First Year Project, Spring 2021

Lecture 2: Command line tools and missing data

Instructor: Michael Szell
Feb 12, 2021

```
> >> awk cat cd
cp cut diff du
grep head ls man
mv nl rm sed sort
tr uniq wc |
```

Today we explore the data with command line tools

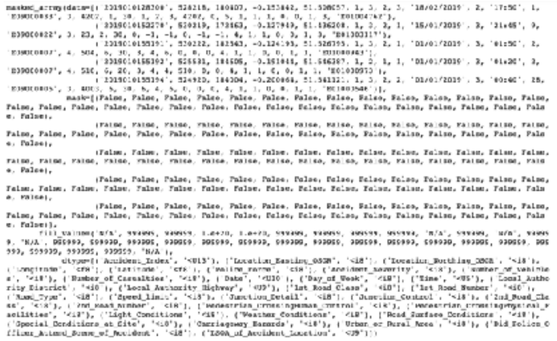
Command line tools

```
> >> awk cat cd
cp cut diff du
grep head ls man
mv nl rm sed sort
tr uniq wc |
```

Project 1
exploratory data analysis

```
1 !awk -F ',' '{print NF}' "../data/raw/Road Safety
  Data - Accidents 2019.csv" | sort | uniq -d
2 !awk -F ',' '{print NF}' "../data/raw/Road Safety
  Data - Casualties 2019.csv" | sort | uniq -d
3 !awk -F ',' '{print NF}' "../data/raw/Road Safety
  Data- Vehicles 2019.csv" | sort | uniq -d
```

How to check missing data



Command line tools are really fast

Agile

Augmenting

Scalable

Extensible

Ubiquitous

https://www.datascienceatthecommandline.com/1e/chapter-1-introduction.html#why-data-science-at-the-command-line

Command line tools are really fast

Agile

Augmenting

Scalable

Extensible

Ubiquitous

Choice between command line, Python, or even Excel depends on size of your data and what you want to do.

https://www.datascienceatthecommandline.com/1e/chapter-1-introduction.html#why-data-science-at-the-command-line

There are 5 types of tools. We focus on builtins and executables

Binary executable

Shell builtin

Interpreted script

Shell function

Alias

`type cmd`

https://www.datascienceatthecommandline.com/1e/chapter-2-getting-started.html#five-types-of-command-line-tools

http://mywiki.wooledge.org/BashGuide/CommandsAndArguments

File system operations you should become familiar with

`cd`	enter directory
`mv`	move
`rm`	remove
`cp`	copy
`ls`	list
`du`	disk usage
`cat`	concatenate, or just output content

Tools you should have heard about

`grep` globally search for a regular expression and print matching lines

`sed` stream editor

good for replacing strings

Getting help

```
man cmd

cmd --help

cmd -h
```

awk is a powerful data extraction tool

Example: print the second column in a CSV file

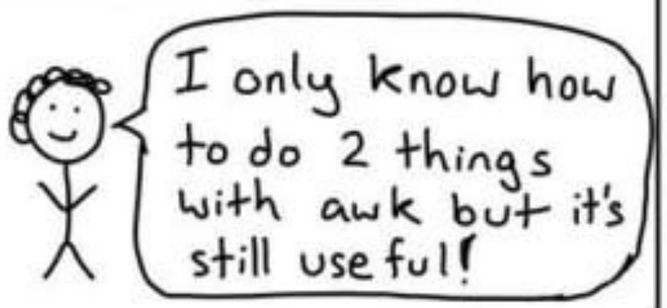
```
awk -F',' '{print $2}' mycsv.csv
```

https://en.wikipedia.org/wiki/AWK

awk

Julia Evans
@b0rk

awk is a tiny programming language for manipulating columns of data



basic awk program structure

```
BEGIN { ... }
CONDITION {action}
CONDITION {action}
END {...}
```

do action on lines matching CONDITION

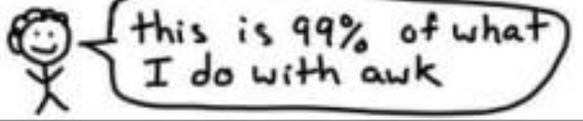
extract a column of text with awk

```
awk -F, '{print $5}'
```

column separator

single quotes!

print the 5th column



SO MANY unix commands print columns of text (ps! ls!)

so being able to get the column you want with awk is GREAT

A few more awk programs →

sum the numbers in the 3rd column

```
{s += $3}
END {print s}
```

action

at the end, print the sum!

print every line over 80 characters

```
length($0) > 80
```

condition

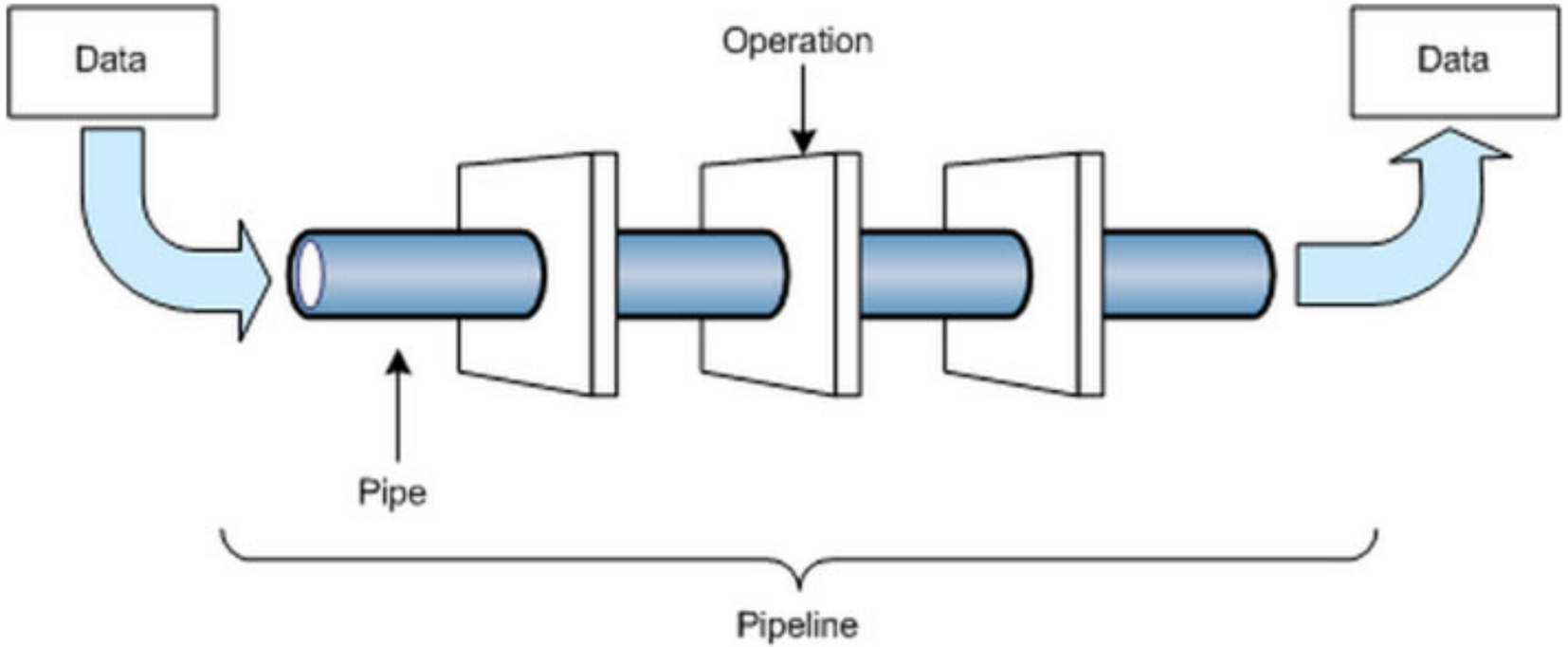
(there's an implicit {print} as the action)

Most important tools for data exploration

`head`	returns top lines
`wc, wc -l`	word count, line count
`tr`	string replacement
`sort`	sort
`uniq`	find unique (adjacent!) lines
`cut`	cut
`diff`	file difference
`nl`	line number

The pipe | combines commands into a pipeline

```
cmd1 | cmd2 | cmd3 | ...
```



https://en.wikipedia.org/wiki/Pipeline_(Unix)

The > redirects your output, >> appends

```
cmd infile.txt > outfile.txt

cmd infile.txt >> outfile.txt
```

Your Data Science command line alphabet

```
cd mv rm cp ls du cat grep sed man awk head wc
tr sort uniq cut diff nl | > >>
```

Your Data Science command line alphabet

```
cd mv rm cp ls du cat grep sed man awk head wc
tr sort uniq cut diff nl | > >>
```

Exercise: Make a pipeline that takes dsalphabet.txt and creates dsalphabet_ordered.txt:

```
> >> awk cat cd cp cut diff du grep head ls
man mv nl rm sed sort tr uniq wc |
```

Get familiar with terminal shortcuts to navigate fast

Must Know Linux Shortcuts

1. Tab
2. Ctrl + C
3. Ctrl + Z
4. Ctrl + D
5. Ctrl + L
6. Ctrl + A
7. Ctrl + E
8. Ctrl + U
9. Ctrl + K
10. Ctrl + W
11. Ctrl + Y
12. Ctrl + P
13. Ctrl + N

Bonus shortcut: Ctrl + R to search in command history

Also: Arrow keys, Option-Arrow keys

https://linuxhandbook.com/linux-shortcuts/

https://hackertyper.net/

In the next 6 months, have a look at all Unix commands

List of Unix commands

From Wikipedia, the free encyclopedia

"Unix command" redirects here. For other uses, see Command (computing).

This is a list of Unix commands as specified by IEEE Std 1003.1-2008, which is part of the Single UNIX Specification (S
can be found on Unix operating systems and most Unix-like operating systems.

Contents [show]

List [edit]

IEEE Std 1003.1-2008 utilities

Name	Category	Status (Option code)	Description
`alias`	Misc	Mandatory	Define or display aliases
`ar`	Misc	Mandatory	Create and maintain library archives
`at`	Process management	Mandatory	Execute commands at a later time
`awk`	Text processing	Mandatory	Pattern scanning and processing language
`basename`	Filesystem	Mandatory	Return non-directory portion of a pathname; see also dirname
`batch`	Process management	Mandatory	Schedule commands to be executed in a batch queue
`bc`	Misc	Mandatory	Arbitrary-precision arithmetic language
`cat`	Filesystem	Mandatory	Concatenate and print files
`cd`	Filesystem	Mandatory	Change the working directory
`chgrp`	Filesystem	Mandatory	Change the file group ownership
`chmod`	Filesystem	Mandatory	Change the file modes/attributes/permissions
`chown`	Filesystem	Mandatory	Change the file ownership
`cksum`	Filesystem	Mandatory	Write file checksums and sizes

https://en.wikipedia.org/wiki/List_of_Unix_commands

Advanced, nonstandard stuff: csvkit


```
head -1000 accidents.csv | csvlook

head -1000 accidents.csv | csvstat
```

https://csvkit.readthedocs.io/en/latest/tutorial.html

https://www.datascienceatthecommandline.com/1e/chapter-7-exploring-data.html

Why we generally do not use Excel in Data Science

Lack of:

- Reproducibility
- Version control
- Testing
- Maintainability
- Accuracy

https://datascience.stackexchange.com/questions/5443/do-data-scientists-use-excel

Why we generally do not use Excel in Data Science

Comment | Open Access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

127k Accesses | **45** Citations | **2567** Altmetric | Metrics

Abstract

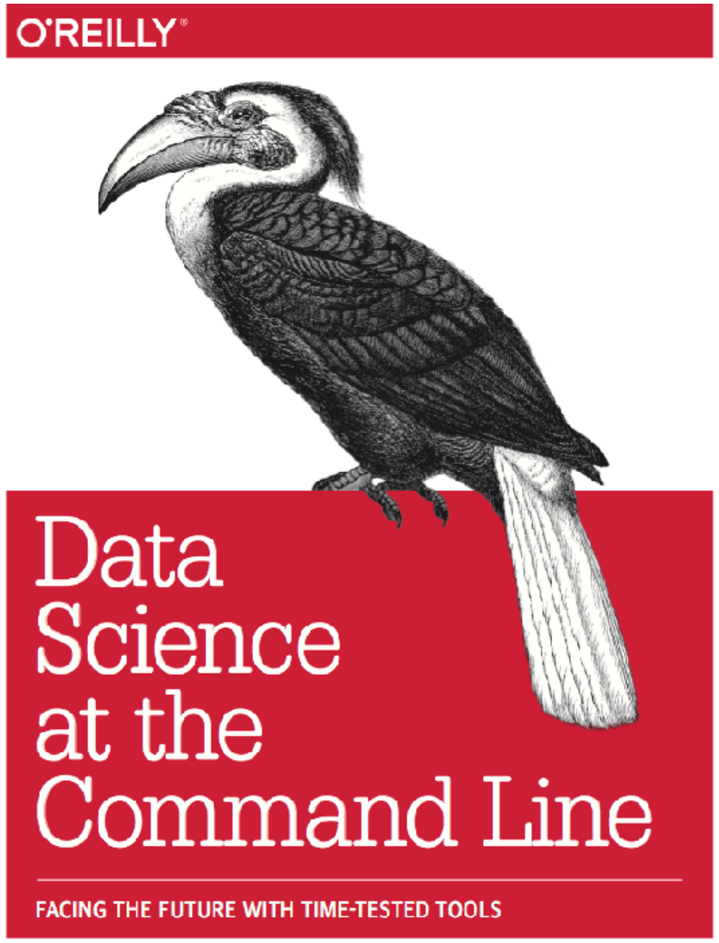
The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. '*SEPT2*' converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are frequently reused. Our aim here is to raise awareness of the problem.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7

Jupyter

Sources and further materials for today's class



https://www.datascienceatthecommandline.com/1e/

http://mywiki.wooledge.org/BashGuide/CommandsAndArguments

https://en.wikipedia.org/wiki/AWK

https://en.wikipedia.org/wiki/Pipeline_(Unix)

https://linuxhandbook.com/linux-shortcuts/

https://datascience.stackexchange.com/questions/5443/do-data-scientists-use-excel

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7