

Lecture 4

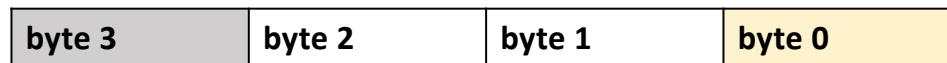
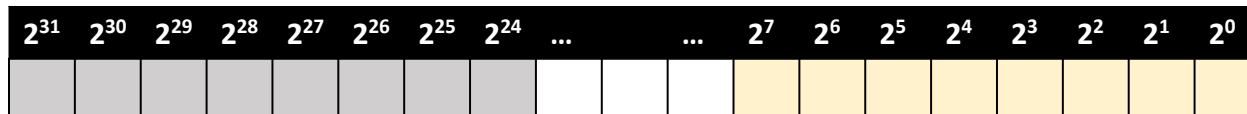
408/508 *Computational Techniques for Linguists*

Today's Topics

- Homework 1 Review
- Finish up with HTML and character encoding
- We're gonna need the **Terminal**
 - but we don't want our mistakes to kill the machine
 - can run stuff inside a guest **Operating System**, or be very careful
 - Example: `rm -rf *`
- Homework 2:
 - for Windows 11 users, install [Microsoft WSL](#)
 - for Apple Silicon users, use existing **Terminal**
 - for Linux users, do nothing

Quick Homework 1 Review

- 2,147,483,659 is a prime number.
- Does the 32 bit binary integer representation have enough bits to represent this prime? Explain.



```
~$ python
>>> 2_147_483_659
2147483659
>>> bin(2_147_483_659)
'0b100000000000000000000000000000001011'
>>> len(bin(2_147_483_659))
34
```

only as an unsigned number
NOT using 2's complement

Quick Homework 1 Review

docs.python.org/3/library/functions.html#len

len(object, /)

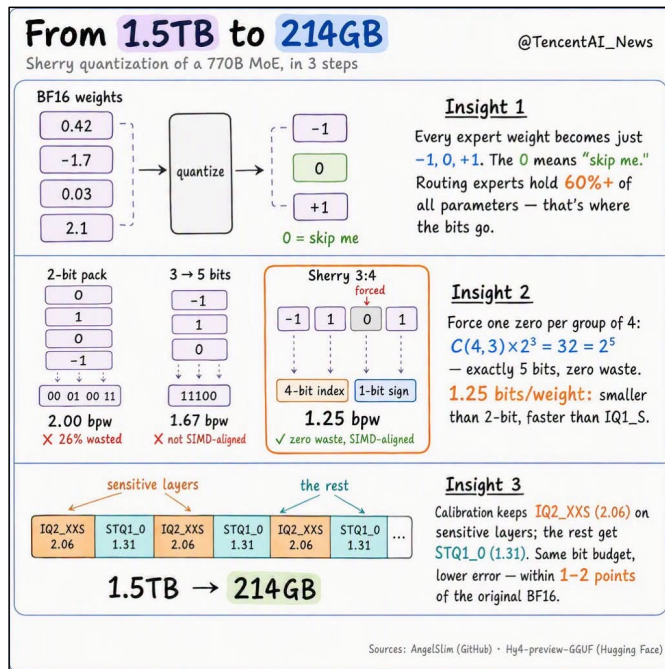
Return the length (the number of items) of an object. The argument may be a sequence (such as a string, bytes, tuple, list, or range) or a collection (such as a dictionary, set, or frozen set).

CPython implementation detail: len raises [OverflowError](#) on lengths larger than [sys.maxsize](#), such as [range\(2 ** 100\)](#).

sys.maxsize

An integer giving the maximum value a variable of type [Py_ssize_t](#) can take. It's usually $2^{31} - 1$ on a 32-bit platform and $2^{63} - 1$ on a 64-bit platform.

Making LLMs smaller



- <https://huggingface.co/AngelSlim/Hy4-preview-GGUF>
- **Idea:** reduce # of bits "needed" per floating number
- **(V)RAM needed:**
 - 1,500GB → 200GB
- Performance on benchmarks:
 - - MCP Atlas 83.7 to 83.2 (AI agents)
 - - SWE-Bench Multi 82.9 to 81.3 (Coding)
 - - MRCR 81.3 to 81.1 (Coreference retrieval task)

Introduction: data types

In your browser, which runs **HTML5**, **UTF-8**, is the default character encoding

ç	231	00E7	ç	LATIN SMALL LETTER C WITH CEDILLA
è	232	00E8	è	LATIN SMALL LETTER E WITH GRAVE
é	233	00E9	é	LATIN SMALL LETTER E WITH ACUTE
ê	234	00EA	ê	LATIN SMALL LETTER E WITH CIRCUMFLEX
ë	235	00EB	ü	LATIN SMALL LETTER E WITH DIAERESIS

[illegible]

Introduction: data types

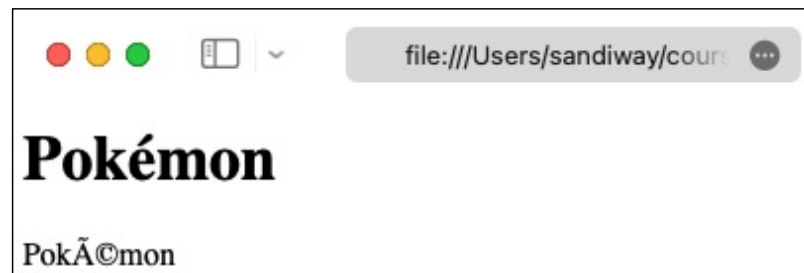
é	233	00E9	é	LATIN SMALL LETTER E WITH ACUTE
---	-----	------	----------	---------------------------------

Pokémon

File: pokemon.html

```
1<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">␣
2<html> <head>␣
3<title></title>␣
4</head>␣
5<body>␣
6<h1>Pok&eacute;mon</h1>␣
7</body>␣
8</html>␣
```

Introduction: data types



line 7: é typed on my mac keyboard using OPT-e e

File: pokemon2.html

```
1<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
2<html> <head>
3<title></title>
4</head>
5<body>
6<h1>Pok&eacute;mon</h1>
7Pokémon
8</body>
9</html>
```


Introduction: data types



File: pokemon3.html

```
1 <!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
2 <html> <head>
3 <meta charset="UTF-8">
4 <title></title>
5 </head>
6 <body>
7 <h1>Pok&eacute;mon</h1>
8 Pokémon
9 </body>
10 </html>
```

Introduction: data types

Bits of code point	First code point	Last code point	Bytes in sequence	Byte 1	Byte 2	Byte 3	Byte 4
7	U+0000	U+007F	1	0xxxxxxx			
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx		
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx	
21	U+10000	U+1FFFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

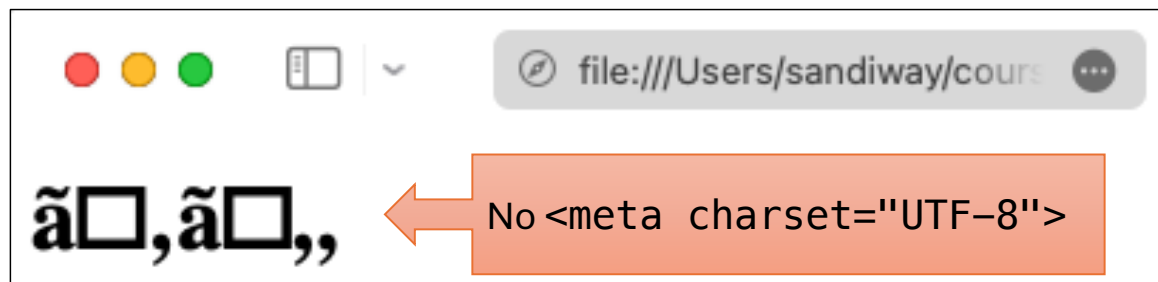
- Example:
 - あ Hiragana letter A: UTF-8: E38182
 - Byte 1: E = 1110, 3 = 0011
 - Byte 2: 8 = 1000, 1 = 0001
 - Byte 3: 8 = 1000, 2 = 0010
 - い Hiragana letter I: UTF-8: E38184

Shift-JIS (Hex):

あ: 82A0

い: 82A2

Introduction: data types



- あ Hiragana letter A: UTF-8: E3 81 82
- い Hiragana letter I: UTF-8: E3 81 84

227	E3	ã
-----	----	---

Latin-1 character set (8 bits)<https://kb.iu.edu/d/aepu>

Introduction: data types

Many [Windows](#) programs (including Windows [Notepad](#)) add the bytes 0xEF, 0xBB, 0xBF at the start of any document saved as UTF-8. This is the UTF-8 encoding of the Unicode [byte order mark](#) (BOM), and is commonly referred to as a UTF-8 BOM, even though it is not relevant to byte order. A BOM can also appear if another encoding with a BOM is translated to UTF-8 without stripping it. Software that is not aware of multibyte encodings will display the BOM as three strange characters (e.g. "ï»¿" in software interpreting the document as [ISO 8859-1](#) or [Windows-1252](#)) at the start of the document.

Introduction: data types

- How can you tell what encoding your file is using?
- Detecting UTF-8
 - Microsoft:
 - 1st three bytes in the file is EF BB BF
 - *(not all software understands this; not everybody uses it)*
 - HTML:
 - `<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">`
 - *(not always present)*
 - Analyze the file:
 - Find non-valid UTF-8 sequences: if found, not UTF-8...
 - Interesting paper:
 - <http://www-archive.mozilla.org/projects/intl/UniversalCharsetDetection.html>

WSL2

- Windows Subsystem for Linux
 - <https://docs.microsoft.com/en-us/windows/wsl/>

Windows Subsystem for Linux Documentation

07/22/2020 • 2 minutes to read • 🗨️ 📧 📄 📖 📑

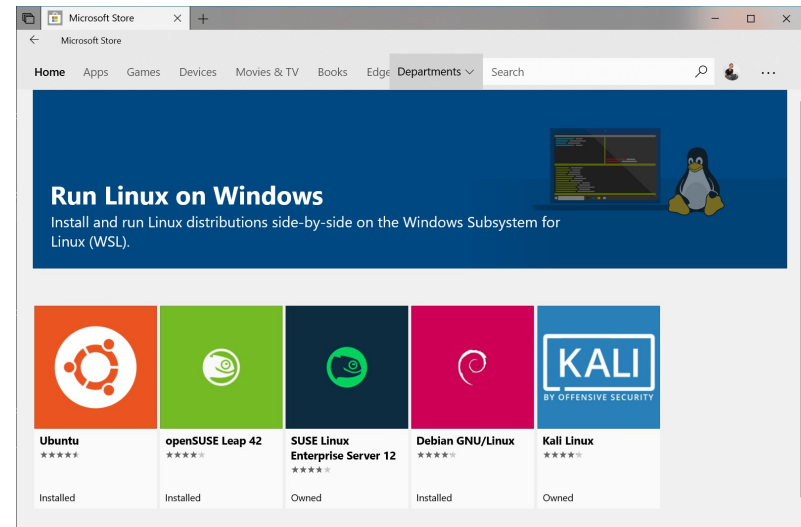
The Windows Subsystem for Linux lets developers run a GNU/Linux environment -- including most command-line tools, utilities, and applications -- directly on Windows, unmodified, without the overhead of a traditional virtual machine or dualboot setup.

Learn more here

- [What is the Windows Subsystem for Linux?](#)
- [What's new with WSL 2?](#)
- [Compare WSL 2 and WSL 1](#)
- [Read frequently asked questions](#)

Get started

- [Install WSL1](#)
- [Update to WSL2](#)



WSL2

- Consider installing WSL2 (Windows Subsystem for Linux 2)
 - <https://learn.microsoft.com/en-us/windows/wsl/>

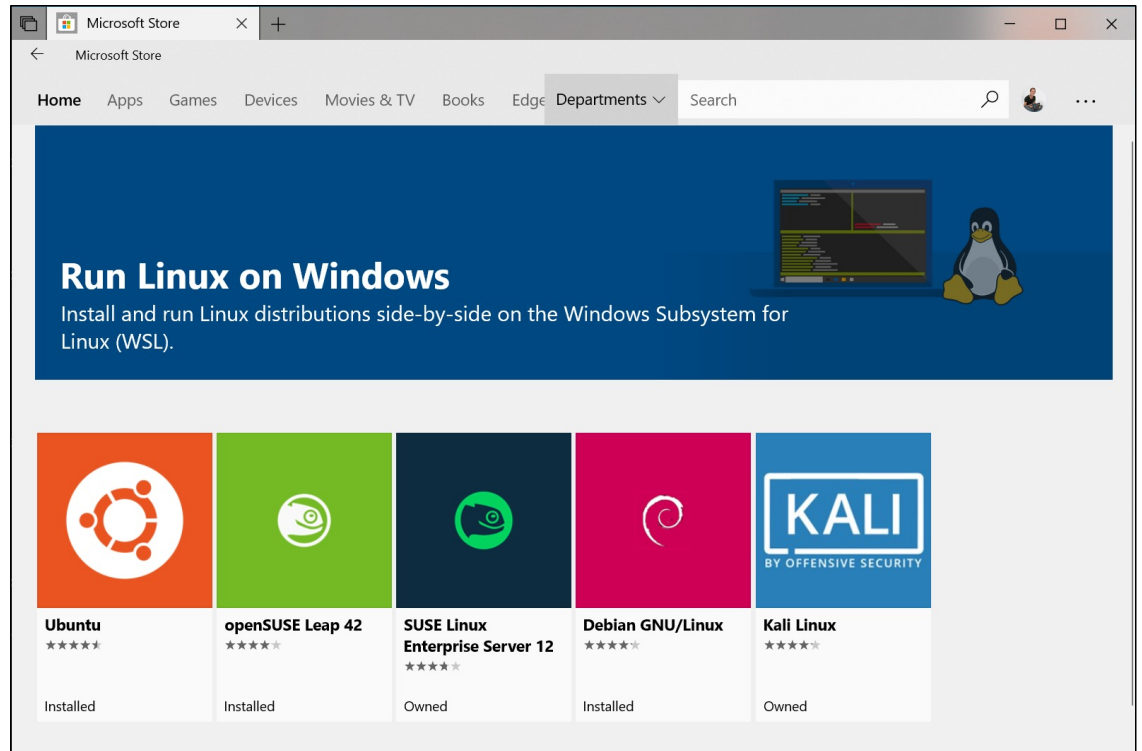
Get started

- [Install WSL](#)
- [Install Linux on Windows Server](#)
- [Manual install steps](#)
- [Best practices for setting up a WSL development environment](#)

- **Why?**
 - gives you a Bash shell (*with quoting consistent with the lecture slides*)
 - (*simultaneously*) use Linux under Windows 10/11 (not dual-booting)
 - can access your Windows C: drive from Ubuntu via directory `/mnt/c`

WSL2

*I installed
Ubuntu*



Ubuntu

Installing, this may take a few minutes...

Please create a default UNIX user account. The username does not need to match your Windows username.

For more information visit: <https://aka.ms/wslusers>

Enter new UNIX username:

WSL2

sandiway@DESKTOP-VEPP64Q: ~

Retype new password:

passwd: password updated successfully

Installation successful!

To run a command as administrator (user "root"), use "sudo <command>".

See "man sudo_root" for details.

Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.10.16.3-microsoft-standard-WSL2 x86_64)

- * Documentation: <https://help.ubuntu.com>
- * Management: <https://landscape.canonical.com>
- * Support: <https://ubuntu.com/advantage>

System information as of Fri Sep 10 10:01:52 MST 2021

System load:	0.13	Processes:	8
Usage of /:	0.5% of 250.98GB	Users logged in:	0
Memory usage:	0%	IPv4 address for eth0:	172.24.13.136
Swap usage:	0%		

1 update can be applied immediately.

To see these additional updates run: `apt list --upgradable`

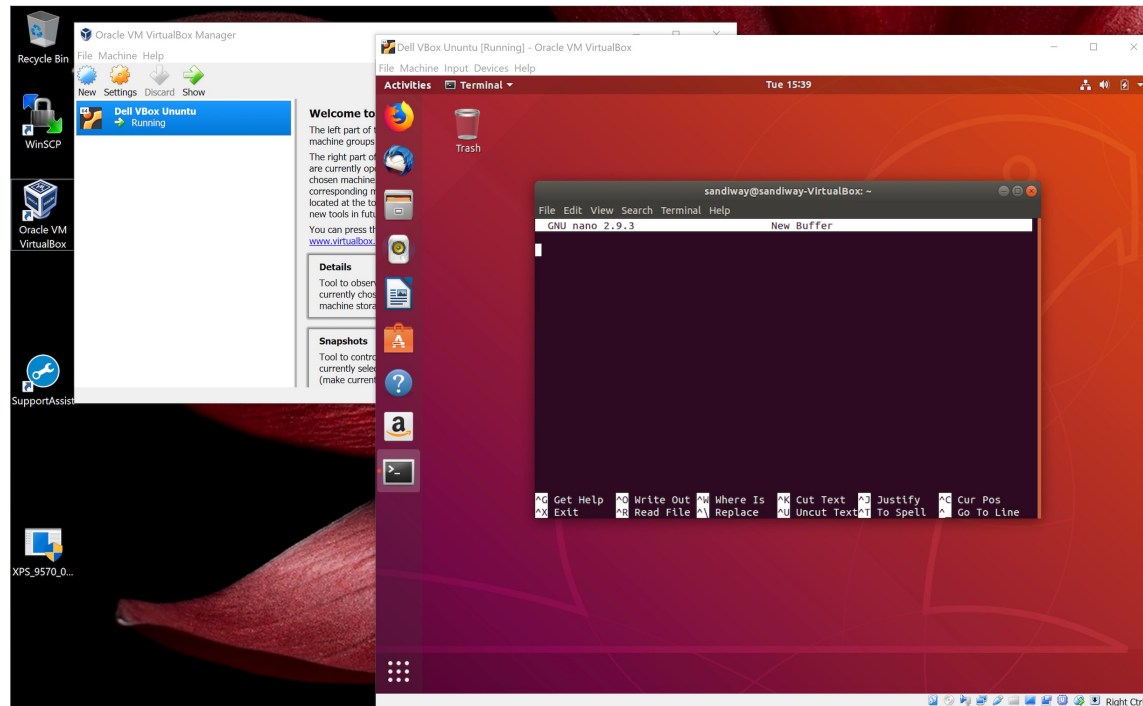
The list of available updates is more than a week old.

To check for new updates run: `sudo apt update`

This message is shown once a day. To disable it please create the
/home/sandiway/.hushlogin file.

sandiway@DESKTOP-VEPP64Q:~\$

Ubuntu Text Editor: nano



- **Text editor (built in)**
 - nano is a decent one for (use inside Terminal)
 - You might prefer it, but I can't stand vi
 - can install others via
 - `sudo apt-get install`
- or
 - `sudo apt install`

Ubuntu: python3 is already installed!

```
sandiway@DESKTOP-CL5DHR x + v

System load: 0.32          Processes:           62
Usage of /:  0.5% of 1006.85GB Users logged in:    0
Memory usage: 2%          IPv4 address for eth0: 172.20.250.158
Swap usage:  0%

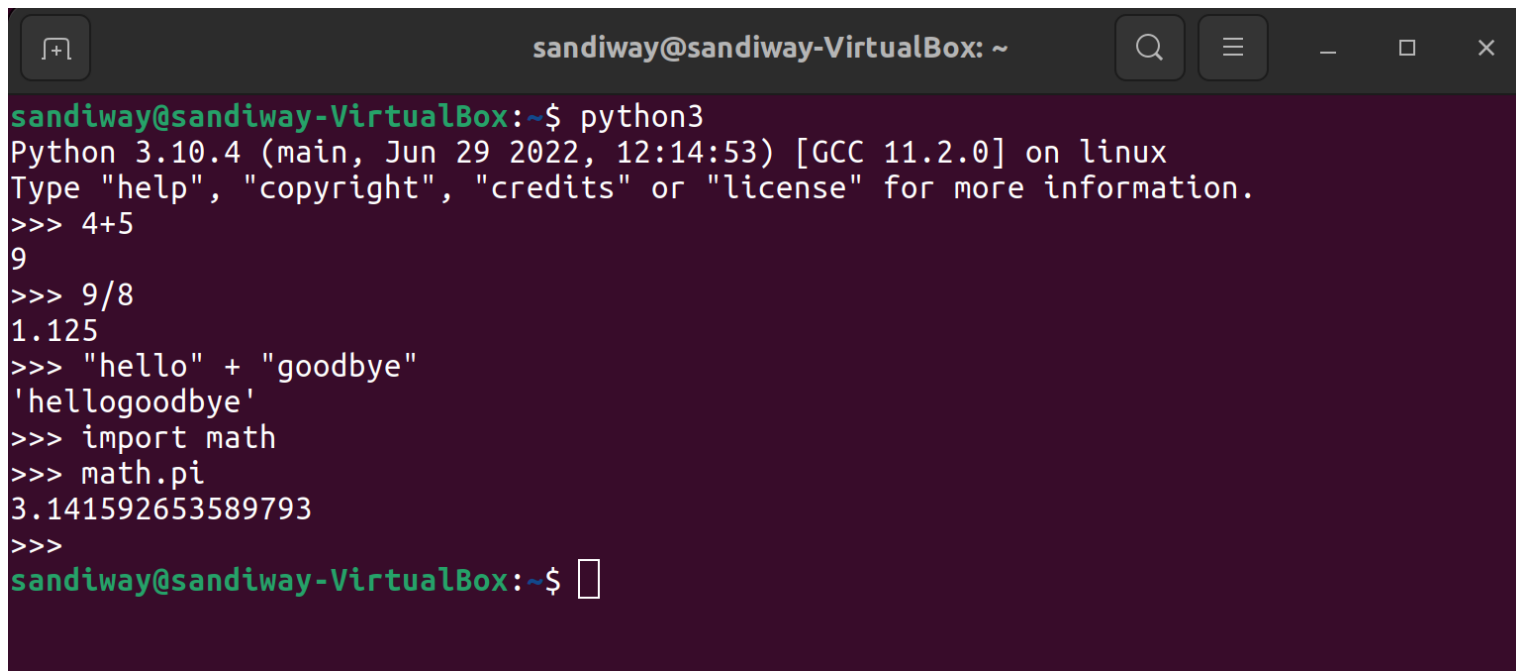
=> There is 1 zombie process.

This message is shown once a day. To disable it please create the
/home/sandiway/.hushlogin file.
sandiway@DESKTOP-CL5DHR:~$ which python
sandiway@DESKTOP-CL5DHR:~$ which python3
/usr/bin/python3
sandiway@DESKTOP-CL5DHR:~$ sudo apt install python
[sudo] password for sandiway:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Package python is not available, but is referred to by another package.
This may mean that the package is missing, has been obsoleted, or
is only available from another source
However the following packages replace it:
  2to3 python-is-python3

E: Package 'python' has no installation candidate
sandiway@DESKTOP-CL5DHR:~$ python3
Python 3.12.3 (main, Feb  4 2025, 14:48:35) [GCC 13.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

- `which python`
returns nothing
- `apt install python`
doesn't work: 1) not administrator, and
2) there's no such package!
- `sudo apt install python3`
if it says it's already there, then
[sudo may ask you for your password]
- `python3`
places you in the interpreter at the
prompt `>>>`
- `<Control>-D`
quits Python

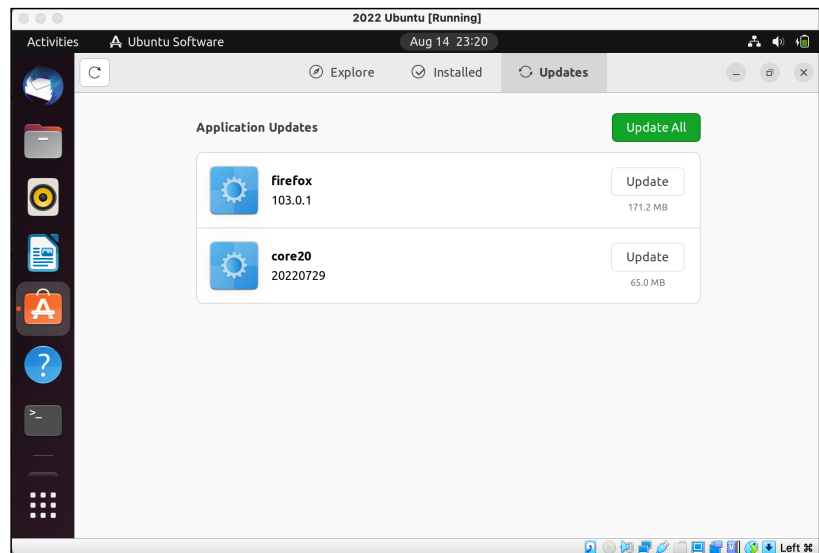
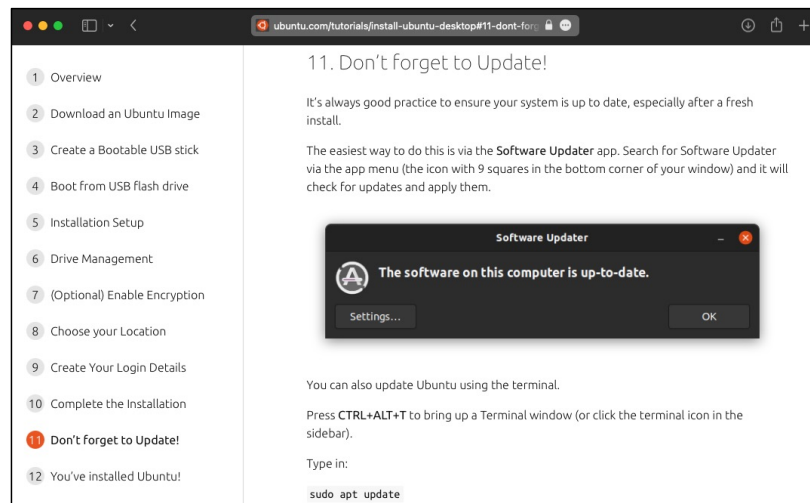
Ubuntu: python3 is already installed!

A terminal window titled 'sandiway@sandiway-VirtualBox: ~' with standard window controls. The terminal shows the execution of 'python3', which outputs the Python version and environment information. Subsequent interactive commands demonstrate basic arithmetic, string concatenation, and module usage.

```
sandiway@sandiway-VirtualBox:~$ python3
Python 3.10.4 (main, Jun 29 2022, 12:14:53) [GCC 11.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> 4+5
9
>>> 9/8
1.125
>>> "hello" + "goodbye"
'hellogoodbye'
>>> import math
>>> math.pi
3.141592653589793
>>>
sandiway@sandiway-VirtualBox:~$
```

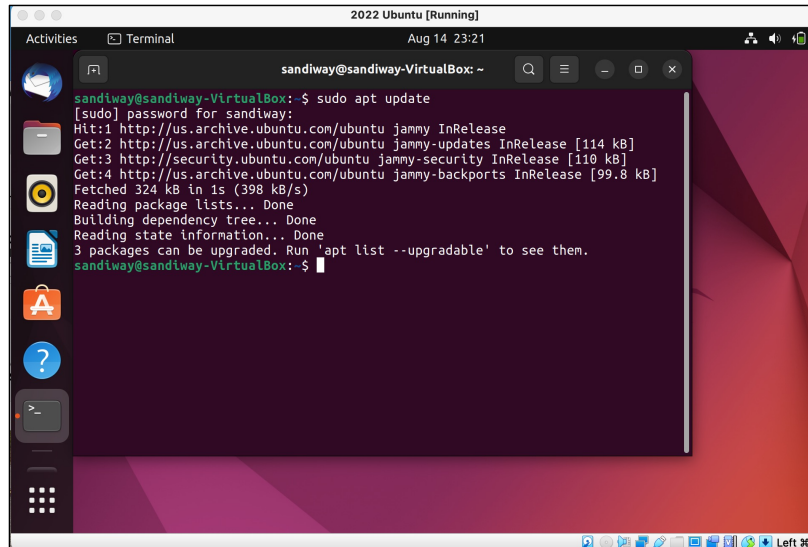
Keep Ubuntu up to date

<https://ubuntu.com/tutorials/install-ubuntu-desktop#1-overview>



Keep Ubuntu up to date

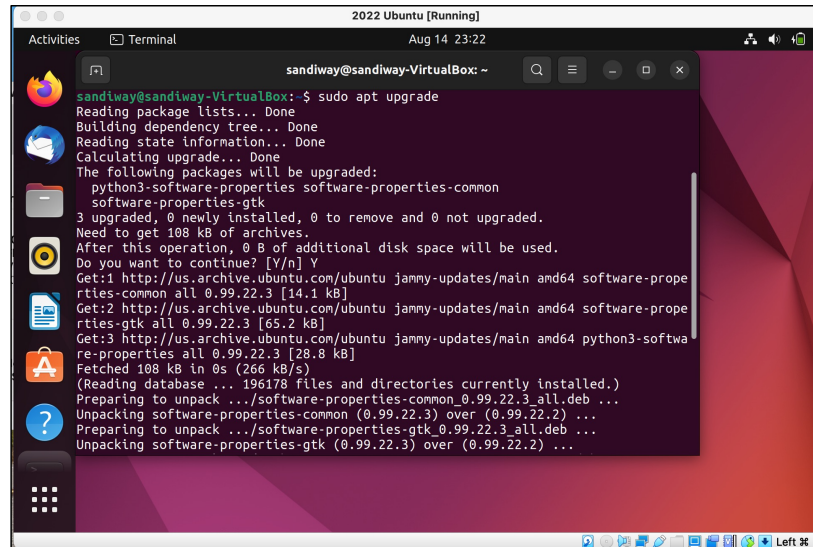
- `sudo apt update`



A terminal window titled "2022 Ubuntu [Running]" showing the output of the `sudo apt update` command. The terminal displays the process of updating package lists from various sources, including the main Ubuntu repository and security updates. It shows the amount of data fetched (324 kB) and the completion of the update process.

```
sandiway@sandiway-VirtualBox: ~  
sandiway@sandiway-VirtualBox:~$ sudo apt update  
[sudo] password for sandiway:  
Hit:1 http://us.archive.ubuntu.com/ubuntu jammy InRelease  
Get:2 http://us.archive.ubuntu.com/ubuntu jammy-updates InRelease [114 kB]  
Get:3 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]  
Get:4 http://us.archive.ubuntu.com/ubuntu jammy-backports InRelease [99.8 kB]  
Fetched 324 kB in 1s (398 kB/s)  
Reading package lists... Done  
Building dependency tree... Done  
Reading state information... Done  
3 packages can be upgraded. Run 'apt list --upgradable' to see them.  
sandiway@sandiway-VirtualBox:~$
```

- `sudo apt upgrade`



A terminal window titled "2022 Ubuntu [Running]" showing the output of the `sudo apt upgrade` command. The terminal displays the process of upgrading installed packages. It lists the packages to be upgraded (python3-software-properties, software-properties-common, and software-properties-gtk) and shows the progress of the upgrade, including the amount of disk space required and the completion of the process.

```
sandiway@sandiway-VirtualBox:~$ sudo apt upgrade  
Reading package lists... Done  
Building dependency tree... Done  
Reading state information... Done  
Calculating upgrade... Done  
The following packages will be upgraded:  
  python3-software-properties software-properties-common  
  software-properties-gtk  
3 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.  
Need to get 108 kB of archives.  
After this operation, 0 B of additional disk space will be used.  
Do you want to continue? [Y/n] Y  
Get:1 http://us.archive.ubuntu.com/ubuntu jammy-updates/main amd64 software-prope  
rties-common all 0.99.22.3 [14.1 kB]  
Get:2 http://us.archive.ubuntu.com/ubuntu jammy-updates/main amd64 software-prope  
rties-gtk all 0.99.22.3 [65.2 kB]  
Get:3 http://us.archive.ubuntu.com/ubuntu jammy-updates/main amd64 python3-softwa  
re-properties all 0.99.22.3 [28.8 kB]  
Fetched 108 kB in 0s (266 kB/s)  
(Reading database ... 196178 files and directories currently installed.)  
Preparing to unpack .../software-properties-common_0.99.22.3_all.deb ...  
Unpacking software-properties-common (0.99.22.3) over (0.99.22.2) ...  
Preparing to unpack .../software-properties-gtk_0.99.22.3_all.deb ...  
Unpacking software-properties-gtk (0.99.22.3) over (0.99.22.2) ...  
Preparing to unpack .../python3-software-properties_0.99.22.3_all.deb ...  
Unpacking python3-software-properties (0.99.22.3) over (0.99.22.2) ...  
Setting up software-properties-common (0.99.22.3) ...  
Setting up software-properties-gtk (0.99.22.3) ...  
Setting up python3-software-properties (0.99.22.3) ...  
sandiway@sandiway-VirtualBox:~$
```

Ubuntu tidbits

- Latest version (LTS = Long Term support)
 - *Ubuntu 24.04 LTS Noble Numbat*
 - `lsb_release -a`
- Tk graphics with Python 3
 - `sudo apt install python3-tk`
- Python's Package Manager PIP
 - `sudo apt install python3-pip`
 - `pip3 --version`



ubuntu[®] packages

» Ubuntu » Packages » plucky (25.04) » python » python3-tk

[Source: [python3-stdlib-extensions](#)] [[jammy](#)]

Package: python3-tk (3.13.3-1)

Tkinter - Writing Tk applications with Python 3.x

Other Packages Related to python3-tk

● depends ◆ recommends ■ suggests • enhances

- [blt](#) (>= 2.4z-9)
graphics extension library for Tcl/Tk - run-time
- [python3](#) (<< 3.14)
interactive high-level object-oriented language (default python3 version)
- [python3](#) (>= 3.13.2-1~)
- [python3.13-tk](#)
Tkinter - Writing Tk applications with Python (v3.13)
- [python3-tk-dbg](#)
Tkinter - Writing Tk applications with Python 3.x (debug extension)
- [tix](#)
library for Tk -- runtime package