Unix - I

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

1/24/2017

Review

- **Flat file**

Two types of line feed (LF, \n) and carriage return (CR, \r)

- **Excel functions (average, vlookup, ...)**

- **Regular expression**

e.g., 1) T{10,12} 2) ^$

- ***vi* has two modes:**

  1. insert mode
  2. command mode

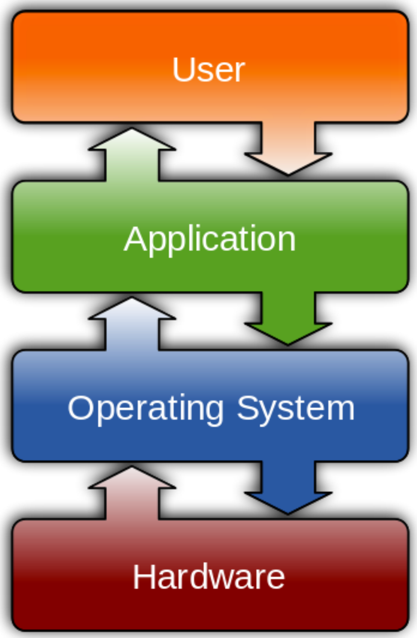
Today

- What is Unix?
- Why do we need to learn Unix?
- Useful commands

Unix is one of Operating Systems (OSs)

www.cj-computers.com

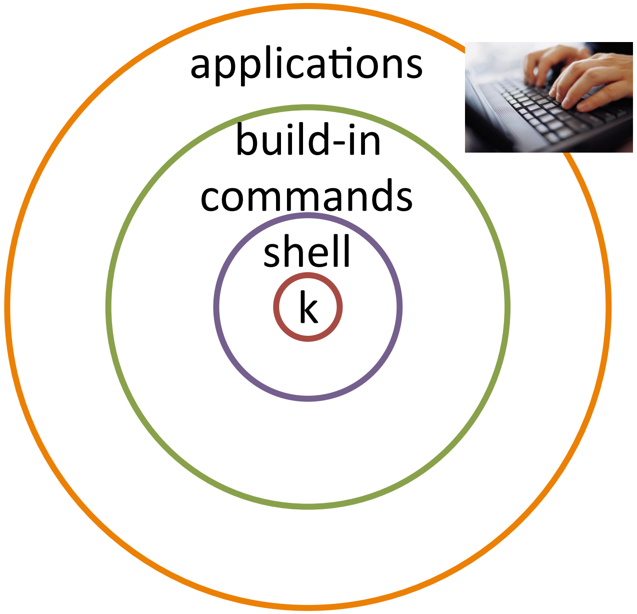
OS, Linux, Window



wikipedia.com

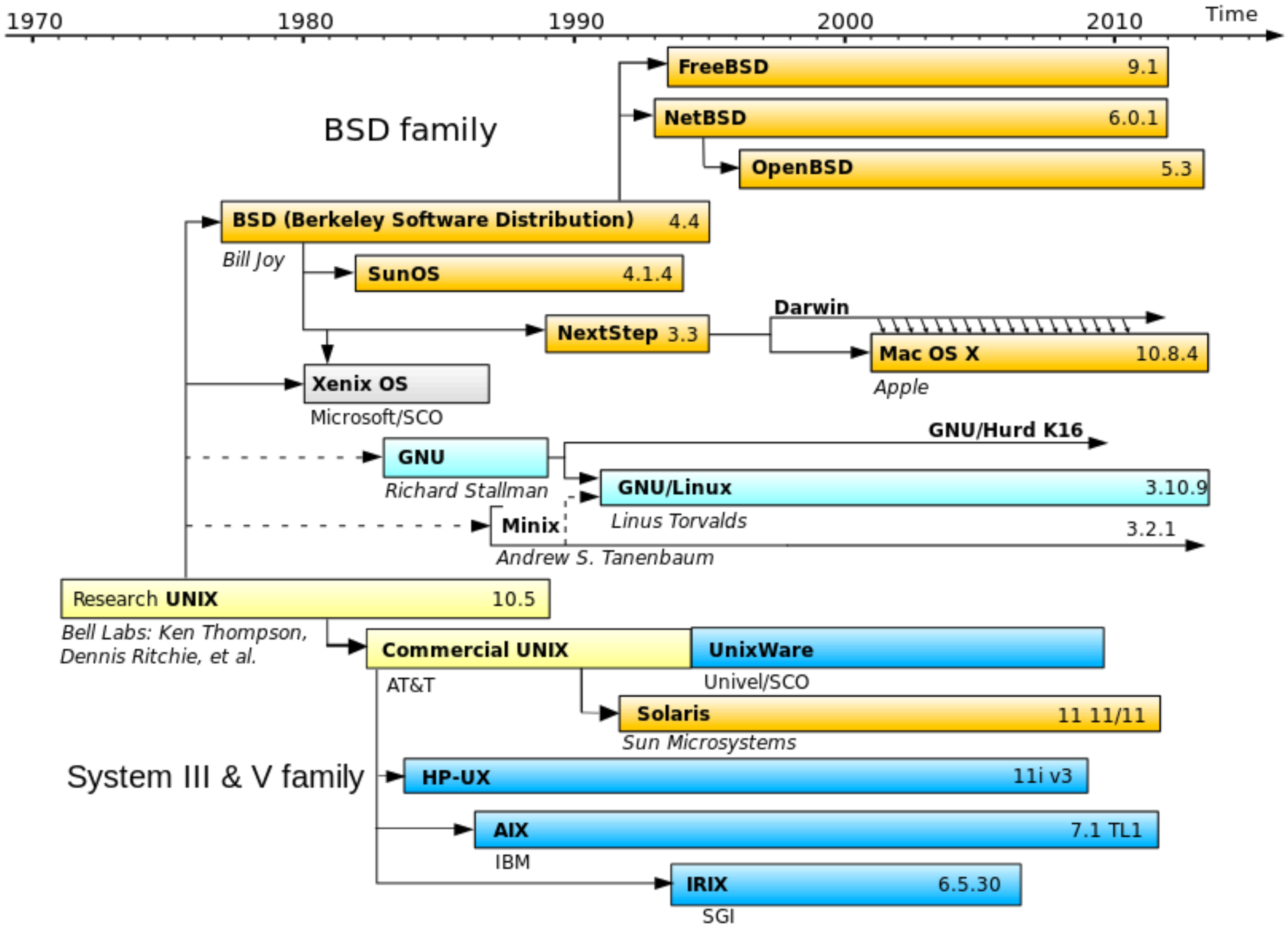
- Control Hardware
- Run Applications
- Manage Data

Parts of Unix OS



- **kernel** (k) - provides the basic software connection to the hardware, managing memory, schedules, and input/output.
- **shell** - as an interpreter to translate commands and pass them to the kernel for execution.
- **build-in commands** - are the built-in system utilities that provide users basic functions, such as content listing (ls), file copying (cp).
- **applications** - are additional application programs.

Evolution of UNIX-based Operating Systems



en.wikipedia.org/wiki/Unix-like

Linux Distributions

cpagnews.com/

Liu lab
DISTRIB_ID=Ubuntu
DISTRIB_RELEASE=14.04

Beocat
Gentoo Base System release 2.2
NAME=Gentoo
ID=gentoo
PRETTY_NAME="Gentoo/Linux”

Change the following three file names to file names ended up with .fasta
a.txt b.txt c.txt

Ubuntu:

```
rename 's/txt/fasta/' *txt
```

gentoo:

```
rename 'txt' 'fasta' *txt
```

Why do we need to learn Unix?

- To perform **efficient** and **reproducible** data analyses
- To use advanced tools in research projects (most genomic software packages are run in the Unix system)
- To access to powerful computer servers

The terminal emulator

A terminal emulator allows users to access to a computer or server.

Mac OS X:

Terminal

iterm2

Linux:

Linux console

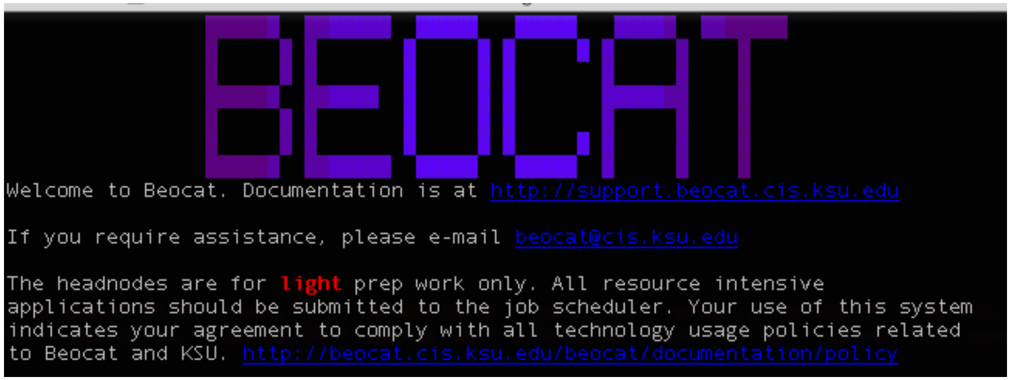
Microsoft Windows:

PuTTY

AbsoluteTelnet

Mintty

xterm



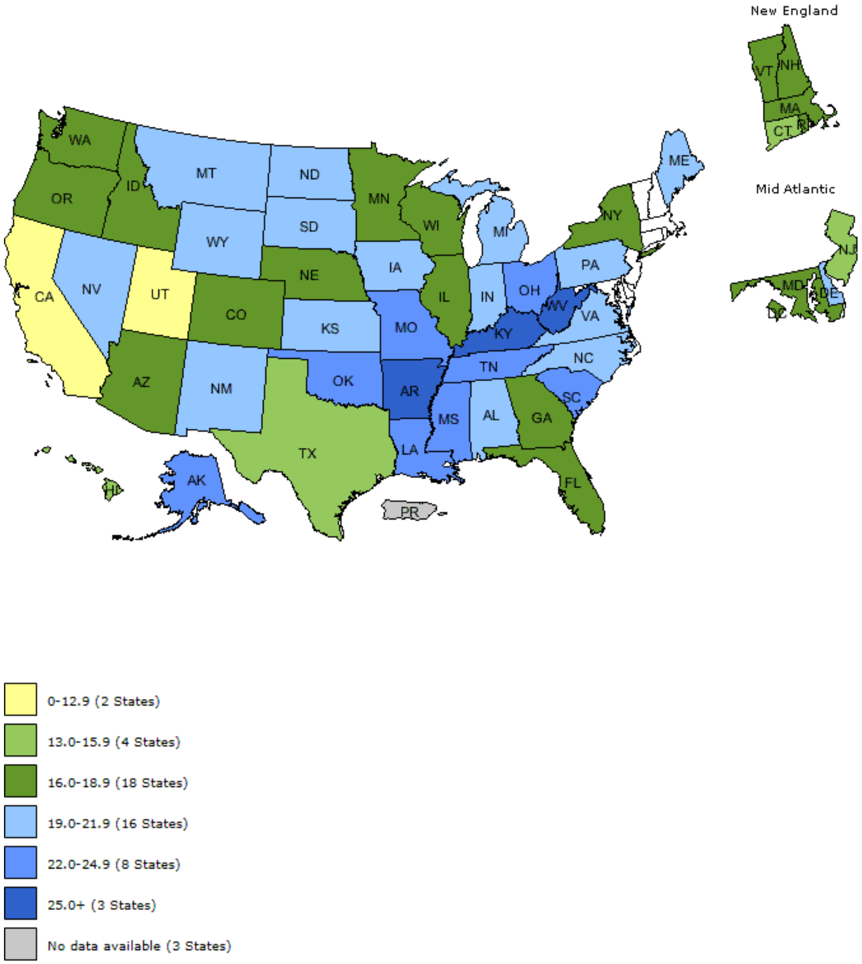
```
Welcome to Beocat. Documentation is at http://support.beocat.cis.ksu.edu

If you require assistance, please e-mail beocat@cis.ksu.edu

The headnodes are for light prep work only. All resource intensive
applications should be submitted to the job scheduler. Your use of this system
indicates your agreement to comply with all technology usage policies related
to Beocat and KSU. http://beocat.cis.ksu.edu/beocat/documentation/policy
```

Imaging ...

If you are working on data using an OS platform, what components and what basic operations are needed?

Example datasets

Cigarette Use (Adults) - BRFSS – 2013



Source: Behavioral Risk Factor Surveillance System (BRFSS)

State	Adult Cigarette Use (%)
Alabama	21.5
Alaska	22.6
…	…

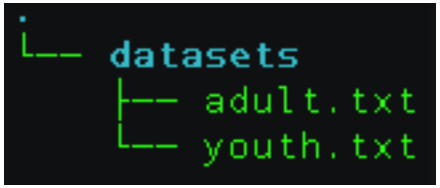
adult.txt

State	Youth Cigarette Use (%)
Alabama	NA
Alaska	10.6
Arizona	14.1
Arkansas	19.1
California	NA
Colorado	NA
Connecticut	13.5
Delaware	14.2
District of Columbia	NA
…	…

youth.txt

Directories and files

Under the directory:
/homes/liu3zhen/teaching/BA17



```
.
└── datasets
    ├── adult.txt
    └── youth.txt
```

Absolute path

- /homes/liu3zhen/teaching/BA17

Relative path

- . (current directory)
- .. (parental directory)
- ~ (home directory, e.g., /home/liu3zhen)

cd, mkdir, pwd

Directory: /homes/liu3zhen/teaching/BA17/datasets

- **cd** - change the working directory

```
% cd /homes/liu3zhen/teaching/
% cd ..
% cd ~
% cd ~/teaching/BA17/datasets/
```

- **mkdir** - make directories

```
% mkdir xxx
```

- **pwd** - print name of current working directory

```
% pwd
```

ls

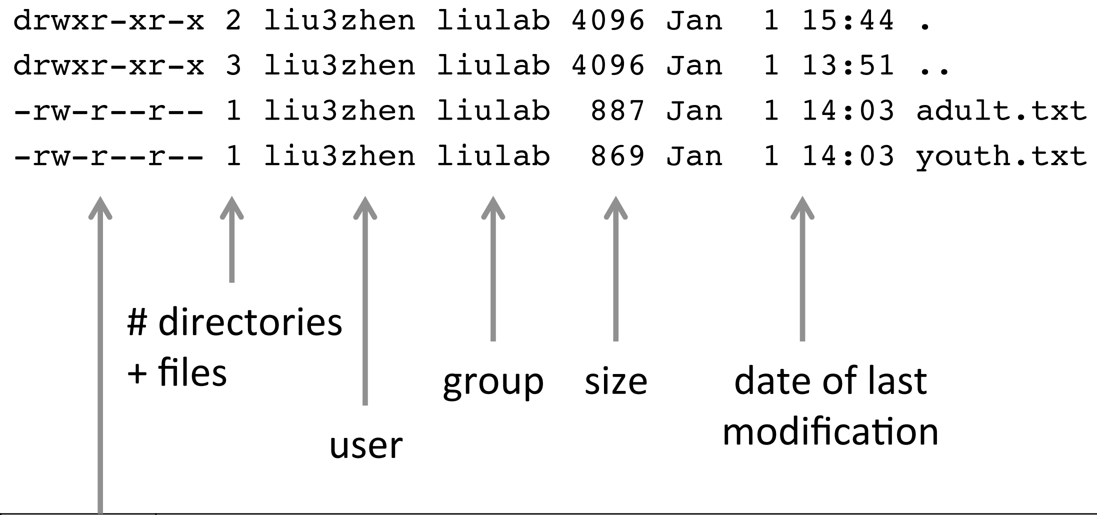
- **ls** – list directory contents

```
% ls
Adult.txt youth.txt
```

```
% ls -1
Adult.txt
Youth.txt
```

```
% ls -la
# -la = -l & -a, long format and list all files
Total 4
drwxr-xr-x 2 liu3zhen liulab 4096 Jan  1 15:44 .
drwxr-xr-x 3 liu3zhen liulab 4096 Jan  1 13:51 ..
-rw-r--r-- 1 liu3zhen liulab  887 Jan  1 14:03 adult.txt
-rw-r--r-- 1 liu3zhen liulab  869 Jan  1 14:03 youth.txt
```

File information



```
drwxr-xr-x 2 liu3zhen liulab 4096 Jan  1 15:44 .
drwxr-xr-x 3 liu3zhen liulab 4096 Jan  1 13:51 ..
-rw-r--r-- 1 liu3zhen liulab  887 Jan  1 14:03 adult.txt
-rw-r--r-- 1 liu3zhen liulab  869 Jan  1 14:03 youth.txt
```

# directories + files

user

group

size

date of last modification

Position	Meaning
1	"d" if a directory, "-" if a normal file
2, 3, 4	read, write, execute permission for **user** (owner) of file
5, 6, 7	read, write, execute permission for **group**
8, 9, 10	read, write, execute permission for **other** (world)

chmod

- **chmod** - change the access permissions to files and directories

```
-rw-r--r-- 1 liu3zhen liu3zhen_users  887 Jan  1 14:03 adult.txt
```

```
% chmod g+w adult.txt
```

```
-rw-rw-r-- 1 liu3zhen liu3zhen_users  887 Jan  1 14:03 adult.txt
```

```
% chmod ug-w adult.txt
```

```
-r--r--r-- 1 liu3zhen liu3zhen_users  887 Jan  1 14:03 adult.txt
```

```
% chmod u+w,go-r adult.txt
```

```
-r-------- 1 liu3zhen liu3zhen_users  887 Jan  1 14:03 adult.txt
```

- *u (user), g (group), o (other), a (all)*
- *Operators: + (add), - (remove), = (specify the exact mode)*

cp, mv, rm

- **cp** - copy files and directories

cp <oldfile> <newfile>

```
% cp adult.txt adult.tmp.txt
```

- **mv** - move (rename) files

mv <oldfile> <newfile>

```
% mv adult.tmp.txt adult.second.txt
```

- **rm** - remove files or directories

rm <filename>

rm <directory> -r

```
% rm adult.second.txt
```

head/tail

head - output the first part of files

```
% head adult.txt
# Cigarette Usage (Adult 2013)
# Source: Behavioral Risk Factor Surveillance System
# http://apps.nccd.cdc.gov/statesystem
State   Adult Cigarette Use (%)
Alabama         21.5
Alaska  22.6
Arizona         16.3
Arkansas        25.9
California      12.5
Colorado        17.7
```

“#” for comments or notes

-n <number of lines>

tail - output the last part of files

```
% tail adult.txt
South Dakota19.6
Tennessee       24.3
Texas   15.9
Utah    10.3
Vermont         16.6
Virginia        19
Washingon       16.1
West Virginia27.3
Wisconsin       18.7
Wyoming         20.6
```

more/less

more and **less** display contents of large files page by page or scroll line by line up and down.

less ("less is more") a bit more smart than more:

```
   % less filename
```

To display line numbers:

```
   % less -N filename
```

```
% more adult.txt
% less adult.txt
```

cat, paste

- **cat** - concatenate files and print on the **standard output**

```
% cat adult.txt youth.txt > two.cat.txt
```

# “>” redirect the output

- **paste** - merge lines of files

```
% paste adult.txt youth.txt > two.merge.txt
```

State	Adult Cigarette Use (%)	State	Youth Cigarette Use (%)
Alabama	21.5	Alabama	NA
Alaska	22.6	Alaska	10.6
…	…	…	…

wc

- **wc** - print line, word, and byte counts for each file

```
% wc adult.txt
55 133 887 adult.txt
```

```
% wc -l adult. txt
55 adult.txt
```

```
% wc -l two.cat.txt
110 two.cat.txt
```

grep

- **grep** - print lines matching a pattern

grep <pattern> filename

```
% grep "Kansas" adult. txt
Kansas 20
```

```
% grep "#" adult. txt
# Cigarette Usage (Adult 2013)
# Source: Behavioral Risk Factor Surveillance System
# http://apps.nccd.cdc.gov/statesystem
```

grep examples using regular expression

```
% grep -e Kansas -e Alaska adult. txt
Alaska   22.6
Kansas   20
```

```
% grep -v -e Kansas -e Alaska adult. txt
```

```
% grep “^>” fasta.file
% grep “^>” fasta.file -c
% grep “^>” fasta.file -v
```

http://www.thegeekstuff.com/2009/03/15-practical-unix-grep-command-examples

grep examples using regular expression

```
% grep -e Kansas -e Alaska adult. txt
Alaska   22.6
Kansas   20
```

```
% grep -v -e Kansas -e Alaska adult. txt
```

```
% grep “^>” fasta.file  # names of sequences
% grep “^>” fasta.file -c  # number of sequences
% grep “^>” fasta.file -v  # sequence lines
```

http://www.thegeekstuff.com/2009/03/15-practical-unix-grep-command-examples

cut

- **cut** - select sections from each line of file

State	Adult Cigarette Use (%)	State	Youth Cigarette Use (%)
Alabama	21.5	Alabama	NA
Alaska	22.6	Alaska	10.6
…	…	…	…

```
% cut two.merge.txt -f 2
```

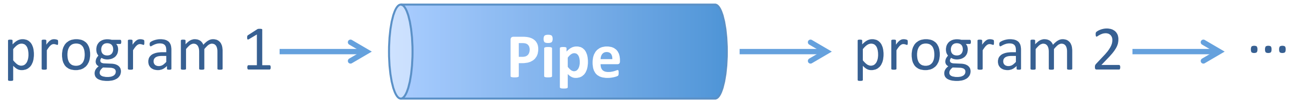
Adult Cigarette Use (%)
21.5
22.6
…

```
% cut two.merge.txt -f 1,2,4
```

State	Adult Cigarette Use (%)	Youth Cigarette Use (%)
Alabama	21.5	NA
Alaska	22.6	10.6
…	…	…

The concept of “pipe”

- Pipe is a method of inter-process communication
- Pipe collects the output of one program on the left side and inputs the collected data to the program on right side
- | is the pipe symbol
- Combining programs with different functions into one to tackle more complicated tasks



program 1 → Pipe → program 2 → …

```
head input -n 10        |        tail -n 1
```

10th line of the input

Problem

Please apply **grep, head,** and **tail** to extract the 3rd line that is not started with “#” from the file of “adult.txt”.

```
% head adult.txt
# Cigarette Usage (Adult 2013)
# Source: Behavioral Risk Factor Surveillance
System
# http://apps.nccd.cdc.gov/statesystem
State Adult Cigarette Use (%)
Alabama     21.5
Alaska      22.6
Arizona     16.3
Arkansas    25.9
California  12.5
Colorado    17.7
```

Pipe example

State	Adult Cigarette Use (%)	State	Youth Cigarette Use (%)
Alabama	21.5	Alabama	NA
Alaska	22.6	Alaska	10.6
…	…	…	…

```
% paste adult.txt youth.txt | grep "#" -v | cut -f 1,2,4 | head
```

```
State                Adult Cigarette Use (%) Youth Cigarette Use (%)
Alabama              21.5                    NA
Alaska               22.6                    10.6
Arizona              16.3                    14.1
Arkansas             25.9                    19.1
California           12.5                    NA
Colorado             17.7                    NA
Connecticut          15.5                    13.5
Delaware             19.6                    14.2
District of Columbia 18.8                    NA
```

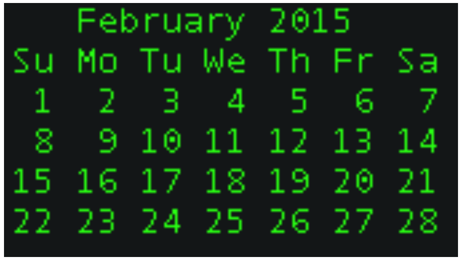
date, cal, sleep

- **date** - print or set the system date and time

```
% date
```

- **cal** - displays a calendar

```
% cal Feb 2014
```



```
   February 2015
Su Mo Tu We Th Fr Sa
 1  2  3  4  5  6  7
 8  9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
```

- **sleep** - delay for a specified amount of time

```
% sleep 2     #2 seconds pause
% sleep 1h
```

history, clear

- **history** - document of command lines

% history

- **clear** - clear the terminal screen

% clear

man

- Manual Pages

```
% man grep
```

- Detailed information about each command
- Could be too detailed to find the answer

Sometime it is more efficient to …

- Google "how-to"
- Ask questions

Unix in Beocat

- login
- create data files
- editing data files
- analyze data files