Linux Shell

DS 5110: Big Data Systems

Spring 2025

Lecture 2

Yue Cheng

Some material taken/derived from:

- Wisconsin CS 544 by Tyler Caraza-Harter.

Learning objectives

- Setting up an EC2 VM instance via AWS Academy
- Navigate a Linux file system
- Operating within a Linux shell
- Automate repeated tasks

What is a Shell?

A shell program

```
Loop:
     print(SOME PROMPT)
     cmd = get_input()
     run(cmd)
```

If you can type it, you can automate it

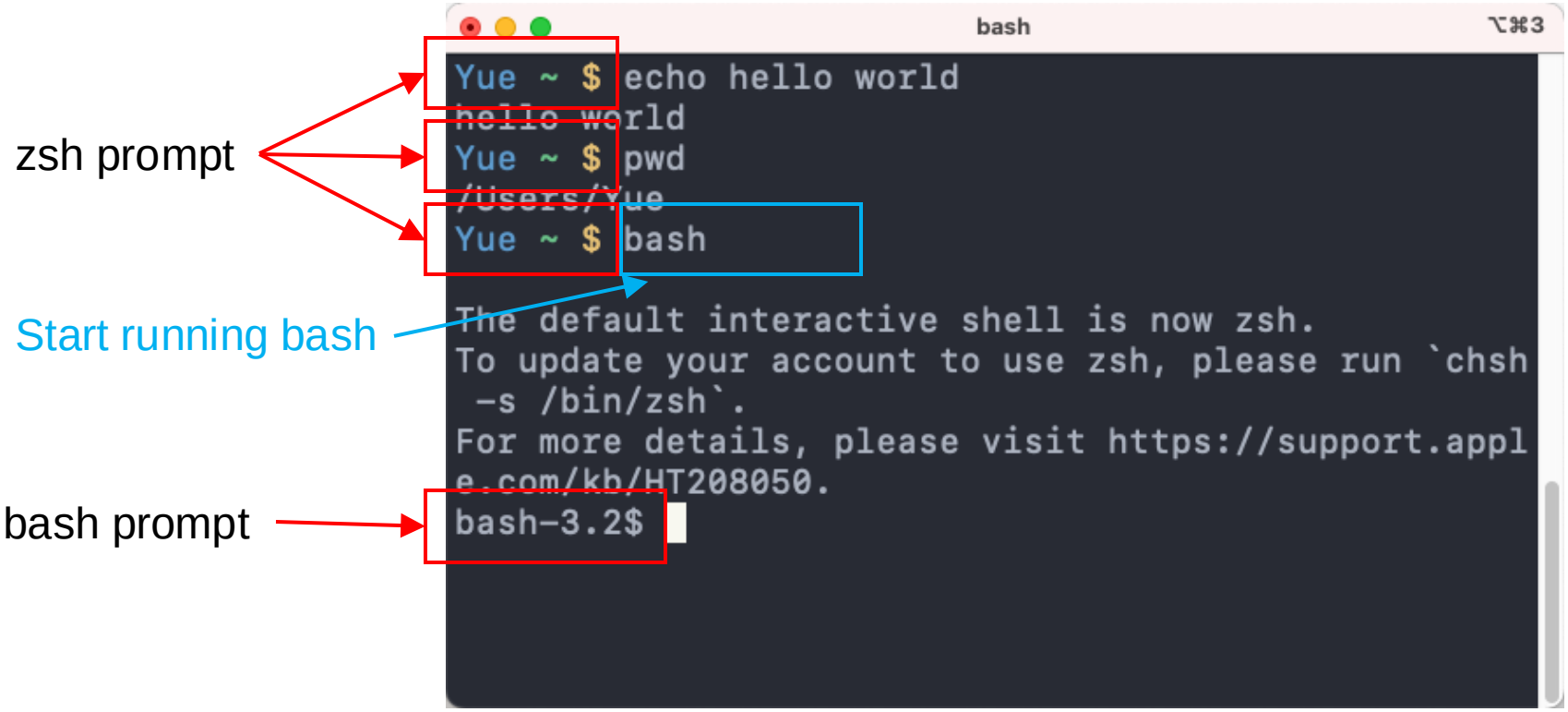
A shell program

```
Loop:
    print(SOME PROMPT)
    cmd = get_input()
    run(cmd)
```

Script.sh

```
Shell commands
…
…
```

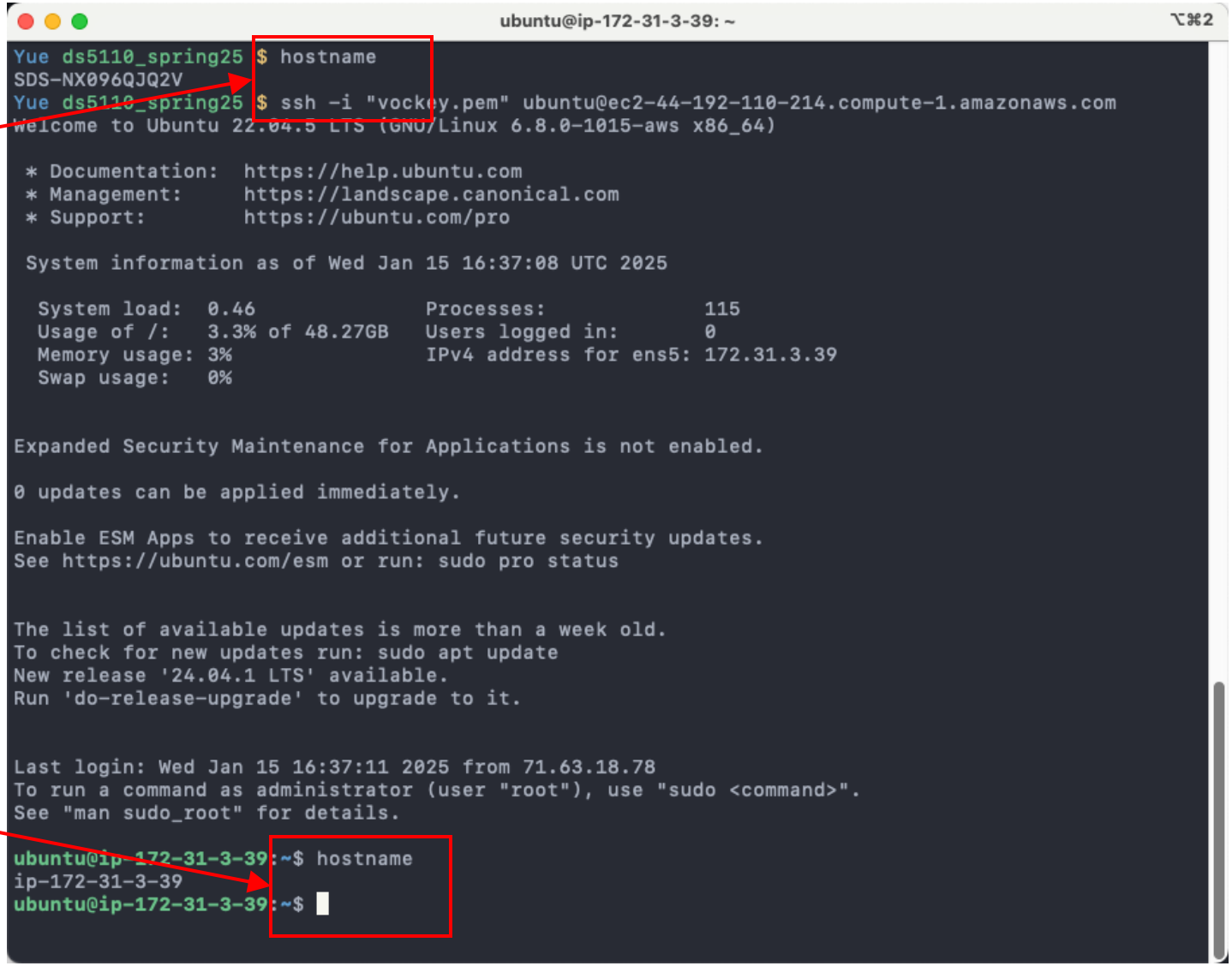
You can run a shell inside of a shell



SSH: Secure shell

Running on my laptop

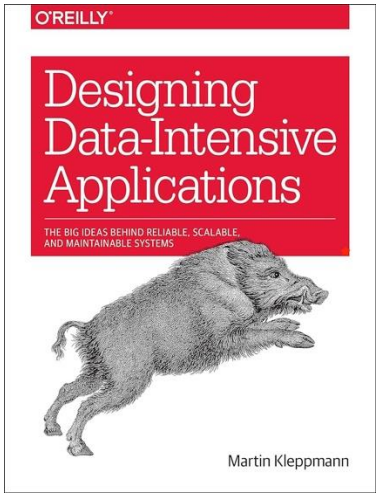
Running on my EC2 VM



Linux pipe

Unix philosophy

- “Make each program do one thing well. To do a new job, build afresh rather than complicate old programs by adding new ‘features’.”
- “Expect the output of every program to become the input of another, as yet unknown, program. Don’t clutter output with extraneous information. Avoid stringently columnar or binary input formats. Don’t insist on interactive input.”



* Designing Data-Intensive Applications (“Batch Processing with Unix Tools” of Chapter 10)

Linux pipe

Simple Log Analysis

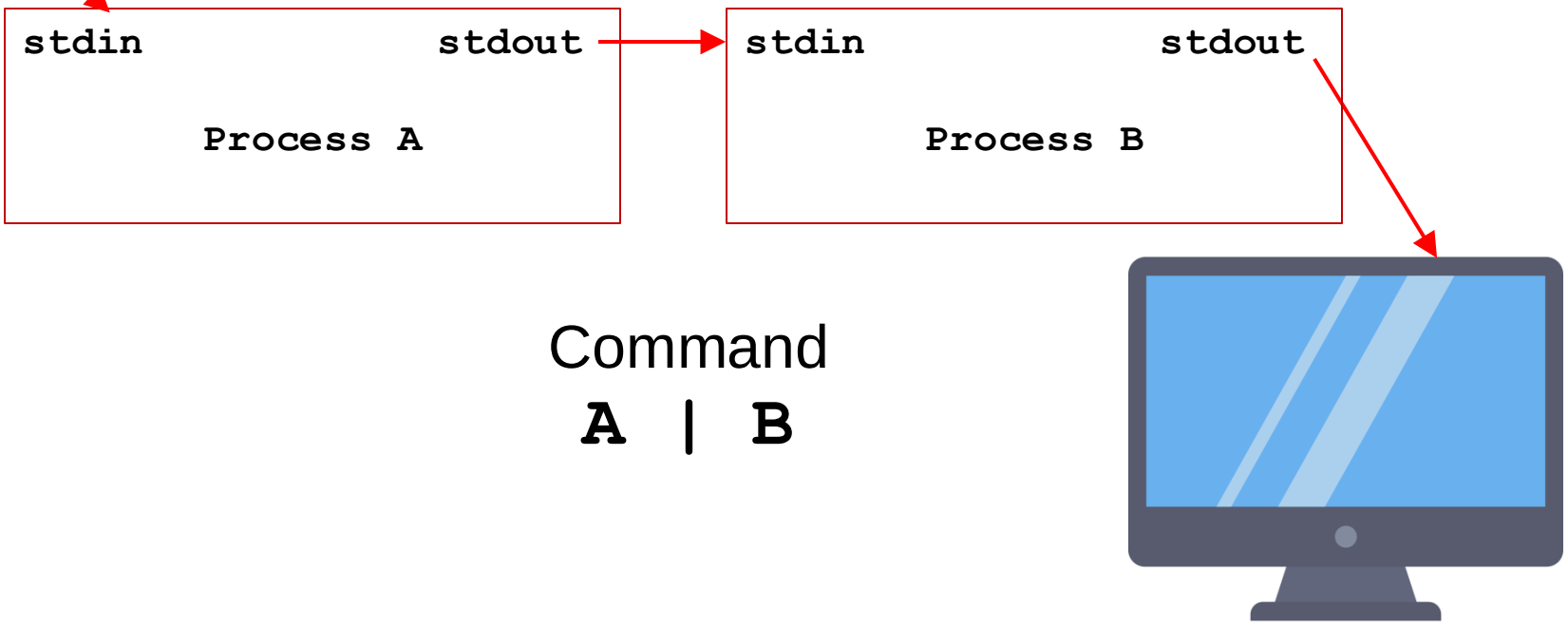
Various tools can take these log files and produce pretty reports about your website traffic, but for the sake of exercise, let's build our own, using basic Unix tools. For example, say you want to find the five most popular pages on your website. You can do this in a Unix shell as follows:

```
cat /var/log/nginx/access.log | ❶
  awk '{print $7}' | ❷
  sort             | ❸
  uniq -c          | ❹
  sort -r -n       | ❺
  head -n 5          ❻
```

The pipe | connects output of one process to input of the next.

Standard input and output (I/O)

stdout => stdin



Command

A | B

Redirection

output.txt

Command

A > output.txt

Announcement

- A0 is out
  - Due by 11am ET, Thursday, Jan 30

Demos ...