DATA SCIENCE AT THE COMMAND LINE - CHEATSHEET

ENVIRONMENTS, HELP, MACROS

alias

```
$ help alias
$ alias ll=’ls -alF’
```

bash - Bourne shell

```
$ sudo apt-get install bash
$ man bash
```

bc - evaluate equation from stdin

```
$ sudo apt-get install bc
$ man bc
$ echo 'e(1)' | bc -l
2.71828182845904523536
```

cols - apply cmnd to subset of cols, merge result

```
$ git clone <github>
$ < iris.csv cols -C species body tapkee --method pca |
header -r x,y,species
```

cowsay - debugging helper

```
$ sudo apt-get install cowsay
$ man cowsay
$ echo 'The command line is awesome!' | cowsay
```

export - set export attribute for shell variables

```
$ help export
$ export WEKAPATH=$HOME/bin
```

for - exec command for each member of list

```
$ help for
$ for i in {A..C} "It's easy as" {1..3}; do echo $i; done
A
B
C
It's easy as
1
2
3
```

man - read reference manuals of cmndline tools

```
$ sudo apt-get install man
$ man man
$ man grep
```

pbc - run bc with parallel

```
$ git clone <github>
$ seq 5 | pbc '{1}^2'
1
4
9
16
25
```

type - display cmndline tool class

```
$ help type
$ type cd
```

sudo - exec cmnd as another user

```
$ sudo apt-get install sudo
$ man sudo
```

FILES & DIRECTORIES

body - apply expression to all but 1st line

```
$ git clone <github.git>
$ echo -e "value\n7\n2\n5\n3" | body sort -n
value
2
3
5
7
```

cd - change working directory

```
$ help cd
$ cd ~; pwd; cd ..; pwd
```

cat - concat files & stdin, print to stdout

```
$ sudo apt-get install coreutils
$ man cat
$ cat results-01 results-02 results-03 > results-all
```

chmod - change file mode bits

```
$ sudo apt-get install coreutils
$ man chmod
$ chmod u+x experiment.sh
```

cp - copy files & directories

```
$ sudo apt-get install coreutils
$ man cp
```

cut - remove sections from each line of files

```
$ sudo apt-get install coreutils
$ man cut
```

echo - display line of text

```
$ sudo apt-get install coreutils
$ man echo
```

env - run program in modified environment

```
$ sudo apt-get install coreutils
```

DATA SCIENCE AT THE COMMAND LINE - CHEATSHEET

```
$ man env
$ #!/usr/bin/env python
```

fieldsplit - split file in multiples based on field value

```
$ # See website for installation instructions
$ fieldsplit --help
```

find - file search in directory

```
$ sudo apt-get install findutils
$ man find
```

head - output first n lines of files

```
$ sudo apt-get install coreutils
$ man head
$ seq 5 | head -n 3
1
2
3
```

header - add / replace / delete header lines

```
$ git clone <github>
$ header -h
```

less - paginate large files

```
$ sudo apt-get install less
$ man less
$ csvlook iris.csv | less
```

ls - list directory contents

```
$ sudo apt-get install coreutils
$ man ls
```

mv - move / rename files & directories

```
$ sudo apt-get install coreutils
$ man mv
```

paste - merge lines of files

```
$ sudo apt-get install coreutils
$ man paste
```

pwd - print working directory name

```
$ man pwd
$ pwd
/home/vagrant
```

rm - remove files & directories

```
$ sudo apt-get install coreutils
$ man rm
```

sort - sort lines of text files

```
$ sudo apt-get install coreutils
$ man sort
```

split - split file into pieces

```
$ sudo apt-get install coreutils",
$ man split
```

tail - output last part of files

```
$ sudo apt-get install coreutils
$ man tail
$ seq 5 | tail -n 3
3
4
5
```

tee - read from stdin, write to stdout and files

```
$ sudo apt-get install coreutils
$ man tee
```

tr - translate or delete characters

```
$ sudo apt-get install coreutils
$ man tr
```

wc - newline, word & byte counts for each file

```
$ sudo apt-get install coreutils
$ man wc
$ echo 'hello world' | wc -c
12
```

PATTERN MATCHING

awk -- pattern scanning & text processing

```
$ sudo apt-get install mawk
$ man awk
$ seq 5 | awk '{sum+=$1} END {print sum}'
15
```

sed - filter & transform text

```
$ sudo apt-get install sed
$ man sed
```

grep - print lines matching pattern

```
$ sudo apt-get install grep
$ man grep
```

DEPLOYMENT

aws -- manage AWS services

```
$ sudo pip install awscli
$ aws help
$ aws ec2 describe-regions | head -n 5
{ "Regions": [ {
"Endpoint": "ec2.eu-west-1.amazonaws.com",
"RegionName": "eu-west-1"
```

DATA SCIENCE AT THE COMMAND LINE - CHEATSHEET

git - manage Git repositories

```
$ sudo apt-get install git
$ man git
```

CSV FILES

csvcut - extract columns from CSV

```
$ sudo pip install csvkit
$ csvcut --help
```

csvgrep - filter CSV where cols=arg or regexp

```
$ sudo pip install csvkit
$ csvgrep --help
```

csvjoin - merge 2+ CSV tables aka SQL JOIN

```
$ sudo pip install csvkit
$ csvjoin --help
```

csvlook - render CSV to readable stdout

```
$ sudo pip install csvkit
$ csvlook --help
$ echo -e "a,b\n1,2\n3,4" | csvlook
```

csvsort - sort CSV

```
$ sudo pip install csvkit
$ csvsort --help
```

csvsql - execute SQL queries on CSV

```
$ sudo pip install csvkit
$ csvsql --help
```

csvstack - stack rows from multiple CSVs

```
$ sudo pip install csvkit
$ csvstack --help
```

csvstat - descriptive stats for all cols in CSV

```
$ sudo pip install csvkit
$ csvstat --help
```

in2csv - convert data formats to CSV

```
$ sudo pip install csvkit
$ in2csv --help
```

json2csv - JSON to CSV

```
$ go get github.com/jehiah/json2csv
$ json2csv --help
```

sql2csv - exec cmnds vs SQL DB, return CSV data

```
$ sudo pip install csvkit
$ sql2csv --help
```

JSON FILES

jq - JSON processor

```
$ man jq
```

xml2json - XML to JSON

```
$ npm install xml2json-command
$ xml2json < input.xml > output.json
```

LOGIN, DOWNLOAD, SCRAPE

curl - download data from URL

```
$ sudo apt-get install curl
$ man curl
```

culique - perform OAuth for curl

```
$ git clone https://github.com/decklin/curlicue.git
```

scp - copy remote files securely

```
$ sudo apt-get install openssh-client
$ man scp
```

scrape - scrape HTML with XPath or CSS3 selector

```
$ git clone <github>
$ curl -sL '<url>' | scrape -e 'head > title'
<title>Data Science Toolbox</title>
```

ssh - login to remote machines

```
$ sudo apt-get install ssh
$ man ssh
```

DISPLAYS

display - display image data, any X server

```
$ sudo apt-get install imagemagick
$ man display
```

feedgnuplot - generate gnuplot script

```
$ sudo apt-get install feedgnuplot
$ man feedgnuplot
```

WORKFLOWS

drake - manage workflow

```
$ # Please see Chapter 6 for installation instructions.
$ drake --help
```

parallel - run shell cmnd lines from stdin in parallel

DATA SCIENCE AT THE COMMAND LINE - CHEATSHEET

```
$ # See website for installation instructions
$ man parallel
$ seq 3 | parallel echo Processing file {}.csv
Processing file 1.csv
Processing file 2.csv
Processing file 3.csv
```

INTEGER / DATE SEQUENCES

dseq - generate date sequence rel to today

```
$ git clone <github>
$ dseq -2 0 # day before yesterday till today
2014-07-15
2014-07-16
2014-07-17
```

seq - print sequence of numbers

```
$ sudo apt-get install coreutils
$ man seq
$ seq 3
1
2
3
```

sample - print from stdout (prob, duration, delay)

```
$ git clone <github>
$ sample --help
```

shuf - generate random permutations

```
$ sudo apt-get install coreutils
$ man shuf
```

PYTHON, R

pip - manage Python packages

```
$ sudo apt-get install python-pip
$ man pip
```

python - exec Python language

```
$ sudo apt-get install python
$ man python
```

R - exec R language

```
$ sudo apt-get install r-base-dev
$ man R
```

Rio - load CSV from stdin, run R script, get output

```
$ git clone <github>
$ Rio -h
$ seq 10 | Rio -nf sum
55
```

Rio-scatter - scatter plot from CSV using Rio

```
$ git clone <github>
$ < iris.csv Rio-scatter sepal_length sepal_width species > iris.png
```

EXTERNAL TOOL APIS

bigmler - prediction API

```
$ sudo pip install bigmler
$ bigmler --help
```

run_experiment - run ML trial with Scikit-Learn

```
$ sudo pip install skll
$ run_experiment --help
```

tapkee - dimensionality reduction API

```
$ # See website for installation instructions
$ tapkee --help
$ < iris.csv cols -C species body tapkee --method pca | header -r x,y,species
```

weka - Weka API command line tool

```
$ git clone <github>
```

FILE EXTRACTION / COMPRESSION

tar - create, list, extract TAR archives

```
$ sudo apt-get install tar
$ man tar
```

tree - list directory contents, tree format

```
$ sudo apt-get install tree
$ man tree
```

uniq - report or omit repeated lines

```
$ sudo apt-get install coreutils
$ man uniq
```

unpack - extract common file formats

```
$ git clone <github>
$ unpack file.tgz
```

unrar - extract files from RAR archives

```
$ sudo apt-get install unrar-free
$ man unrar
```

unzip - list, test, extract compressed ZIP files

```
$ sudo apt-get install unzip
$ man unzip
```