

LING 408/508: Programming for Linguists

Lecture 4

Today's Topics

- Homework 1 graded
- Homework 1 Review
- Unicode
- File Systems
- Special characters: end of file
- Homework 2:
 - Install VirtualBox on your computer



Homework 1 Review

`math.pi` in Python 3 reports the decimal value of PI to the best of its ability

```
ling538-20$ python3
Python 3.8.3 (v3.8.3:6f8c8320e9, May 13 2020, 16:29:34)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import math
>>> math.pi
3.141592653589793
>>> █
```

Homework 1 Review

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	Sign	Exponent									Fraction																						
2	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	
4	binary point @ left of bit 22																																
5	Sval	Exponent Value (bits 30 to 23)									Fraction Value (1 + bits 22 to 0)																						
6	1	1									1.5																						
7	Decimal Value																																
8	3																																
9	ExpVal	Exponent									Decimal Value																						
10	-3	0	1	1	1	1	1	0	0	0.125																							
11	-2	0	1	1	1	1	1	0	1	0.25																							
12	-1	0	1	1	1	1	1	1	0	0.5																							
13	0	0	1	1	1	1	1	1	1	1																							
14	1	1	0	0	0	0	0	0	0	2																							
15	2	1	0	0	0	0	0	0	1	4																							
16	3	1	0	0	0	0	0	1	0	8																							

Homework 1 Review

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	Sign	Exponent																	
2	0	1	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0
3	31	30	29	28	27	26	25	24	23		22	21	20	19	18	17	16	15	
4										binary point @ left of bit 22									
5	Sval	Exponent Value (bits 30 to 23)									Fraction Value (1 + bits 22 to 0)								
6	1	1									1.5625								
7	Decimal Value																		
8	3.125																		

Homework 1 Review

- Continuing from left to right flipping bits, we should converge on:

7	Decimal Value
8	3.141592503

```
[~$ python3 -c 'import math; print(math.pi)'  
3.141592653589793
```

Decimal Value
3.141592741

Unicode

<https://home.unicode.org>

The screenshot shows the Unicode.org homepage. On the left is a navigation menu with the Uni logo and links for Adopt a Character, Emoji, Basic Info, News, Events, Connect, Membership, and Press. The main content area features a grid of adopted characters with their Unicode points. A central message states: "Everyone in the world should be able to use their own language on phones and computers." Below this is a "LEARN MORE ABOUT UNICODE" link and a "ADOPT A CHARACTER" button. At the bottom, there are links for "Tableaux des caractères Unicode" and "Unicode Regular Expressions v21".

U+4E14	U+05D3	U+3007	U+0E25	U+1F5A4	U+00D3	U+2015	U+263C
U+270C	U+65E6	U+FF74	U+0398	U+1F49B	U+0918	U+3064	U+2191
U+FF03	U+141B	U+30FC	U+2690	U+1F91E	U+76BF	U+26A2	U+0923
U+2752	U+0296	U+1F490	U+4E09	U+1F970	U+0CA3	U+2193	U+2666

Tableaux des caractères Unicode

Unicode Regular Expressions v21

Unicode

- <https://www.unicode.org/versions/Unicode13.0.0/UnicodeStandard-13.0.pdf>

UTF-8

To meet the requirements of byte-oriented, ASCII-based systems, a third encoding form is specified by the Unicode Standard: UTF-8. This variable-width encoding form preserves ASCII transparency by making use of 8-bit code units.

Preferred Usage. UTF-8 is typically the preferred encoding form for HTML and similar protocols, particularly for the Internet. The ASCII transparency helps migration. UTF-8 also has the advantage that it is already inherently byte-serialized, as for most existing 8-bit character sets; strings of UTF-8 work easily with the C standard library, and many existing APIs that work for typical East Asian multibyte character sets adapt to UTF-8 as well with little or no change required.

On the Mac

The screenshot shows the Mac Character Viewer window. On the left is a sidebar with categories like 'Frequently Used', 'Favorites', 'Arrows', 'Bullets/Stars', 'Currency Symbols', 'Latin', 'Letterlike Symbols', 'Math Symbols', 'Parentheses', 'Pictographs', 'Punctuation', 'Enclose...aracters', 'Sign/St...Symbols', 'Technic...Symbols', 'Digits - All', 'Unicode', and 'Phonetic Alphabet'. The main area is a table with columns for 'Unicode', 'Title', and 'Category'. Below the table, there are two sections: 'Basic Latin' and 'Latin-1 Supplement'. The 'Basic Latin' section shows characters from U+0000 to U+007F, including the standard ASCII set. The 'Latin-1 Supplement' section shows characters from U+0080 to U+00FF, including accented letters and special characters.

Unicode	Title	Category
00000000	Basic Latin	Latin
00000080	Latin-1 Supplement	Latin
00000100	Latin Extended-A	Latin
00000180	Latin Extended-B	Latin
00000250	IPA Extensions	IPA Extensions

Basic Latin

Unicode	Character
0000	[] [] [] [] [] [] [] [] [] [] [] [] [] [] [] []
0010	[] [] [] [] [] [] [] [] [] [] [] [] [] [] [] []
0020	! " # \$ % & ' () * + , - . /
0030	0 1 2 3 4 5 6 7 8 9 : ; < = > ?
0040	@ A B C D E F G H I J K L M N O
0050	P Q R S T U V W X Y Z [\] ^ _
0060	` a b c d e f g h i j k l m n o
0070	p q r s t u v w x y z { } ~ []

Latin-1 Supplement

0080	[] [] [] [] [] [] [] [] [] [] [] [] [] [] [] []
0090	[] [] [] [] [] [] [] [] [] [] [] [] [] [] [] []
00A0	¡ ¢ £ ¤ ¥ ¦ § ¨ © ª « ¬ ® ¯
00B0	° ± ² ³ ´ µ ¶ · ¸ ¹ º » ¼ ½ ¾ ¿
00C0	À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï
00D0	Ð Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß
00E0	à á â ã ä å æ ç è é ê ë ì í î ï
00F0	ð ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ÿ

← ASCII

← Latin-1 supplement

Introduction: data types

- UTF-8
 - standard in the post-ASCII world
 - backwards compatible with ASCII
 - *(previously, different languages had multi-byte character sets that clashed)*
 - Universal Character Set (UCS) Transformation Format 8-bits

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

(Wikipedia)

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

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>

Introduction: data types

- **Filesystem:**

- different on different computers: *sometimes a problem if you mount filesystems across different systems*

- **Examples:**

- FAT32 (File Allocation Table)

limited to 4GB max file size

- ExFAT (Extended FAT)
- NTFS (New Technology File System)
- ext4 (Fourth Extended Filesystem)
- HFS+ (Hierarchical File System Plus)

DOS, Windows,

memory cards

SD cards (> 4GB files)

Windows

Linux

Macs

Introduction: data types

- **Filesystem:**

- different on different computers: *sometimes a problem if you mount filesystems across different systems*

- **Files:**

- Name (Path from / root)
- Type (e.g. .docx, .pptx, .pdf, .html, .txt)
- Owner (*usually the Creator*)
- Permissions (for the Owner, Group, or Everyone)
- need to be opened (*to read from or write to*)
- Mode: read/write/append
- Binary/Text

in all programming languages:
open command

Introduction: data types

- Text files:
 - text files have lines: *how do we mark the end of a line?*
 - End of line (EOL) control character(s):
 - LF 0x0A (Mac/Linux),
 - CR 0x0D (Old Macs),
 - CR+LF 0x0D0A (Windows)
 - End of file (EOF) control character:
 - EOT 0x04 (*aka* Control-D)

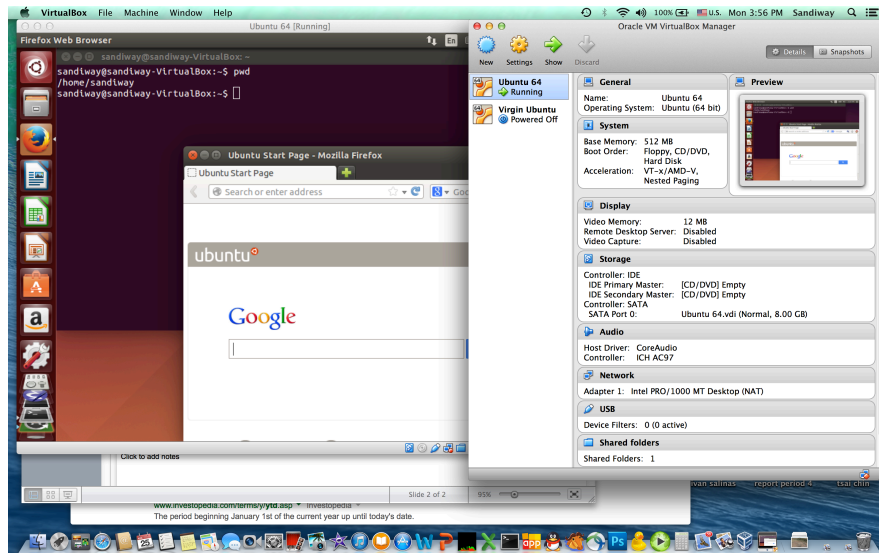
programming languages:
NUL used to mark
the end of a string

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

VirtualBox

- Virtual x86 machine(s)
 - install other operating systems (OSs) running inside a window, we'll install Ubuntu (Linux) as a Guest OS
- Free application at <https://www.virtualbox.org>



Ubuntu 14.04 LTS*
under
VirtualBox
under
OS X 10.10.5

*LTS = Long Term Support

Homework 2

- Install VirtualBox on your laptop:



VirtualBox

Welcome to VirtualBox.org!

VirtualBox is a powerful x86 and AMD64/Intel64 [virtualization](#) product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. See "[About VirtualBox](#)" for an introduction.

Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of [guest operating systems](#) including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

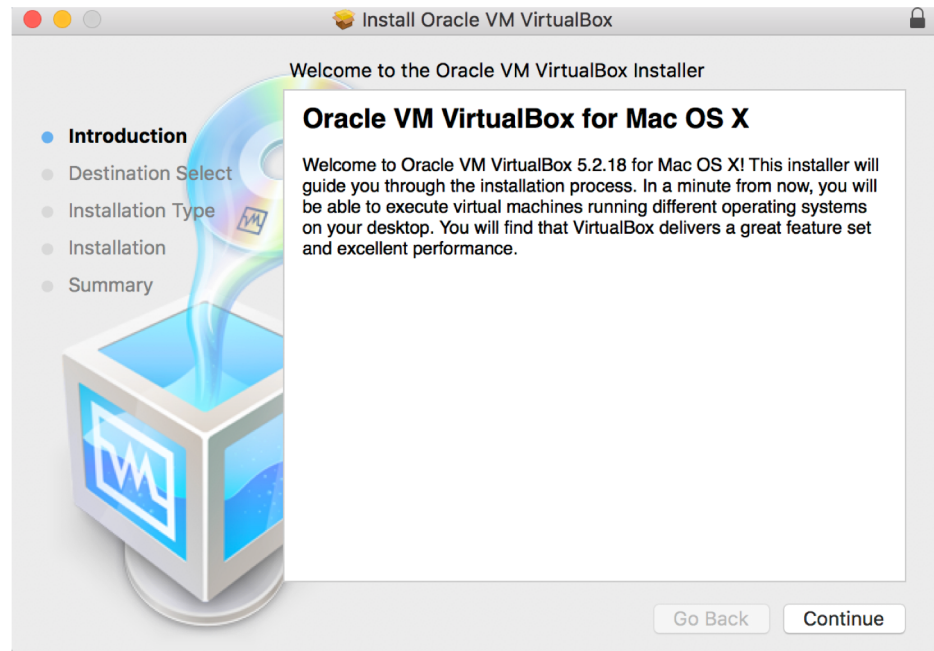
[Download VirtualBox 6.1](#)

VirtualBox 6.1.12 platform packages

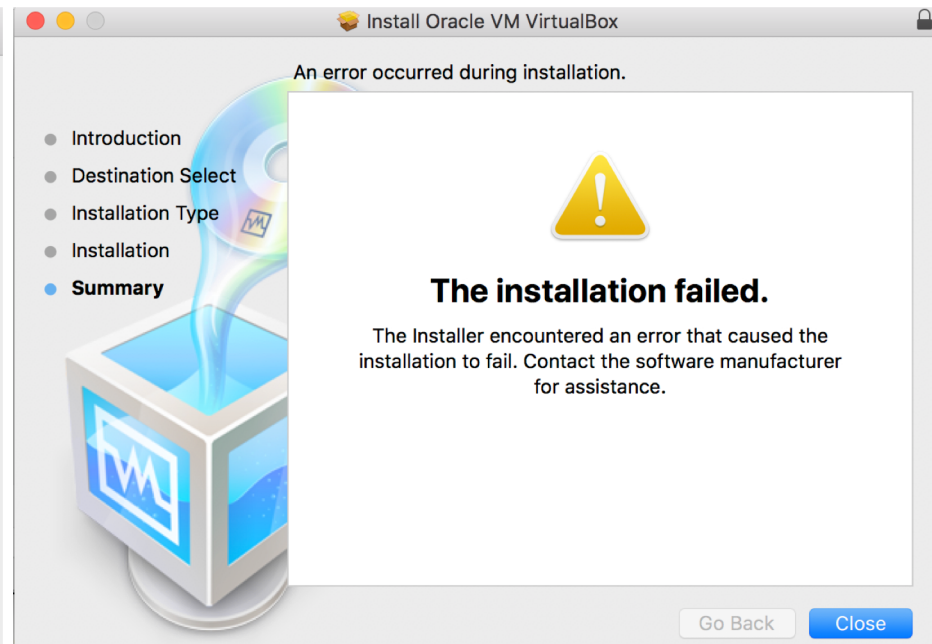
- [Windows hosts](#)
- [OS X hosts](#)
- [Linux distributions](#)
- [Solaris hosts](#)

[About](#)
[Screenshots](#)
[Downloads](#)
[Documentation](#)
 [End-user docs](#)
 [Technical docs](#)
[Contribute](#)
[Community](#)

Homework 2

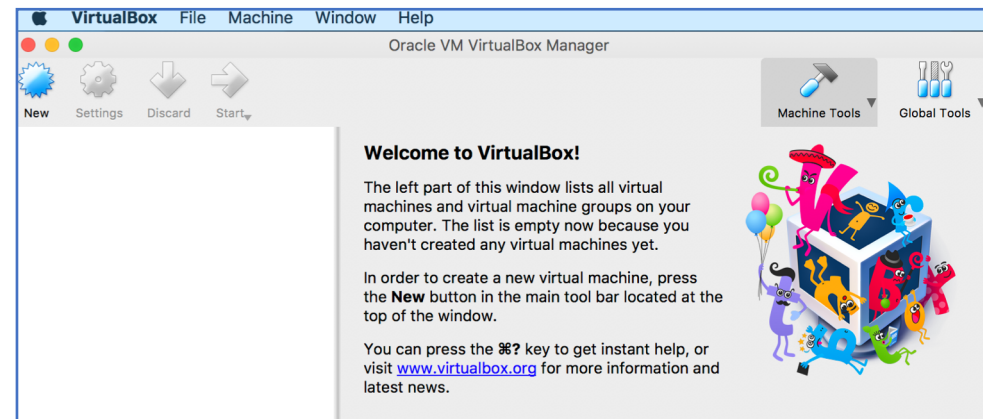
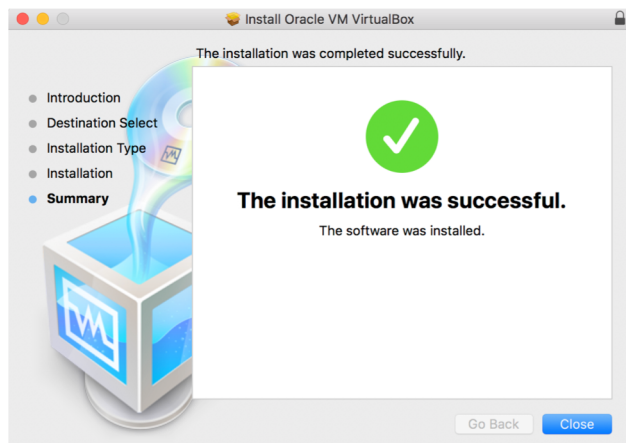


Homework 2



Homework 2

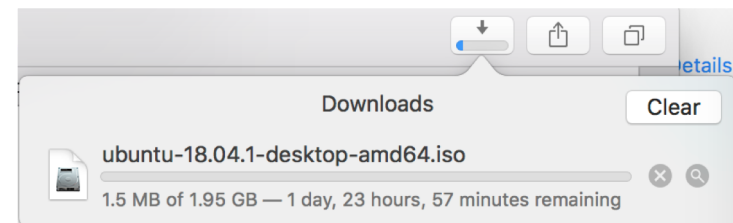
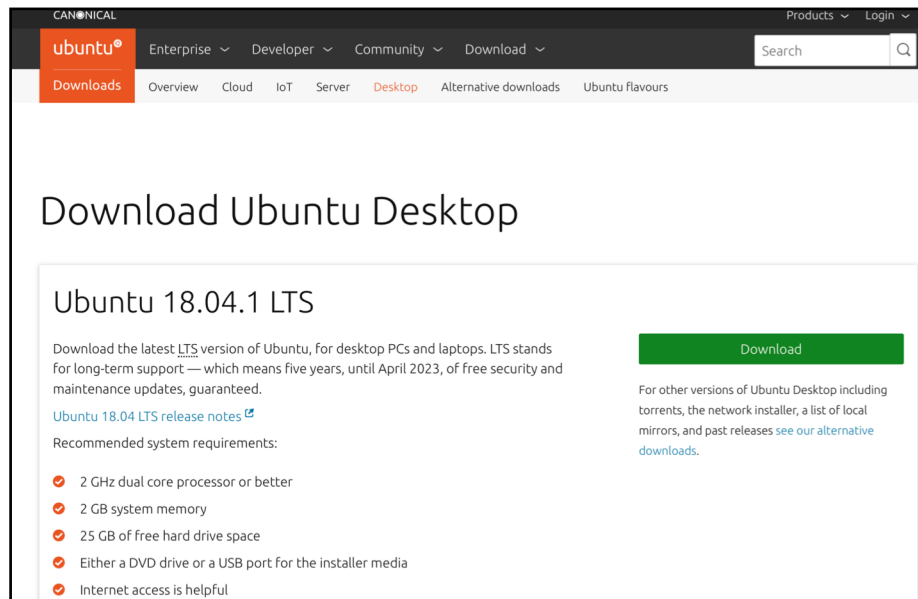
- Re-run installer after giving Oracle Computer permission



Read <https://www.virtualbox.org/manual/ch01.html#idm272>

Homework 2

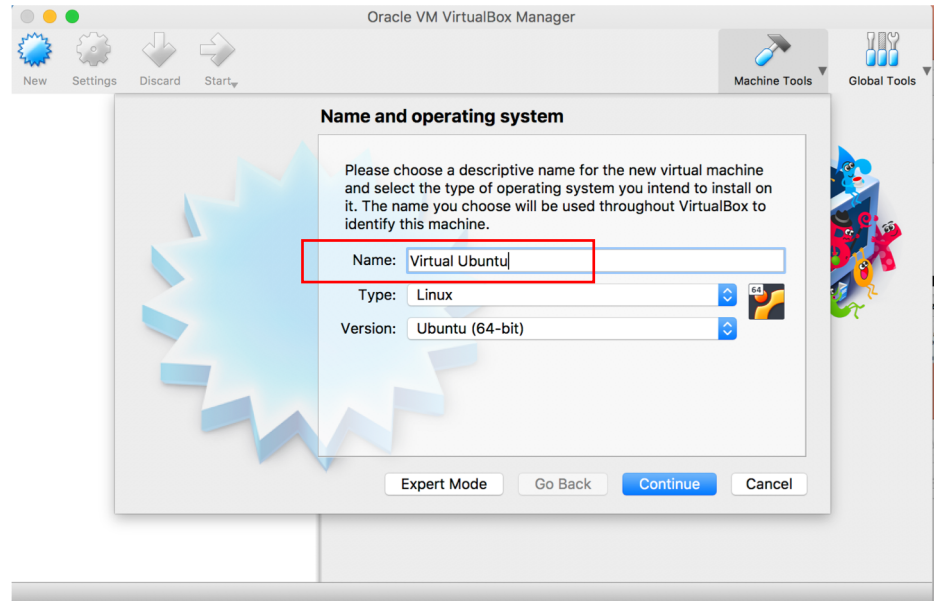
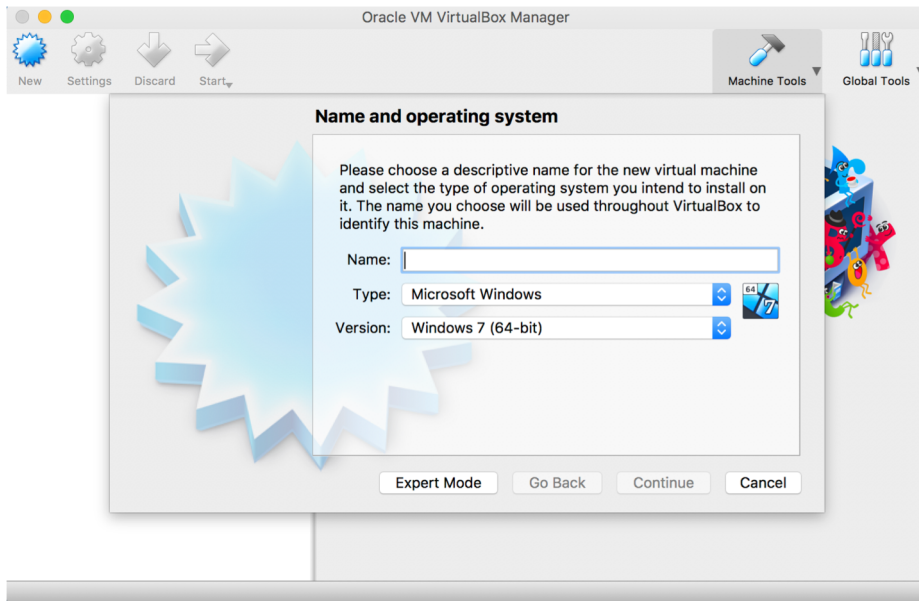
- Now we need a guest operating system: we'll use Ubuntu (Linux)
- <http://www.ubuntu.com/download/desktop>



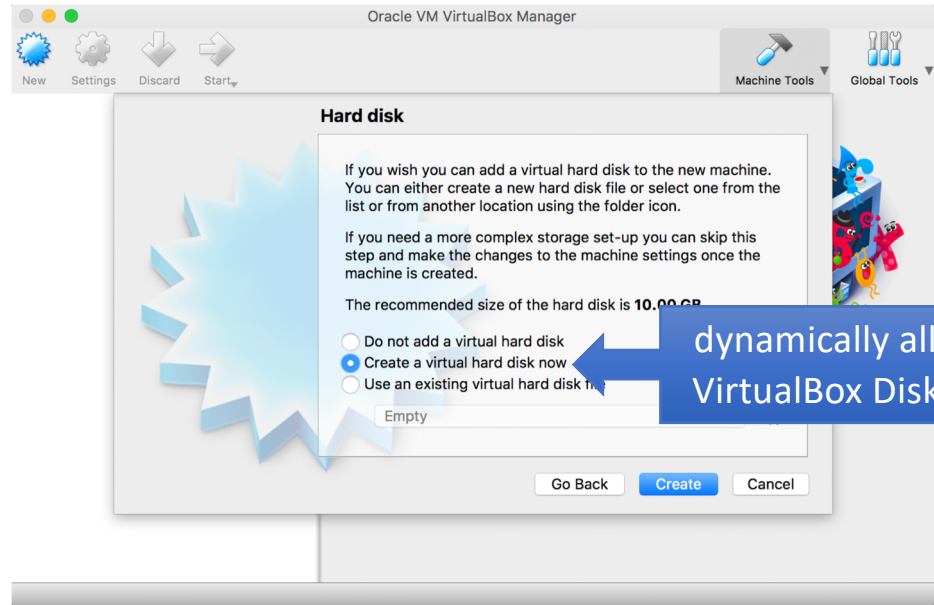
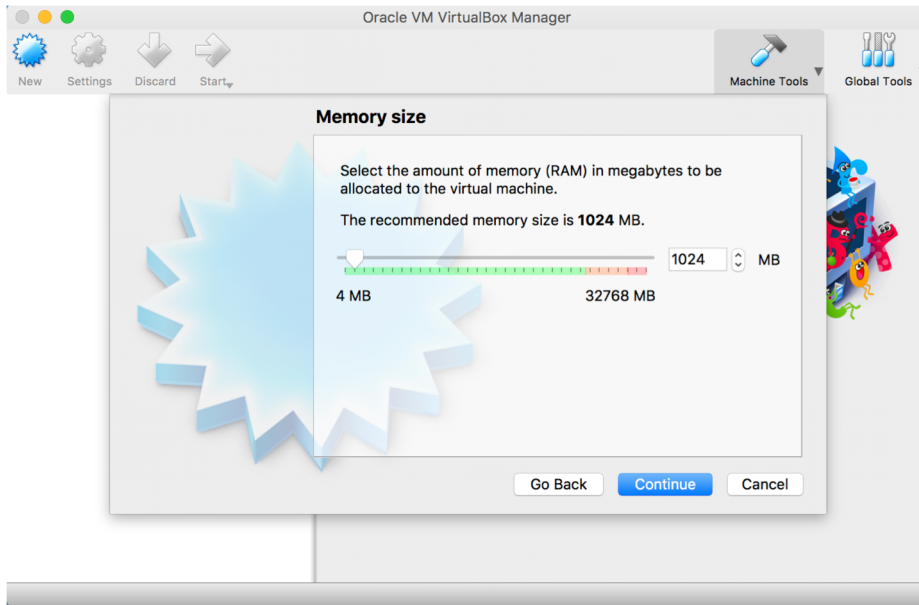
The .iso file takes considerable time to download
The .iso file is a special file that is a virtual cd
ISO images:

- Macs can mount ISO images.
- Macs can't boot off a multitrack ISO image
- some versions of Windows can't mount an ISO image (without extra software).
- Install Microsoft's Virtual CD-ROM Control Panel.

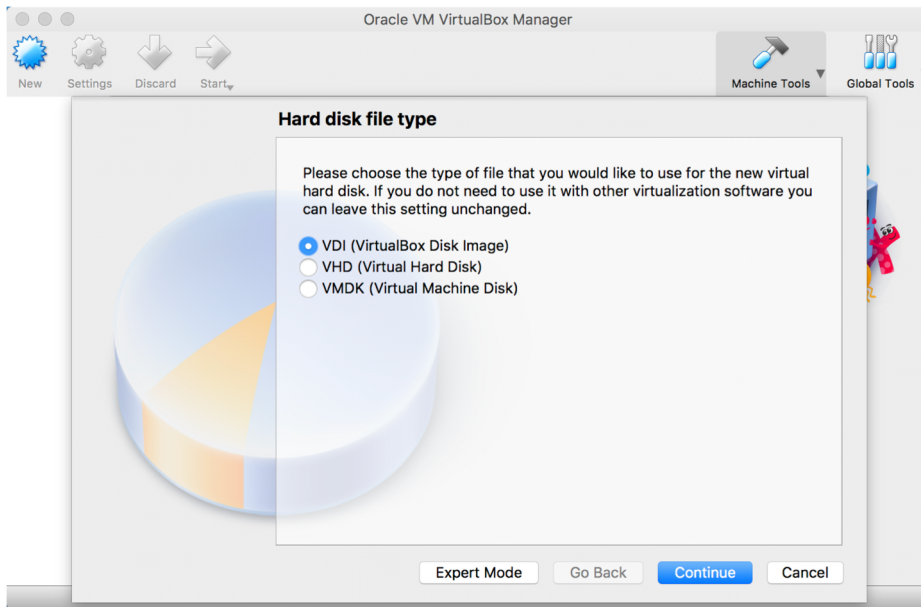
VirtualBox



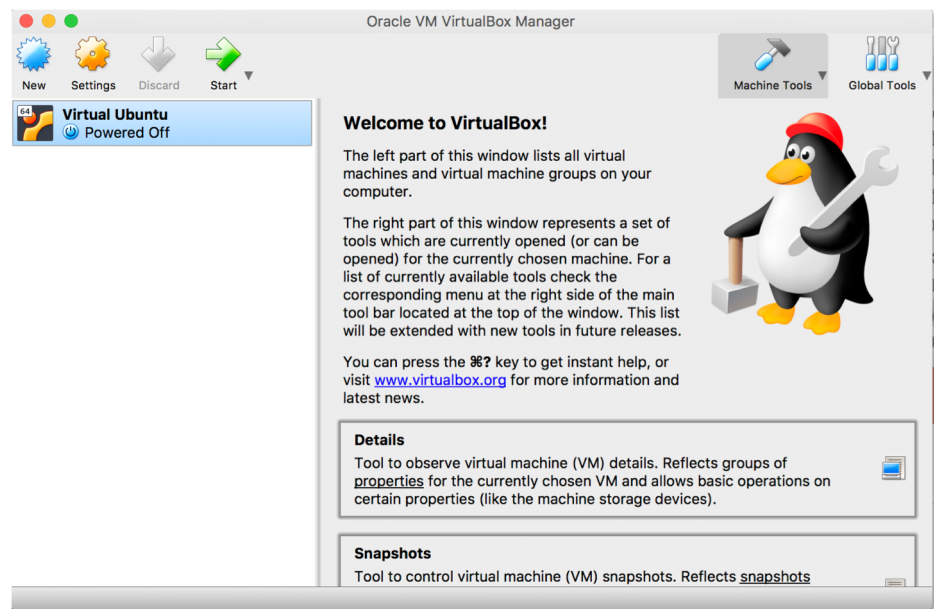
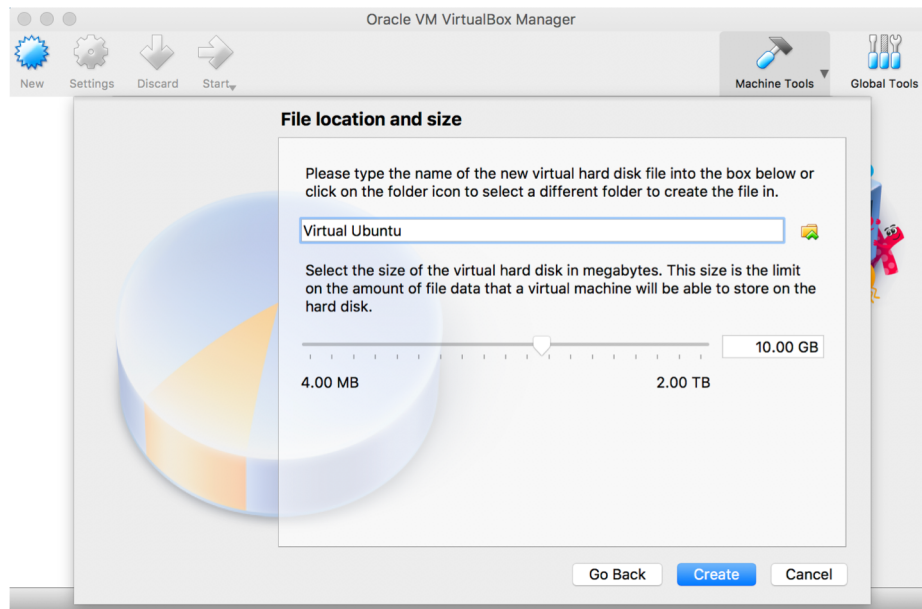
VirtualBox



VirtualBox

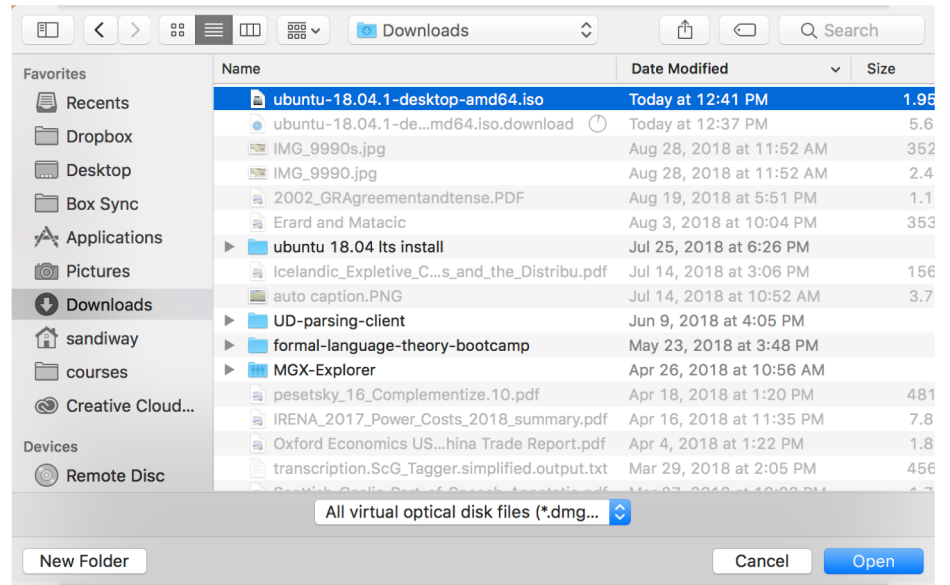
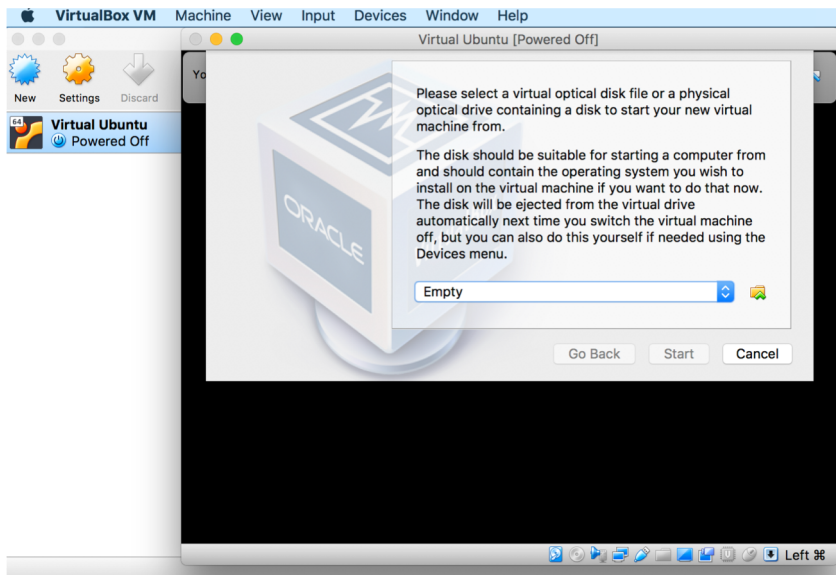


VirtualBox

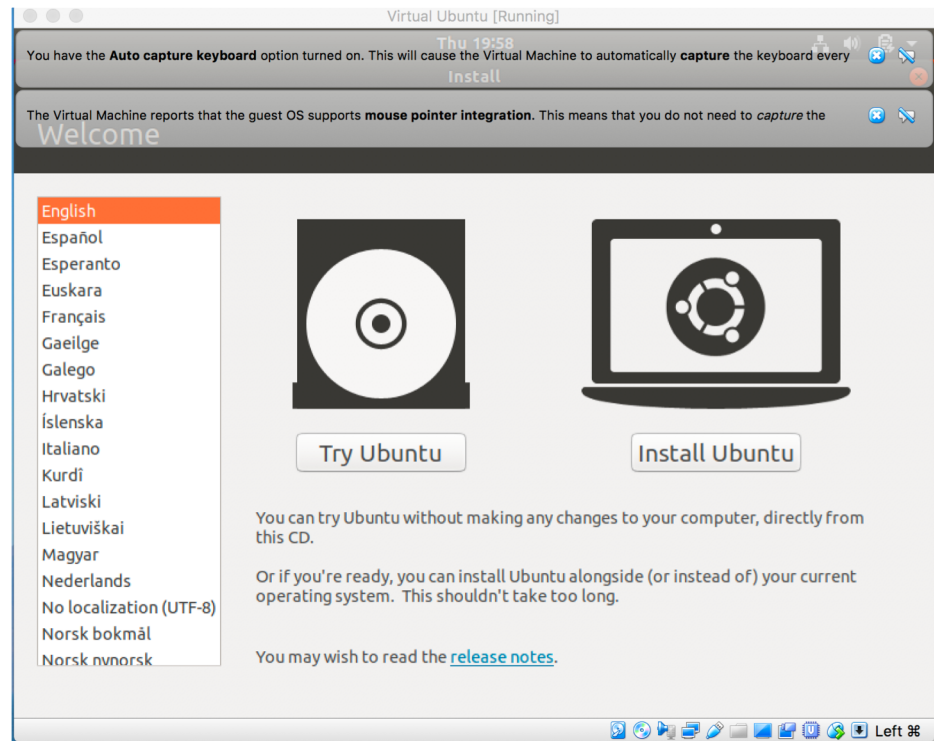
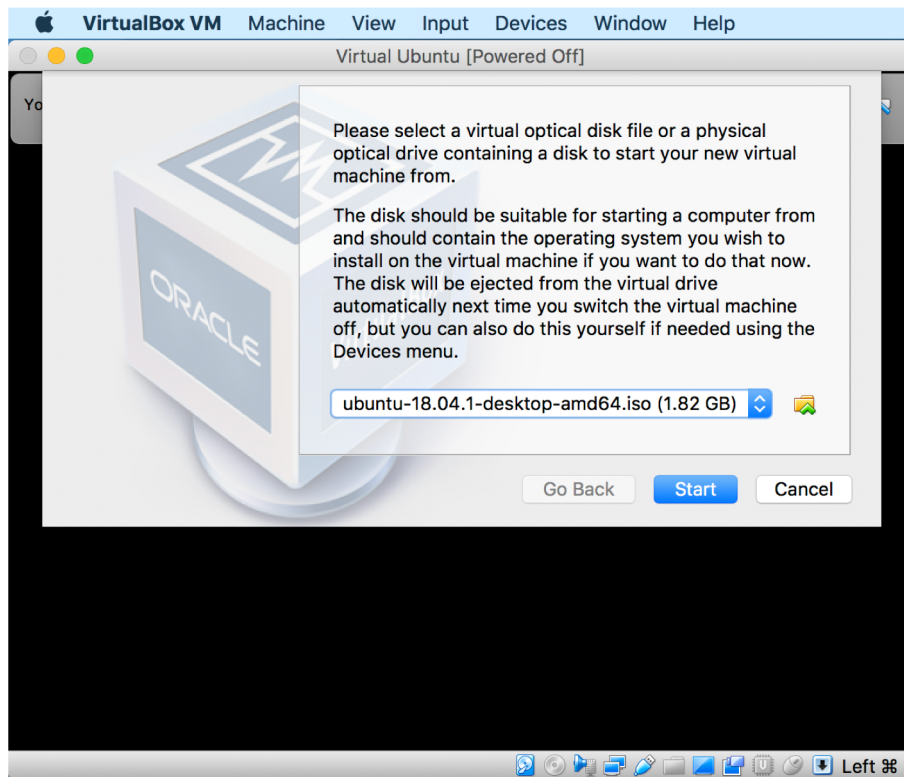


VirtualBox

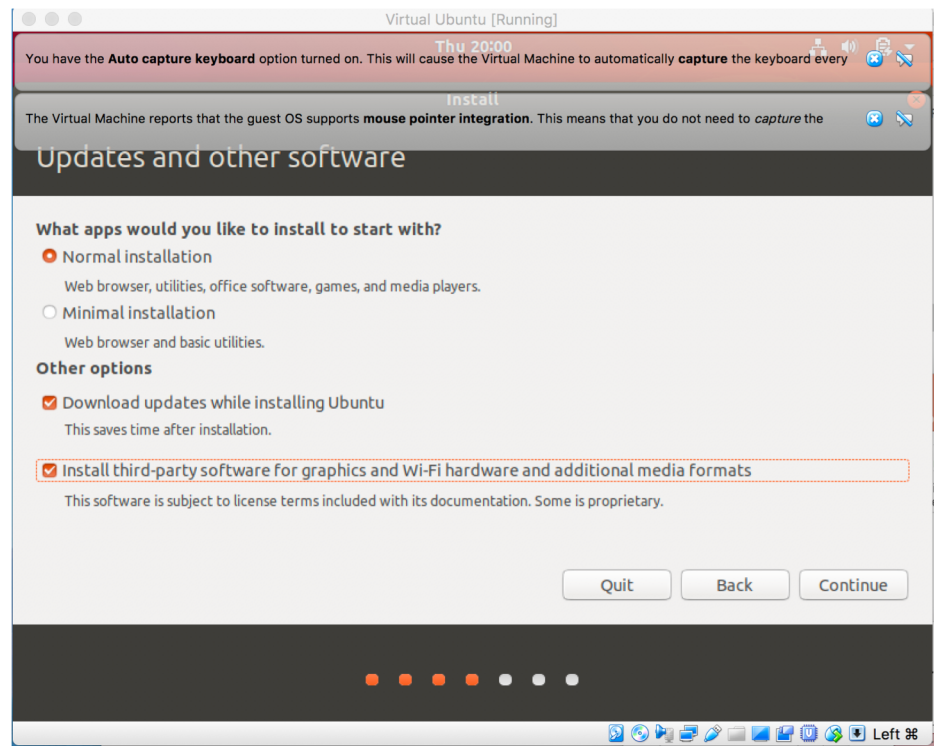
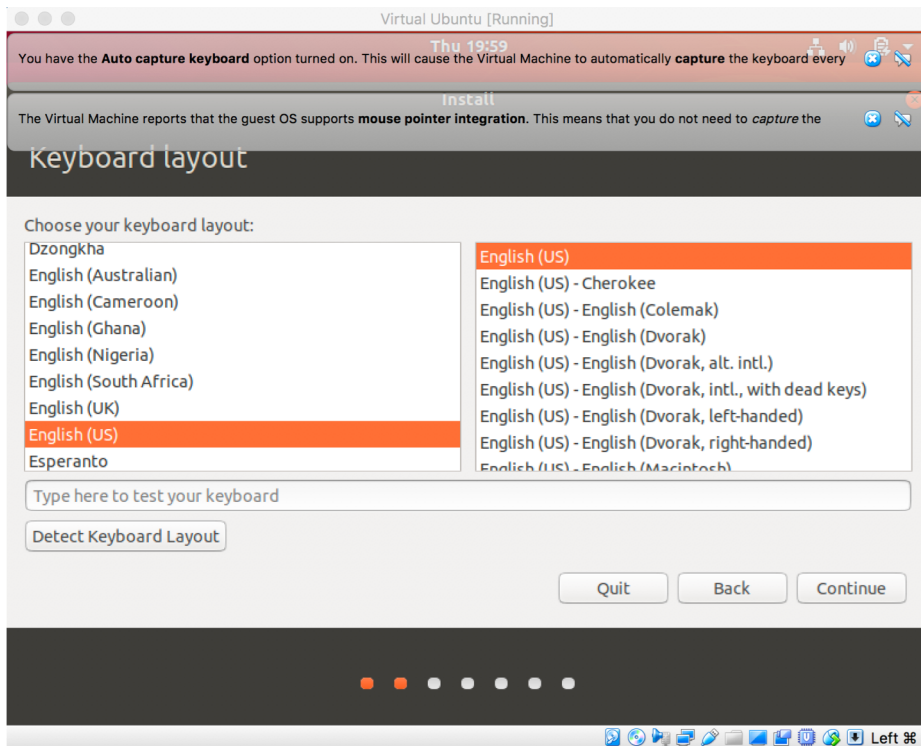
Start your virtual machine (double-click or Start)



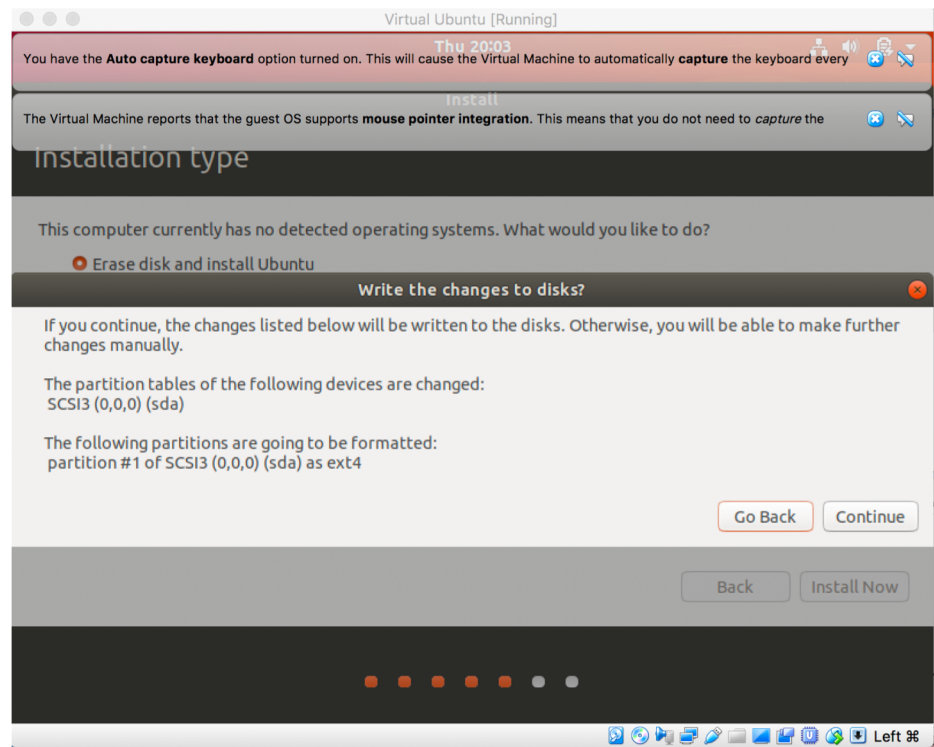
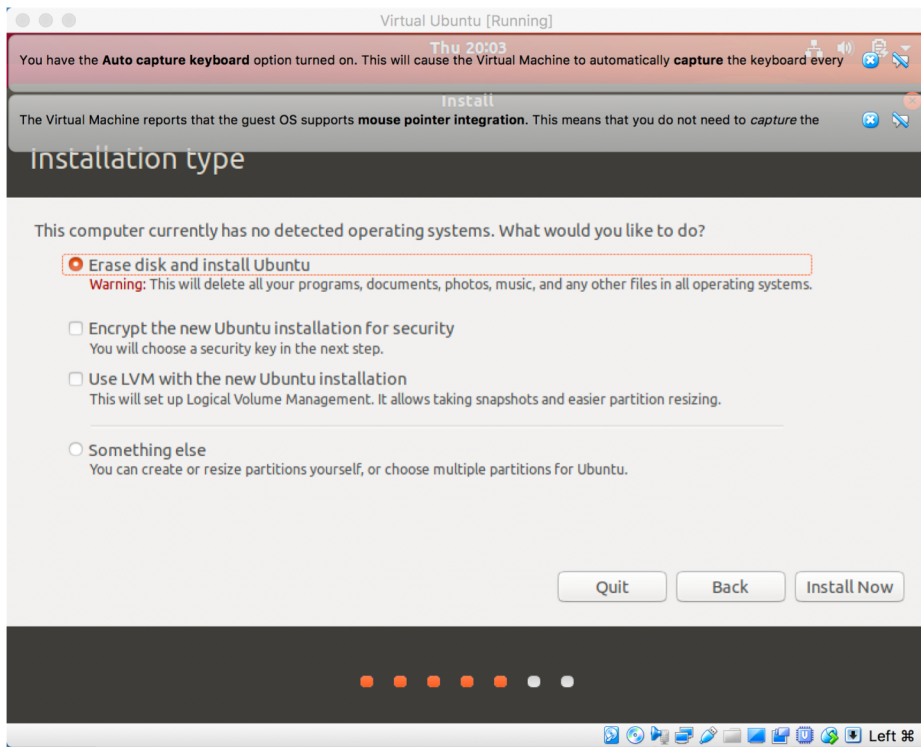
VirtualBox: installing Ubuntu



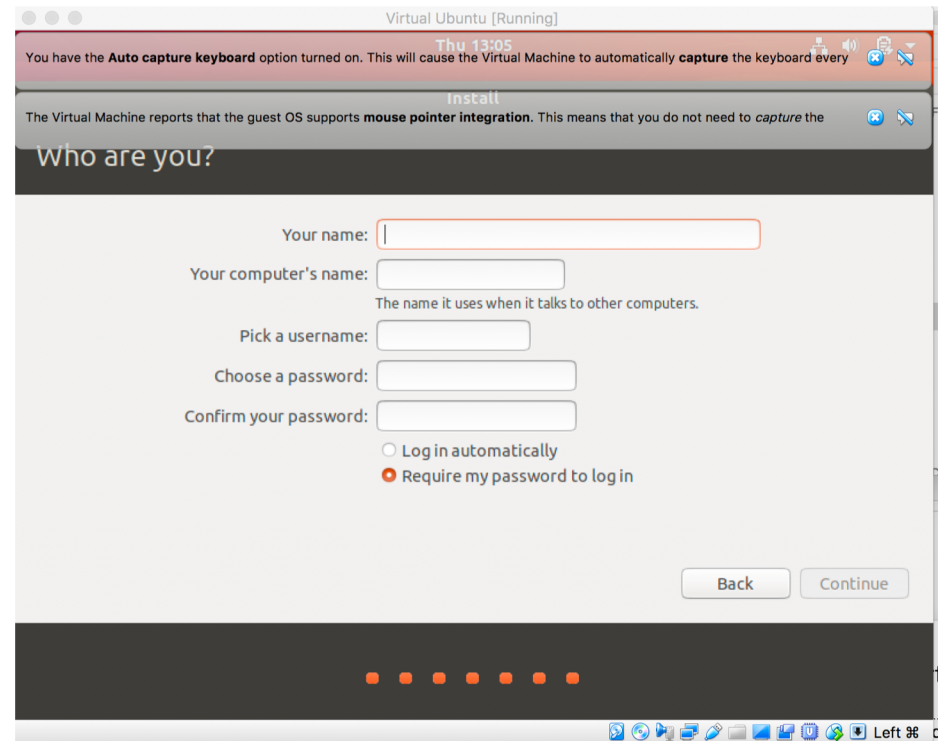
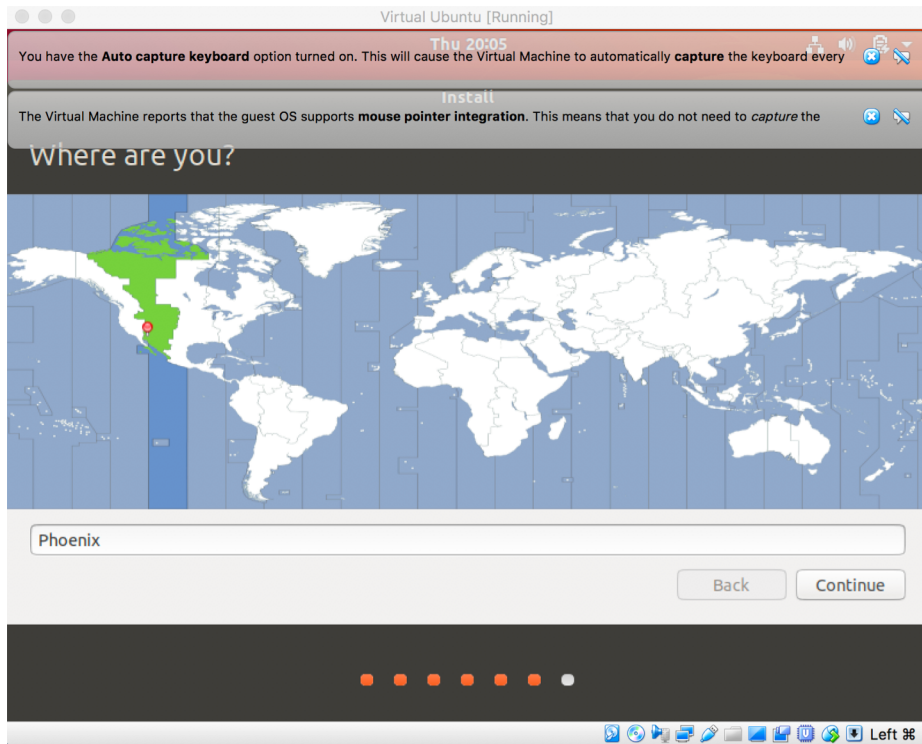
VirtualBox: installing Ubuntu



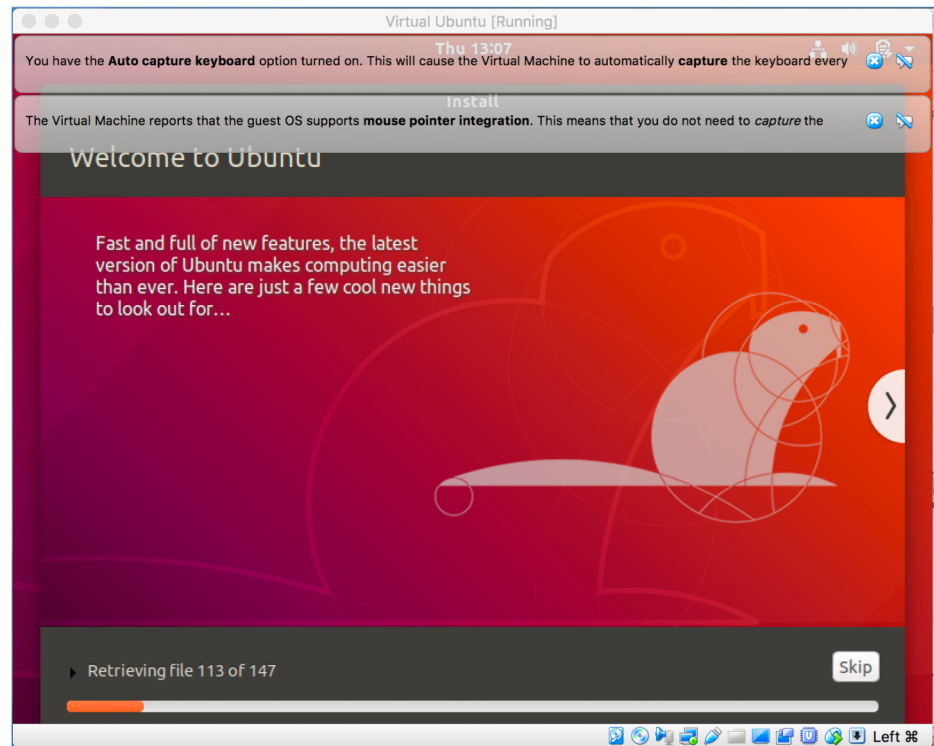
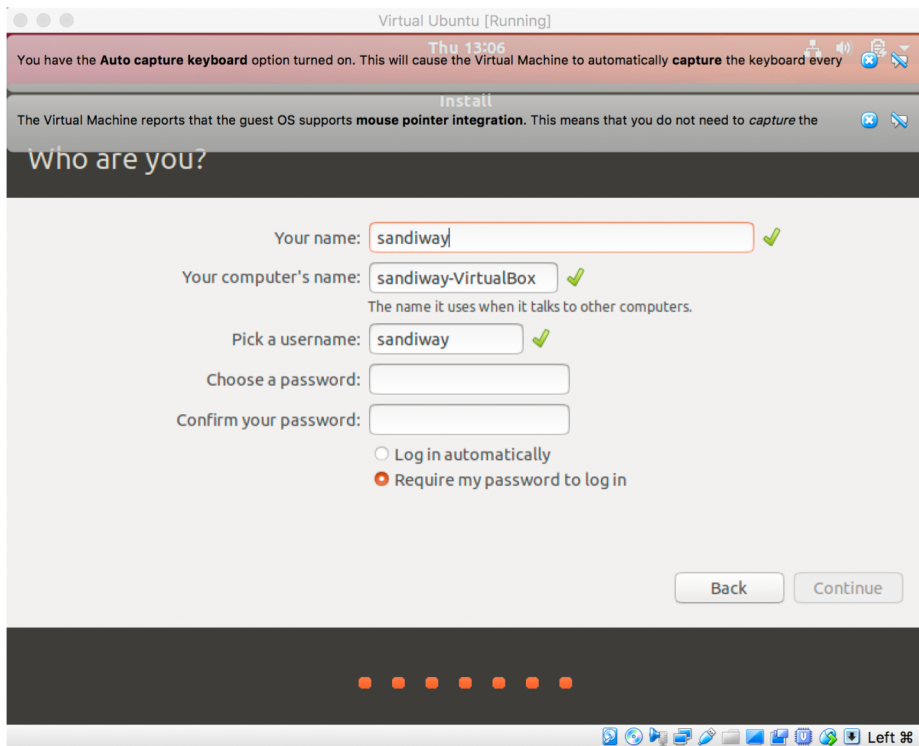
VirtualBox: installing Ubuntu



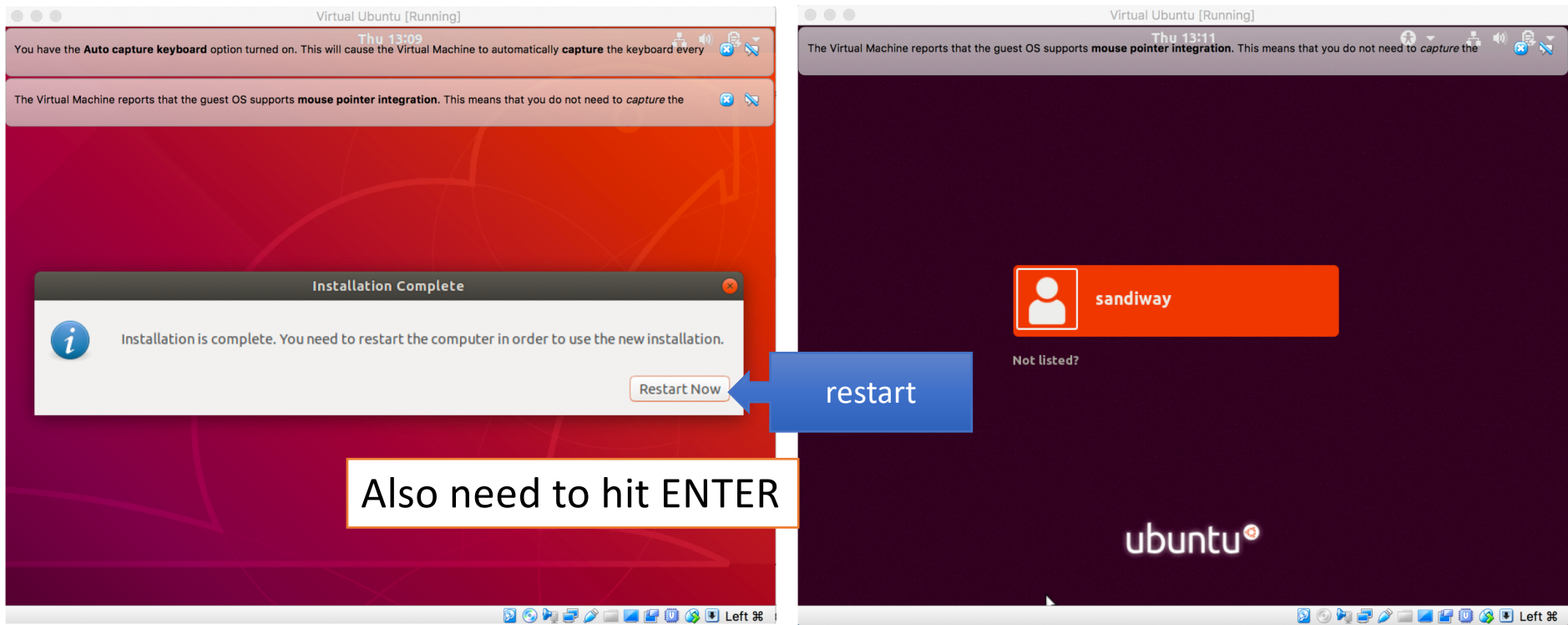
VirtualBox: installing Ubuntu



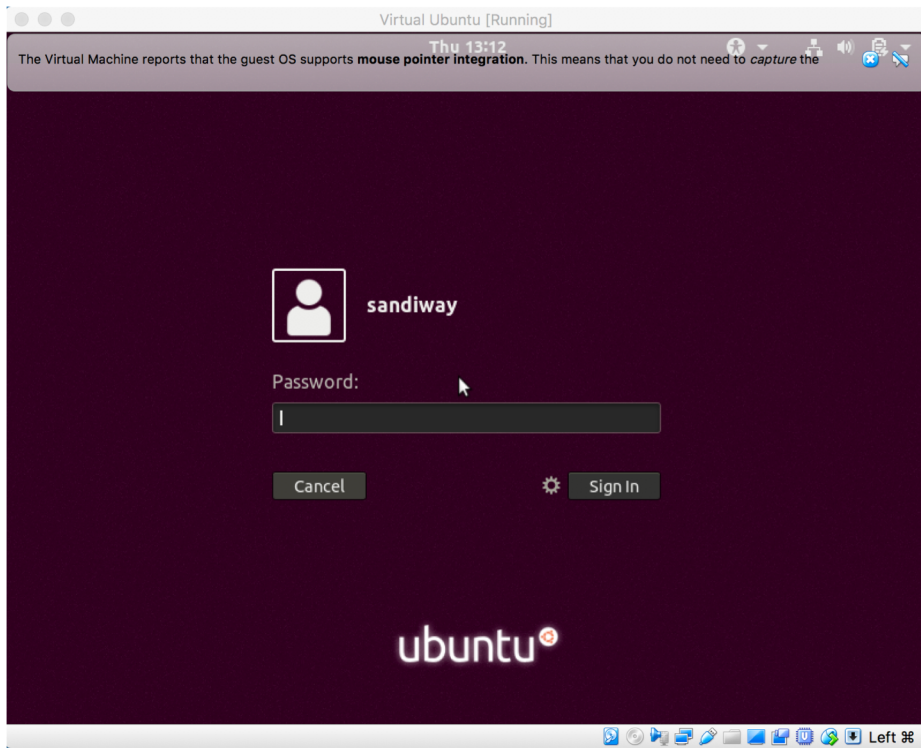
VirtualBox: installing Ubuntu



VirtualBox: installing Ubuntu

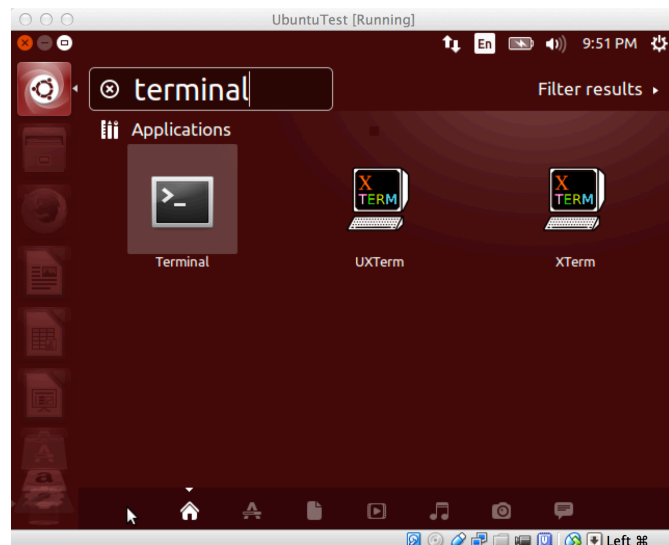


VirtualBox: running Ubuntu

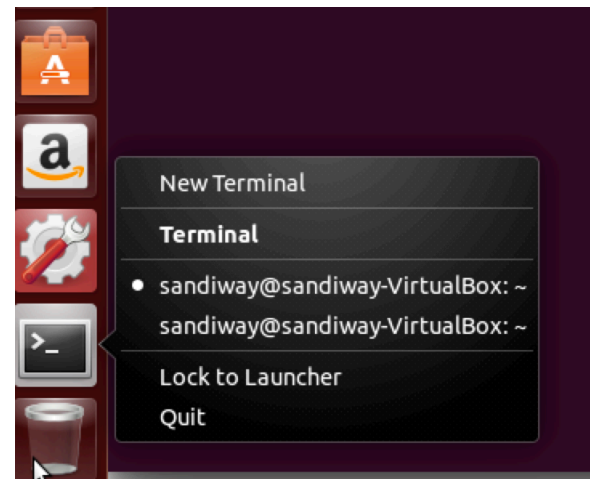


VirtualBox: running Ubuntu

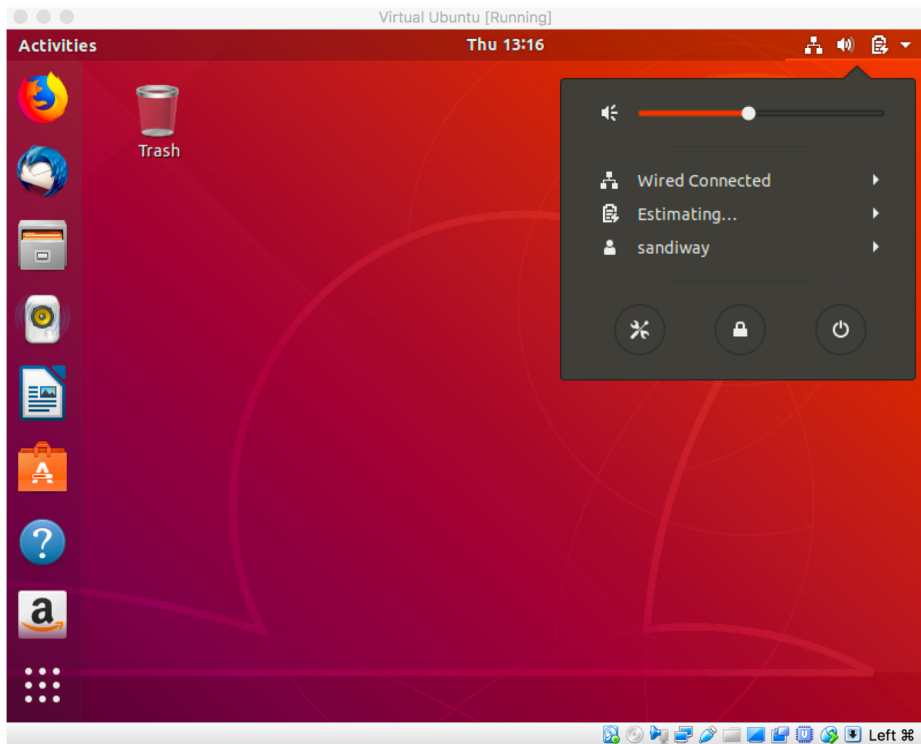
- Ubuntu Software Center
 - App store
 - (full screen to see Search box)
- Software packages
 - Terminal: **sudo** apt-get install <pkg-name>
 - **sudo** prefix: means execute the apt-get command with superuser privileges (typically needed for packages)
- How to find Terminal: use search



Lock to Launcher



VirtualBox: running Ubuntu



- Click right bottom icon to power off