

MODULE 03 – HIGH PERFORMANCE COMPUTING
PROCESSING BIG DATA IN UNIX ENVIRONMENT

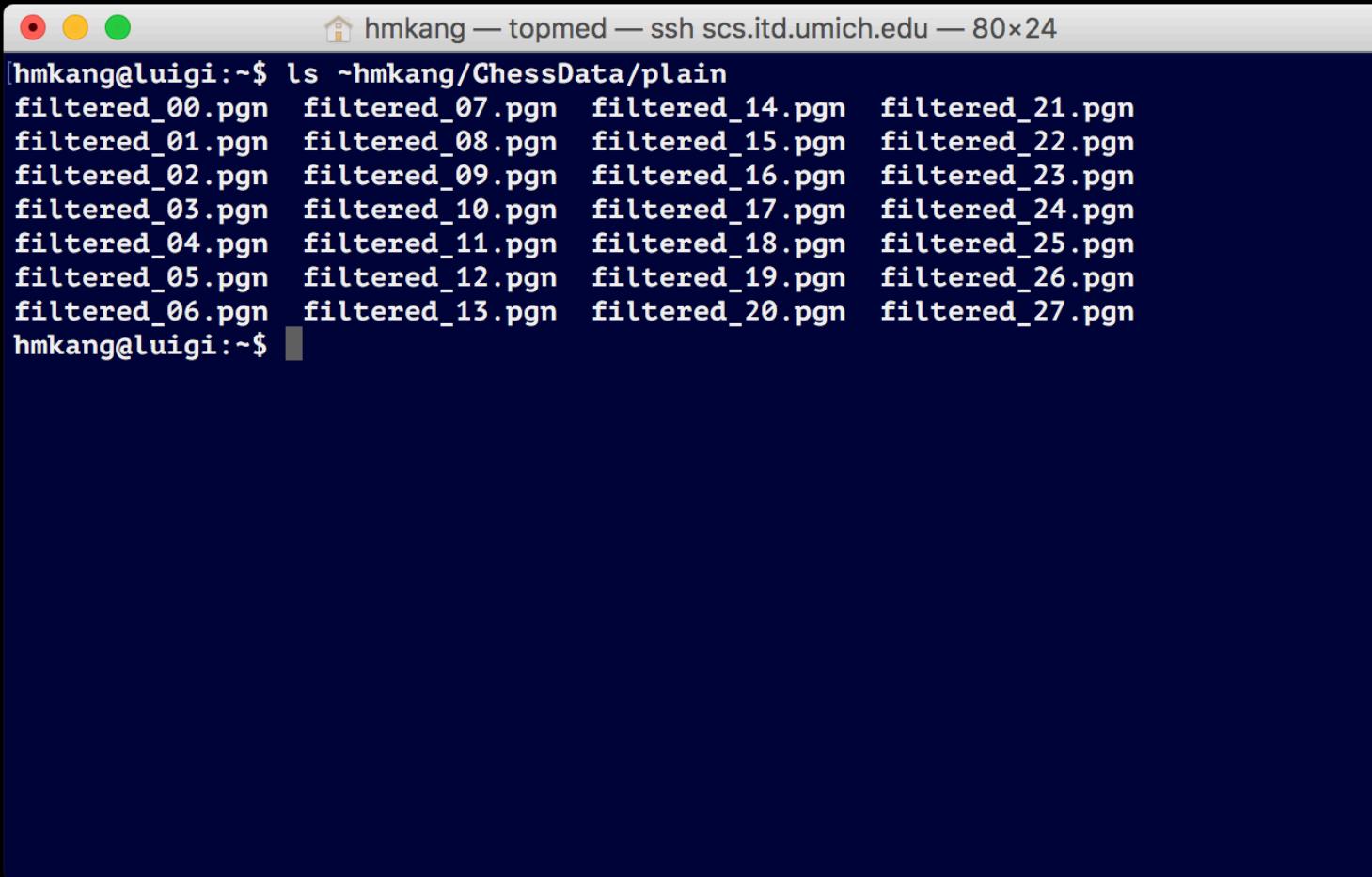
Raw Data is Messy



Raw Data is Messy – FASTQ files

```
1.TCA.454Reads.fastq
@HC9D00P01AN1VB rank=0000246 x=156.0 y=3301.0 length=309
ACACATACGCACTGGCGTAAAGGGCGCGCAGGCGGTAGAGGCGTCGGTGCTCAAAGTCCACCGCTTAACGGTGGAGGCGTG
+HC9D00P01AN1VB
FFFFFFFFFFFFGD554A6911144442AAABDFFFIIIIIIIIIIHHHFFFFFFA@@CFFDFFFDFC???CCFFFFFFF
@HC9D00P01AWYAE rank=0000402 x=258.0 y=772.0 length=373
ACACATACGCACTGGGCATAAAGGGCACGTAGGCGGATTGTAAGTCAGGGGTGAATCCGGGGCGTCAACCTCGGAAC TGCT
+HC9D00P01AWYAE
IIIIIIIIIIHHHII;666HHHIIIIIIIIIIICCIIEEEFDC2//.<-//93.....--9?CCCCEFECCIIIID
@HC9D00P01A3C8R rank=0000675 x=331.0 y=1081.0 length=373
ACACATACGCACTGGGTAAAGGGTGCCTAGGCGGGCTTTAAGTCAGGGGTGAAATCCTGGAGCTCAACTCCAGAACTGCCT
+HC9D00P01A3C8R
IIIIIIIIII3...//...--4AIIIECCE466GH974EEIAC@.0004.000>9@CEEI IIIIIIIIIIIIIIHHH
@HC9D00P01AW8TJ rank=0000926 x=261.5 y=2133.0 length=373
ACACATACGCACTGGGTGTAAAGCGCACGTAGGCGGATTGCTAAGTCAGGGGTGAAATCCTGGAGCTCAACTCCAGAACTGCCT
+HC9D00P01AW8TJ
IIIIHHHIIIIHHHII;;IIIIIIIIIIIIIIIIIIII@@@@H4;;?IIIIIIIIIIIIIIIIIIIIHHH
@HC9D00P01AUI8Y rank=0000952 x=230.0 y=2656.0 length=372
ACACATACGCACTGGGCATAAAGAGAGCGCGTAGGCGGGCTTGTAGTCGAGTGTGAAAGCCCTGGCTTAACCCGGGAAGCGCGC
+HC9D00P01AUI8Y
IIIIIIIIIIHHHII;;IIIIIIIIIIIIIIIIIIII?666DHIIHFEIIIIIC;;555994?FIGI
@HC9D00P01AUAIW rank=0000977 x=228.0 y=226.0 length=372
```

Today's Example Data



A screenshot of a macOS terminal window titled "hmkang — topmed — ssh scs.itd.umich.edu — 80x24". The window shows the command "ls ~hmkang/ChessData/plain" being run, listing 28 files named "filtered_00.pgn" through "filtered_27.pgn".

```
[hmkang@luigi:~$ ls ~hmkang/ChessData/plain
filtered_00.pgn  filtered_07.pgn  filtered_14.pgn  filtered_21.pgn
filtered_01.pgn  filtered_08.pgn  filtered_15.pgn  filtered_22.pgn
filtered_02.pgn  filtered_09.pgn  filtered_16.pgn  filtered_23.pgn
filtered_03.pgn  filtered_10.pgn  filtered_17.pgn  filtered_24.pgn
filtered_04.pgn  filtered_11.pgn  filtered_18.pgn  filtered_25.pgn
filtered_05.pgn  filtered_12.pgn  filtered_19.pgn  filtered_26.pgn
filtered_06.pgn  filtered_13.pgn  filtered_20.pgn  filtered_27.pgn
hmkang@luigi:~$ ]
```

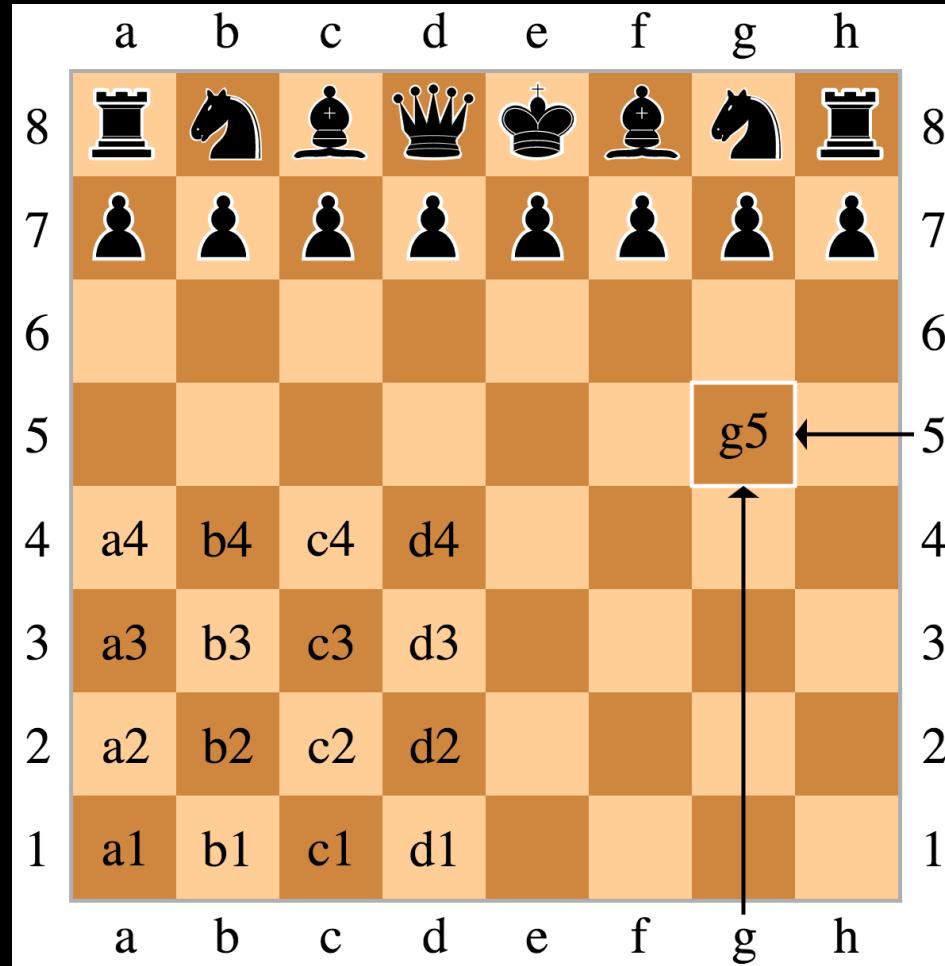
```
zless ~hmkang/ChessData/plain/filtered_00.pgn
```

```
hmkang — topmed — ssh scs.itd.umich.edu — 80x24
[Event "BL 0708 SK Zehlendorf - OSC Baden Baden"]
[Site "?"]
[Date "2007.10.20"]
[Round "1.1"]
[White "Svidler, Peter"]
[Black "Maksimenko, Andrei"]
[Result "1-0"]
[WhiteElo "2735"]
[BlackElo "2505"]
[ECO "C63"]

1. e4 e5 2. Nf3 Nc6 3. Bb5 f5 4. d3 d6 5. exf5 Bxf5 6. d4 exd4 7. 0-0 Bd7 8. Re1
+ Be7 9. c3
Nf6 10. cxd4 0-0 11. Nc3 Bg4 12. Be2 Qd7 13. h3 Bh5 14. d5 Bxf3 15. Bxf3 Ne5 16.
Be2 c6
17. Be3 Kh8 18. a4 a5 19. Rb1 Rfe8 20. Bf1 Bf8 21. b4 axb4 22. Rxb4 Qf7 23. Qb3
Nxd5 24. Rxb7
Qg6 25. Nxd5 cxd5 26. Qxd5 Rec8 27. a5 1-0

[Event "BL 0708 SK Zehlendorf - OSC Baden Baden"]
[Site "?"]
[Date "2007.10.20"]
[Round "1.2"]
/afs/umich.edu/user/h/m/hmkang/ChessData/plain/filtered_00.pgn
```

Chessboard Notation



Questions

- How big is each file?
- How many lines does each file have?
- What is the overall size of the directory?

Questions

- How big is each file?

```
$ ls -lh ~hmkang/ChessData/plain/
```

- How many lines does each file have?

```
$ wc -l ~hmkang/ChessData/plain/
```

- What is the overall size of the directory?

```
$ du -sh ~hmkang/ChessData/plain/
```

More **sophisticated** questions

- How many games in total?
- How many times white and black won?
- How many unique last names of players exists in the files?

More sophisticated questions

- How many games in total?

```
cat ~hmkang/ChessData/plain/* | grep Result | wc -l
```

- How many times white and black won?

```
cat ~hmkang/ChessData/plain/* | grep Result | sort | uniq -c
```

- How many unique last names of players exists in the files?

```
cat ~hmkang/ChessData/plain/* | grep -E '^\\[(White|Black) " | cut -d '"' -f 2 | sort | uniq | wc -l
```

Impact of data **compression**

```
[hmkang@luigi:~$ du -sh ~hmkang/ChessData/*
276M  /afs/umich.edu/user/h/m/hmkang/ChessData/bz2
429M  /afs/umich.edu/user/h/m/hmkang/ChessData/gz
1.4G  /afs/umich.edu/user/h/m/hmkang/ChessData/plain
hmkang@luigi:~$ ]
```

Space-time tradeoff

```
hmkang — topmed — ssh scs.itd.umich.edu — 80x24
[hmkang@luigi:~$ cat ~hmkang/ChessData/plain/* | grep -E "^\[(White|Black) " | cut -d '"' -f 2 | sort | uniq | wc -l
265264
0:19.40 elapsed, 22.903 u, 1.087 s, cpu 123.6%, 0 swaps, 0 rds, 134904 wrts, pgs : 0 avg., 25546 max.
[hmkang@luigi:~$ zcat ~hmkang/ChessData/gz/* | grep -E "^\[(White|Black) " | cut -d '"' -f 2 | sort | uniq | wc -l
grep: Unmatched [ or [
0
[hmkang@luigi:~$ zcat ~hmkang/ChessData/gz/* | grep -E "^\[(White|Black) " | cut -d '"' -f 2 | sort | uniq | wc -l
265244
0:39.77 elapsed, 53.083 u, 1.552 s, cpu 137.3%, 0 swaps, 0 rds, 134904 wrts, pgs : 0 avg., 25546 max.
[hmkang@luigi:~$ bzcat ~hmkang/ChessData/bz2/* | grep -E "^\[(White|Black) " | cut -d '"' -f 2 | sort | uniq | wc -l
265264
1:42.75 elapsed, 114.869 u, 3.590 s, cpu 115.2%, 0 swaps, 0 rds, 149640 wrts, pgs : 0 avg., 25546 max.
hmkang@luigi:~$
```

Iterating over files using `xargs`

- How many games does each file have?

```
ls ~hmkang/ChessData/bz2 | xargs -t -I {} sh -c  
"bzcat ~hmkang/ChessData/bz2/{} | grep Result |  
wc -l"
```

```
seq -w 0 27 | xargs -t -I {} sh -c "bzcat  
~hmkang/ChessData/bz2/filtered_{}.pgn.bz2 | grep  
Result | wc -l"
```

Even more sophisticated questions

- Count the number of wins, draws, and losses for each player. Who is the best player based on the record?
- Resolve variations of player's name by taking the last name and the first initial in a case insensitive way.
- What is the most frequent first move?

You may need to learn script languages..

- **bash script**

<http://www.tldp.org/LDP/Bash-Beginners-Guide/html/>

- **sed and awk script**

<http://www.tldp.org/LDP/abs/html/sedawk.html>

- **python**

<https://www.codecademy.com/learn/python>

- **perl, php, or javascript..**

**Ask more questions
about the data!**