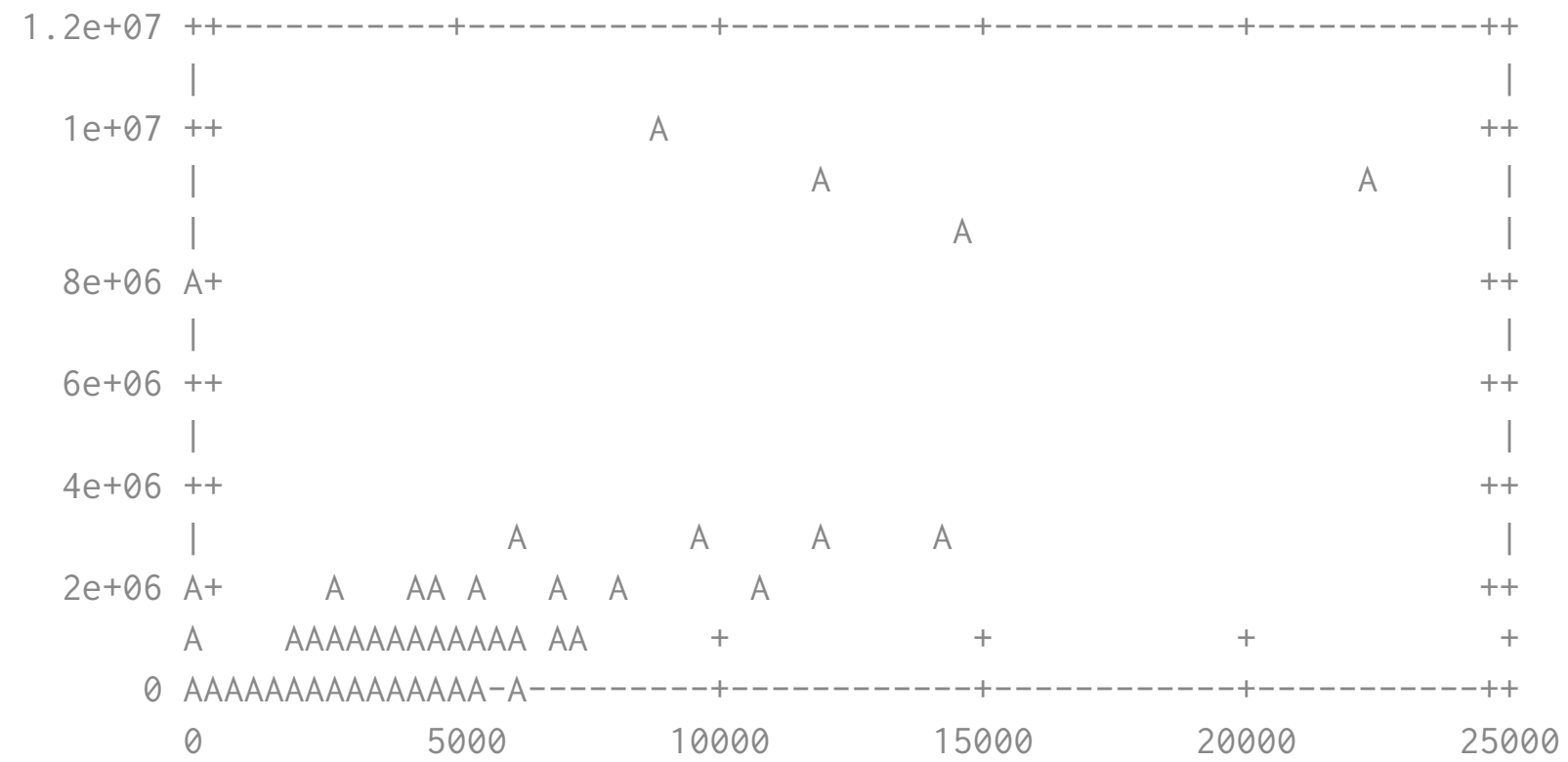
Behold, the beast

```
$ curl -s 'http://en.wikipedia.org/wiki/List_of_countries
> _and_territories_by_border/area_ratio' |
> scrape -be 'table.wikitable > tr:not(:first-child)' |
> xml2json | jq -c '.html.body.tr[] | {country: .td[1][],
> border: .td[2][], surface: .td[3][], ratio: .td[4][]}' |
> json2csv -p -k=border,surface | csvlook
```

Exploration

Gnuplot

```
$ < d.csv gnuplot -e 'set term dumb; set datafile separator ","; plot "-"'
```



Statistics at the command line

```
$ < tips.csv tail -n +2 | cut -d, -f2 | qstats
Min.      1
1st Qu.   2
Median    2.9
Mean      2.99828
3rd Qu.   3.575
Max.      10
Range     9
Std Dev.  1.3808
Length    244

$ < tips.csv | tail -n +2 | cut -d, -f2 | qstats -m
2.99828
```

Statistics at the command line

```
$ < tips.csv tail -n +2 | cut -d, -f2 | histogram.py -b10
NumSamples = 244; Min = 1.00; Max = 10.00
Mean = 2.998279; Variance = 1.906609; SD = 1.380800
each * represents a count of 1
 1.0000 -  1.9000 [41]: ****************************************
 1.9000 -  2.8000 [79]: ********************************************
 2.8000 -  3.7000 [66]: ************************************************
 3.7000 -  4.6000 [27]: ***************************
 4.6000 -  5.5000 [19]: *******************
 5.5000 -  6.4000 [ 5]: *****
 6.4000 -  7.3000 [ 4]: ****
 7.3000 -  8.2000 [ 1]: *
 8.2000 -  9.1000 [ 1]: *
 9.1000 - 10.0000 [ 1]: *
```

Rio: Making R part of the pipeline

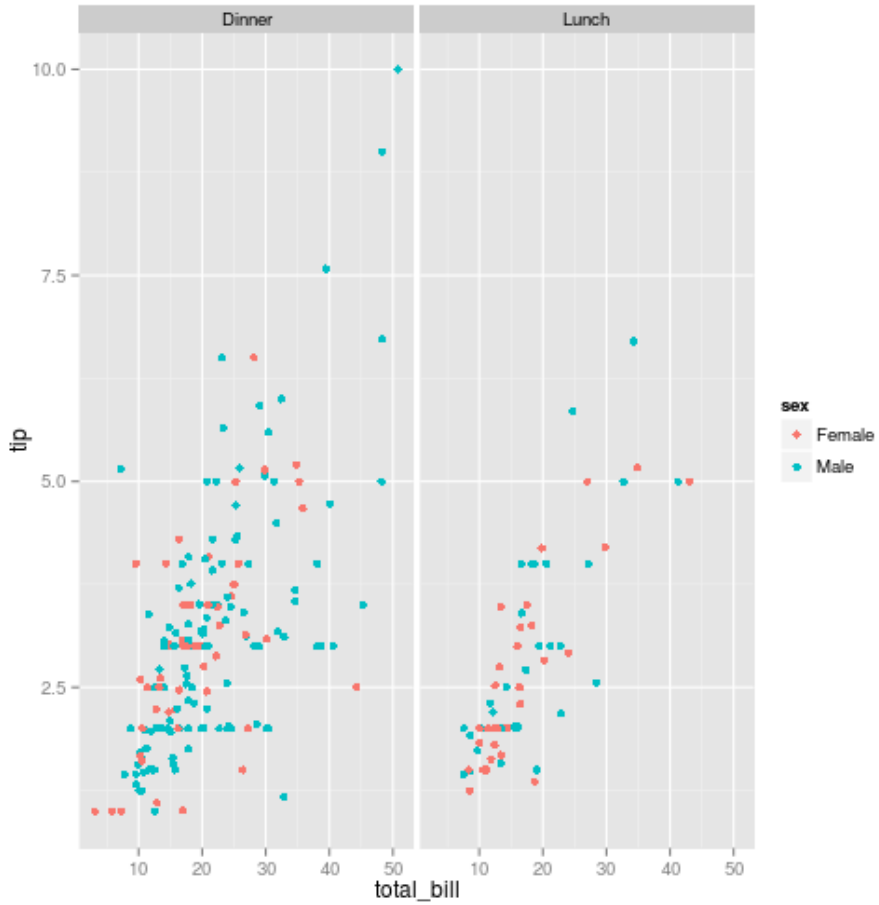
```
$ < tips.csv Rio -se 'sqldf("select time,count(*) from
> df group by time;")'
time,count(*)
Dinner,176
Lunch,68
```

Rio: Making R part of the pipeline

```
$ < tips.csv Rio -se 'sqldf("select time,count(*) from
> df group by time;")'
time,count(*)
Dinner,176
Lunch,68

$ < tips.csv | csvcut -c time | tail -n+2 | sort | uniq -c
    176 Dinner
     68 Lunch
```

ggplot at the command line

```
$ < tips.csv Rio -ge 'g+geom_point(aes(total_bill,tip,
> colour=sex))+facet_wrap(~ time)' | display
```

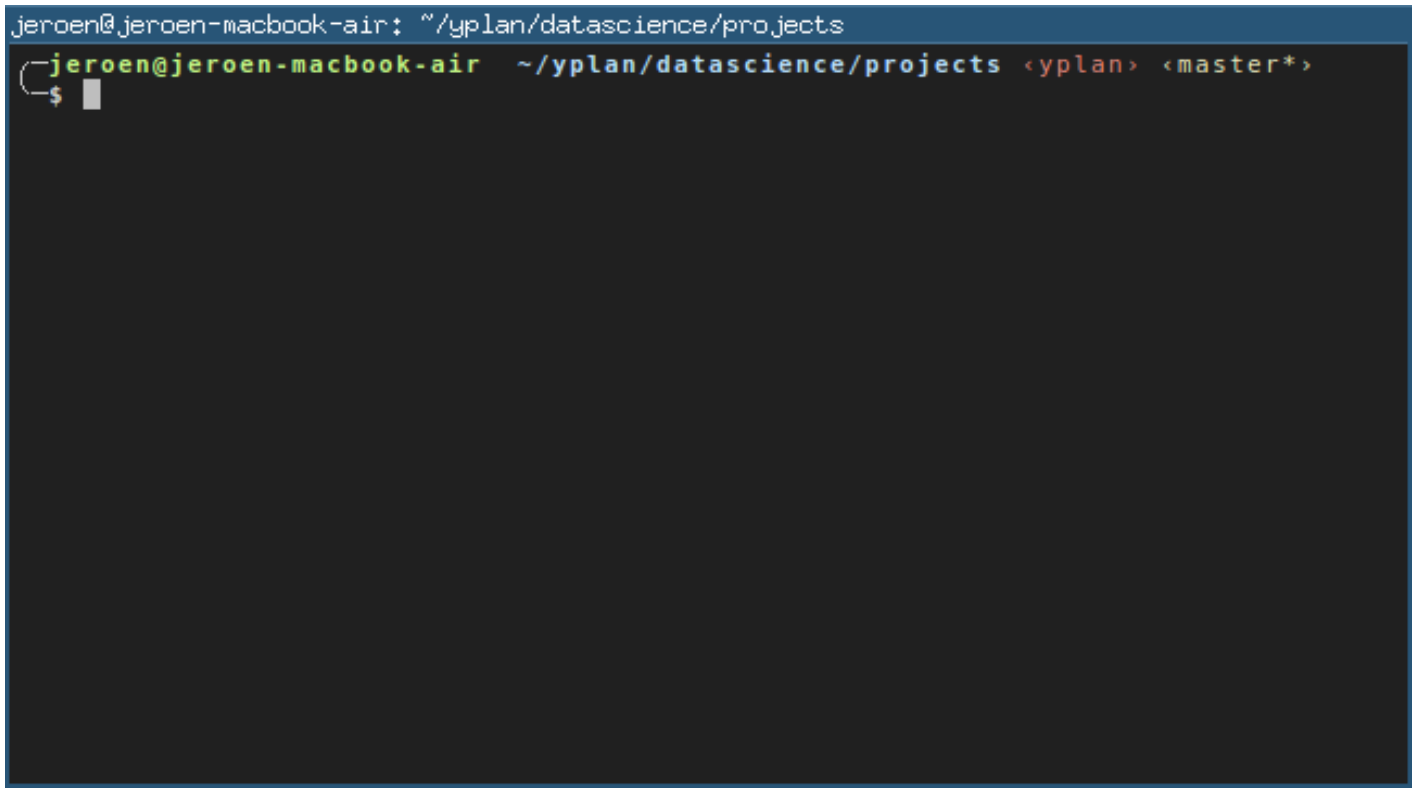


Data Science Toolbox

Optimizing your environment

- Terminal, shell, and prompt
- Aliases, functions, and scripts
- Shortcuts

Custom terminal, shell, and prompt



Aliases

```
alias l '/bin/ls -ltrFsA'
alias mi 'mv -i'
alias up "cd .."
alias fox "open -a 'Firefox' \!:*"

# spelling while typing is hard
alias alais alias
alias moer more
alias mroe more
alias pu up

#alias onion 'open http://www.theonion.com/content/index'
alias onion echo "back to work"
```

Shortcuts

```
$ cd ~/some/very/deep/often-used/directory
$ mark deep

$ jump deep

$ unmark deep

$ marks
deep    -> /home/jeroen/some/very/deep/often-used/directory
foo     -> /usr/bin/foo/bar
```

Shortcuts

```
export MARKPATH=$HOME/.marks
function mark {
    mkdir -p "$MARKPATH"; ln -s "$(pwd)" "$MARKPATH/$1"
}
function jump {
    cd -P "$MARKPATH/$1" 2>/dev/null ||
    echo "No such mark: $1"
}
function unmark {
    rm -i "$MARKPATH/$1"
}
function marks {
    ls -l "$MARKPATH" | sed 's/  / /g' |
    cut -d' ' -f9- | sed 's/ -/\t-/g' && echo
}
```

From one-liners to reusable tools

- Shebang: #!/usr/bin/env bash
- Permission: chmod +x
- Arguments: $1, $2, $@
- Exit codes: 0, 1, 2
- Extension is not important
- Add to PATH

Example: CLI for explainshell.com

Example: CLI for explainshell.com

```
#!/usr/bin/env bash
# explain: Command-line wrapper for explainshell.com
#
# Example usage: explain tar xzvf
# Dependency: scrape
# Author: http://jeroenjanssens.com

COMMAND="$@"
URL="http://explainshell.com/explain?cmd=${COMMAND}"
curl -s "${URL}" |
scrape -e 'span.dropdown > a, pre' |
sed -re 's/<(\/?)[^>]*>//g'
```

Example: CLI for explainshell.com

```
$ explain tar xzvf
The GNU version of the tar archiving utility

-x, --extract, --get
      extract files from an archive

-z, --gzip, --gunzip --ungzip

-v, --verbose
      verbosely list files processed

-f, --file ARCHIVE
      use archive file or device ARCHIVE
```

Command-line tools from existing code

- Accept standard input
- Write to standard output
- Write to standard error
- Parse command-line arguments
- Provide help
- Take Unix philosophy into account

Parsing command-line arguments with docopt

```
#!/usr/bin/env python
"""Usage: pycho [-hnv] [STRING ...]

-h --help     Show this screen.
-n            Do not output trailing newline.
-v --version  Show version.
"""
from docopt import docopt
from sys import stdout
if __name__ == "__main__":
    args = docopt(__doc__, version="Pycho 1.0")
    stdout.write(" ".join(args["STRING"]))
    if not args["-n"]:
        stdout.write("\n")
```

Parsing command-line arguments with docopt

```
$ pycho -h
Usage: pycho [-hnv] [STRING ...]

-h --help     Show this screen.
-n            Do not output trailing newline.
-v --version  Show version.

$ pycho --version
Pycho 1.0

$ pycho -n COMMAND LINE REPRESENT
COMMAND LINE REPRESENT%

$
```

DataScienceToolbox.org

- Collection of command-line tools
- Vagrant environment
- Linux, Mac OS X, and Windows

Parallelization

Looping in serial

```
$ echo "4^2" | bc
16

$ for i in {0..100..2}
> do
> echo "$i^2" | bc
> done | tail -n 5
8464
8836
9216
9604
10000
```

Looping in serial

```
$ curl -s "http://api.randomuser.me/?results=5" |
> jq -r '.results[].user.email' > data/emails.txt

$ while read line
> do
> echo "Sending invitation to ${line}."
> done < data/emails.txt
Sending invitation to kaylee.anderson64@example.com.
Sending invitation to arthur.baker92@example.com.
Sending invitation to chloe.graham66@example.com.
Sending invitation to wyatt.nelson80@example.com.
Sending invitation to peter.coleman75@example.com.
```

Looping in serial

```
$ cat slow.sh
#!/bin/bash
echo "Starting job $1"
duration=$((1+RANDOM%5))
sleep $duration
echo "Job $1 took ${duration} seconds"
```

Looping in serial

```
$ for i in {1..4}; do
> slow.sh $i &
> done
[1] 1776
[2] 1777
[3] 1778
[4] 1780
Starting job 4
Starting job 1
Starting job 3
Starting job 2
Job 3 took 2 seconds
Job 2 took 2 seconds
Job 4 took 4 seconds
Job 1 took 5 seconds
```

Introducing GNU Parallel

- Parallelize existing tools
- For loop in its simplest form
- More than 100 options
- Distributed computation
- Drop-in replacement for xargs

Giving input

```
$ seq 5 | parallel "echo {}"
1
2
3
4
5

$ seq 5 | parallel -N0 "echo Hi"
Hi
Hi
Hi
Hi
Hi
```

Giving input

```
$ < input.csv | parallel -C, "mv {1} {2}"

$ < input.csv | parallel -C, --header : "invite.sh
> {name} {email}"
```

Controlling number of concurrent jobs

```
$ seq 5 | parallel --jobs 1 echo
$ seq 5 | parallel -j0 echo
$ seq 5 | parallel -j100% echo
$ seq 5 | parallel -j200% echo
$ seq 5 | parallel -j-1 echo
$ seq 5 | parallel -j+1 echo

$ parallel --number-of-cpus
1
$ parallel --number-of-cores
4

$ seq 5 | parallel --noswap echo
$ seq 5 | parallel --nice 17 echo
```

Logging and output

```
$ seq 5 | parallel --results data/outdir "echo Hi {}"
$ find data/outdir
data/outdir
data/outdir/1
data/outdir/1/1
data/outdir/1/1/stderr
data/outdir/1/1/stdout
data/outdir/1/3
data/outdir/1/3/stderr
data/outdir/1/3/stdout
data/outdir/1/5
data/outdir/1/5/stderr
data/outdir/1/5/stdout
data/outdir/1/2
```

Resuming and remote execution

```
$ seq 5 | parallel --joblog /tmp/log echo
$ seq 5 | parallel --resume --joblog /tmp/log echo

$ cat /tmp/log
Seq Host Starttime Runtime Send Receive Exitval Signal Command
1  :  1391027365.223  0.006  0  0  0  0  echo 1
2  :  1391027365.225  0.007  0  0  0  0  echo 2
3  :  1391027365.227  0.006  0  0  0  0  echo 3
4  :  1391027365.229  0.003  0  0  0  0  echo 4
5  :  1391027365.232  0.002  0  0  0  0  echo 5



$ seq 5 | parallel -S $SERVERS echo
```

Workflow Management

Working with the command line can be chaotic

- Invoke many different commands
- Create custom command-line tools
- Obtain and generate many (intermediate) files

Top e-books from Project Gutenberg

```
curl -s 'http://www.gutenberg.org/browse/scores/top' |
grep -E '^<li>' |
head -n 5 |
sed -E "s/.*ebooks\/([0-9]+)\">([^<]+)<.*/\\1,\\2/" > top-5
```

Top e-books from Project Gutenberg

```
$ cat top-5
119,A Tramp Abroad by Mark Twain (7065)
76,Adventures of Huckleberry Finn by Mark Twain (1814)
1342,Pride and Prejudice by Jane Austen (1299)
1661,The Adventures of Sherlock Holmes by Arthur Conan Doyle (
11,Alice's Adventures in Wonderland by Lewis Carroll (951)
```

Introducing Drake

- Formalize steps in terms of dependencies
- Run specific steps from the command line
- Use inline code
- Store and retrieve data from external sources

Every workflow starts with the first step

```
top-5 <- [-timecheck]
    curl -s 'http://www.gutenberg.org/browse/scores/top' |
    grep -E '^<li>' |
    head -n 5 |
    sed -E "s/.*ebooks\/([0-9]+)\">([^<]+)<.*/\\1,\\2/" >
    top-5
```

Every workflow starts with the first step

```
$ drake
The following steps will be run, in order:
  1: top-5 <-  [missing output]
Confirm? [y/n] y
Running 1 steps with concurrence of 1...

--- 0. Running (missing output): top-5 <-
--- 0: top-5 <-  -> done in 0.35s
Done (1 steps run).
```

Dependencies, variables, and data separation

```
NUM=5
BASE=../../data/

top.html <- [-timecheck]
    curl -s 'http://www.gutenberg.org/browse/scores/top' >
    $OUTPUT

top-$[NUM] <- top.html
    < $INPUT grep -E '^<li>' |
    head -n $[NUM] |
    sed -E "s/.*ebooks\/([0-9]+)\">([^<]+)<.*/\\1,\\2/" >
    $OUTPUT
```

Dependencies, variables, and data separation

```
$ drake -w 02.drake
The following steps will be run, in order:
  1: ../../data/top.html <-  [missing output]
  2: ../../data/top-5 <- ../../data/top.html [projected timesta
Confirm? [y/n] y
Running 2 steps with concurrence of 1...

--- 0. Running (missing output): ../../data/top.html <-
--- 0: ../../data/top.html <-  -> done in 0.89s

--- 1. Running (missing output): ../../data/top-5 <- ../../data
--- 1: ../../data/top-5 <- ../../data/top.html -> done in 0.02s
Done (2 steps run).
```

Dependencies, variables, and data separation

```
$ NUM=10 drake -w 02.drake
The following steps will be run, in order:
  1: ../../data/top-10 <- ../../data/top.html [missing output]
Confirm? [y/n] y
Running 1 steps with concurrence of 1...

--- 1. Running (missing output): ../../data/top-10 <- ../../dat
--- 1: ../../data/top-10 <- ../../data/top.html -> done in 0.02
Done (1 steps run).
```

Tags and running certain steps

```
NUM:=5
BASE=../../data/

top.html, %html <- [-timecheck]
    curl -s 'http://www.gutenberg.org/browse/scores/top' >
    $OUTPUT

top-$[NUM], %filter <- %html
    < $INPUT grep -E '^<li>' |
    head -n $[NUM] |
    sed -E "s/.*ebooks\/([0-9]+)\">([^<]+)<.*/\\1,\\2/" >
    $OUTPUT
```

```
$ NUM=10 drake -w 03.drake +^%html
```

Inline code

```
somefile.out <- somefile.csv [python]
    from drakeutil import *
    with open(INPUT) as istream:
        with open(OUTPUT) as ostream:
            for l in istream:
                for word in l.split():
                    print >> ostream, word
```

Conclusion

- Command line is great for doing data science
- Does not solve all your problems
- OK to continue with R / IPython / ...

Where to go from here?

- Install Data Science Toolbox
- Do a tutorial
- Practice your one-liners
- Give (feed)back

References

- http://datasciencetoolbox.org
- http://cli.learncodethehardway.org/book/
- https://github.com/tonyfischetti/qstats
- https://github.com/jehiah/json2csv
- https://github.com/bitly/data_hacks
- https://github.com/Factual/drake
- https://github.com/chrishwiggins/mise
- http://csvkit.readthedocs.org/en/latest/
- http://stedolan.github.io/jq/
- http://www.gnu.org/software/parallel

Thank you!

jeroen@jeroenjanssens.com
http://jeroenjanssens.com
@jeroenhjanssens