APPENDIX A

List of Command-Line Tools

This is an overview of all the command-line tools discussed in this book. This includes binary executables, interpreted scripts, and Bash builtins and keywords. For each command-line tool, the following information, when available and appropriate, is provided:

- The actual command to type at the commandline
- A description
- The name of the package it belongs to
- The version used in the book
- The year that version was released
- The primary author(s)
- A website to find more information
- How to install it
- How to obtain help
- An example usage

All command-line tools listed here are included in the Data Science Toolbox for *Data Science at the Command Line*. See Chapter 2 for instructions on how to set it up. The install commands assume that you're running Ubuntu 14.04. Please note that citing open source software is not trivial, and that some information may be missing or incorrect.

alias

Define or display aliases. Alias is a Bash builtin.

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

awk

Pattern scanning and text processing language. Mawk (version 1.3.3) by Mike Brennan (1994). *http://invisible-island.net/mawk*.

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

aws

Manage AWS Services such as EC2 and S3 from the command line. AWS Command Line Interface (version 1.3.24) by Amazon Web Services (2014). *http://aws.amazon.com/cli*.

```
$ 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"
```

bash

GNU Bourne-Again SHell. Bash (version 4.3) by Brian Fox and Chet Ramey (2010). *http://www.gnu.org/software/bash*.

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

bc

Evaluate equation from standard input. Bc (version 1.06.95) by Philip A. Nelson (2006). *http://www.gnu.org/software/bc*.

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

bigmler

Access BigML's prediction API. BigMLer (version 1.12.2) by BigML (2014). *http://bigmler.readthedocs.org*.

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

body

Apply an expression to all but the first line. Useful if you want to apply classic command-line tools to CSV files with a header. Body by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ echo -e "value\n7\n2\n5\n3" | body sort -n
value
2
3
5
7
```

cat

Concatenate files and standard input, and print on standard output. Cat (version 8.21) by Torbjorn Granlund and Richard M. Stallman (2012). *http://www.gnu.org/software/coreutils*.

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

cd

Change the shell working directory. Cd is a Bash builtin.

```
$ help cd
$ cd ~; pwd; cd ..; pwd
/home/vagrant
/home
```

chmod

Change file mode bits. We use it to make our command-line tools executable. Chmod (version 8.21) by David MacKenzie and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

cols

Apply a command to a subset of the columns and merge the result back with the remaining columns. Cols by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ < iris.csv cols -C species body tapkee --method pca | header -r x,y,species
```

cowsay

Generate an ASCII picture of a cow with a message. Useful for when building up a particular pipeline is starting to frustrate you a bit too much. Cowsay (version 3.03+dfsg1) by Tony Monroe (1999).

```
$ sudo apt-get install cowsay
$ man cowsay
$ echo 'The command line is awesome!' | cowsay
 _____________________________
< The command line is awesome! >
 -----------------------------
        \   ^__^
         \  (oo)\_______
            (__)\       )\/\
                ||----w |
                ||     ||
```

cp

Copy files and directories. Cp (version 8.21) by Torbjorn Granlund, David MacKenzie, and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

csvcut

Extract columns from CSV data. Like `cut` command-line tool, but for tabular data. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

csvgrep

Filter tabular data to only those rows where certain columns contain a given value or match a regular expression. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

csvjoin

Merge two or more CSV tables together using a method analogous to a SQL JOIN operation. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

csvlook

Renders a CSV file to the command line in a readable, fixed-width format. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

```
$ sudo pip install csvkit
$ csvlook --help
$ echo -e "a,b\n1,2\n3,4" | csvlook
|----+----|
|  a | b  |
|----+----|
|  1 | 2  |
|  3 | 4  |
|----+----|
```

csvsort

Sort CSV files. Like the `sort` command-line tool, but for tabular data. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

csvsql

Execute SQL queries directly on CSV data or insert CSV into a database. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

csvstack

Stack up the rows from multiple CSV files, optionally adding a grouping value to each row. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

csvstat

Print descriptive statistics for all columns in a CSV file. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

curl

Download data from a URL. cURL (version 7.35.0) by Daniel Stenberg (2012). *http://curl.haxx.se*.

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

curlicue

Perform OAuth dance for `curl`. Curlicue by Decklin Foster (2014). *https://github.com/decklin/curlicue*.

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

cut

Remove sections from each line of files. Cut (version 8.21) by David M. Ihnat, David MacKenzie, and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

display

Display an image or image sequence on any X server. Can read image data from standard input. Display (version 8:6.7.7.10) by ImageMagick Studio LLC (2009). *http://www.imagemagick.org*.

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

drake

Manage a data workflow. Drake (version 0.1.6) by Factual (2014). *https://github.com/Factual/drake*.

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

dseq

Generate sequence of dates relative to today. Dseq by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ dseq -2 0 # day before yesterday till today
2014-07-15
2014-07-16
2014-07-17
```

echo

Display a line of text. Echo (version 8.21) by Brian Fox and Chet Ramey (2012). *http://www.gnu.org/software/coreutils*.

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

env

Run a program in a modified environment. It's often used to specify which interpreter should run our script. Env (version 8.21) by Richard Mlynarik and David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

```
$ sudo apt-get install coreutils
$ man env
$ #!/usr/bin/env python
```

export

Set export attribute for shell variables. Useful for making shell variables available to other command-line tools. Export is a Bash builtin.

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

feedgnuplot

Generate a script for `gnuplot` while passing data to standard input. Feedgnuplot (version 1.32) by Dima Kogan (2014). *http://search.cpan.org/perldoc?feedgnuplot*.

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

fieldsplit

Splits a file into multiple files according to a particular field value. Fieldsplit (version 2010-01) by Jeremy Hinds, Jason Gessner, Jim Renwick, Norman Gocke, Rodofo Granata, and Tobias Wolff (2010). *http://code.google.com/p/crush-tools*.

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

find

Search for files in a directory hierarchy. Find (version 4.4.2) by James Youngman (2008). *http://www.gnu.org/software/findutils*.

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

for

Execute commands for each member in a list. In Chapter 8, we discuss the advantages of using `parallel` instead of `for`. For is a Bash keyword.

```
$ 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
```

git

Manage repositories for Git, which is a distributed version control system. Git (version 1:1.9.1) by Linus Torvalds and Junio C. Hamano (2014). *http://git-scm.com*.

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

grep

Print lines matching a pattern. Grep (version 2.16) by Jim Meyering (2012). *http://www.gnu.org/software/grep*.

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

head

Output the first part of files. Head (version 8.21) by David MacKenzie and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

header

Add, replace, and delete header lines. Header by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ header -h
```

in2csv

Convert common, but less awesome, tabular data formats to CSV. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

jq

Process JSON. Jq (version jq-1.4) by Stephen Dolan (2014). *http://stedolan.github.com/jq*.

```
$ # See website for installation instructions
$ # See website for documentation
```

json2csv

Convert JSON to CSV. Json2Csv (version 1.1) by Jehiah Czebotar (2014). *https://github.com/jehiah/json2csv*.

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

less

Paginate large files. Less (version 458) by Mark Nudelman (2013). *http://www.greenwoodsoftware.com/less*.

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

ls

List directory contents. Ls (version 8.21) by Richard M. Stallman and David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

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

man

Read reference manuals of command-line tools. Man (version 2.6.7.1) by John W. Eaton and Colin Watson (2014).

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

mkdir

Make directories. Mkdir (version 8.21) by David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

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

mv

Move or rename files and directories. Mv (version 8.21) by Mike Parker, David MacKenzie, and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

parallel

Build and execute shell command lines from standard input in parallel. GNU Parallel (version 20140622) by Ole Tange (2014). *http://www.gnu.org/software/parallel*.

```
$ # 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
```

paste

Merge lines of files. Paste (version 8.21) by David M. Ihnat and David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

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

pbc

Run `bc` with `parallel`. First column of input CSV is mapped to `{1}`, second to `{2}`, and so forth. Pbc by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ seq 5 | pbc '{1}^2'
1
4
9
16
25
```

pip

Install and manage Python packages. Pip (version 1.5.4) by PyPA (2014). *https://pip.pypa.io*.

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

pwd

Print name of current working directory. Pwd (version 8.21) is a Bash builtin by Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

python

Execute Python, which is an interpreted, interactive, and object-oriented programming language. Python (version 2.7.5) by Python Software Foundation (2014). *http://www.python.org*.

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

R

Analyze data and create visualizations with the `R` programming language. To install the latest version of `R` on Ubuntu, follow the instructions on *http://bit.ly/ubuntu_packages_for_R*. R (version 3.1.1) by R Foundation for Statistical Computing (2014). *http://www.r-project.org*.

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

Rio

Load CSV from standard input into R as a data.frame, execute given commands, and get the output as CSV or PNG. Rio by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ Rio -h
$ seq 10 | Rio -nf sum
55
```

Rio-scatter

Create a scatter plot from CSV using `Rio`. Rio-scatter by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ < iris.csv Rio-scatter sepal_length sepal_width species > iris.png
```

rm

Remove files or directories. Rm (version 8.21) by Paul Rubin, David MacKenzie, Richard M. Stallman, and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

run_experiment

Run machine learning experiments with the Python package scikit-learn. SciKit-Learn Laboratory (version 0.26.0) by Educational Testing Service (2014). *https://skll.readthedocs.org*.

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

sample

Print lines from standard output with a given probability, for a given duration, and with a given delay between lines. Sample by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ sample --help
```

scp

Copy remote files securely. Scp (version 1:6.6p1) by Timo Rinne and Tatu Ylonen (2014). *http://www.openssh.com*.

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

scrape

Extract HTML elements using an XPath query or CSS3 selector. Scrape by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ curl -sL 'http://datasciencetoolbox.org' | scrape -e 'head > title'
<title>Data Science Toolbox</title>
```

sed

Filter and transform text. Sed (version 4.2.2) by Jay Fenlason, Tom Lord, Ken Pizzini, and Paolo Bonzini (2012). *http://www.gnu.org/software/sed*.

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

seq

Print a sequence of numbers. Seq (version 8.21) by Ulrich Drepper (2012). *http://www.gnu.org/software/coreutils*.

```
$ sudo apt-get install coreutils
$ man seq
$ seq 5
1
2
3
4
5
```

shuf

Generate random permutations. Shuf (version 8.21) by Paul Eggert (2012). *http://www.gnu.org/software/coreutils*.

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

sort

Sort lines of text files. Sort (version 8.21) by Mike Haertel and Paul Eggert (2012). *http://www.gnu.org/software/coreutils*.

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

split

Split a file into pieces. Split (version 8.21) by Torbjorn Granlund and Richard M. Stallman (2012). *http://www.gnu.org/software/coreutils*.

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

sql2csv

Executes arbitrary commands against an SQL database and outputs the results as a CSV. Csvkit (version 0.8.0) by Christopher Groskopf (2014). *http://csvkit.readthedocs.org*.

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

ssh

Login to remote machines. OpenSSH client (version 1.8.9) by Tatu Ylonen, Aaron Campbell, Bob Beck, Markus Friedl, Niels Provos, Theo de Raadt, Dug Song, and Markus Friedl (2014). *http://www.openssh.com*.

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

sudo

Execute a command as another user. Sudo (version 1.8.9p5) by Todd C. Miller (2013). *http://www.sudo.ws/sudo*.

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

tail

Output the last part of files. Tail (version 8.21) by Paul Rubin, David MacKenzie, Ian Lance Taylor, and Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

tapkee

Reduce dimensionality of a data set using various algorithms. Tapkee by Sergey Lisitsyn and Fernando Iglesias (2014). *http://tapkee.lisitsyn.me*.

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

tar

Create, list, and extract TAR archives. Tar (version 1.27.1) by Jeff Bailey, Paul Eggert, and Sergey Poznyakoff (2014). *http://www.gnu.org/software/tar*.

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

tee

Read from standard input and write to standard output and files. Tee (version 8.21) by Mike Parker, Richard M. Stallman, and David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

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

tr

Translate or delete characters. Tr (version 8.21) by Jim Meyering (2012). *http://www.gnu.org/software/coreutils*.

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

tree

List contents of directories in a tree-like format. Tree (version 1.6.0) by Steve Baker (2014). *https://launchpad.net/ubuntu/+source/tree*.

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

type

Display the type of a command-line tool. Type is a Bash builtin.

```
$ help type
$ type cd
cd is a shell builtin
```

uniq

Report or omit repeated lines. Uniq (version 8.21) by Richard M. Stallman and David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

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

unpack

Extract common file formats. Unpack by Patrick Brisbin (2013). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
$ unpack file.tgz
```

unrar

Extract files from RAR archives. Unrar (version 1:0.0.1+cvs20071127) by Ben Asselstine, Christian Scheurer, and Johannes Winkelmann (2014). *http://home.gna.org/unrar*.

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

unzip

List, test and extract compressed files in a ZIP archive. Unzip (version 6.0) by Samuel H. Smith (2009).

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

wc

Print newline, word, and byte counts for each file. Wc (version 8.21) by Paul Rubin and David MacKenzie (2012). *http://www.gnu.org/software/coreutils*.

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

weka

Weka is a collection of machine learning algorithms for data mining tasks by Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. This command-line tool allows you to run Weka from the command line. Weka command-line tool by Jeroen H.M. Janssens (2014). *https://github.com/jeroenjanssens/data-science-at-the-command-line*.

```
$ git clone https://github.com/jeroenjanssens/data-science-at-the-command-
line.git
```

which

Locate a command-line tool. Does not work for Bash builtins. Which by unknown (2009).

```
$ man which
$ which man
/usr/bin/man
```

xml2json

Convert XML to JSON. Xml2Json (version 0.0.2) by Francois Parmentier (2014). *https://github.com/parmentf/xml2json*.

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