50 reasons to learn the shell for doing data science
Jeroen Janssens
@jeroenhjanssens
CEO at Data Science Workshops B.V.
Author of Data Science at the Command Line
The shell makes you look like a 1337 hacker.
When it comes to **hacking**, the shell is indispensable.

Source: Drew Conway
Data science is **OSEMN:**

- Obtaining data
- Scrubbing data
- Exploring data
- Modelling data
- Interpreting data

*Source: Mason & Wiggins (2010)*
$ pip install scikit-learn
Requirement already satisfied: scikit-learn in /usr/lib/python3.6/site-packages

$ cd ~/.ssh
$ ssh-keygen
$ cat ~/.ssh/id_rsa.pub | pbcopy

$ curl 'http://api.citybik.es/v2/networks/santander-cycles' |
  jq '.network.stations[].free_bikes' |
  paste -sd+ | bc
9525
The shell, with its `read-eval-print-loop`, enables you to **play** with your data.
The shell is very close to the filesystem, which makes it very convenient to work with files on a large scale.
Velociraptors.

It's a UNIX system.
I know this.
Plenty of great **resources** are available to learn the shell.
There's a fantastic book about using the shell for doing data science.

Read it for free at: data science at the command line .com
The shell has a vast and interesting history.
Like *wine*, the shell takes time to be appreciated. Good thing the shell also ages like wine.
There's always something new to learn about the shell and its many tools.

And learning is fun.
Docker containers are great for safely learning the shell.
The shell gives you access to man pages, which is like an offline Stack Overflow.
explainshell.com explains a given command line by matching each argument to the relevant help text in the man page.
The shell is free.
The shell doesn't care whether a tool has been implemented in Bash, C, Go, Java, JavaScript, Lisp, Perl, Python, R, Rust, or Scala.
You can customize the hell out of the shell.
The shell uses **text** as the universal interface, which enables tools from all over the world to **work together** and solve problems.
Most command-line tools do one thing and do it well.

The shell is there to let these tools work together in various ways.
The shell *never bothers you* about software updates.

Unless you want it to.
The shell gives you great **control** over your system.
When *shit* hits the fan with *git*, the shell is the only interface that can clean up the mess.
You can also **program** in the shell. A simple **for**-loop can do miracles.
Want to **parallelize** or **distribute** your task to multiple cores or machines? Use the shell with a pinch of **parallel**.
The shell: come for the tools, stay for the environment.
By default, the shell comes with many great tools such as `find`, `grep`, and `cut`. 
Package managers such as apt-get, brew, and pacman make it a pleasure to install additional command-line tools.
New tools are being developed every day for the shell.
The shell keeps a **history**.
You can easily extend the shell with your own tools, making you a more efficient and effective data scientist.
The shell lets you quickly find out things like: the size of a directory, the encoding of a CSV file, and the resolution of an image.
The shell lets you query databases, access APIs, open remote sheets, and even scrape websites.
With tools like `csvkit`, `jq`, and `xmlstarlet`, you can easily wrangle CSV, JSON, and XML in the shell.
csvsql allows you to perform SQL queries directly on CSV files in the shell.
telnet towel.blinkenlights.nl lets you watch Star Wars IV.

Use the shell, Luke.
The shell isn’t just available on UNIX machines and supercomputers. It can also be found on macOS, Raspberry Pi, and even Windows 10.
Sometimes the shell **outperforms** fancy **big data** technologies.
You can easily invoke Python and R from the shell.
Want to continue working in your favourite programming language or statistical environment? The shell is totally cool with that.
You can easily invoke the shell from Jupyter Notebook and RStudio.
jeroen at strata in ~ $ learn-shell-for-data-science --reason 41

$ echo data science at the command line | cowsay
$ echo data science at the command line | cowsay

< data science at the command line >
These days, many frontend developers also use the shell.
Invoke **sudo** and the shell will make you a **sandwich**.

Source: XKCD

*Note: Do not try on frontend developers*
You can **automate** just about everything using the shell.
Good luck managing a gazillion instances on AWS, Azure, and Google Cloud using the mouse.
The shell often requires less typing than a programming language.
The shell allows you to rename 750 files with just **three lines** of code.

Or **one**, if you have the right tool.
Your *wrists* will thank you for using the shell.
The shell has been around for almost 50 years, and probably will be around for the rest of your career.
Because Tim says so.

Anyone who does not have the command line at their beck-and-call is really missing something.
Jeroen Janssens
@jeroenhjanssens
CEO at Data Science Workshops B.V.
Author of Data Science at the Command Line