MODULE 03 – HIGH PERFORMANCE COMPUTING

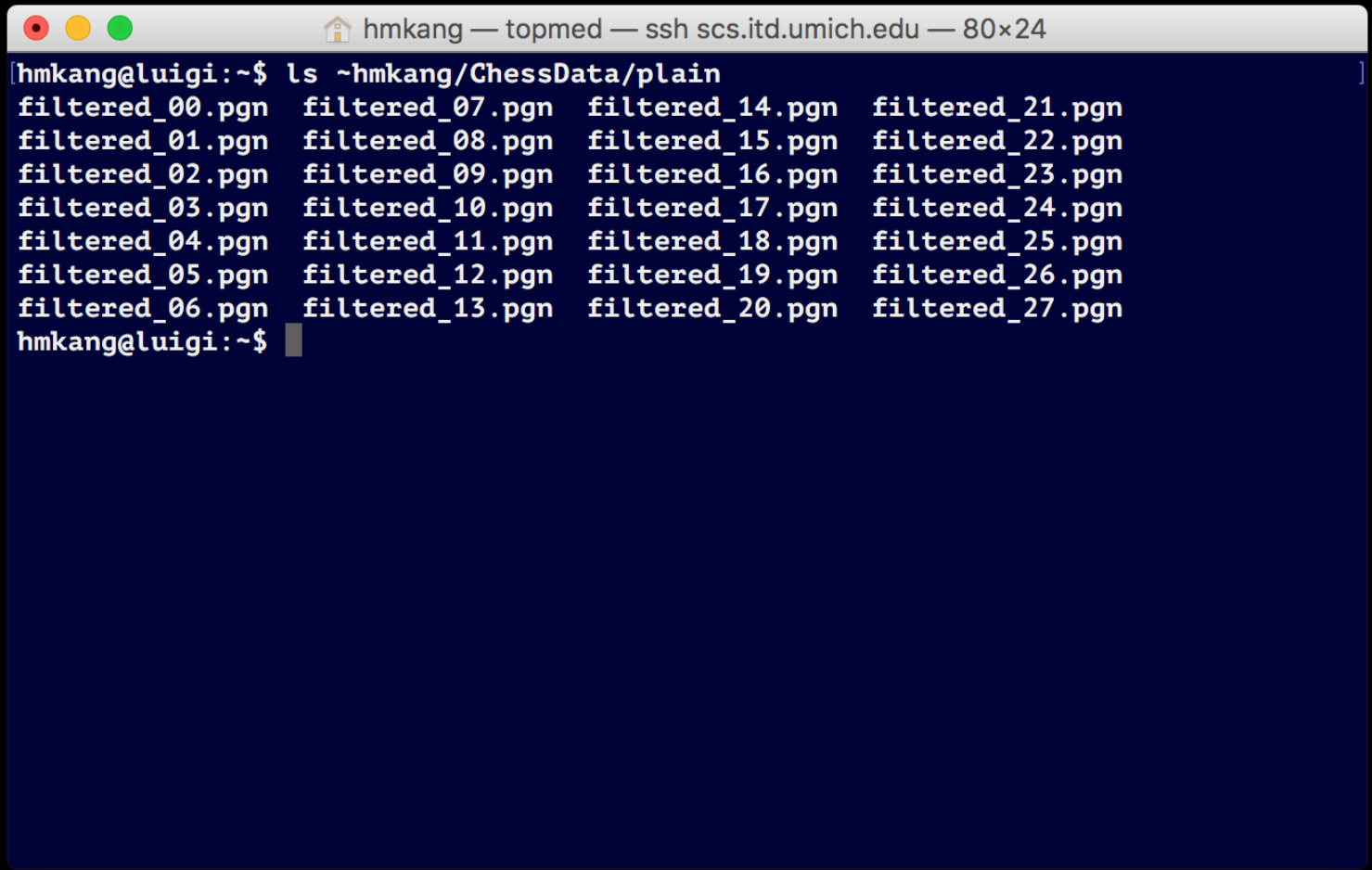
PROCESSING BIG DATA IN UNIX ENVIRONMENT

HYUN MIN KANG @ UNIVERSITY OF MICHIGAN

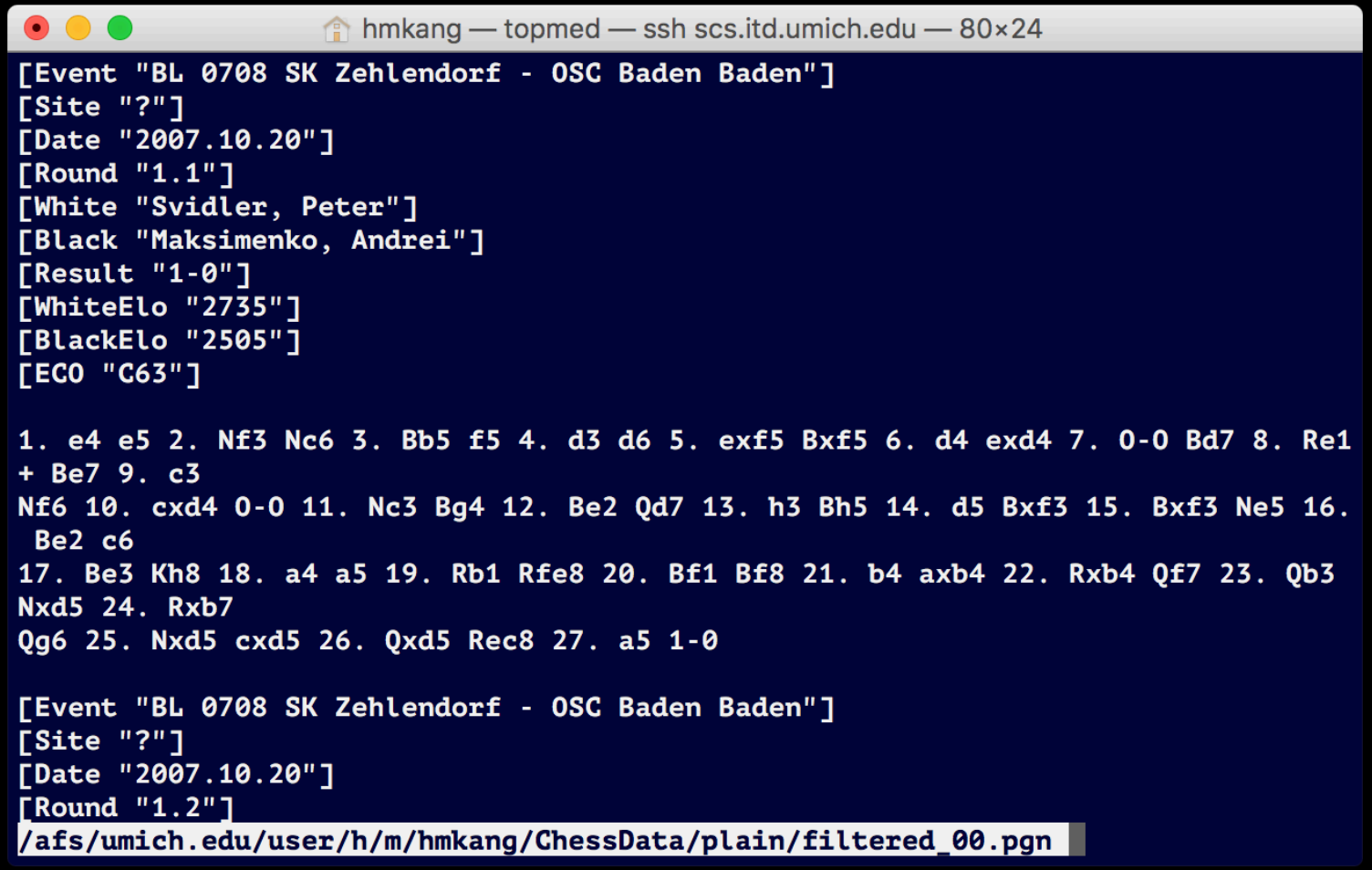
Raw Data is Messy

Raw Data is Messy – FASTQ files

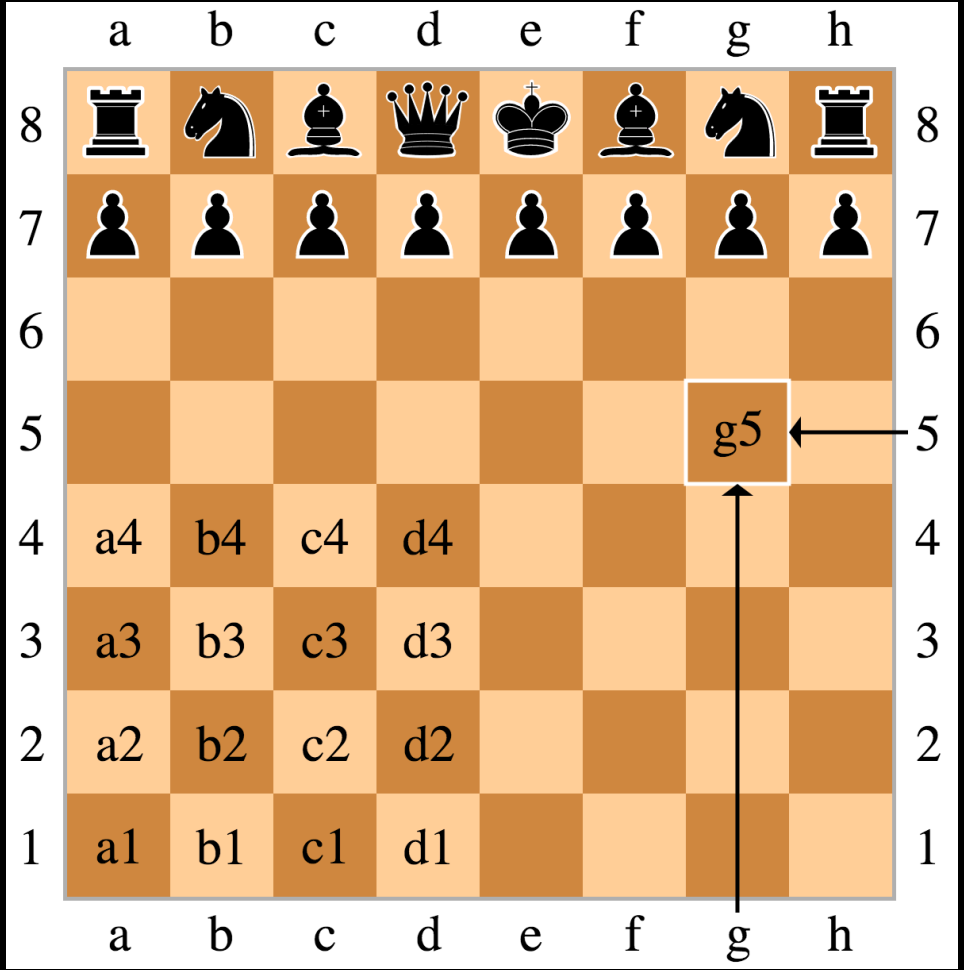
Today's Example Data



zless ~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
```

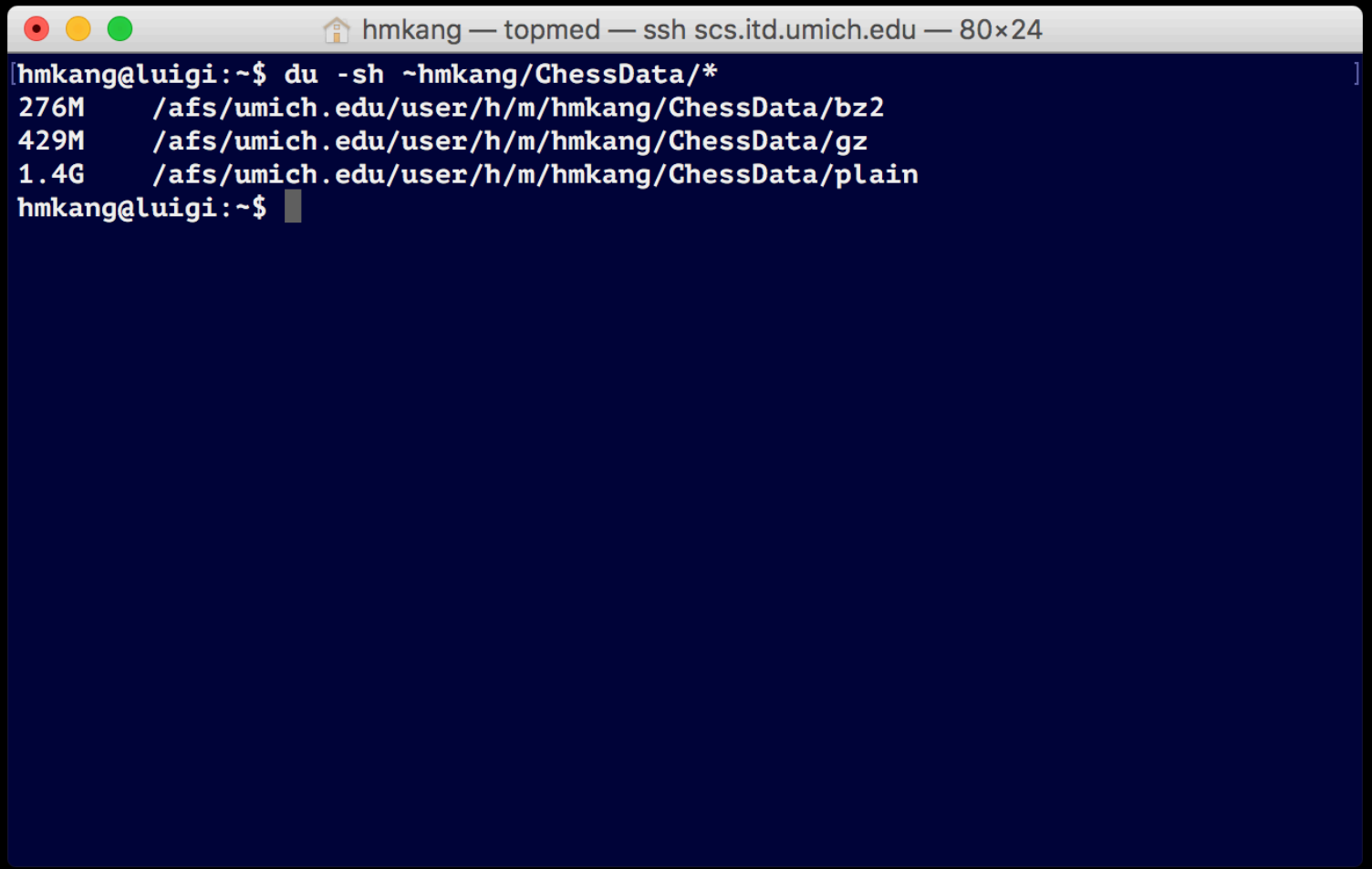
- How many times white and black won?

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

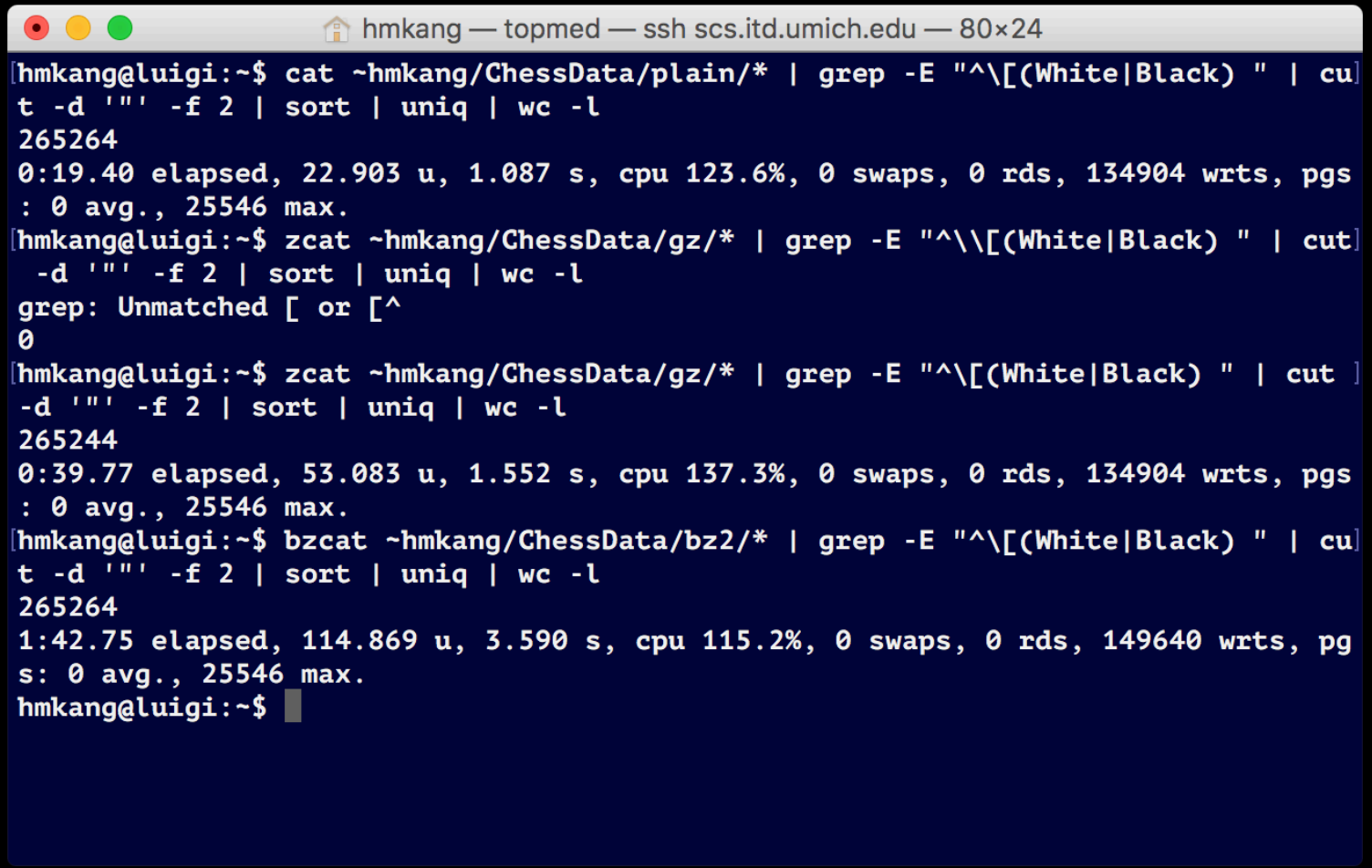
- How many unique last names of players exists in the files?

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

Impact of data compression



Space-time tradeoff



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!