Screenshots in virtualbox

As you get settled in, run the following lines in virtualbox terminal:

```
sudo apt-get update                    # to install system updates
sudo apt-get install gnome-screenshot  # to install screenshot utility
sudo reboot                            # to restart computer
```

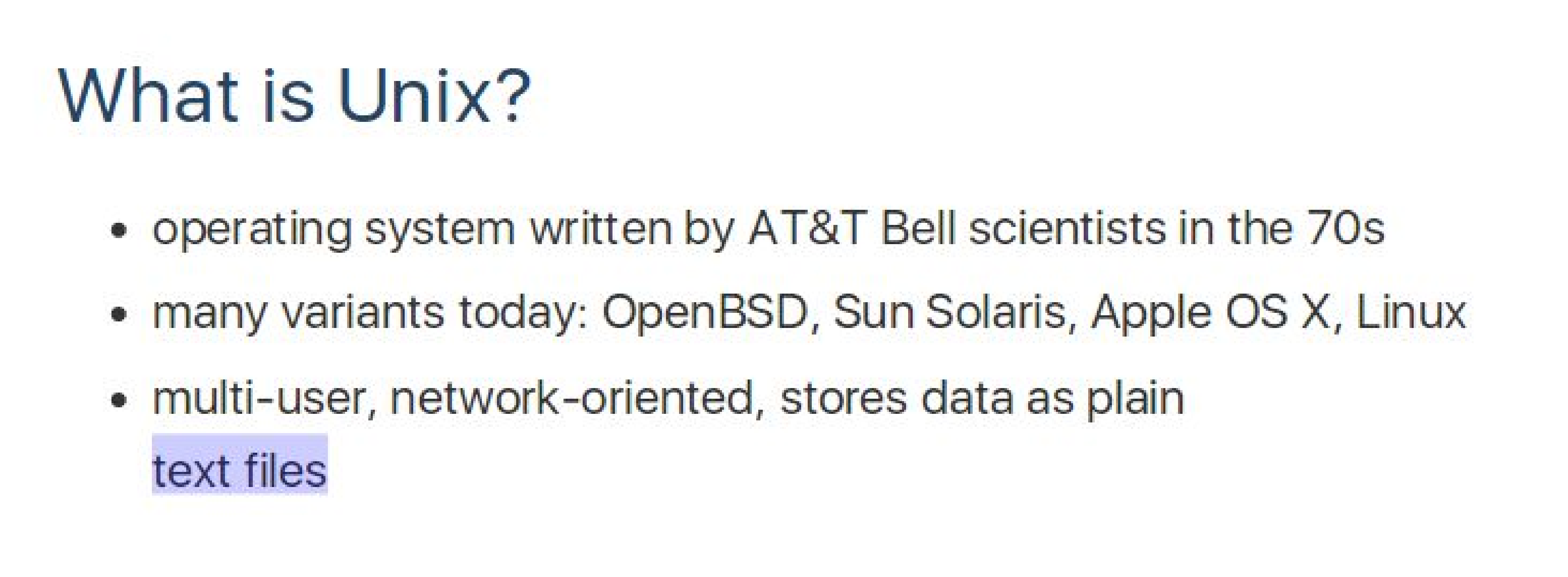
Then, pin screenshot utility to taskbar (Gaurav will demo this)

New class website: **https://eeb177-w17.github.io/**

Extracting information with **regular expressions**

18 January 2017

Text files are central to scientific computing



What is Unix?

- operating system written by AT&T Bell scientists in the 70s
- many variants today: OpenBSD, Sun Solaris, Apple OS X, Linux
- multi-user, network-oriented, stores data as plain text files

FASTQ format for sequence data

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Extract the 10 nucleotides that come before and after each instance of “ATAA”

KML format for geographical data

```
   <Placemark>
      <name>Entity references example</name>
      <description>
            &lt;h1&gt;Entity references are hard to type!&lt;/h1&gt;
            &lt;p&gt;&lt;font color="green"&gt;Text is
          &lt;i&gt;more readable&lt;/i&gt;
          and &lt;b&gt;easier to write&lt;/b&gt;
          when you can avoid using entity references.&lt;/font&gt;&lt;/p&gt;
      </description>
      <Point>
        <coordinates>102.594411,14.998518</coordinates>
      </Point>
    </Placemark>
```

Extract the name of each entity with longitude 102.XXX

Text files are central to scientific computing, so **Text editors** are central to scientific computing.

There’s a lot of text editors out there, and people have very strong opinions.

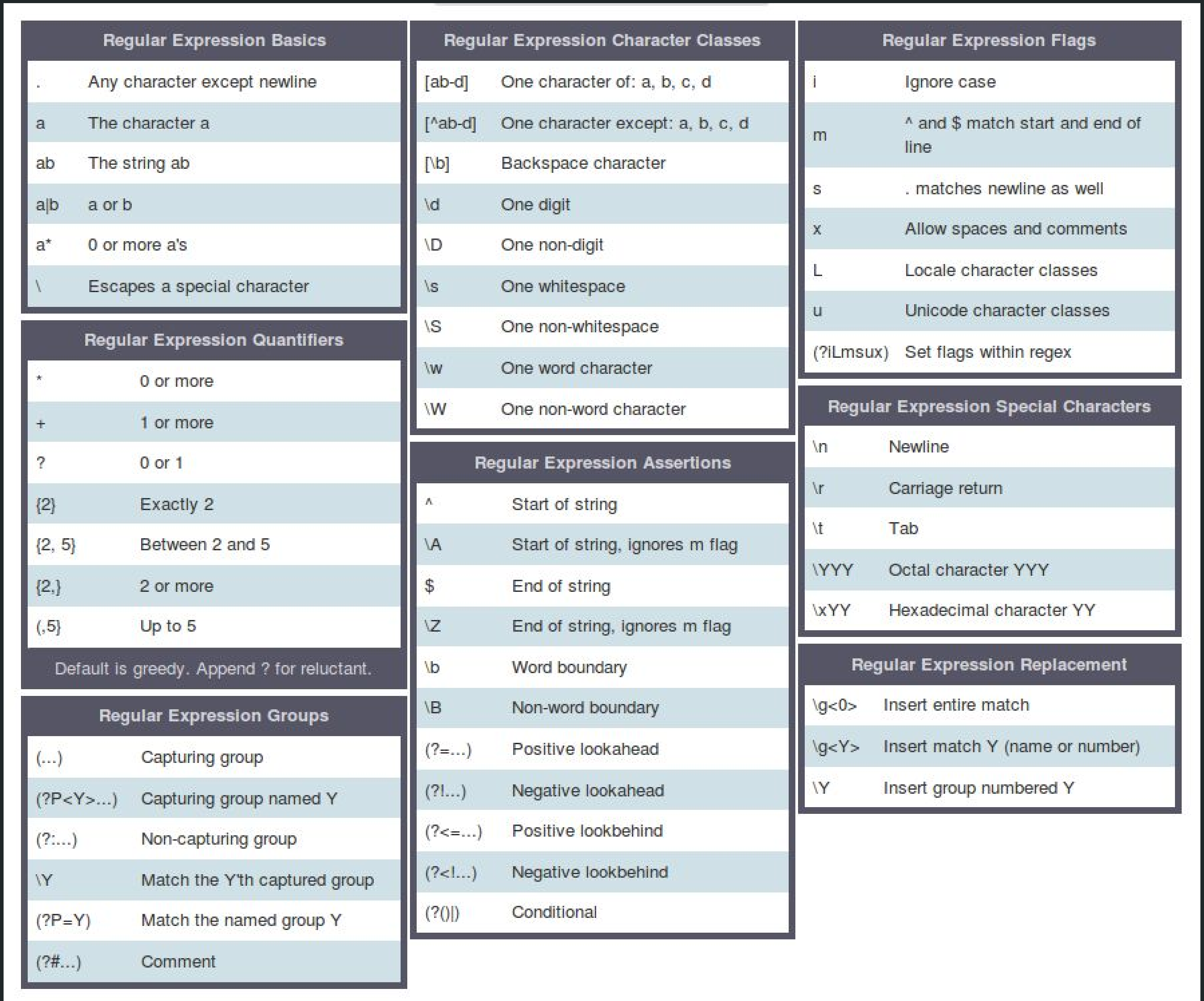
I think that gedit will work just fine for all of your needs in this course.

Mac users may be interested in using TextWrangler, and Windows users in Notepad++ (“plus plus”) for native text editors.

The terminal has its own text editors- nano and vim. You may encounter them, but I won’t introduce them formally here.

We use text editors (or Python scripts that do the job of text editors) to achieve the tasks outlined on the previous slide.

Regular expressions ("regexes") help us search for complicated patterns.



Regular Expression Basics

.	Any character except newline
a	The character a
ab	The string ab
a|b	a or b
a*	0 or more a's
\	Escapes a special character

Regular Expression Quantifiers

*	0 or more
+	1 or more
?	0 or 1
{2}	Exactly 2
{2, 5}	Between 2 and 5
{2,}	2 or more
(,5}	Up to 5

Default is greedy. Append ? for reluctant.

Regular Expression Groups

(...)	Capturing group
(?P<Y>...)	Capturing group named Y
(?:...)	Non-capturing group
\Y	Match the Y'th captured group
(?P=Y)	Match the named group Y
(?#...)	Comment

Regular Expression Character Classes

[ab-d]	One character of: a, b, c, d
[^ab-d]	One character except: a, b, c, d
[\b]	Backspace character
\d	One digit
\D	One non-digit
\s	One whitespace
\S	One non-whitespace
\w	One word character
\W	One non-word character

Regular Expression Assertions

^	Start of string
\A	Start of string, ignores m flag
$	End of string
\Z	End of string, ignores m flag
\b	Word boundary
\B	Non-word boundary
(?=...)	Positive lookahead
(?!...)	Negative lookahead
(?<=...)	Positive lookbehind
(?<!...)	Negative lookbehind
(?()|)	Conditional

Regular Expression Flags

i	Ignore case
m	^ and $ match start and end of line
s	. matches newline as well
x	Allow spaces and comments
L	Locale character classes
u	Unicode character classes
(?iLmsux)	Set flags within regex

Regular Expression Special Characters

\n	Newline
\r	Carriage return
\t	Tab
\YYY	Octal character YYY
\xYY	Hexadecimal character YY

Regular Expression Replacement

\g<0>	Insert entire match
\g<Y>	Insert match Y (name or number)
\Y	Insert group numbered Y

```
denovo105;   Orthopteroidea; 100; Orthoptera; 100; Ensifera;
100; Grylloidea; 100; Gryllidae; 100; Gryllinae; 100;
Gryllodes; 50; Gryllodes sigillatus; 50;

denovo105;   Orthopteroidea; 100; Orthoptera; 100; Ensifera;
100; Grylloidea; 100; Gryllidae; 100; Gryllinae; 100;
Gryllodes; 50; Gryllodes sigillatus; 50;

denovo105; ; Not assigned; 100;

denovo105;   Endopterygota; 100; Coleoptera; 100; Polyphaga
<Coleoptera>; 100; Scarabaeiformia; 58; Scarabaeoidea; 58;
Scarabaeidae; 58; Melolonthinae; 58; Phyllophaga; 32;
Phyllophaga densata; 5;

denovo105;   Endopterygota; 100; Coleoptera; 100; Polyphaga
<Coleoptera>; 100; Scarabaeiformia; 37; Scarabaeoidea; 37;
Scarabaeidae; 37; Melolonthinae; 37; Astenopholis; 11;
Astenopholis sp. DA-2012; 11;

..... (thousands of lines)
```

denovo105; Orthopteroidea; 100; Orthoptera; 100; Ensifera; 100; Grylloidea; 100; Gryllidae; 100; Gryllinae; 100~~; Gryllodes; 50; Gryllodes sigillatus; 50;~~

denovo105; Orthopteroidea; 100; Orthoptera; 100; Ensifera; 100; Grylloidea; 100; Gryllidae; 100; Gryllinae; 100~~; Gryllodes; 50; Gryllodes sigillatus; 50;~~

denovo105; ; Not assigned; 100~~;~~

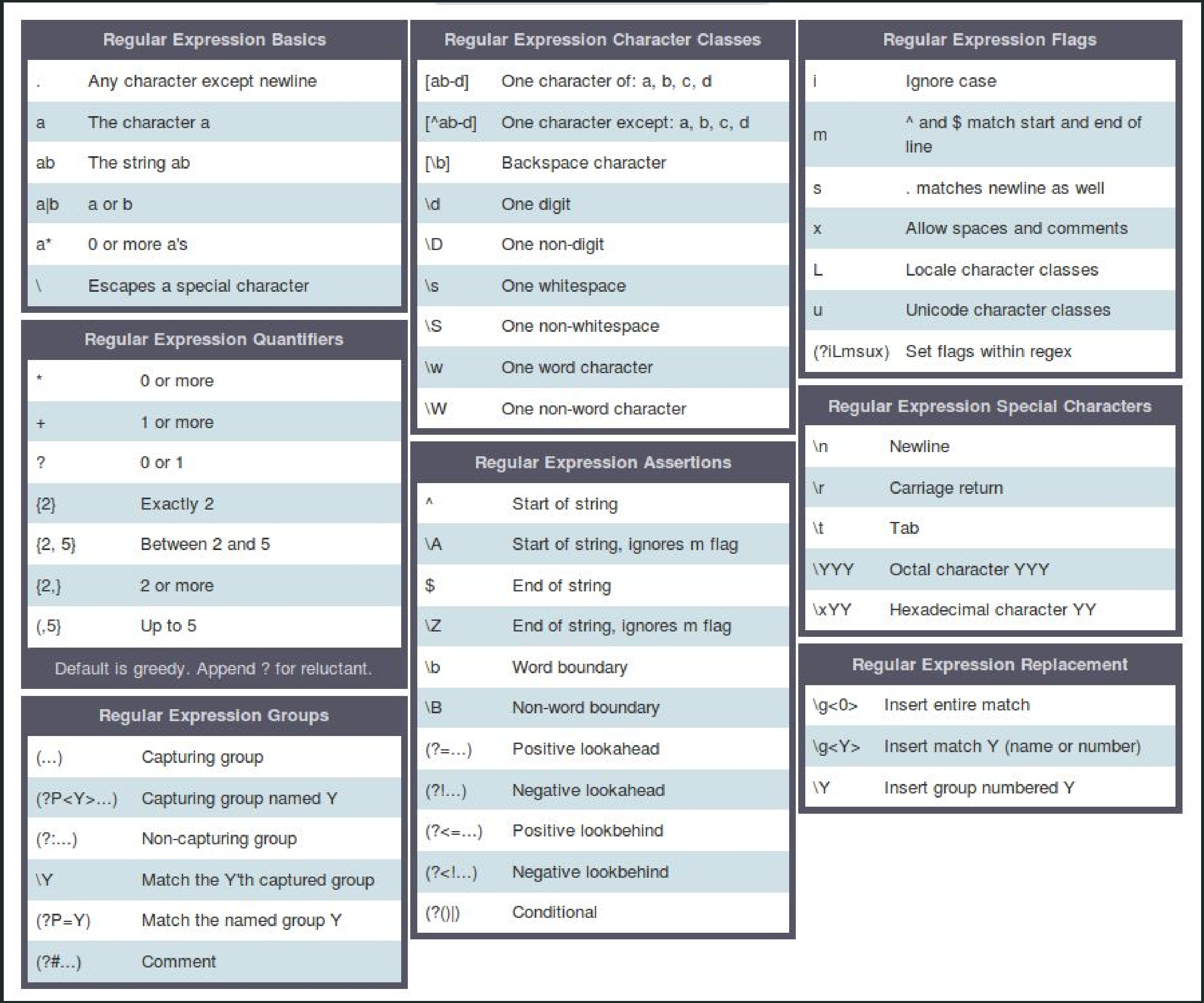
denovo105; Endopterygota; 100; Coleoptera; 100; Polyphaga <Coleoptera>; 100~~; Scarabaeiformia; 58; Scarabaeoidea; 58; Scarabaeidae; 58; Melolonthinae; 58; Phyllophaga; 32; Phyllophaga densata; 5;~~

denovo105; Endopterygota; 100; Coleoptera; 100; Polyphaga <Coleoptera>; 100~~; Scarabaeiformia; 37; Scarabaeoidea; 37; Scarabaeidae; 37; Melolonthinae; 37; Astenopholis; 11; Astenopholis sp. DA-2012; 11;~~

..... (thousands of lines)

```
denovo105;   Orthopteroidea; 100; Orthoptera; 100; Ensifera;
100; Grylloidea; 100; Gryllidae; 100; Gryllinae; 100;
Gryllodes; 50; Gryllodes sigillatus; 50;

denovo105;   Orthopteroidea; 100; Orthoptera; 100; Ensifera;
100; Grylloidea; 100; Gryllidae; 100; Gryllinae; 100;
Gryllodes; 50; Gryllodes sigillatus; 50;

denovo105; ; Not assigned; 100;

denovo105;   Endopterygota; 100; Coleoptera; 100; Polyphaga
<Coleoptera>; 100; Scarabaeiformia; 58; Scarabaeoidea; 58;
Scarabaeidae; 58; Melolonthinae; 58; Phyllophaga; 32;
Phyllophaga densata; 5;

denovo105;   Endopterygota; 100; Coleoptera; 100; Polyphaga
<Coleoptera>; 100; Scarabaeiformia; 37; Scarabaeoidea; 37;
Scarabaeidae; 37; Melolonthinae; 37; Astenopholis; 11;
Astenopholis sp. DA-2012; 11;

..... (thousands of lines)
```

Search for almost any pattern with regular expressions.



Regular Expression Basics

.	Any character except newline
a	The character a
ab	The string ab
a|b	a or b
a*	0 or more a's
\	Escapes a special character

Regular Expression Quantifiers

*	0 or more
+	1 or more
?	0 or 1
{2}	Exactly 2
{2, 5}	Between 2 and 5
{2,}	2 or more
{,5}	Up to 5

Default is greedy. Append ? for reluctant.

Regular Expression Groups

(...)	Capturing group
(?P<Y>...)	Capturing group named Y
(?:...)	Non-capturing group
\Y	Match the Y'th captured group
(?P=Y)	Match the named group Y
(?#...)	Comment

Regular Expression Character Classes

[ab-d]	One character of: a, b, c, d
[^ab-d]	One character except: a, b, c, d
[\b]	Backspace character
\d	One digit
\D	One non-digit
\s	One whitespace
\S	One non-whitespace
\w	One word character
\W	One non-word character

Regular Expression Assertions

^	Start of string
\A	Start of string, ignores m flag
$	End of string
\Z	End of string, ignores m flag
\b	Word boundary
\B	Non-word boundary
(?=...)	Positive lookahead
(?!...)	Negative lookahead
(?<=...)	Positive lookbehind
(?<!...)	Negative lookbehind
(?()|)	Conditional

Regular Expression Flags

i	Ignore case
m	^ and $ match start and end of line
s	. matches newline as well
x	Allow spaces and comments
L	Locale character classes
u	Unicode character classes
(?iLmsux)	Set flags within regex

Regular Expression Special Characters

\n	Newline
\r	Carriage return
\t	Tab
\YYY	Octal character YYY
\xYY	Hexadecimal character YY

Regular Expression Replacement

\g<0>	Insert entire match
\g<Y>	Insert match Y (name or number)
\Y	Insert group numbered Y

demonstration of regular expressions in gedit

today’s exercise and this week’s homework

Complete an online regular expressions tutorial, and commit screenshots of your progress to a git repository.

> Note: This is a very thorough tutorial. Take your time with it.

This week’s homework (due one week from today) asks you to write regular expressions that search for certain patterns in a paragraph of text and in a data table. You are to download a text file, add your answers to that text file, and commit the text file to a git repository.