Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Bioinformatics

Jean-Claude Walser

jean-claude.walser@env.ethz.ch

HS2020

REPRODUCIBLE RESEARCH

Recap

Bioinformatics - Reproducible Research

Repeatability is a measure of the likelihood that, having produced one result from an experiment, you can try the same experiment, with the same setup, and produce that same result. It is a way for researchers to verify that their own results are true and are not just **chance artefacts**.

The **reproducibility** of data is a measure of whether a **different research team can attain results published in a paper using the same methods**. This shows that the results are **not artefacts of the unique setup in one research lab**. It is easy to see why reproducibility is desirable, as it reinforces findings and protects against rare cases of fraud, or less rare cases of human error, in the production of significant results.

Replicability - Different team, different experimental setup. If an observation is replicable it should be able to be made by a different team, using a different measuring system and dataset, in a different location, on multiple trials. This would therefore involve collecting data anew.

Source: https://www.technologynetworks.com/informatics/articles/repeatability-vs-reproducibility

Bioinformatics - Reproducible Research

Material and Methods

In a first step, all paired-end raw reads were successfully merged using FLASh (version 1.2.9, Magoc and Salzberg 2011) with minimum overlap of 5nt and maximal mismatch ration of 0.8.

Supplementary Data

```
## (a) Merging overlapping paired-end reads
# -v Version (1.2.9)
# -m minimum overlap (default 10bp)
# -x max mismatch ration (default 0.25)

flash -m 5 -x 0.8 random_1000_R1.fq random_1000_R2.fq -o
merged | tee flash.log
```

CODE / SCRIPTS

Recap

A Quick Recap

1 **Code Style**
Learn and use a common style.

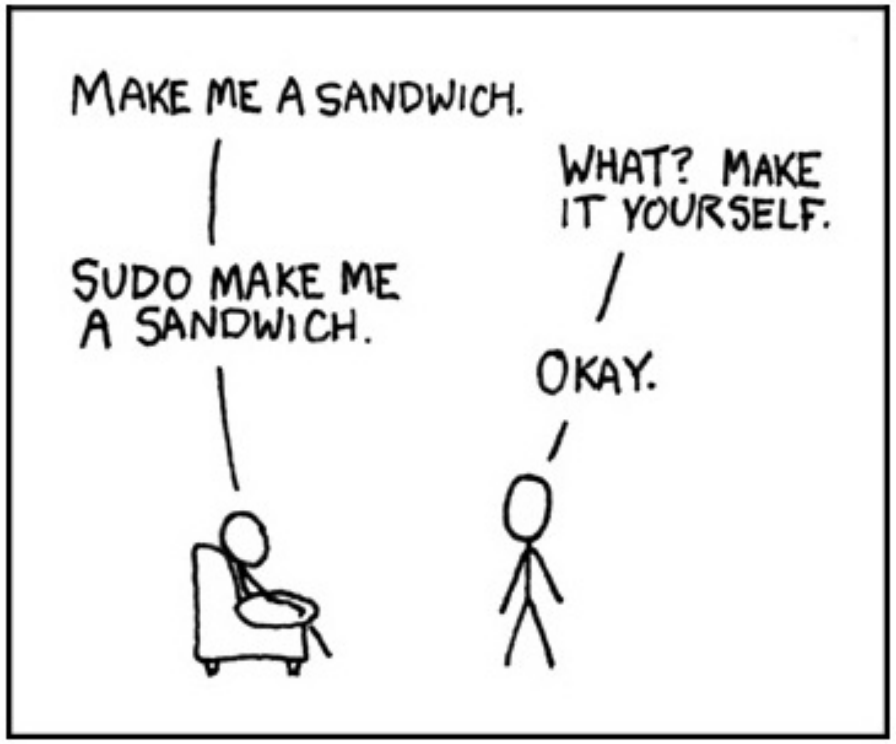
2 **#**
Comment your code and do it generously.

3 **Code Editor**
Make use of the different features like sintax highlighting, code folding or auto- completion.

4 **Version Control**
For bigger or collaborative projects use version control and /or cloud based solutions.

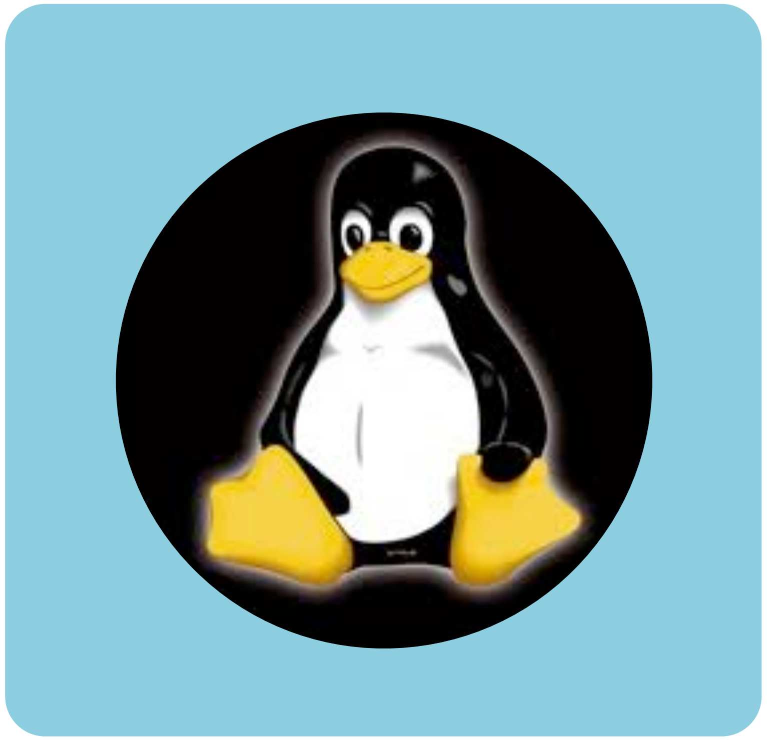
Bioinformatics - Terminal

TERMINAL



The war of the OS and the conflict of the V

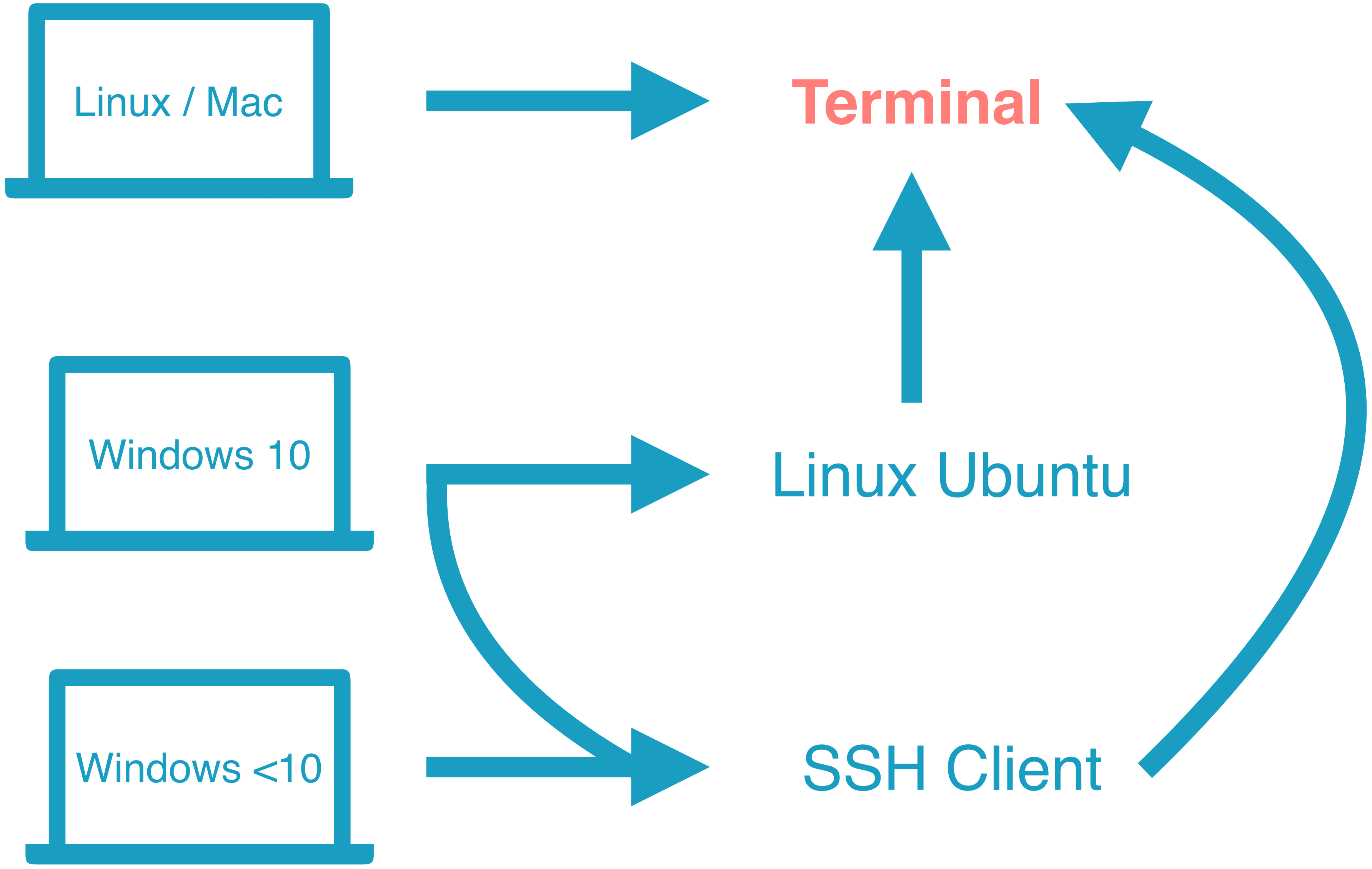
Mac



Linux

Windows

Bioinformatics - Digital Safety

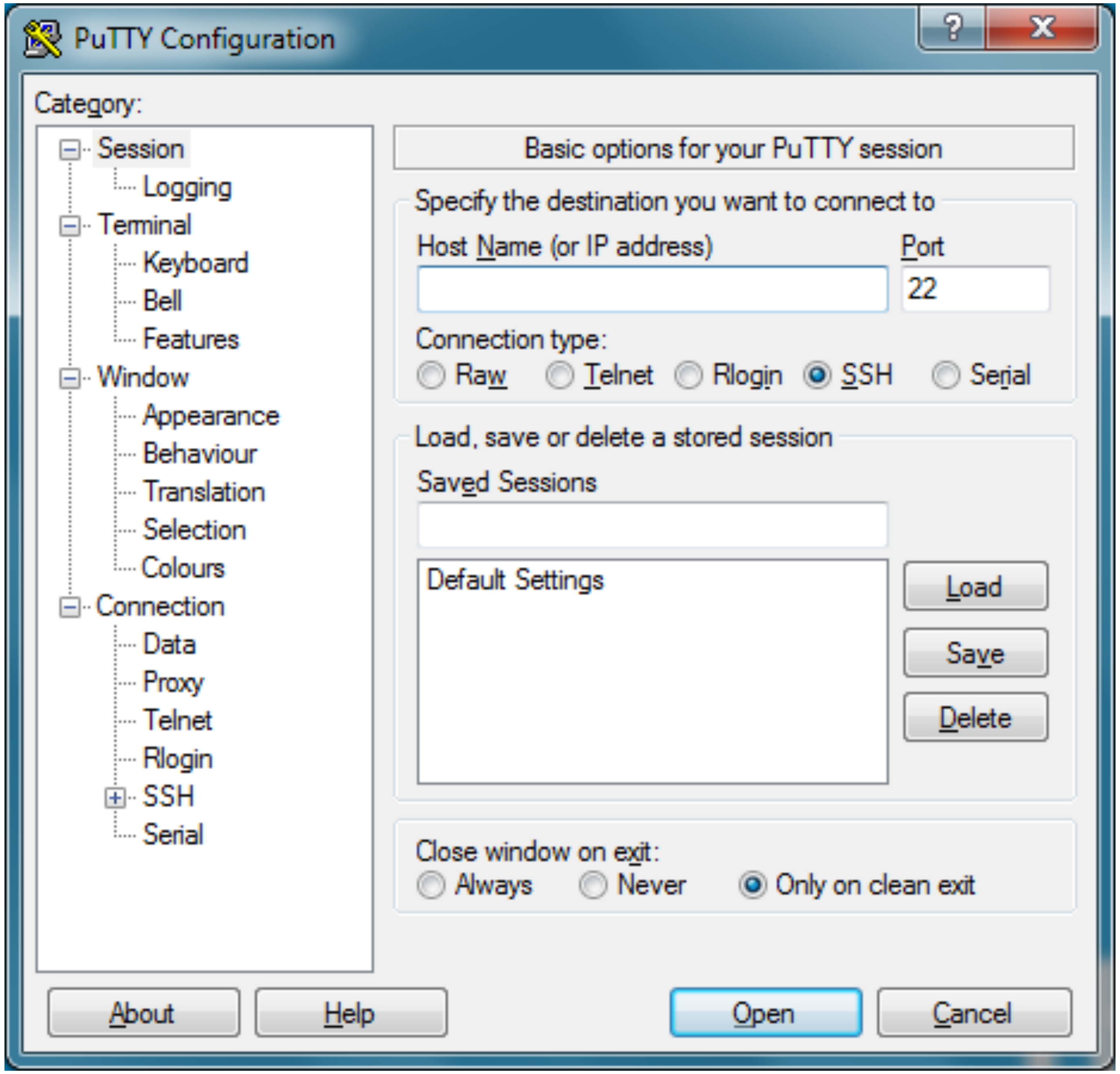


Putty

PuTTY is an SSH and telnet client, developed originally by Simon Tatham for the Windows platform. PuTTY is open source software that is available with source code and is developed and supported by a group of volunteers.

http://www.putty.org/

Bioinformatics - Introduction



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address):

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions: Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

Host Name:

student0X@gdcsrv2.ethz.ch

Note: Change X accordingly.

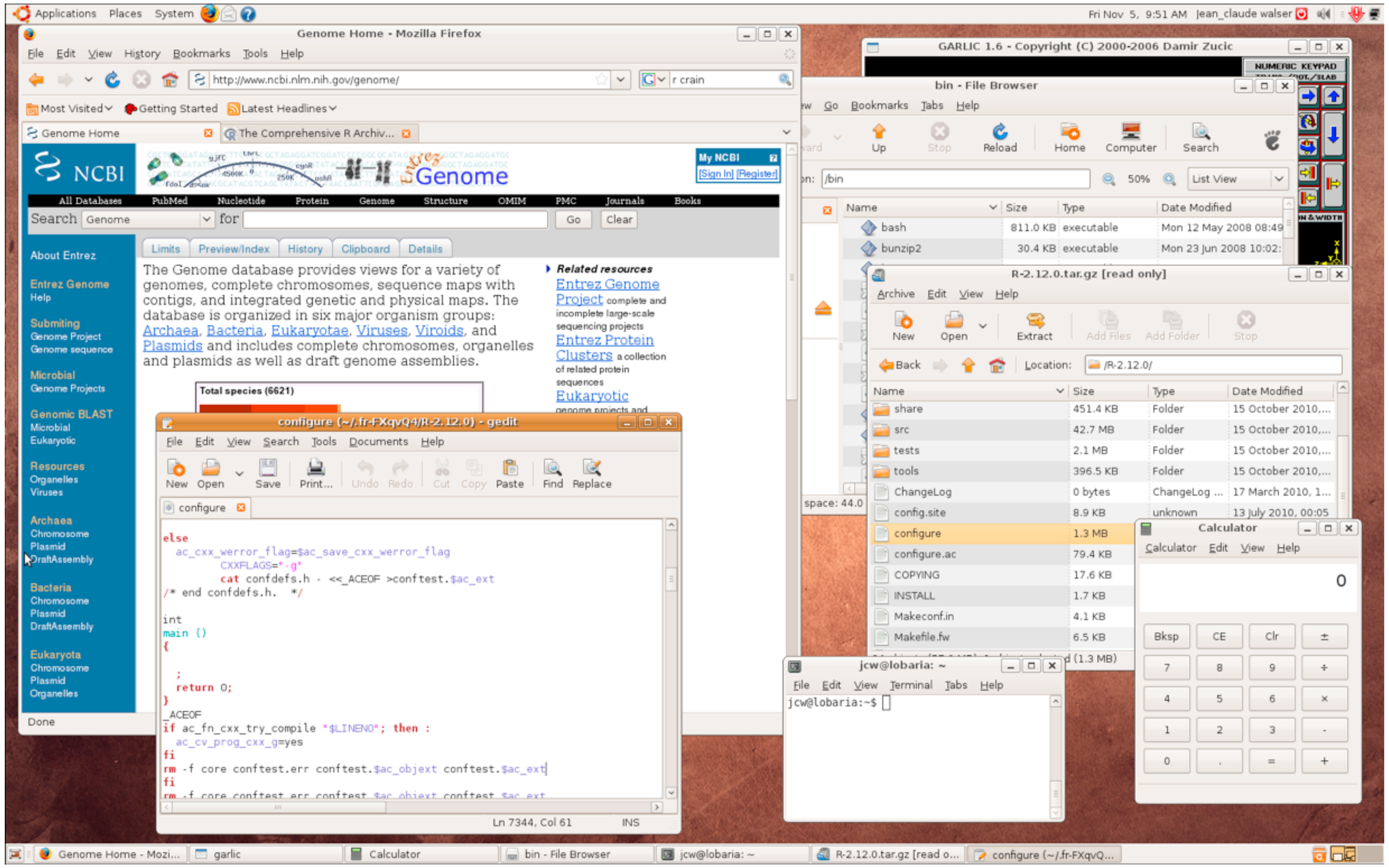
Bioinformatics - Introduction

I CAME FROM A DISTANT PLANET TO BRING YOU ADVANCED TECHNOLOGY, BUT NO ONE HERE WILL LISTEN!
scottadams@aol.com
www.dilbert.com
I AM A SUPERIOR BEING, YOU MORON! LISTEN TO WHAT I TELL YOU AND THEN DO IT!
1-25-07 ©2007 Scott Adams, Inc./Dist. by UFS, Inc.
I FIRED HIM BEFORE HE STARTED YAMMERING ABOUT LINUX.
EASY COME, EASY GO.

© Scott Adams, Inc./Dist. by UFS, Inc.

Bioinformatics - **Terminal**

A **graphical user interface (GUI)** - often pronounced gooey - an interface that allows the user (you) to interact with programs in more ways than typing.


Applications Places System
Fri Nov 5, 9:51 AM jean_claude walser
Genome Home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/genome/
Most Visited Getting Started Latest Headlines
Genome Home
The Comprehensive R Archiv...
NCBI
Entrez Genome
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search Genome for Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Genome
Help
Submiting
Genome Project
Genome sequence
Microbial
Genome Projects
Genomic BLAST
Microbial
Eukaryotic
Resources
Organelles
Viruses
Archaea
Chromosome
Plasmid
DraftAssembly
Bacteria
Chromosome
Plasmid
DraftAssembly
Eukaryota
Chromosome
Plasmid
Organelles
Done
GARLIC 1.6 - Copyright (C) 2000-2006 Damir Zucic
NUMERIC KEYPAD
bin - File Browser
Go Bookmarks Tabs Help
Up Stop Reload Home Computer Search
/bin
50% List View
space: 44.0
R-2.12.0.tar.gz [read only]
Archive Edit View Help
New Open Extract Add Files Add Folder Stop
Back Location: /R-2.12.0/
configure (~/.fr-FXqvQ4/R-2.12.0) - gedit
File Edit View Search Tools Documents Help
New Open Save Print... Undo Redo Cut Copy Paste Find Replace
configure
Ln 7344, Col 61 INS
jcw@lobaria: ~
File Edit View Terminal Tabs Help
Calculator
Calculator Edit View Help
0
Bksp CE Clr ±
7 8 9 ÷
4 5 6 ×
1 2 3 -
0 . = +
Genome Home - Mozi... garlic Calculator bin - File Browser jcw@lobaria: ~ R-2.12.0.tar.gz [read o... configure (~/.fr-FXqvQ...

The Genome database provides views for a variety of genomes, complete chromosomes, sequence maps with contigs, and integrated genetic and physical maps. The database is organized in six major organism groups: Archaea, Bacteria, Eukaryotae, Viruses, Viroids, and Plasmids and includes complete chromosomes, organelles and plasmids as well as draft genome assemblies.

Related resources

Entrez Genome Project complete and incomplete large-scale sequencing projects

Entrez Protein Clusters a collection of related protein sequences

Eukaryotic genome projects and

Total species (6621)

```
else
  ac_cxx_werror_flag=$ac_save_cxx_werror_flag
        CXXFLAGS="-g"
        cat confdefs.h - <<_ACEOF >conftest.$ac_ext
/* end confdefs.h.  */

int
main ()
{

  ;
  return 0;
}
_ACEOF
if ac_fn_cxx_try_compile "$LINENO"; then :
  ac_cv_prog_cxx_g=yes
fi
rm -f core conftest.err conftest.$ac_objext conftest.$ac_ext
fi
rm -f core conftest.err conftest.$ac_objext conftest.$ac_ext
```

Name	Size	Type	Date Modified
bash	811.0 KB	executable	Mon 12 May 2008 08:49
bunzip2	30.4 KB	executable	Mon 23 Jun 2008 10:02:

Name	Size	Type	Date Modified
share	451.4 KB	Folder	15 October 2010,...
src	42.7 MB	Folder	15 October 2010,...
tests	2.1 MB	Folder	15 October 2010,...
tools	396.5 KB	Folder	15 October 2010,...
ChangeLog	0 bytes	ChangeLog ...	17 March 2010, 1...
config.site	8.9 KB	unknown	13 July 2010, 00:05
configure	1.3 MB		
configure.ac	79.4 KB		
COPYING	17.6 KB		
INSTALL	1.7 KB		
Makeconf.in	4.1 KB		
Makefile.fw	6.5 KB		

d (1.3 MB)

```
jcw@lobaria:~$
```

GUIs are nice but have limitations. It is also difficult to describe what you exactly did and therefore it is difficult to reproduce.

Bioinformatics - Terminal

GUIs were introduced in reaction to the steep learning curve of **command-line interfaces (CLI)**, which require commands to be typed on the keyboard. Since the commands available in command line interfaces can be numerous, complicated operations can be completed using a short sequence of words and symbols. This allows for greater efficiency and productivity once many commands are learned.

Shell - Terminal

X

Shell is a UNIX term for the interactive user interface with an operating system. The shell is the layer of programming that understands and executes the commands a user enters.

Bourne-Shell (sh)

Korn-Shell (ksh)

C-Shell (csh)

TC-Shell (tcsh)

Bourne-Again-Shell (bash)

Debian Almquist Shell (dash)

Z-Shell (zsh)

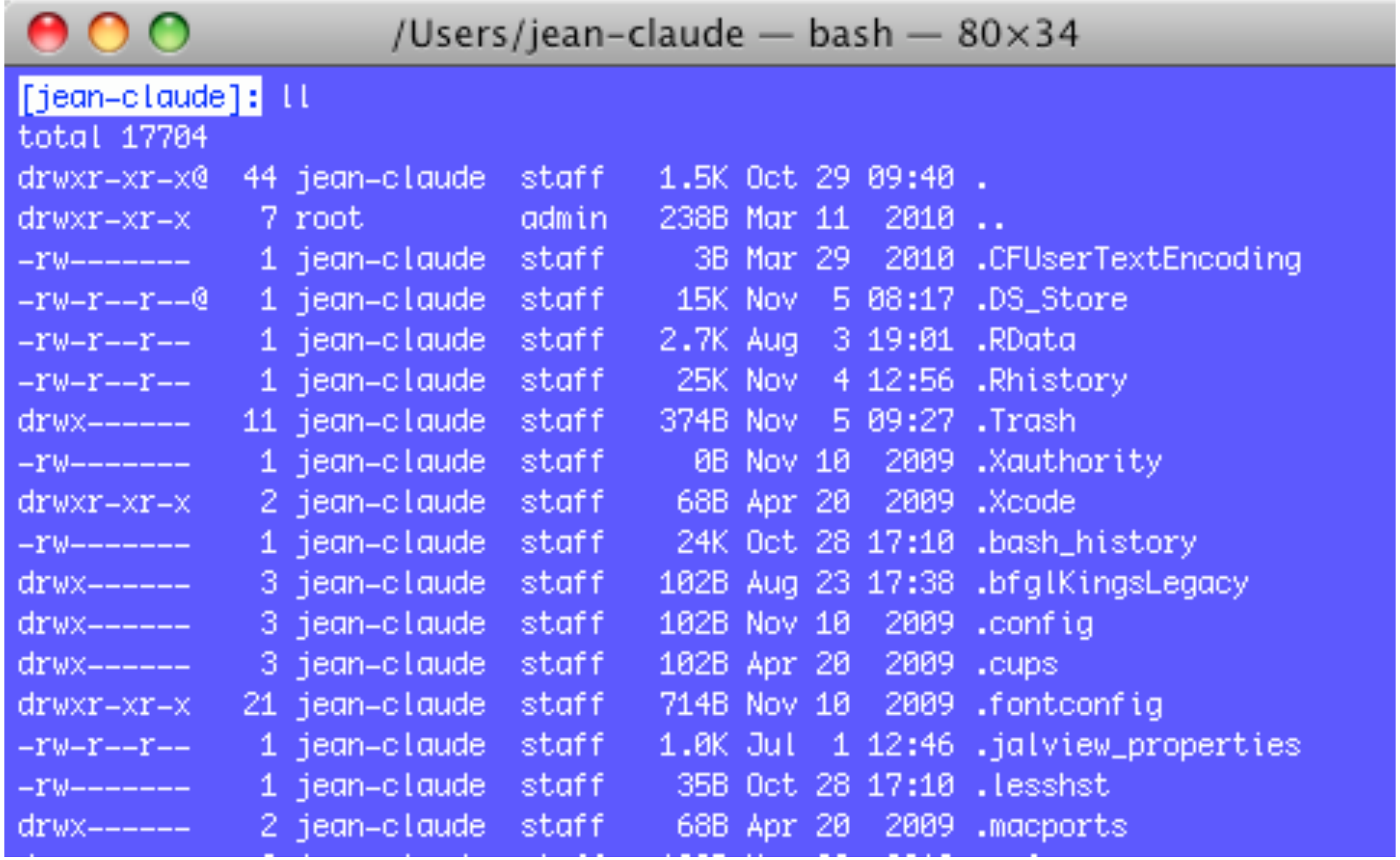
A-Shell (ash)

PowerShell / cmd.exe

```
What do I have?
$> echo ${SHELL}
```

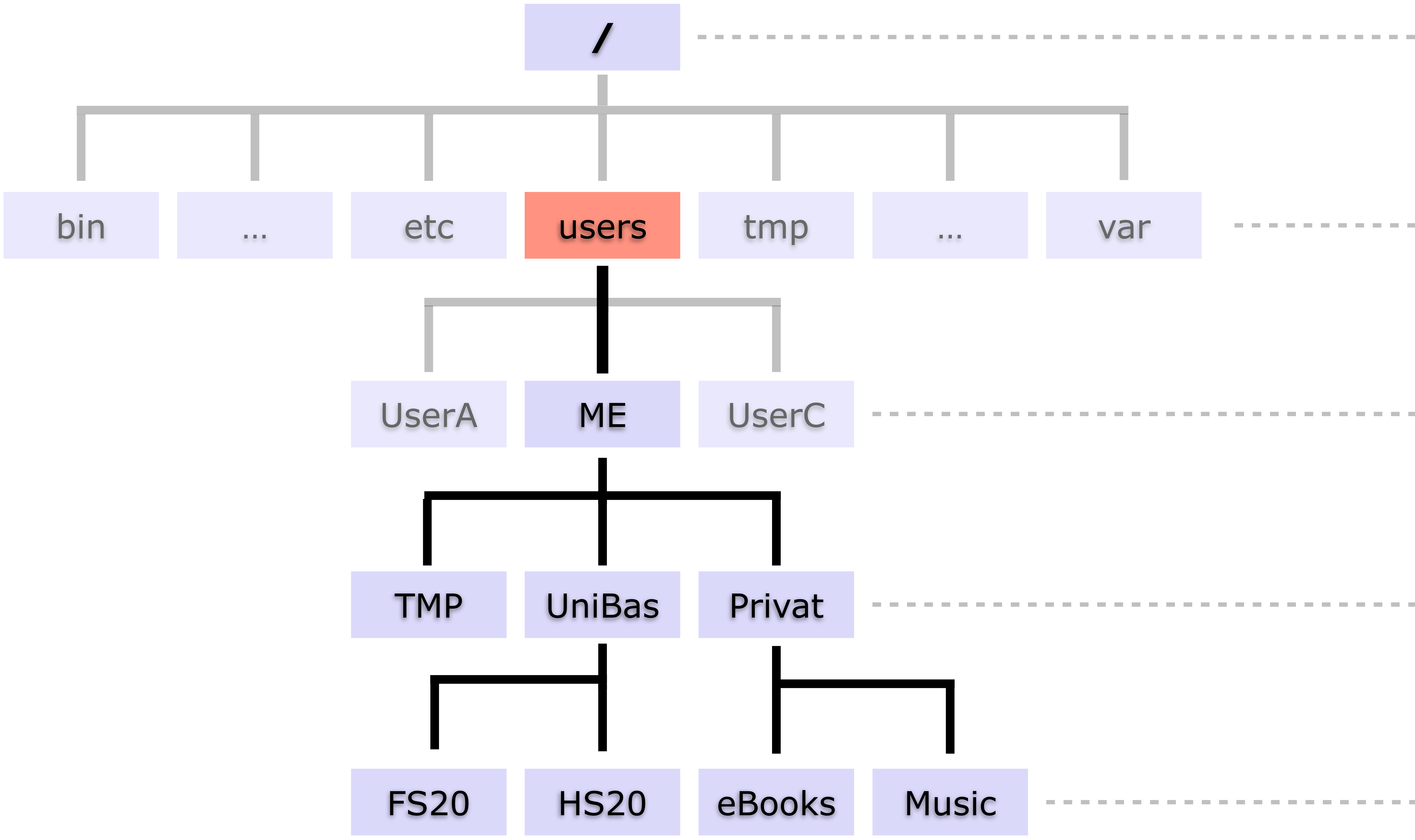
Bioinformatics - Terminal

➜ Open your (local) terminal

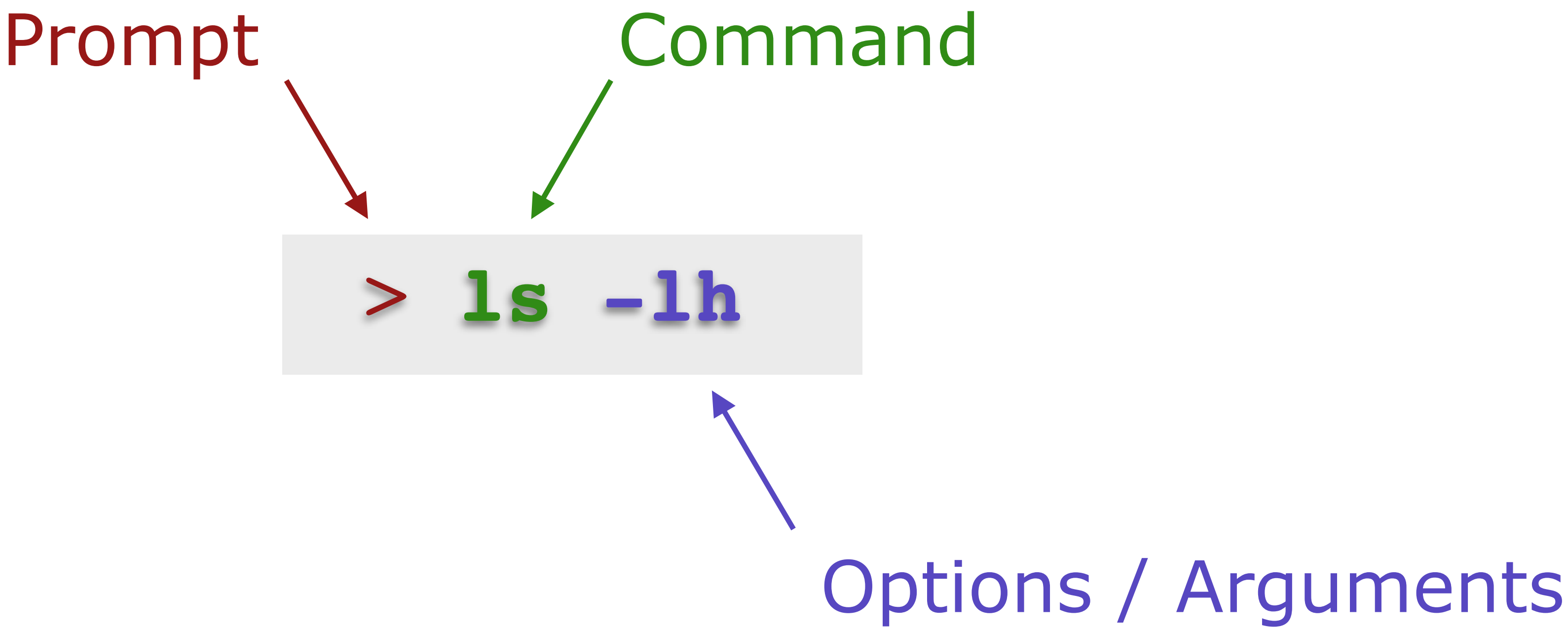


```
[jean-claude]: ll
total 17704
drwxr-xr-x@  44 jean-claude  staff   1.5K Oct 29 09:40 .
drwxr-xr-x    7 root         admin   238B Mar 11  2010 ..
-rw-------    1 jean-claude  staff     3B Mar 29  2010 .CFUserTextEncoding
-rw-r--r--@   1 jean-claude  staff    15K Nov  5 08:17 .DS_Store
-rw-r--r--    1 jean-claude  staff   2.7K Aug  3 19:01 .RData
-rw-r--r--    1 jean-claude  staff    25K Nov  4 12:56 .Rhistory
drwx------   11 jean-claude  staff   374B Nov  5 09:27 .Trash
-rw-------    1 jean-claude  staff     0B Nov 10  2009 .Xauthority
drwxr-xr-x    2 jean-claude  staff    68B Apr 20  2009 .Xcode
-rw-------    1 jean-claude  staff    24K Oct 28 17:10 .bash_history
drwx------    3 jean-claude  staff   102B Aug 23 17:38 .bfglKingsLegacy
drwx------    3 jean-claude  staff   102B Nov 10  2009 .config
drwx------    3 jean-claude  staff   102B Apr 20  2009 .cups
drwxr-xr-x   21 jean-claude  staff   714B Nov 10  2009 .fontconfig
-rw-r--r--    1 jean-claude  staff   1.0K Jul  1 12:46 .jalview_properties
-rw-------    1 jean-claude  staff    35B Oct 28 17:10 .lesshst
drwx------    2 jean-claude  staff    68B Apr 20  2009 .macports
```

Bioinformatics - Terminal



Bioinformatics - Terminal



Prompt

Command

```
> ls -lh
```

Options / Arguments

Bioinformatics - Terminal

Get Help

```
> info <command>
```

```
> info ls
```

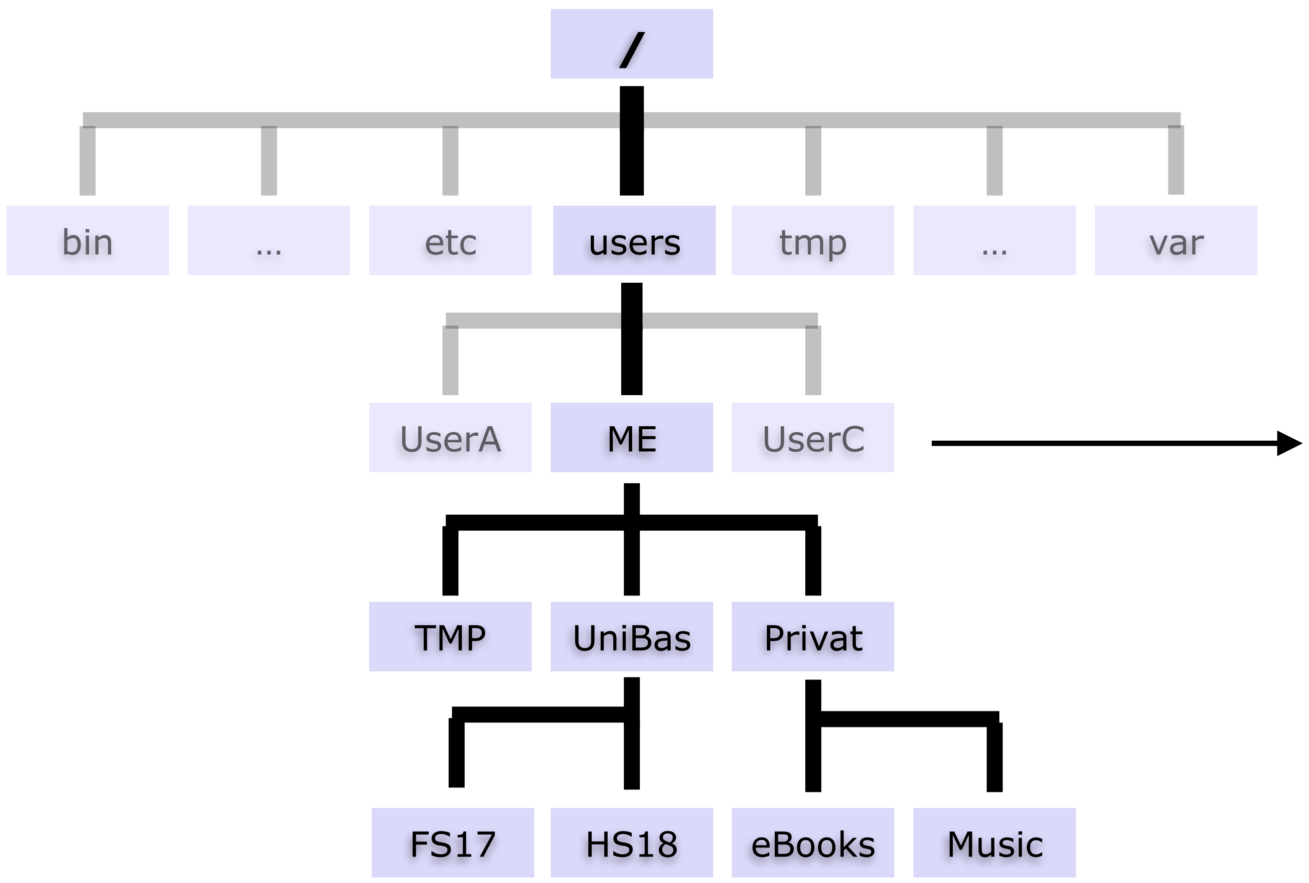
```
> man <command>
```

```
> man ls
```

* press [Q] key to leave info or help

Bioinformatics - Terminal

```
pwd - print name of current/working directory
```



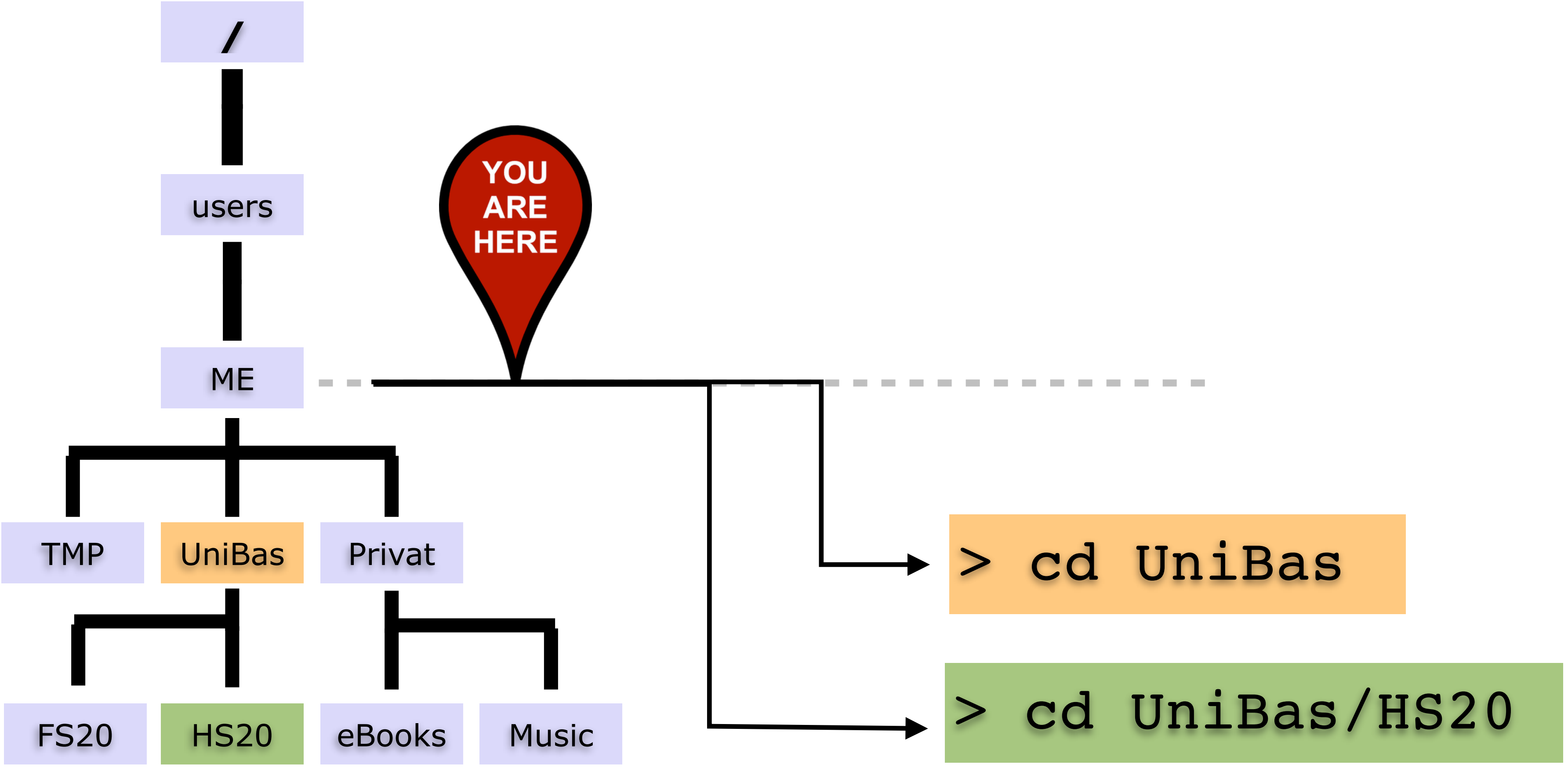
directory with users homes

```
> pwd
/home/ME
```

user homes (~)

Bioinformatics - Terminal

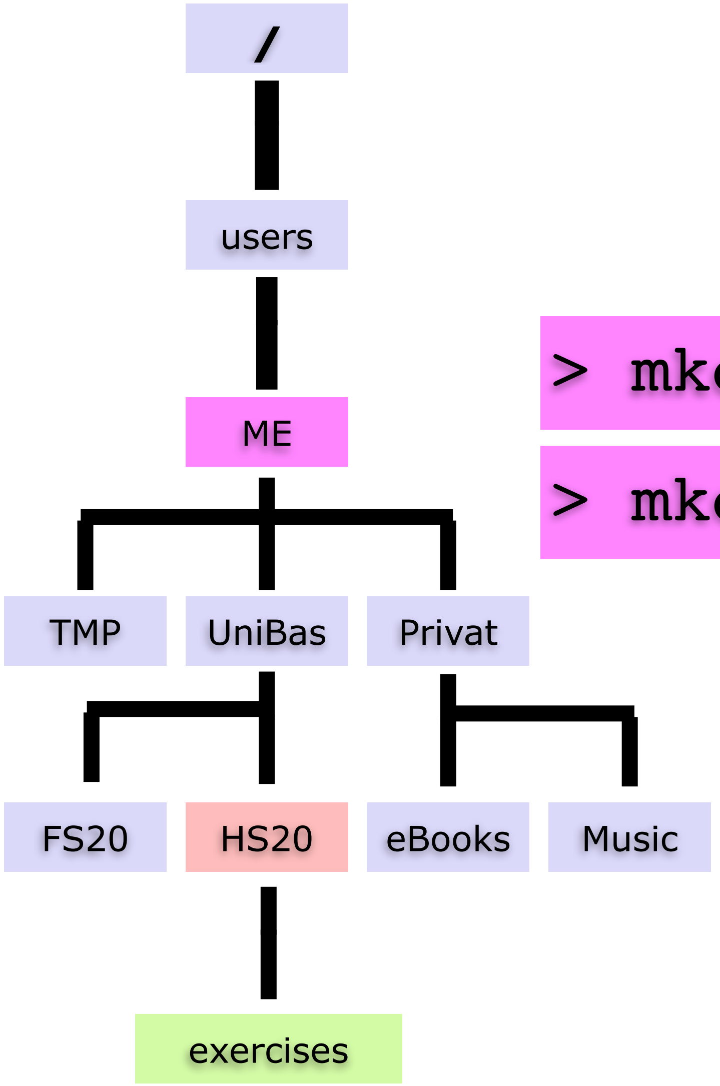
cd - change directory



```
> cd UniBas
```

```
> cd UniBas/HS20
```

mkdir - creating directories

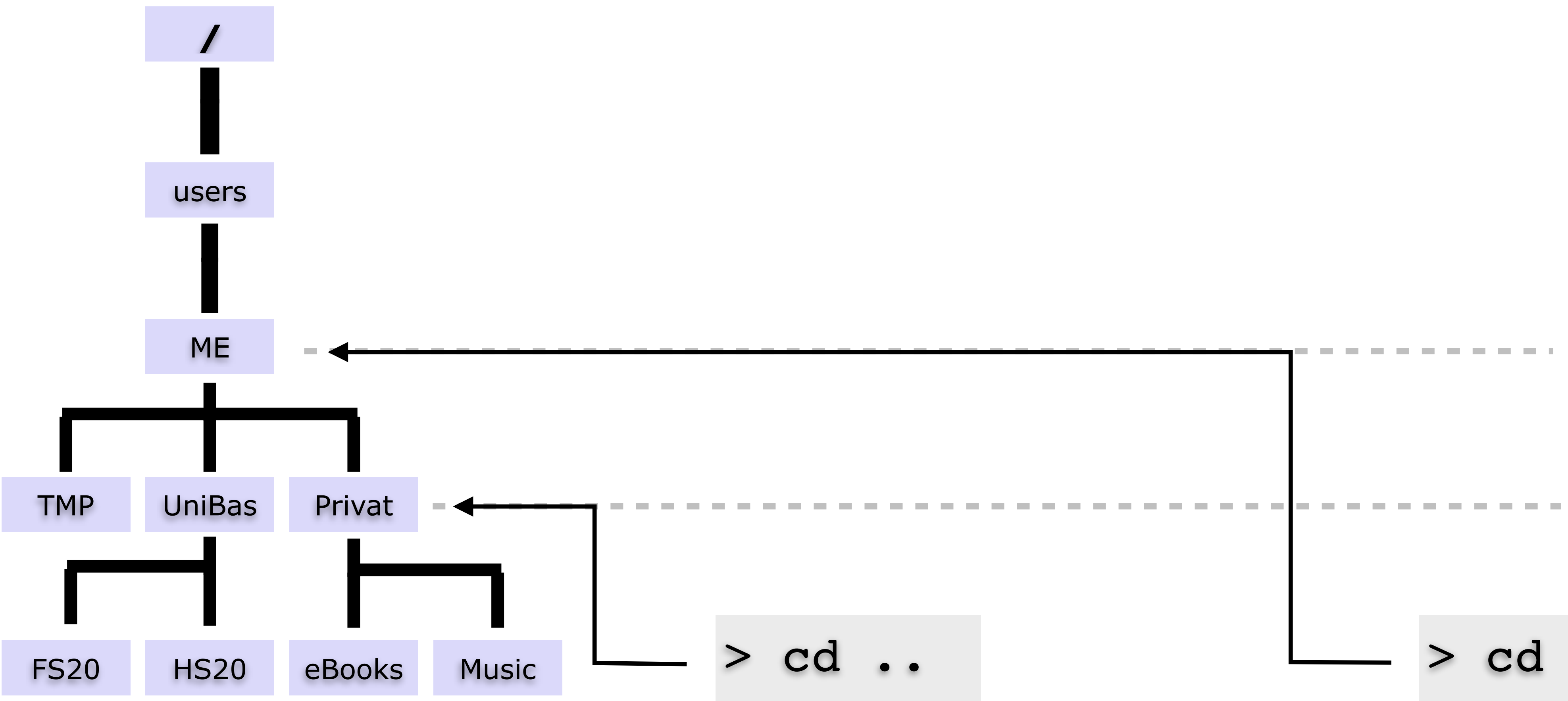


```
> mkdir ME/UniBas/HS20/exercises
```

```
> mkdir -p ME/UniBas/HS20/exercises
```

```
> mkdir exercises
```

cd – change directory (going up)



```
> cd ..
```

```
> cd ../..
```

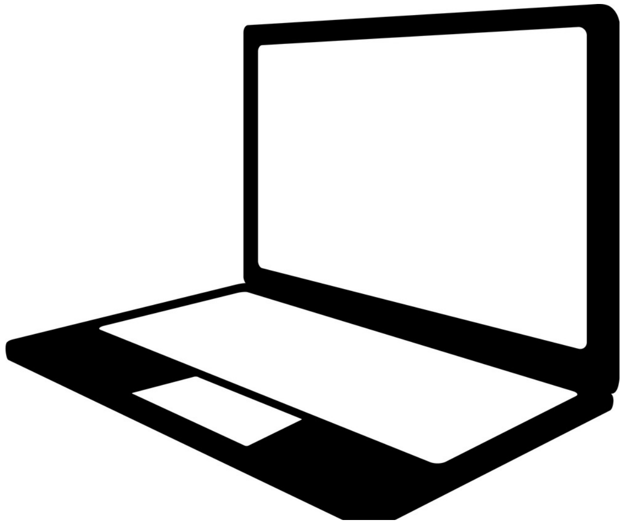
```
> cd /home/ME/UniBas
```

```
> cd /home/ME
```

Bioinformatics - Terminal

```
# Open Application
> open /Volumes/Mac*/Applications/Firefox.app
# Open Text File with Editor
> edit ${HOME}/Documents/TMP/test.txt
```

Bioinformatics - Terminal



Local
Software
Managment

Software
Dependencies

Parallel
Versions

Version Control

Application **A**
Requirments: Python 2.7

Application **B**
Requirments: Python >3.0

Bioinformatics - Terminal

CONDA

*Package, **dependency and environment management** for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN. Conda as a package manager helps you find and install packages. If you need a package that requires a different version of Python, you do not need to switch to a different environment manager, because conda is also an environment manager. With just a few commands, you can set up a totally separate environment to run that different version of Python, while continuing to run your usual version of Python in your normal environment.*

Bioinformatics - Terminal

BIOCONDA®

Bioconda is a channel for the conda package manager specializing in bioinformatics software. The conda package manager makes installing software a vastly more streamlined process. Conda is a combination of other package managers you may have encountered, such as pip, CPAN, CRAN, Bioconductor, apt-get, and homebrew. Conda is both language- and OS-agnostic, and can be used to install C/C++, Fortran, Go, R, Python, Java etc programs on Linux, Mac OSX, and Windows.

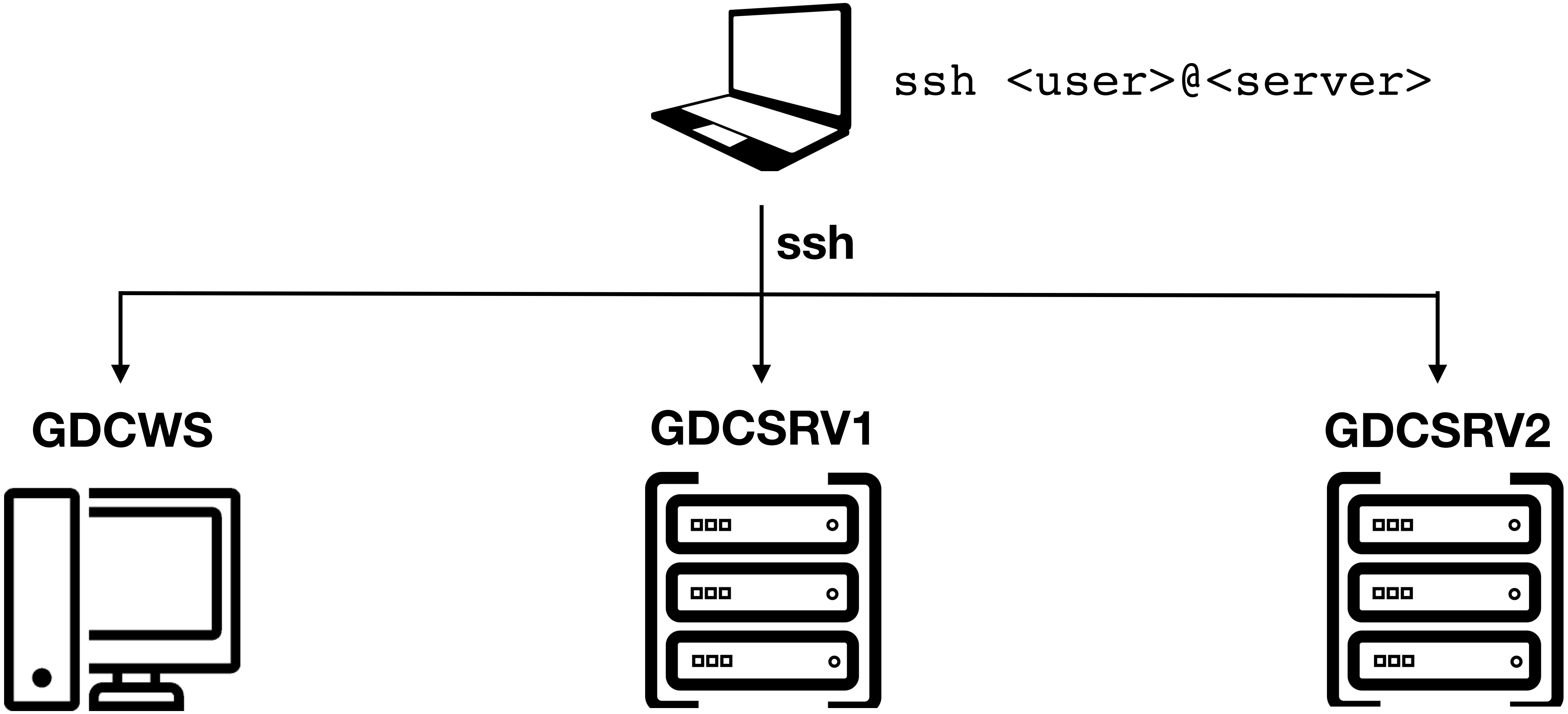
Bioinformatics - Terminal

```
python --version
# Python 2.7.15
bwa
# -bash: bwa: command not found
blast -help
# -bash: blast: command not found

conda info --envs
source activate aligners
conda info --envs
python --version
# Python 3.6.7
bwa
blastn -help
```

Bioinformatics - SSH

Secure Shell

Bioinformatics - SSH



```
ssh <user>@<server>
```

ssh

GDCWS

GDCSRV1

GDCSRV2

```
lscpu; free -m
# Architecture: x86_64
# Model name: AMD Opteron(TM)
# CPU(s): 21
# Mem: 161’008k
# NUMA node0 CPU(s): 0-5
# NUMA node1 CPU(s): 6-11
# NUMA node2 CPU(s): 12-17
# NUMA node3 CPU(s): 18-23
```

```
lscpu; free -m
# Architecture: x86_64
# Model name: Intel(R) Xeon
# CPU(s): 160
# Mem: 926’346’512k
# NUMA node0 CPU(s): 0-9,80-89
# NUMA node1 CPU(s): 10-19,90-99
# NUMA node2 CPU(s): 20-29,100-109
# NUMA node3 CPU(s): 30-39,110-119
# NUMA node4 CPU(s): 40-49,120-129
# NUMA node5 CPU(s): 50-59,130-139
# NUMA node6 CPU(s): 60-69,140-149
# NUMA node7 CPU(s): 70-79,150-159
```

```
lscpu; free -m
# Architecture: x86_64
# Model name: Intel(R) Xeon
# CPU(s): 48
# Mem: 775’363’000k
# NUMA node0 CPU(s):     0-11
# NUMA node1 CPU(s):     12-23
# NUMA node2 CPU(s):     24-35
# NUMA node3 CPU(s):     36-47
```

Bioinformatics - SSH

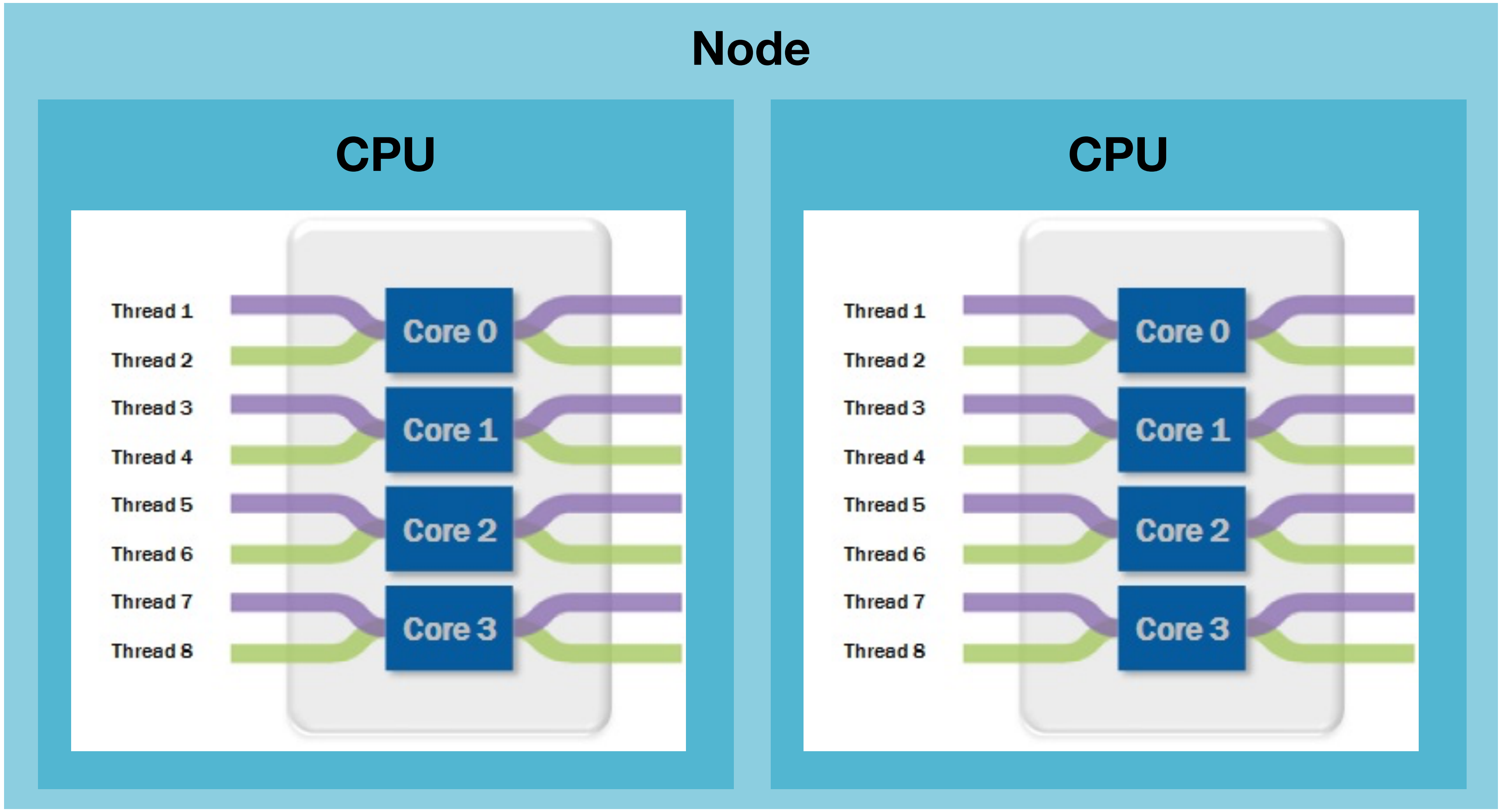
Local:

```
$ df -h

Filesystem        Size    Used   Avail  Capacity
/dev/disk1s1     932Gi   253Gi   676Gi     28%
```

Remote:

```
$ df -h

# Filesystem                 Size   Used  Avail  Use% Mounted on
# /dev/sdb                   5.3T   2.7T   2.6T   51% /data/local
# /data/gdc_home              11T   4.6T   6.1T   43% /gdc_home
# /data2/gdc_home2            22T   7.4T    15T   34% /gdc_home2
# /data3/gdc_home3            28T   5.2T    23T   19% /gdc_home3
# /data4/gdc_home4            37T    25T    13T   67% /gdc_home4
# /data5/gdc_home5            50T    27T    23T   55% /gdc_home5
```

Bioinformatics - SSH



Node

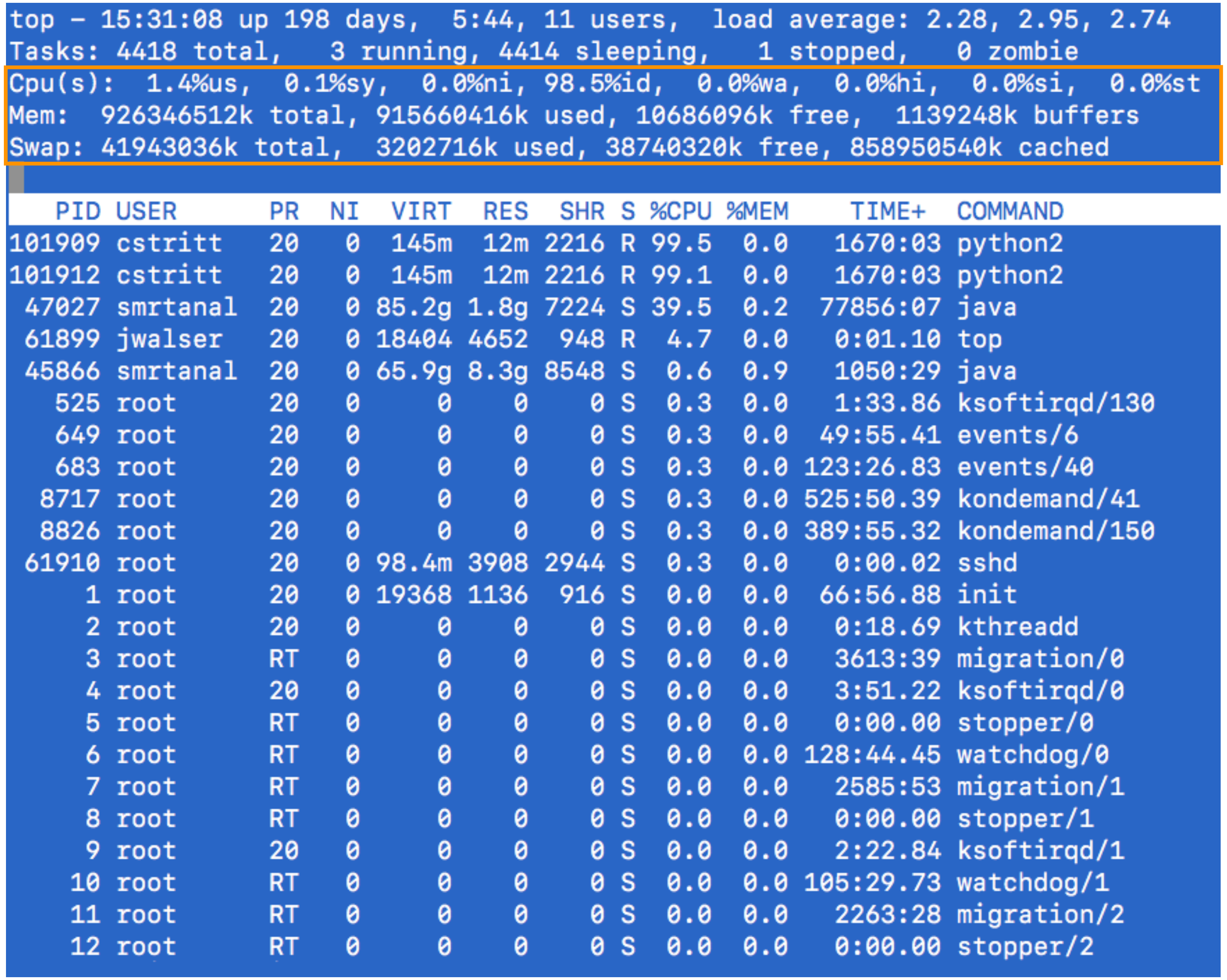
CPU

CPU

Bioinformatics - **SSH**

- **Compute node**: Currently most compute node have two sockets, each with a single CPU, volatile working memory (RAM), a hard drive, typically small, and only used to store temporary files, and a network card.
- **CPU**: Central Processing Unit, the chip that performs the actual computation in a compute node. A modern CPU is composed of numerous cores, typically 8 or 10. It has also several cache levels that help in data reuse.
- **Core**: part of a modern CPU. A core is capable of running processes, and has its own processing logic and floating point unit. Each core has its own level 1 and level 2 cache for data and instructions. Cores share last level cache.
- **Threads**: a process can perform multiple computations, i.e., program flows, concurrently. In scientific applications, threads typically process their own subset of data, or a subset of loop iterations.

Bioinformatics - SSH



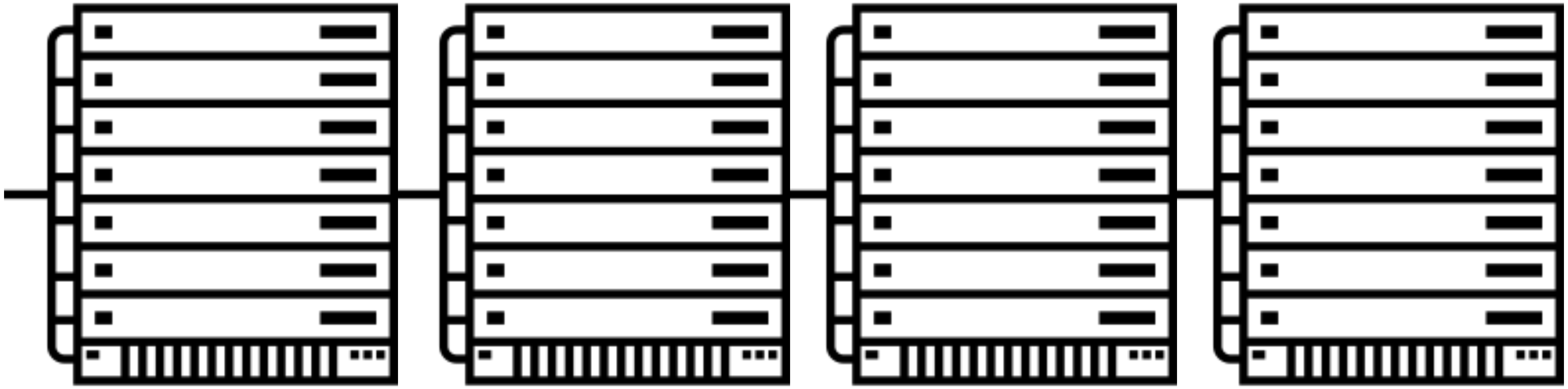
```
top - 15:31:08 up 198 days,  5:44, 11 users,  load average: 2.28, 2.95, 2.74
Tasks: 4418 total,   3 running, 4414 sleeping,   1 stopped,   0 zombie
Cpu(s):  1.4%us,  0.1%sy,  0.0%ni, 98.5%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  926346512k total, 915660416k used, 10686096k free,  1139248k buffers
Swap: 41943036k total,  3202716k used, 38740320k free, 858950540k cached

   PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
101909 cstritt   20   0  145m  12m 2216 R 99.5  0.0   1670:03 python2
101912 cstritt   20   0  145m  12m 2216 R 99.1  0.0   1670:03 python2
 47027 smrtanal  20   0 85.2g 1.8g 7224 S 39.5  0.2  77856:07 java
 61899 jwalser   20   0 18404 4652  948 R  4.7  0.0   0:01.10 top
 45866 smrtanal  20   0 65.9g 8.3g 8548 S  0.6  0.9   1050:29 java
   525 root      20   0     0    0    0 S  0.3  0.0   1:33.86 ksoftirqd/130
   649 root      20   0     0    0    0 S  0.3  0.0  49:55.41 events/6
   683 root      20   0     0    0    0 S  0.3  0.0 123:26.83 events/40
  8717 root      20   0     0    0    0 S  0.3  0.0 525:50.39 kondemand/41
  8826 root      20   0     0    0    0 S  0.3  0.0 389:55.32 kondemand/150
 61910 root      20   0 98.4m 3908 2944 S  0.3  0.0   0:00.02 sshd
     1 root      20   0 19368 1136  916 S  0.0  0.0  66:56.88 init
     2 root      20   0     0    0    0 S  0.0  0.0   0:18.69 kthreadd
     3 root      RT   0     0    0    0 S  0.0  0.0   3613:39 migration/0
     4 root      20   0     0    0    0 S  0.0  0.0   3:51.22 ksoftirqd/0
     5 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 stopper/0
     6 root      RT   0     0    0    0 S  0.0  0.0 128:44.45 watchdog/0
     7 root      RT   0     0    0    0 S  0.0  0.0   2585:53 migration/1
     8 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 stopper/1
     9 root      20   0     0    0    0 S  0.0  0.0   2:22.84 ksoftirqd/1
    10 root      RT   0     0    0    0 S  0.0  0.0 105:29.73 watchdog/1
    11 root      RT   0     0    0    0 S  0.0  0.0   2263:28 migration/2
    12 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 stopper/2
```

CPU state percentages

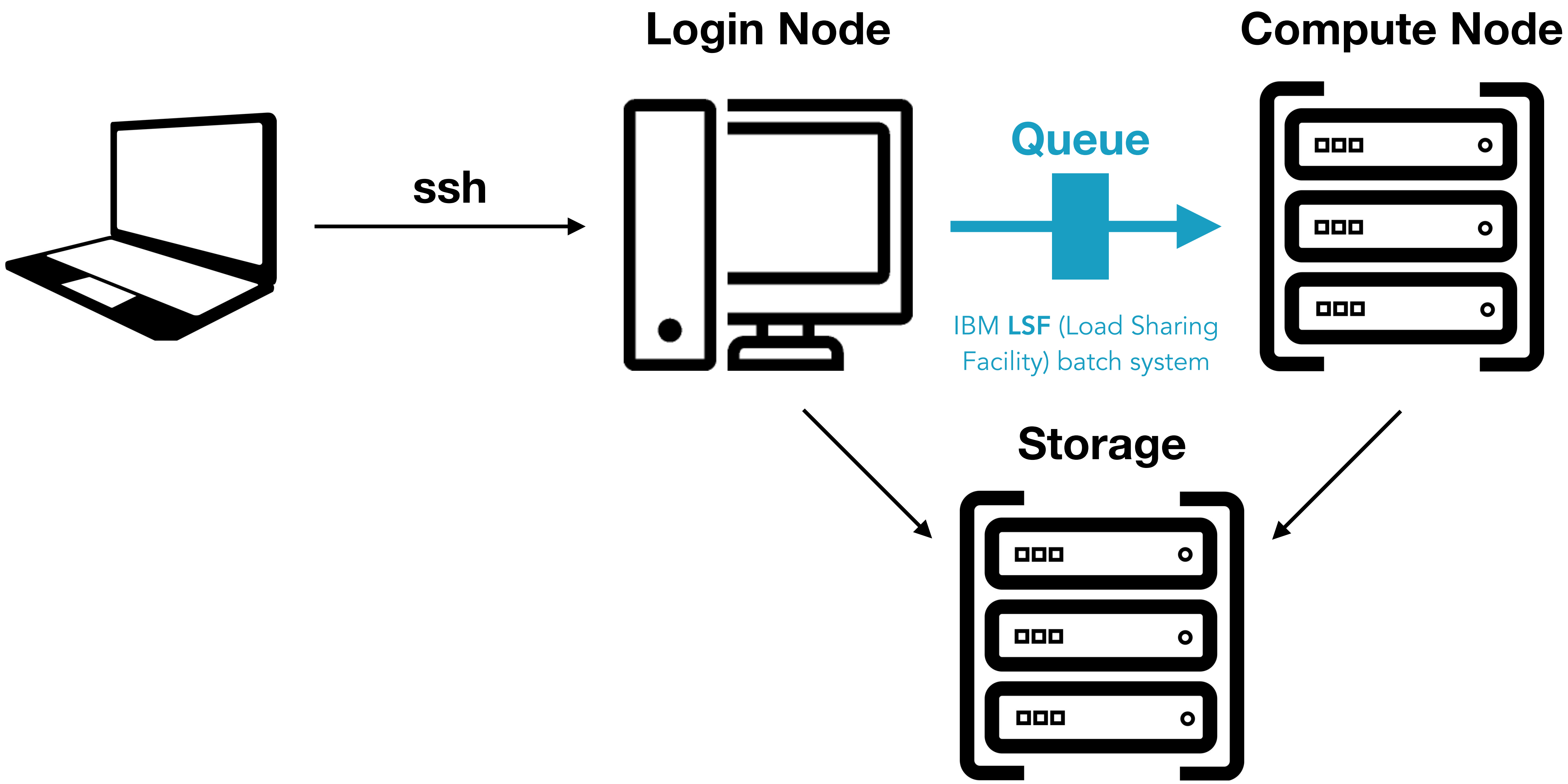
us: user
sy: system
ni: nice
wa: IO-wait
hi: hardware interrupts
si: software interrupts

Bioinformatics - SSH

High Performance Cluster



Bioinformatics - SSH

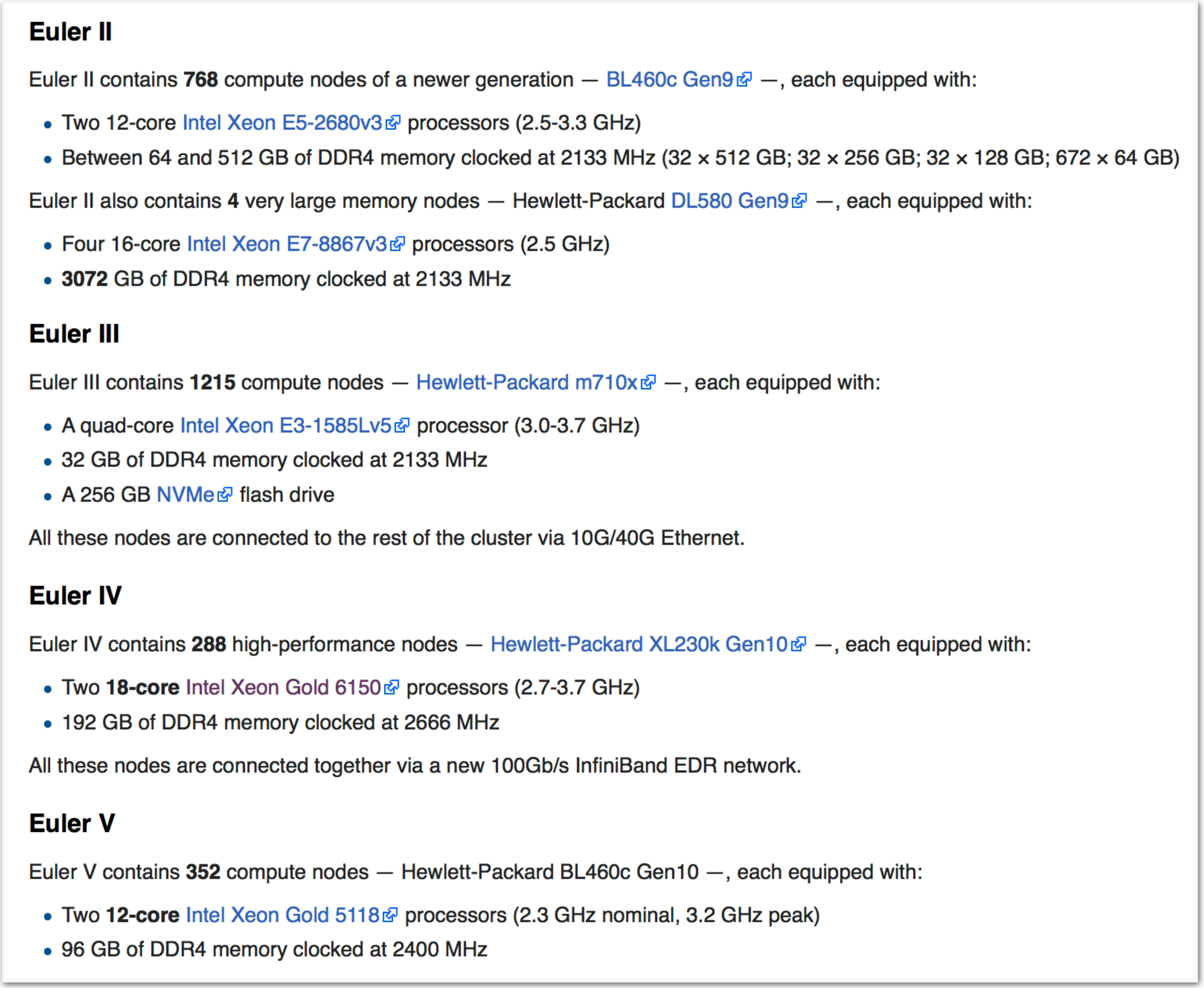


Bioinformatics - SSH

- **HPC cluster**: relatively tightly coupled collection of compute nodes. Access to the cluster is provided through a login node. A resource manager and scheduler provide the logic to schedule jobs efficiently on the cluster.
- **Compute node**: an individual computer, part of an HPC cluster. Currently most compute node have two sockets, each with a single CPU, volatile working memory (RAM), a hard drive, typically small, and only used to store temporary files, and a network card.
- **CPU**: Central Processing Unit, the chip that performs the actual computation in a compute node. A modern CPU is composed of numerous cores, typically 8 or 10. It has also several cache levels that help in data reuse.
- **Core**: part of a modern CPU. A core is capable of running processes, and has its own processing logic and floating point unit. Each core has its own level 1 and level 2 cache for data and instructions. Cores share last level cache.
- **Threads**: a process can perform multiple computations, i.e., program flows, concurrently. In scientific applications, threads typically process their own subset of data, or a subset of loop iterations.

Bioinformatics - SSH

Bioinformatics - SSH



Euler II

Euler II contains **768** compute nodes of a newer generation — BL460c Gen9 —, each equipped with:

- Two 12-core Intel Xeon E5-2680v3 processors (2.5-3.3 GHz)
- Between 64 and 512 GB of DDR4 memory clocked at 2133 MHz (32 × 512 GB; 32 × 256 GB; 32 × 128 GB; 672 × 64 GB)

Euler II also contains **4** very large memory nodes — Hewlett-Packard DL580 Gen9 —, each equipped with:

- Four 16-core Intel Xeon E7-8867v3 processors (2.5 GHz)
- **3072** GB of DDR4 memory clocked at 2133 MHz

Euler III

Euler III contains **1215** compute nodes — Hewlett-Packard m710x —, each equipped with:

- A quad-core Intel Xeon E3-1585Lv5 processor (3.0-3.7 GHz)
- 32 GB of DDR4 memory clocked at 2133 MHz
- A 256 GB NVMe flash drive

All these nodes are connected to the rest of the cluster via 10G/40G Ethernet.

Euler IV

Euler IV contains **288** high-performance nodes — Hewlett-Packard XL230k Gen10 —, each equipped with:

- Two **18-core** Intel Xeon Gold 6150 processors (2.7-3.7 GHz)
- 192 GB of DDR4 memory clocked at 2666 MHz

All these nodes are connected together via a new 100Gb/s InfiniBand EDR network.

Euler V

Euler V contains **352** compute nodes — Hewlett-Packard BL460c Gen10 —, each equipped with:

- Two **12-core** Intel Xeon Gold 5118 processors (2.3 GHz nominal, 3.2 GHz peak)
- 96 GB of DDR4 memory clocked at 2400 MHz

Bioinformatics - SSH

Basic job submission

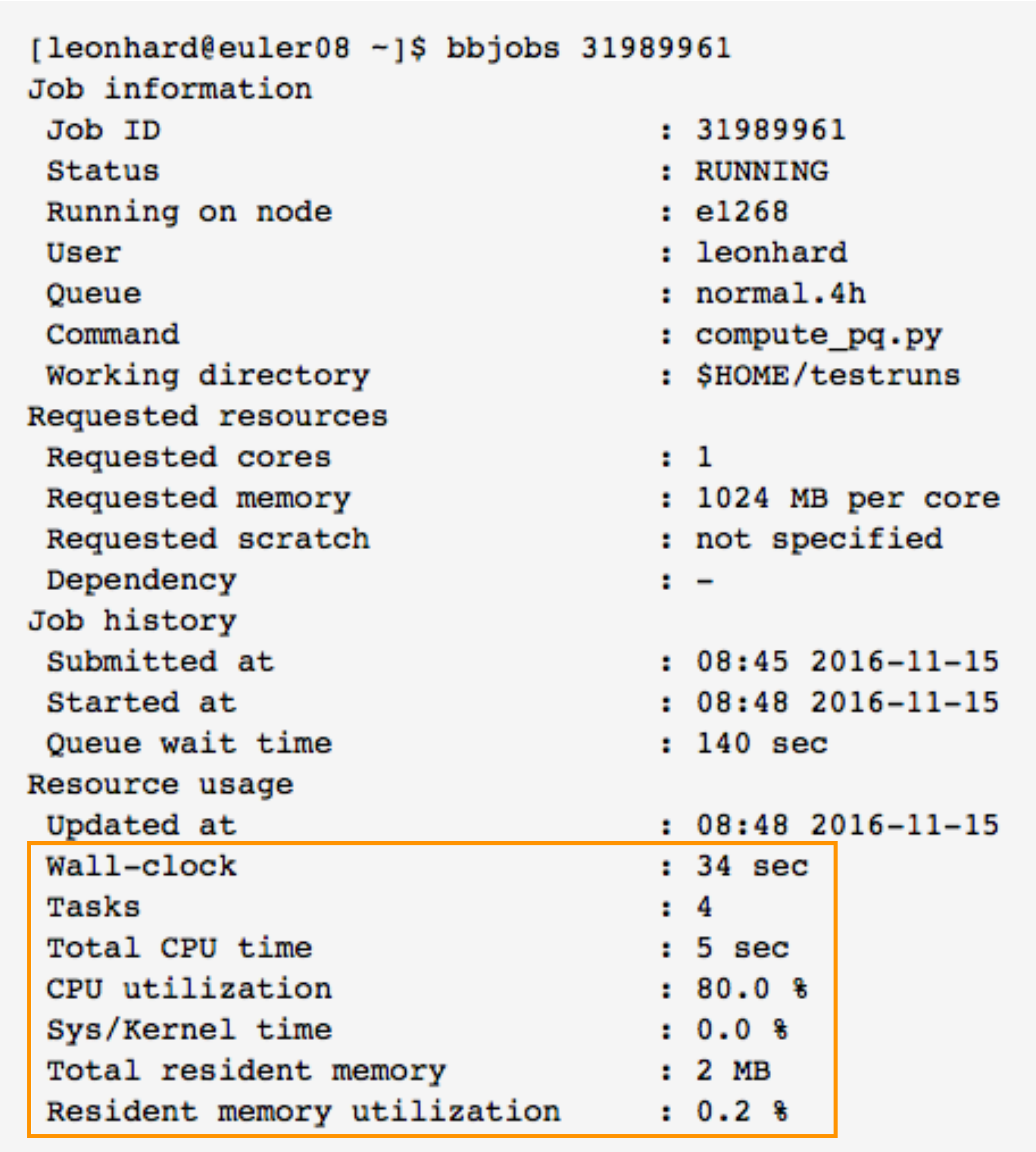
```
bsub -W 2:00 -n number_of_procs -R "rusage[mem=2048,scratch=5000]" <command>
<parameters>

# -n request multiple cores (or threads)
# -R mem default the batch system allocates 1024 MB (1 GB) of memory per
processor core
# -R scratch for temporary data
```

Submission script

```
#!/bin/bash
#BSUB -J "MyScript"          ## Job Title
#BSUB -n 10                  ## Number of Cores
#BSUB -R "rusage[mem=2048]"  ## Memory Request
#BSUB -W 2:00                ## Running Time

## Load environment
module load gcc/4.8.2 gdc perl/5.18.4

## ...
```

Bioinformatics - **SSH**

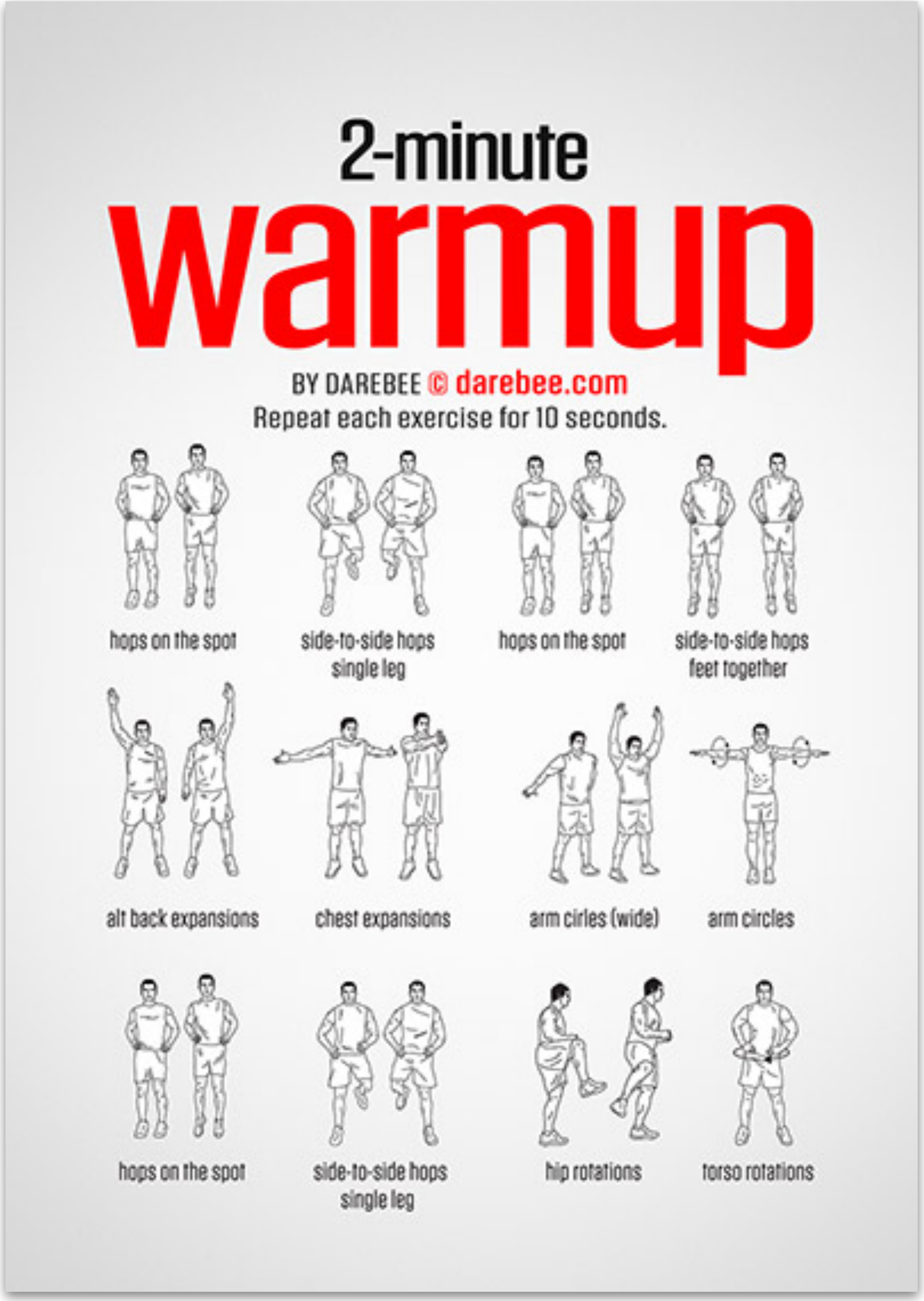
Job monitoring



```
[leonhard@euler08 ~]$ bbjobs 31989961
Job information
 Job ID                          : 31989961
 Status                          : RUNNING
 Running on node                 : e1268
 User                            : leonhard
 Queue                           : normal.4h
 Command                         : compute_pq.py
 Working directory               : $HOME/testruns
Requested resources
 Requested cores                 : 1
 Requested memory                : 1024 MB per core
 Requested scratch               : not specified
 Dependency                      : -
Job history
 Submitted at                    : 08:45 2016-11-15
 Started at                      : 08:48 2016-11-15
 Queue wait time                 : 140 sec
Resource usage
 Updated at                      : 08:48 2016-11-15
 Wall-clock                      : 34 sec
 Tasks                           : 4
 Total CPU time                  : 5 sec
 CPU utilization                 : 80.0 %
 Sys/Kernel time                 : 0.0 %
 Total resident memory           : 2 MB
 Resident memory utilization     : 0.2 %
```

Bioinformatics - Introduction

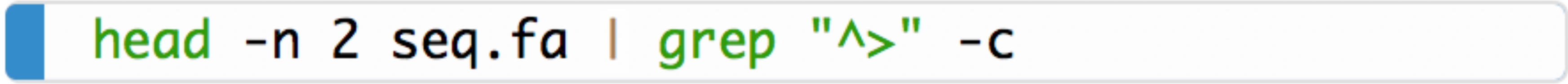
Bioinformatics - RegEx



Bioinformatics - Example Questions

Q1

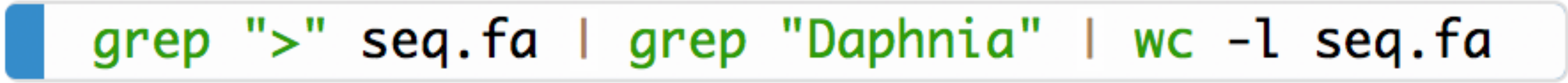
What is the outcome of the following terminal command?



```
head -n 2 seq.fa | grep "^>" -c
```

Q2

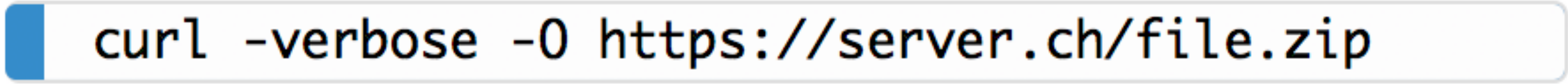
What does the following command line do?



```
grep ">" seq.fa | grep "Daphnia" | wc -l seq.fa
```

Q3

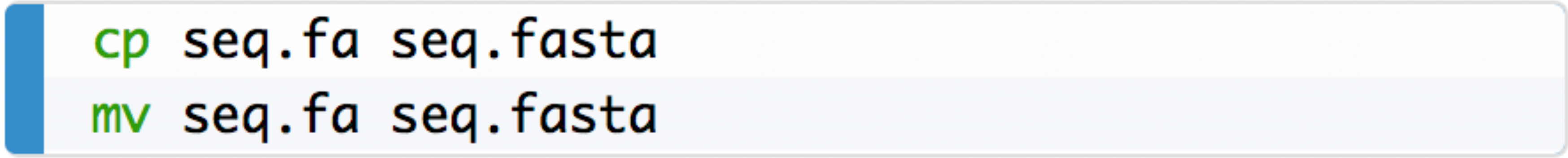
What does verbose mean and why would it be important to use?



```
curl -verbose -O https://server.ch/file.zip
```

Q4

With both commands, a new file called seq.fasta is created. What is the difference?



```
cp seq.fa seq.fasta
mv seq.fa seq.fasta
```

Bioinformatics - Example Questions

Q5 What is the conection between comments and reproducibility?

Q6 What is missing in the description below to make it reproducible?

"USEARCH (Edgar 2017) with default parameters was used to quality filter the sequences?"

Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Extension

Digital Safety

Jean-Claude Walser

jean-claude.walser@env.ethz.ch

HS2020

GOOD SAFETY HABITS

You are being tracked by:

Cookies

Evercookies

IP address

Flash cookies

HTML 5 storage

Fingerprinting

and more …

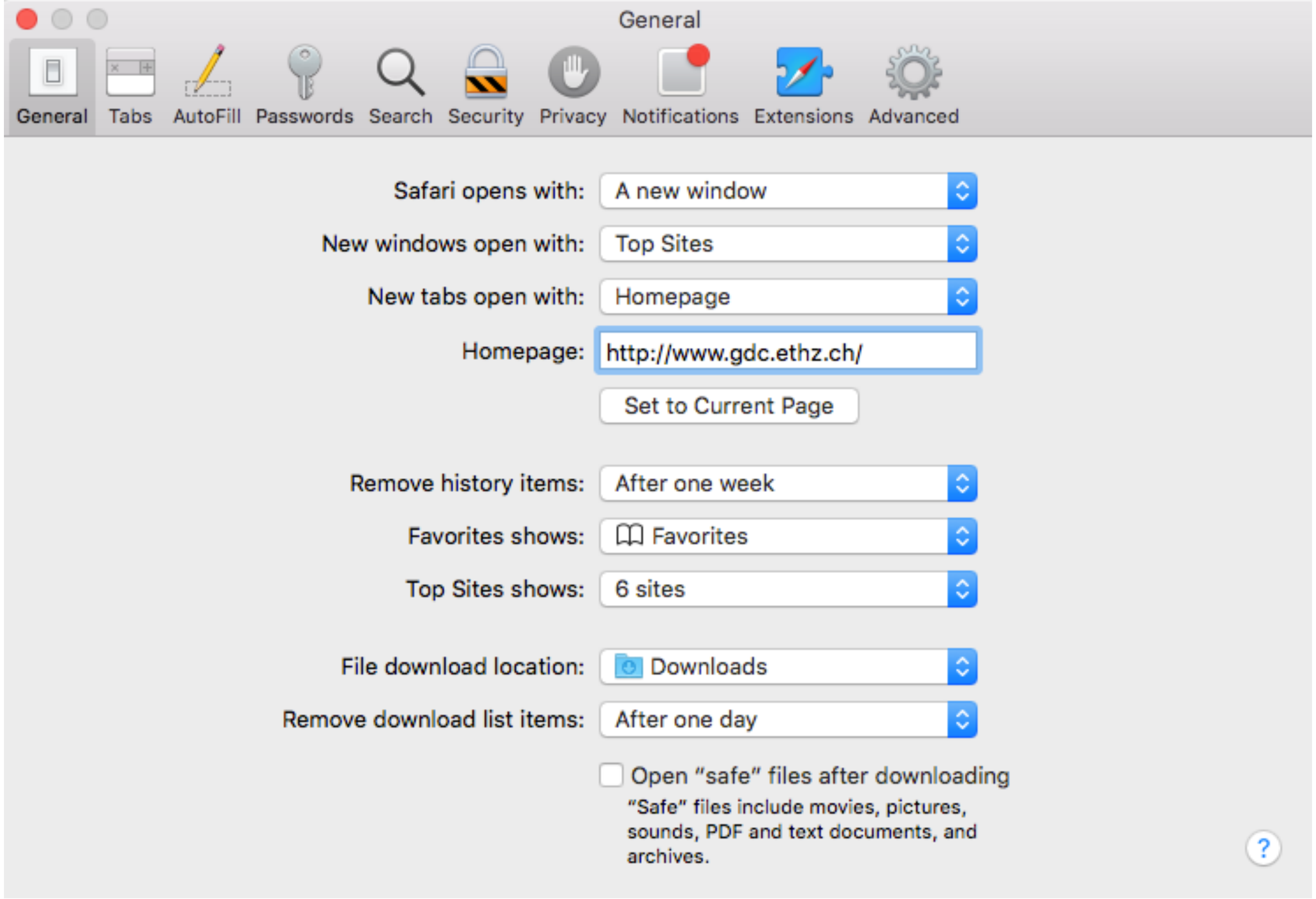
Bioinformatics - Digital Safety

- Visiting the top 50 websites will install over 3,000 tracking files on your computer.
- Over 80% of websites use one or more tracking tools.
- The average number of tracking tools is six per website. Bigger websites can have up to 20 trackers.

Webbrowser Safety

- Security settings
- Browser extensions
- VPN
- Incognito Mode

Change / Check Settings



Safari opens with: A new window

New windows open with: Top Sites

New tabs open with: Homepage

Homepage: http://www.gdc.ethz.ch/

Remove history items: After one week

Favorites shows: Favorites

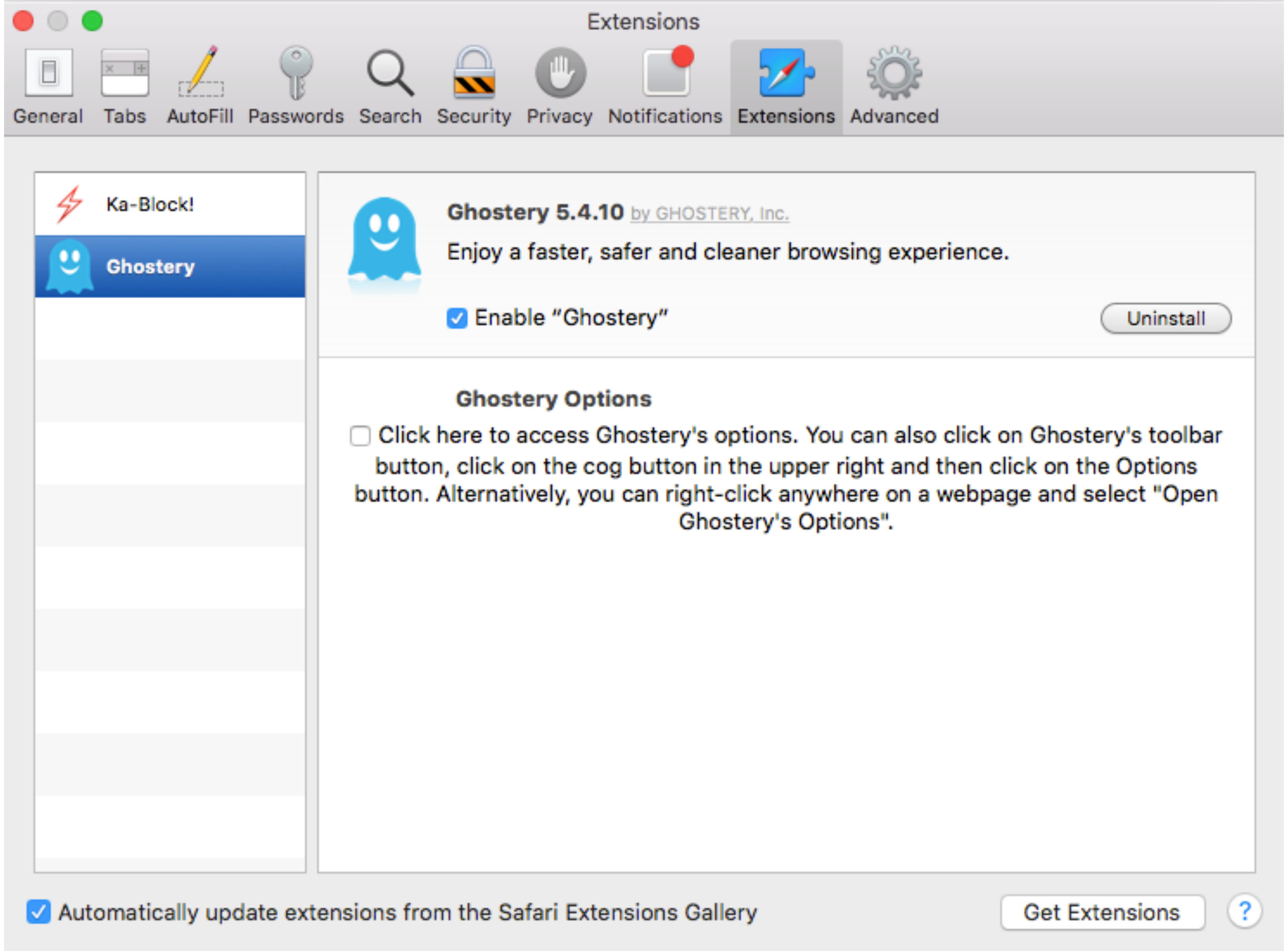
Top Sites shows: 6 sites

File download location: Downloads

Remove download list items: After one day

□ Open "safe" files after downloading
"Safe" files include movies, pictures, sounds, PDF and text documents, and archives.

Install Extensions



Ghostery 5.4.10 by GHOSTERY, Inc.

Enjoy a faster, safer and cleaner browsing experience.

☑ Enable "Ghostery"

Ghostery Options

□ Click here to access Ghostery's options. You can also click on Ghostery's toolbar button, click on the cog button in the upper right and then click on the Options button. Alternatively, you can right-click anywhere on a webpage and select "Open Ghostery's Options".

☑ Automatically update extensions from the Safari Extensions Gallery

Bioinformatics - Digital Safety

How to Enable **Private Browsing** on Any Web Browser

Privacy Browser

- Epic Privacy Browser
- Comodo Dragon
- Brave
- Tor

Bioinformatics - Digital Safety

VPN is a **virtual private network** that enables you to have a secure connection between your device and an Internet server that no one can monitor or access the data that you’re exchanging. A VPN connection establishes a safe passageway through all the insecurities of public networks. You can benefit from a VPN connection for a number of reasons, including, security and privacy.

When you’re connected to the Internet through a VPN connection, this private internet access ensures that you’re not exposed to phishing, malware, viruses and other cyber threats. Your privacy is also guaranteed, as no one will be able to detect your online behaviour. Anyone can benefit from an added safety measure that the VPN connection feature provides.

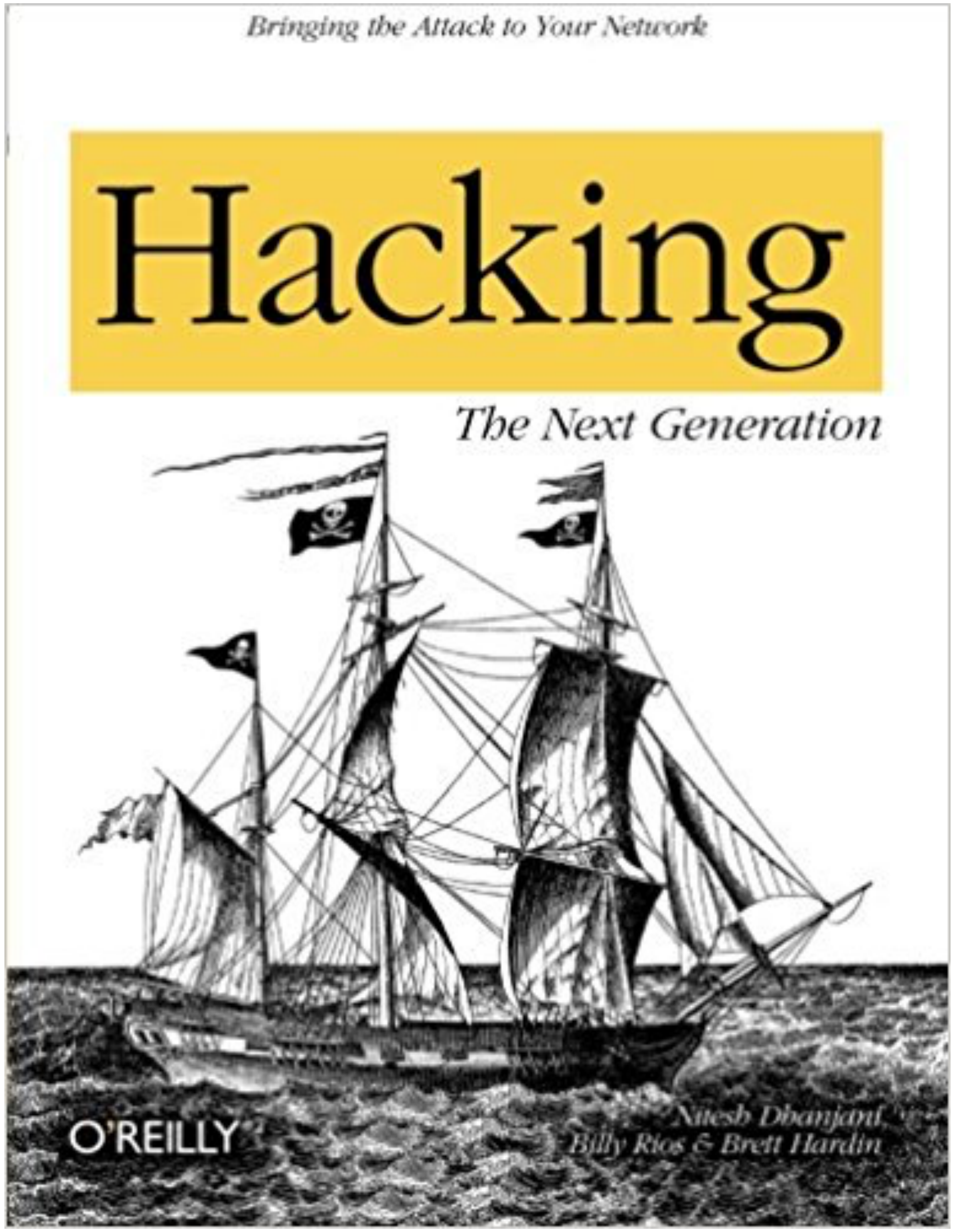
> https://mobile.unibas.ch/home.html

Bioinformatics - Digital Safety

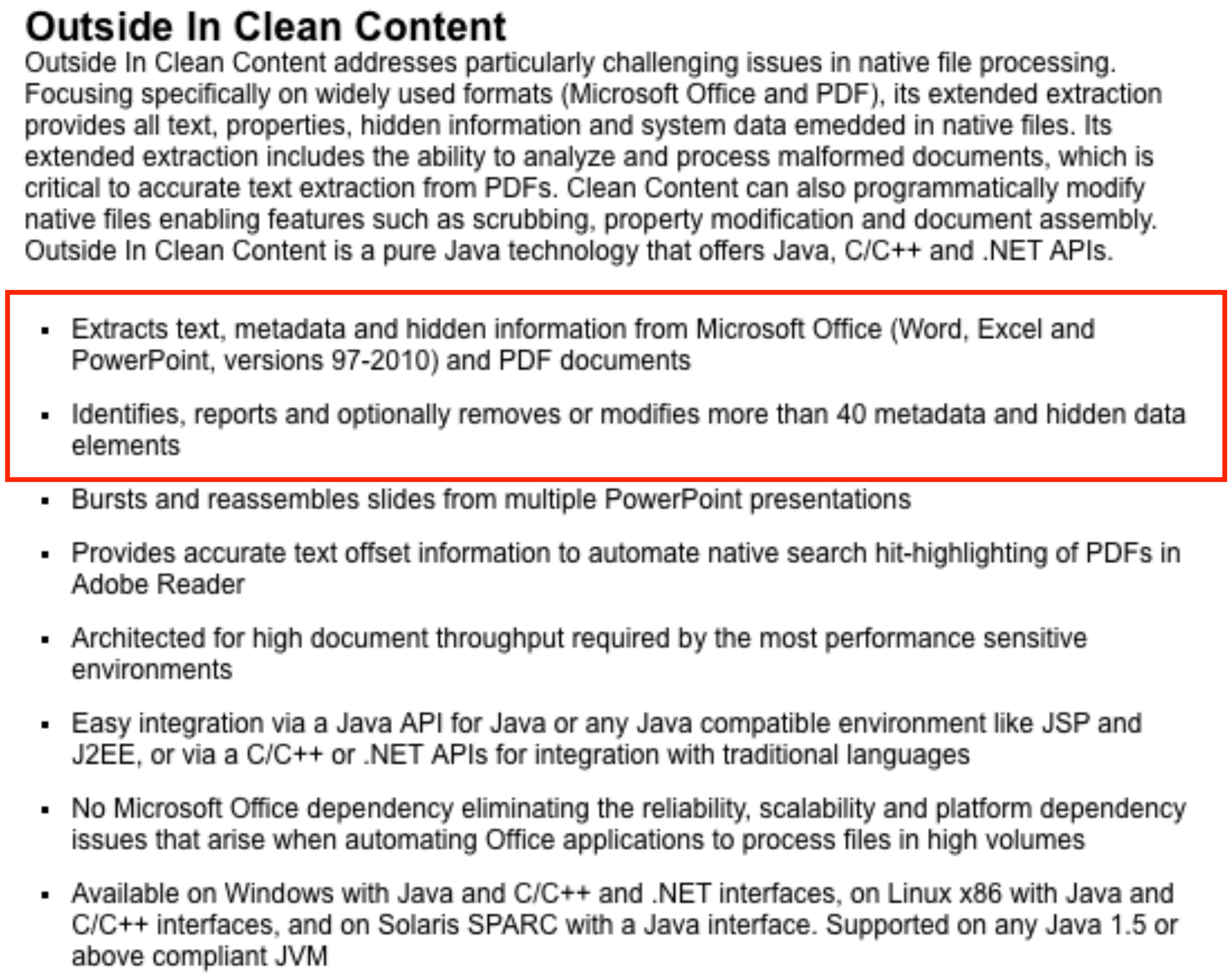
Dark Data

Bioinformatics - Digital Safety

Metadata is data about data. It usually provides additional information. Metadata is essential for most data formats. You can't render a JPEG, send an email, or visit a web site without metadata. Metadata could be used for surveillance purposes, but it also has many common everyday purposes. **Metadata is not scary**.


Bringing the Attack to Your Network
Hacking
The Next Generation
O'REILLY
Nitesh Dhanjani
Billy Rios & Brett Hardin

Bioinformatics - Digital Safety



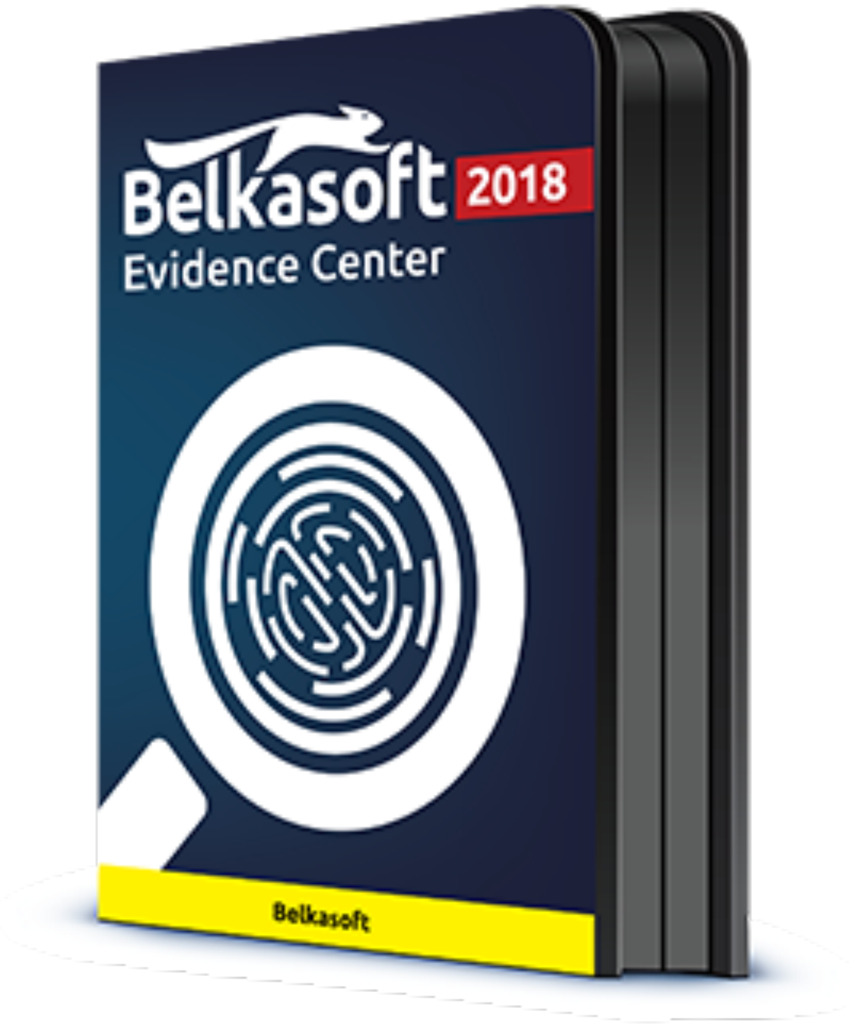
Outside In Clean Content

Outside In Clean Content addresses particularly challenging issues in native file processing. Focusing specifically on widely used formats (Microsoft Office and PDF), its extended extraction provides all text, properties, hidden information and system data emeded in native files. Its extended extraction includes the ability to analyze and process malformed documents, which is critical to accurate text extraction from PDFs. Clean Content can also programmatically modify native files enabling features such as scrubbing, property modification and document assembly. Outside In Clean Content is a pure Java technology that offers Java, C/C++ and .NET APIs.

- Extracts text, metadata and hidden information from Microsoft Office (Word, Excel and PowerPoint, versions 97-2010) and PDF documents
- Identifies, reports and optionally removes or modifies more than 40 metadata and hidden data elements
- Bursts and reassembles slides from multiple PowerPoint presentations
- Provides accurate text offset information to automate native search hit-highlighting of PDFs in Adobe Reader
- Architected for high document throughput required by the most performance sensitive environments
- Easy integration via a Java API for Java or any Java compatible environment like JSP and J2EE, or via a C/C++ or .NET APIs for integration with traditional languages
- No Microsoft Office dependency eliminating the reliability, scalability and platform dependency issues that arise when automating Office applications to process files in high volumes
- Available on Windows with Java and C/C++ and .NET interfaces, on Linux x86 with Java and C/C++ interfaces, and on Solaris SPARC with a Java interface. Supported on any Java 1.5 or above compliant JVM

Bioinformatics - Digital Safety

Belkasoft Evidence Center makes it easy for an investigator to acquire, search, analyse, store and share digital evidence found inside computer and mobile devices. The toolkit will quickly extract digital evidence from multiple sources by analysing hard drives, drive images, cloud, memory dumps, iOS, Blackberry and Android backups, UFED, JTAG and chip-off dumps. Evidence Center will automatically analyse the data source and lay out the most forensically important artefacts for investigator to review, examine more closely or add to report.



Pop Corn Concept

Bioinformatics - **Concepts**

ICYWW*, you totally can make popcorn with a hair straightener!

*In case your were wondering.

Dent Corn

(Zea mays var. indentata)

Flint Corn

(Zea mays var. indurata)

Popcorn

(Zea mays var. everta)

Sweet Corn

(Zea mays convar. saccharata var. rugosa)

Flour corn

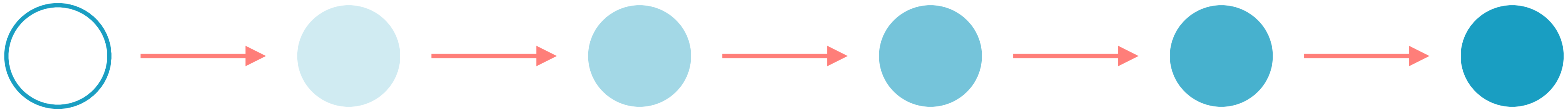
(Zea mays var. amylacea)

Bioinformatics - Concepts

Each kernel of popcorn contains a certain amount of moisture and oil. Unlike most other grains, the outer hull of the popcorn kernel is both strong and impervious to moisture and the starch inside consists almost entirely of a hard type. As the oil and water within the kernel are heated, they turn the moisture in the kernel into pressurized steam. Under these conditions, the starch inside the kernel gelatinizes, softens, and becomes pliable. The internal pressure of the entrapped steam continues to increase until the breaking point of the hull is reached: a pressure of approximately **930 kPa** and a temperature of **180 °C**. The hull thereupon ruptures rapidly and explodes, causing a sudden drop in pressure inside the kernel and a corresponding rapid expansion of the steam, which expands the starch and proteins of the endosperm into airy foam. As the foam rapidly cools, the starch and protein polymers set into the familiar crispy puff. Special varieties are grown to give improved popping yield. Though the kernels of some wild types will pop, the cultivated strain is Zea mays everta, which is a special kind of flint corn.

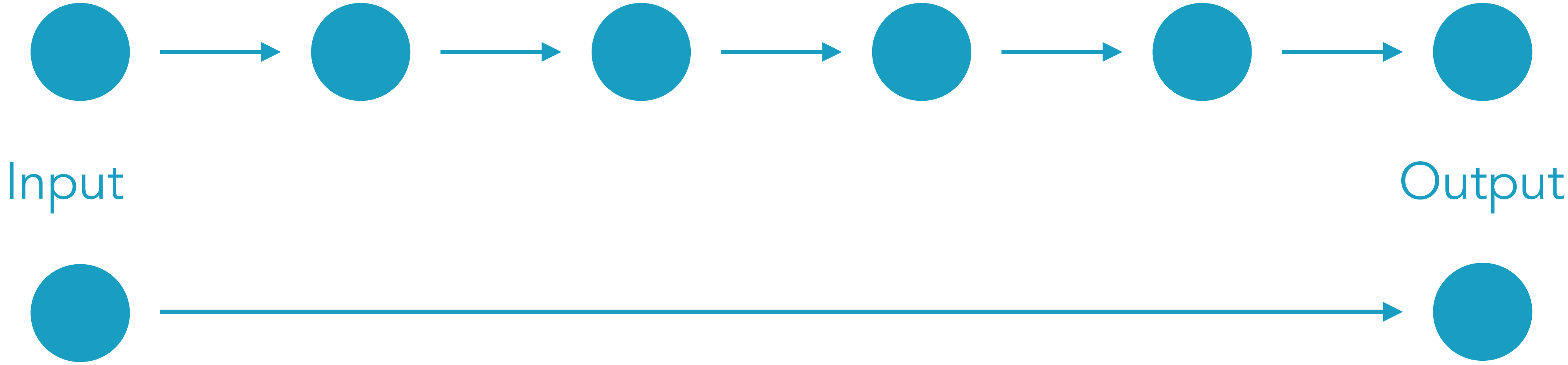
Source: Wikipedia

Bioinformatics - Concepts

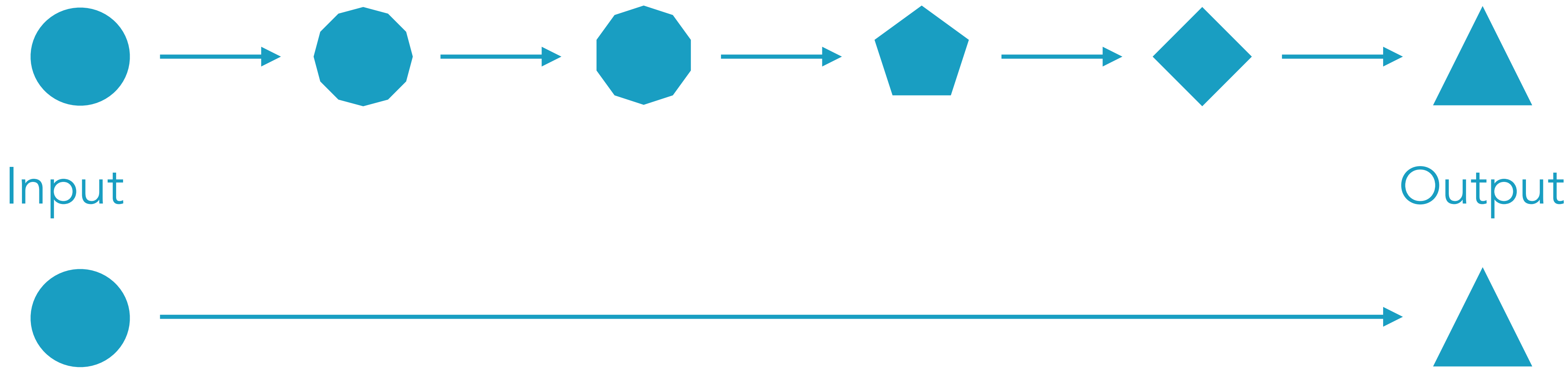


Chain Concept

Bioinformatics - Concepts



Bioinformatics - Concepts


Input
Output