CEG2722: Data Analysis II

Command Line Data Processing

- Lecture 1 -

Achraf Koulali

Geospatial Engineering

November 15, 2021

Objectives

- To prepare the skills and knowledge to manipulate data through scripts and programs for final-year projects.

Objectives

- To prepare the skills and knowledge to manipulate data through scripts and programs for final-year projects.
- To introduce batch processing concepts and tools.

Objectives

- To prepare the skills and knowledge to manipulate data through scripts and programs for final-year projects.
- To introduce batch processing concepts and tools.
- To develop subject-specific programming and scripting skills within current software and tools.

Activities & Assessment

- Four lectures: Monday/Tuesday.

Activities & Assessment

- Four lectures: Monday/Tuesday.
- One assesment: 50% module evaluation.

Activities & Assessment

- Four lectures: Monday/Tuesday.
- One assesment: 50% module evaluation.
- Submission deadline: 10^{th} January 2022 14:00pm.

Activities

15th November	Session 1	Otaining data
22nd November	Session 2	Scrubbing data
23nd November	Practical 1	...
29th November	Session 4	Exploring data
6th December	Session 5	Towards data modelling
7th December	Practical 2	...
13th December	Session 6	Coursework

“Data science” & “Command line”

Figure 1: Practical definition by Mason and Wiggins (2010).

Obtaining data

- Download data (e.g., a webpage or server).
- Extract data from another file (e.g., an HTML file or spreadsheet).
- Generate data yourself (e.g., GPS surveys).

Scrubbing data

- Filtering lines
- Extracting certain columns
- Replacing values
- Extracting words
- Handling missing values
- Converting data from one format to another (e.g. conerting csv to shapefile)

Exploring Data

- Look at your data
- Derive statistics from your data
- Create interesting visualizations (e.g. plot locations lat,lon)

Modelling Data

- Techniques to create a models, include: Clustering, classification, regression, and dimensionality reduction...

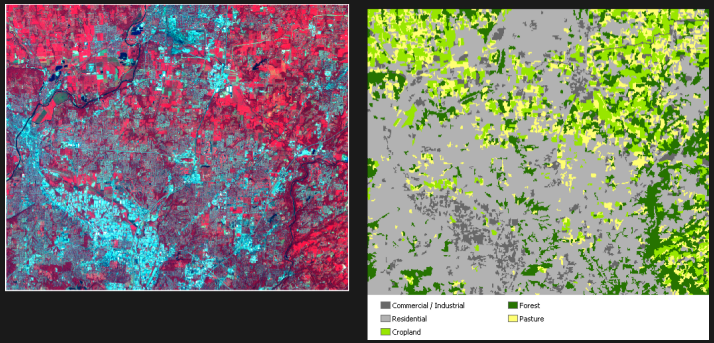

Figure 2: Example of Landsat TM image classification (source: ESRI)

Interpreting Data

- Drawing conclusions from your data

Interpreting Data

- Drawing conclusions from your data
- Evaluating what your results mean

Interpreting Data

- Drawing conclusions from your data
- Evaluating what your results mean
- Communicating your result

What is the Command Line?

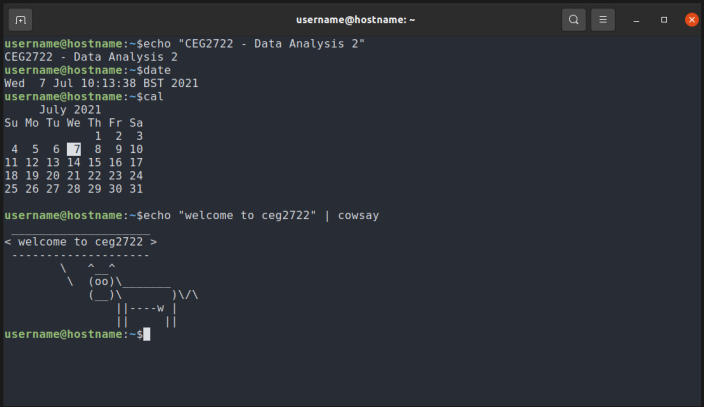

Figure 3: Command line on Ubuntu

What is the Command Line?

Linux/Unix command line:

```
$ whoami
username
$ hostname
Mymachine
$ date
Mon 17 May 15:13:25 BST 2021
```

What is Linux?

- Linux is an Operating System (OS) distributed under an open-source license.
- An OS is the software that directly manages a system's hardware and resources, like CPU, memory, and storage.

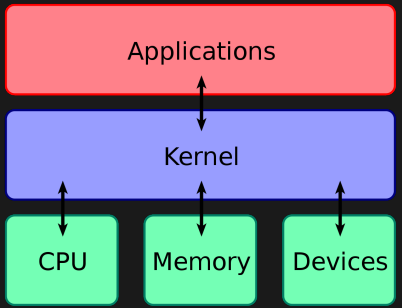

Figure 4: Linux Kernel

What does Linux include?

- Kernel: The kernel manages the system's resources and communicates with the hardware.
- System user space: The administrative layer for system level tasks (inclues command line/shell).
- Applications: A type of software that lets you perform a task (Desktop, Apps,...).

Linux distributions

- Desktops/laptops with Linux do have nice graphical user interfaces (KDE, Gnome, ...).
- HPC systems use the Linux command line.

Figure 5: Linux distributions

Why Data Science using Linux?

- Free Software / Open Source
- Safe & Secure, Linux is renowned for its security.
- Efficient: modular, extensible software development.

Why Data Science using Linux?

Command line vs Graphical User Interface (GUI)

- Typing.
- Very easy to re-run (mistakes, change in data input).
- Scriptable $\rightarrow$ the ability to automate tasks.

- Clicking mouse.
- It is not straightforward to automate pointing and clicking.
- GUIs are less suitable for doing scalable and repeatable data science.

The Command Line is Extensible

- Command line tools can work together: You can also create your own tools.

The Command Line is Extensible

- Command line tools can work together: You can also create your own tools.
- The open source community provides new tools on daily basis.

The Command Line is Extensible

- Command line tools can work together: You can also create your own tools.
- The open source community provides new tools on daily basis.
- Unix-like operating systems an be found in many places.

The Command Line is Extensible

- Command line tools can work together: You can also create your own tools.
- The open source community provides new tools on daily basis.
- Unix-like operating systems an be found in many places.
- 95% of the top 500 supercomputers are running GNU/Linux [Janssens, 2020].

Geospatial data analysis: real-world example

The percentage of GPS monuments with stainless steel pole in the Ordonance Network?

Step 1: download GPS logfiles from the OS archive.

```
$ cd ~/os_analysis/
$ url="https://www.ordnancesurvey.co.uk"
$ dir="/gps/rinex/station_log_files/"
$ wget -A log -r -l 1 -nd ${archive}${dir}
```

Geospatial data analysis: real-world example

Example of site logfile:

```
     EASI Site Information Form (site log)
     International GPS Service
     See Instructions at:
       ftp://igscb.jpl.nasa.gov/pub/station/general/sitelog_instr.txt

0.   Form

     Prepared by (full name)  : Colin Fane
     Date Prepared            : 2009-03-06
     Report Type              : Update
     If Update:
      Previous Site Log       : easi_20080429.log
      Modified/Added Sections : 3.2, 4.1, 11


1.   Site Identification of the GNSS Monument

     Site Name                : Easington
     Four Character ID        : EASI
     Monument Inscription     :
     IERS DOMES Number        :
     CDP Number               :
     Monument Description     : Stainless Steel
```

Geospatial data analysis: real-world example

Step 2: search sites with “stainless steel poles”

```
$ total_sites=`ls *.log | wc -l`
$ sites=`grep "^\s\s\s\s\s\Monument Description" *.log\
| grep -i "Stainless Steel pole" | wc -l`
$ percentage=$(( 10**3 * $sites*100 / $total_sites ))e-3
$ printf "Percentage of monuments with Stainless Steel pole : %.2f%%" "$percentage"
```

- Percentage of monuments with Stainless Steel pole : 24%

Geospatial data analysis: real-world example

Using python to plot the distribution of monuments styles:

Step 3: visualization

```
fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')
plt.savefig('ex1.png')
plt.show()
```

Geospatial data analysis: real-world example

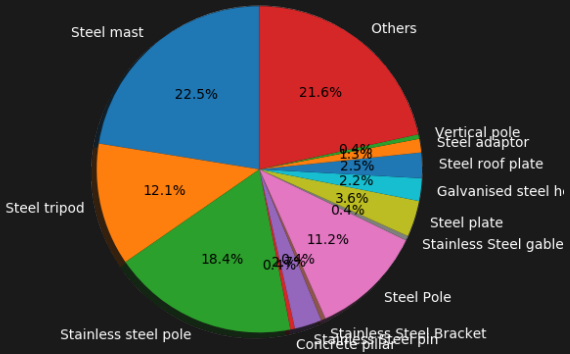

Figure 6: Monument description for all sites

Summary

- We introduced some important concepts of the command line in Linux.

Summary

- We introduced some important concepts of the command line in Linux.
- We showed an example of how to use command line tools for geospatial data analysis.

Summary

- We introduced some important concepts of the command line in Linux.
- We showed an example of how to use command line tools for geospatial data analysis.
- More details and new tools will be introduced during the next sessions.