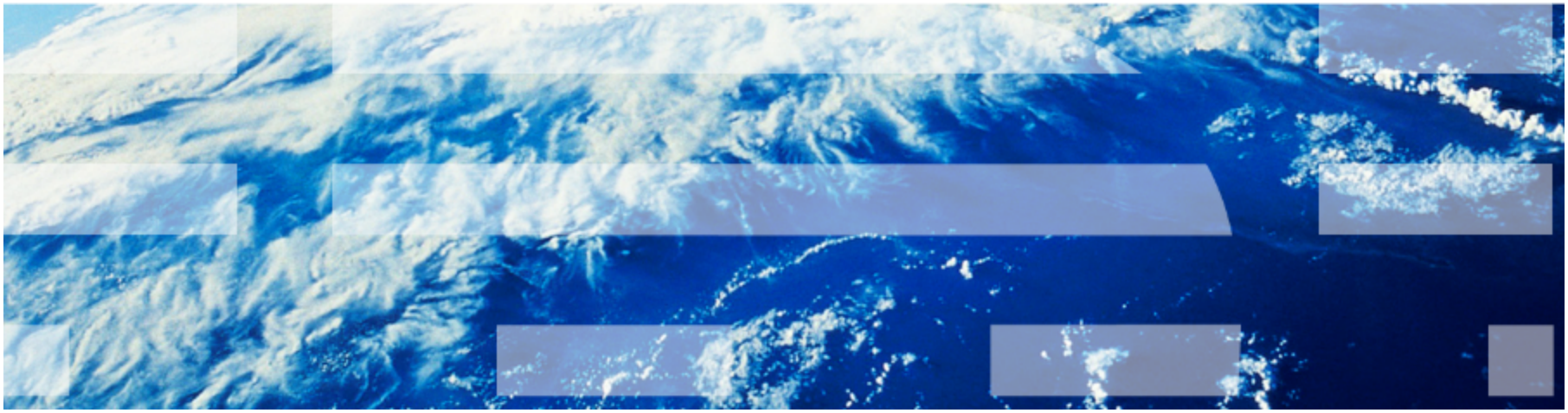
COLUMBIA UNIVERSITY

E6893 Big Data Analytics:

Game Outcome Analysis

Team Members: **Raymond Barker** (rjb2150)



December 9, 2014

Motivation

COLUMBIA UNIVERSITY

The goal of this project is to be able to answer to following question:

Given a game state, who will win?

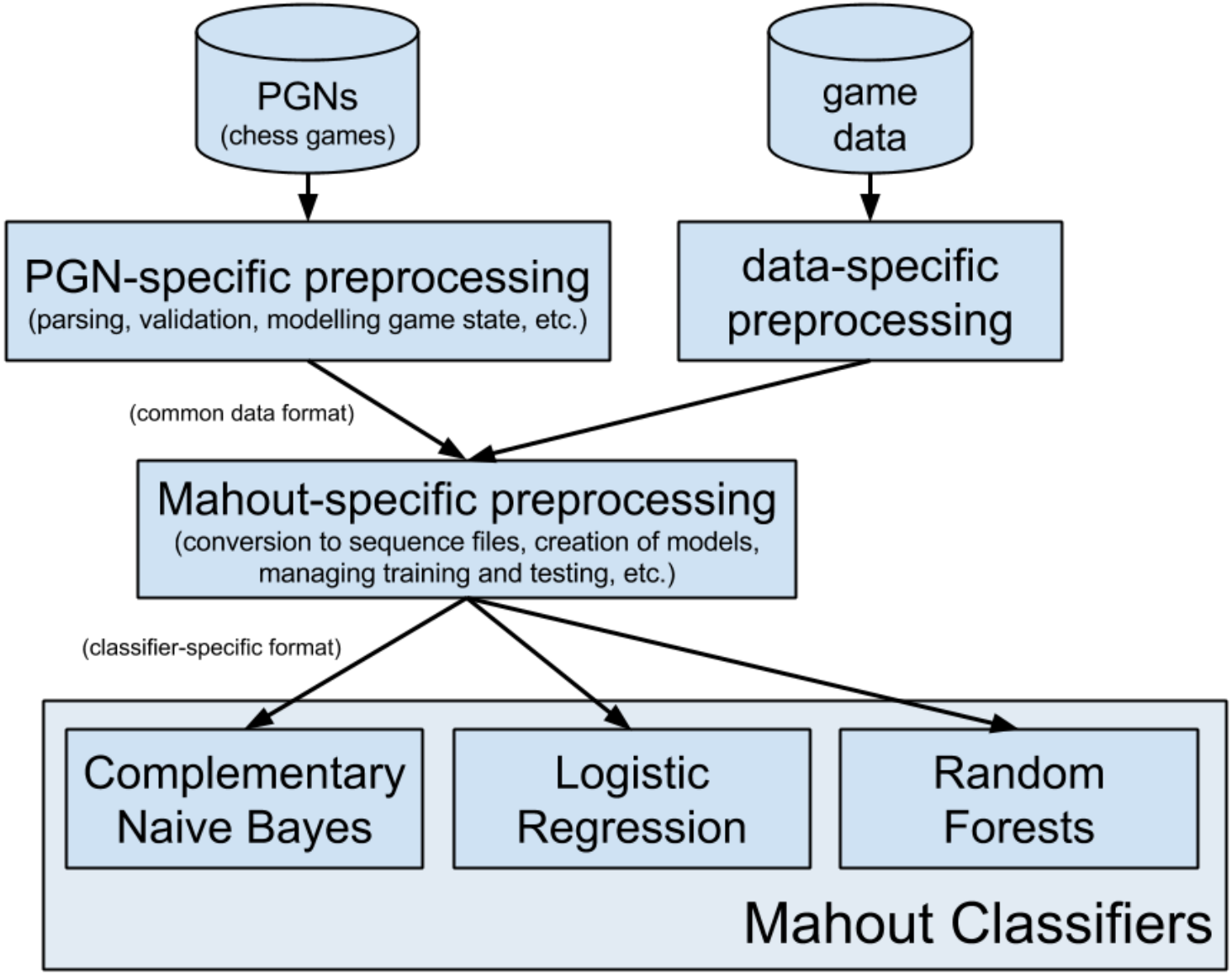
Specifically, this project focuses on games that are:

- two-team
- each team can be represented as a multiset of predefined members

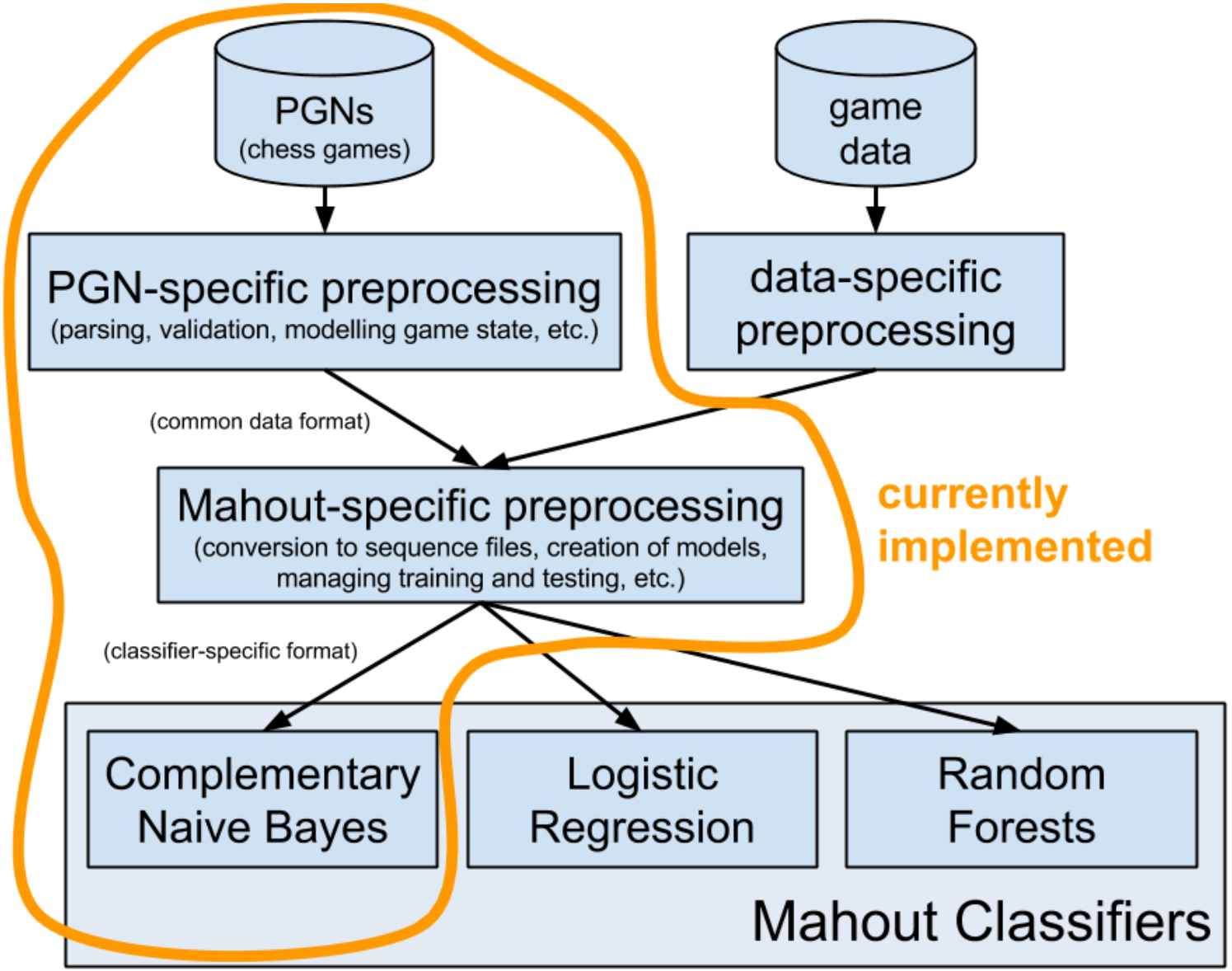
Many games fall into this category:

- chess (each team is composed of N pawns, M rooks, etc.)
- deck-building games (each team is composed of predefined cards)
- MMORPGs (each team is composed of various “classes”)

System Overview



System Overview



Dataset

COLUMBIA UNIVERSITY

This project focuses primarily on analyzing chess game data:

- chess games are widely available in the public domain
- there is a huge amount of existing chess theory to compare to

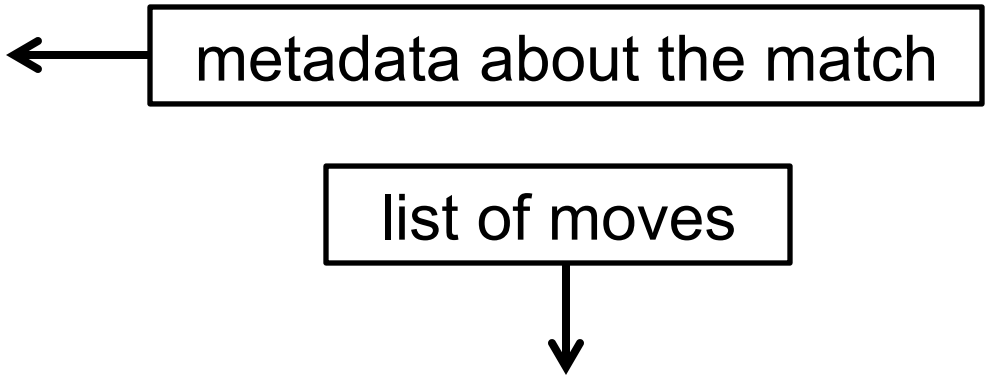
Specifically, I'll be using a corpus of chess games in the public domain and compiled by Norman Pollock. (see http://hoflink.com/~npollock/chess.html)

dataset id	gm2006.pgn
number of games	74,726
number of players	1,227
minimum player Elo rating	2475
years included	2006 - 2014
gameplay restrictions	no blitz or correspondence games

Portable Game Notation: Example File

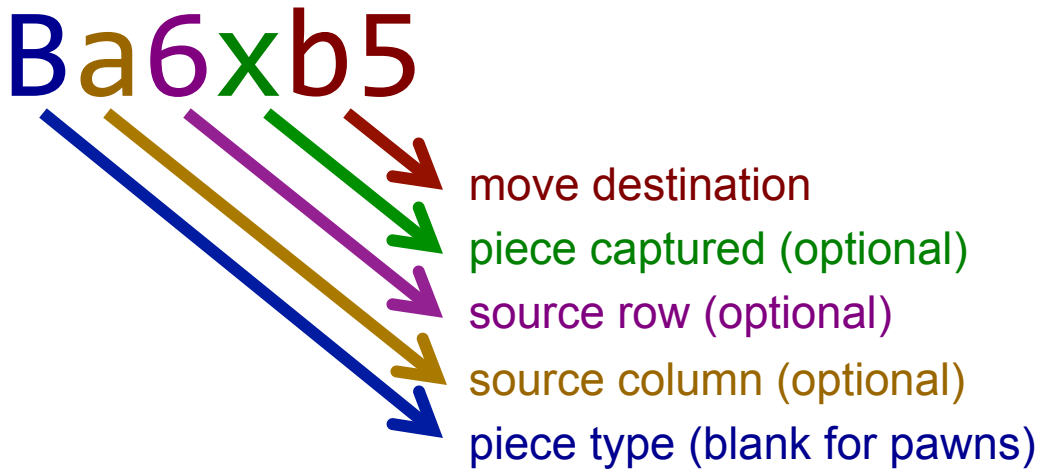
There is a standardized format for storing chess games known as Portable Game Notation (PGN), for example:

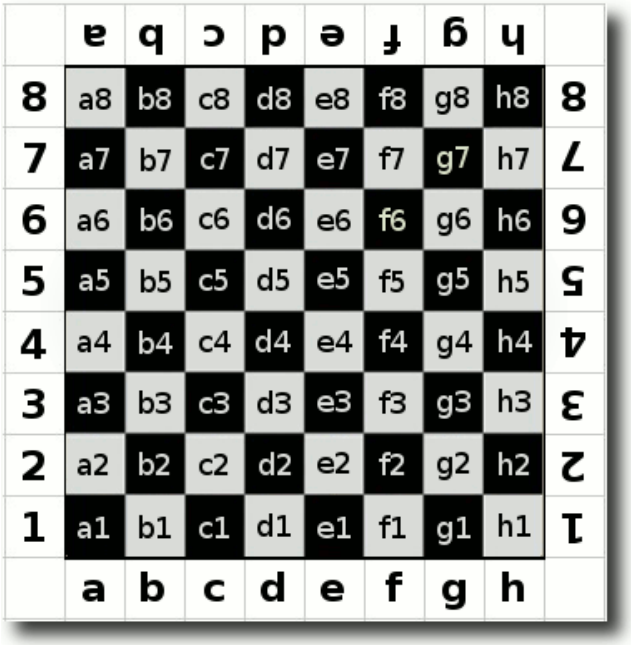
```
[Event "1st Grand Europe Open"]
[Date "2012.06.10"]
[White "Tikkanen, H."]
[Black "Grigoryan, K2"]
[WhiteElo "2566"]
[BlackElo "2517"]
[Result "1-0"]
```




```
1. c4 Nf6 2. Nc3 g6 3. d4 Bg7 4. e4 d6 5. Nf3 O-O 6. Be2 e5 7. O-O Nc6
8. d5 Ne7 9. Ne1 Nd7 10. Bd2 f5 11. Nd3 Nf6 12. f3 c6 13. b4 cxd5
14. cxd5 Qb6+ 15. Nf2 Bd7 16. Qb3 Rac8 17. Nd1 f4 18. Nb2 Qd8 19. Rfc1 g5
20. Nc4 Ne8 21. b5 Ng6 22. b6 Ra8 23. bxa7 Rxa7 24. Nb6 h5 25. Nxd7 Qxd7
26. Qb6 Ra8 27. Bb5 Qe7 28. Bxe8 Rfxe8 29. Rc7 Qd8 30. a4 g4 31. fxg4 hxg4
32. Nxg4 Nh4 33. h3 Qg5 34. Qxd6 Nxg2 35. Kxg2 f3+ 36. Kh1 Qxd2 37. Rg1 Ra6
38. Qd7 1-0
```

Portable Game Notation: Parsing Moves





"use the bishop at A6 to capture a piece at B5"

- Note that many of these fields are optional
  - e.g., the move "e4" means "move [the only pawn that's able to] to E4"
- There is additional syntax for less common moves
  - "+" for check, "0-0" for castling, "=Q" for pawn promotion, etc.
- PGN is designed to be convenient for humans as well, so "obvious" information about moves is omitted; however, this means that PGN parsers must understand the rules of chess

Preliminary Results

The dataset contains 74,726 total games:

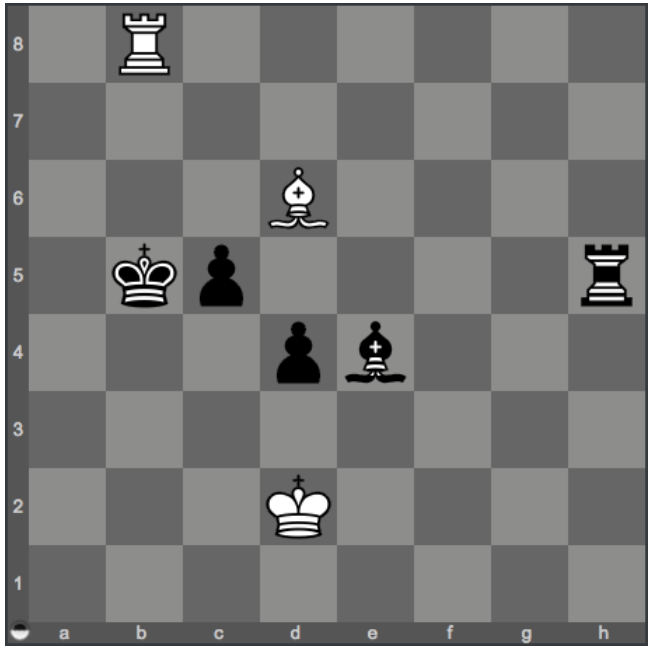
- 25,089 wins for white
- 15,110 wins for black
- 34,527 ties

Of the non-tie games, 32,039 (~80%) were used as training points and 8,160 (20%) were used as testing points with a Naïve Bayes classifier:

	correctly classified	total	accuracy
all	5,755	8,160	70.5%
white	3,590	5,097	70.4%
black	2,165	3,063	70.7%

Difficulties

- Naïve Bayes features are treated independently, so strong combinations of pieces are not reflected in the model
    - e.g., bishops are worth more when used together
- positional data not captured in features
    - a player can be losing despite having more pieces
- the dataset contains only games between highly-ranked players
    - skilled players won't throw "expensive" pieces away, and so many games are differentiated by pawns
    - pawns are therefore over-valued in the model



Next Steps

COLUMBIA UNIVERSITY

- Adding support for logistic regression
- Adding support for random forests
- adding support for another source of game data

or

- removing the generality and optimizing for chess
  - *e.g., adding features to capture piece positions*

Thank you!

Any questions?