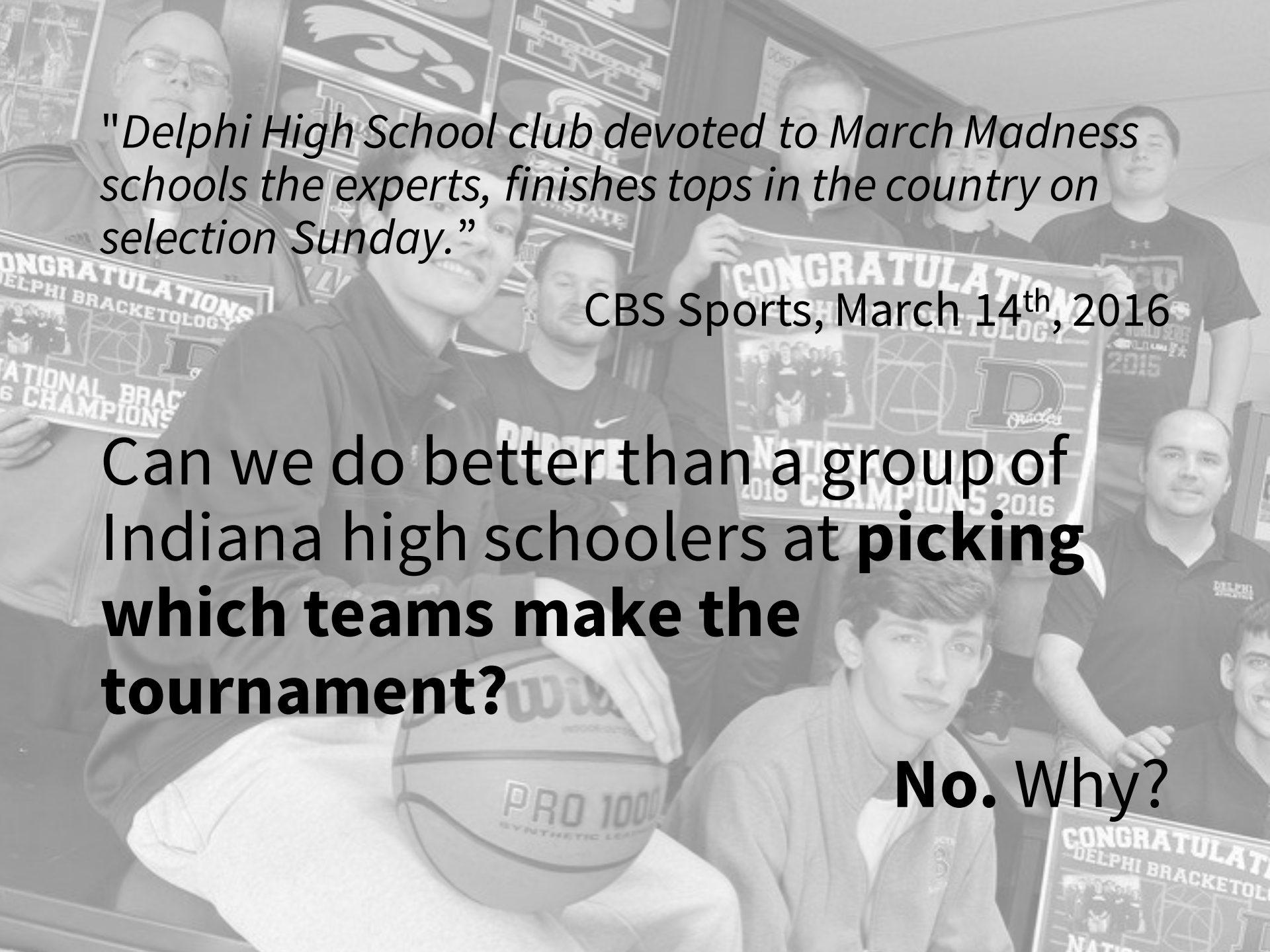
Bracketology's Black Box: Can We Predict the Selection Sunday Committee?

Dan Trunzo and Libby Scholz

May 27, 2016



"Delphi High School club devoted to March Madness schools the experts, finishes tops in the country on selection Sunday."

CBS Sports, March 14th, 2016

Can we do better than a group of Indiana high schoolers at **picking which teams make the tournament?**

No. Why?

Selection Sunday

32 teams get automatic bids as conference champions

36 teams are selected for "at-large" berths through…

an initial ballot

multiple rounds where progressively fewer votes are needed to get a berth

The committee follows rules for **seeding** teams

Selection Criteria: Our Best Guess

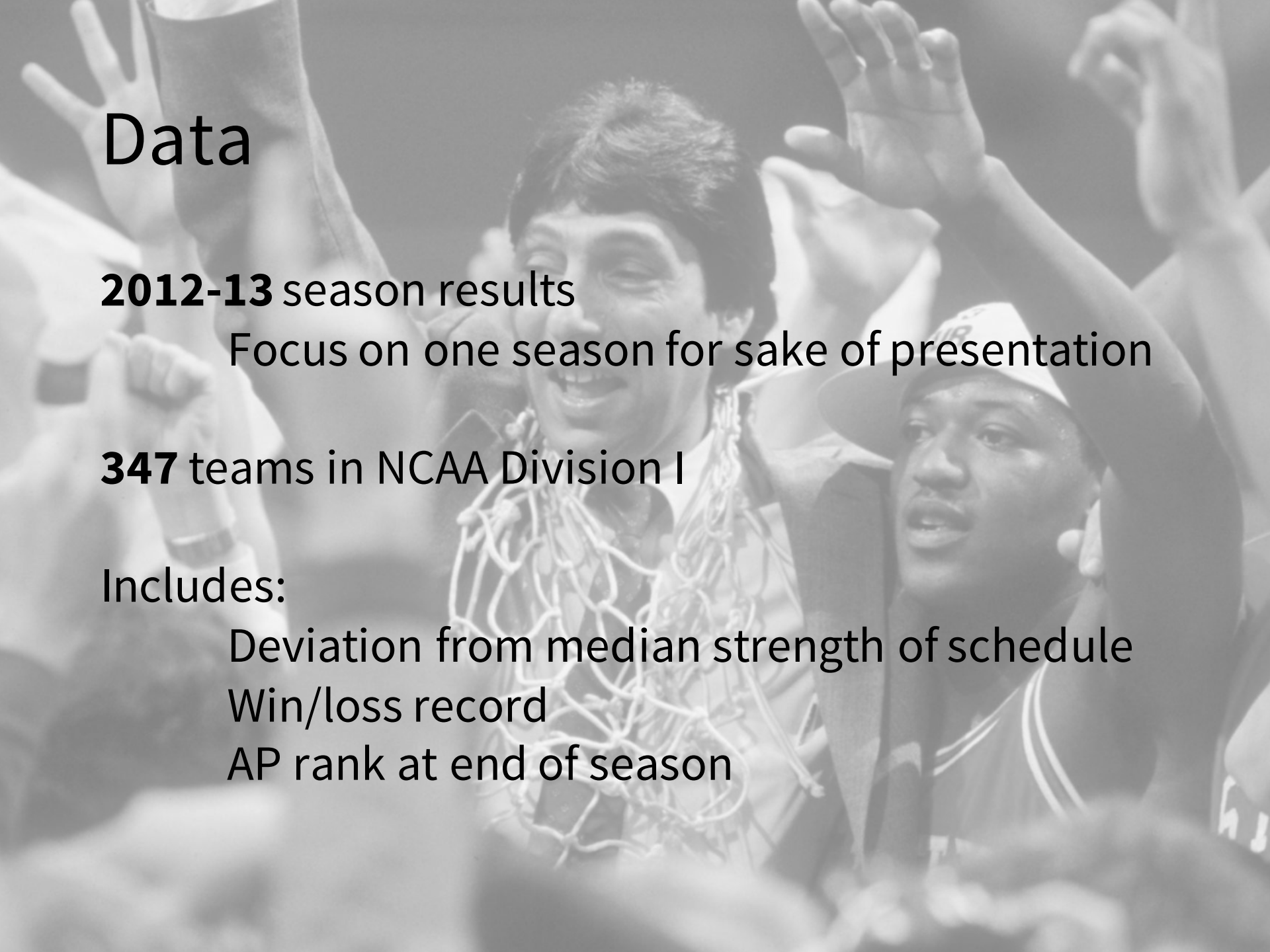
Quality wins matter more than losses to good teams

"Ratings Percentage Index:"

0.25(your winning percentage) + 0.5(average opponents' winning percentage) + 0.25(your average opponents' opponents' winning percentage)

Strength of conference isn't considered separately from strength of schedule

Geography matters most in tournament placement



Data

2012-13 season results

Focus on one season for sake of presentation

347 teams in NCAA Division I

Includes:

Deviation from median strength of schedule

Win/loss record

AP rank at end of season

Methodology

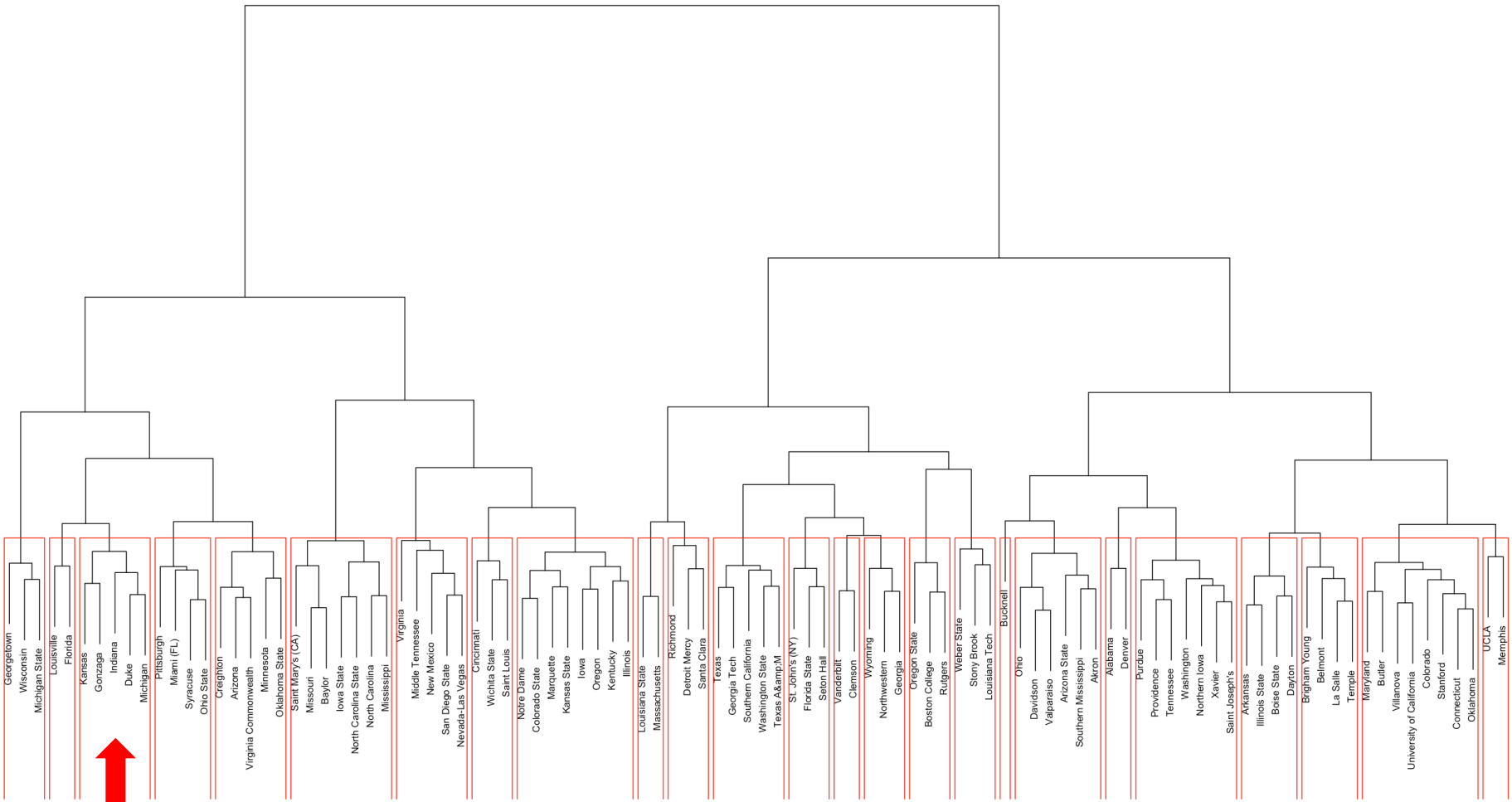
Tried **Bradley-Terry**, **logistic regression**

- Arbitrary betas
- Not specific enough

Use **k-means** and a **dendrogram** where each cluster is a group of teams with roughly similar seeds based on Euclidian distance of various statistics

- Applying method from *Machine Learning for Social Scientists*!

Seeding Dendrogram



Teams by Cluster

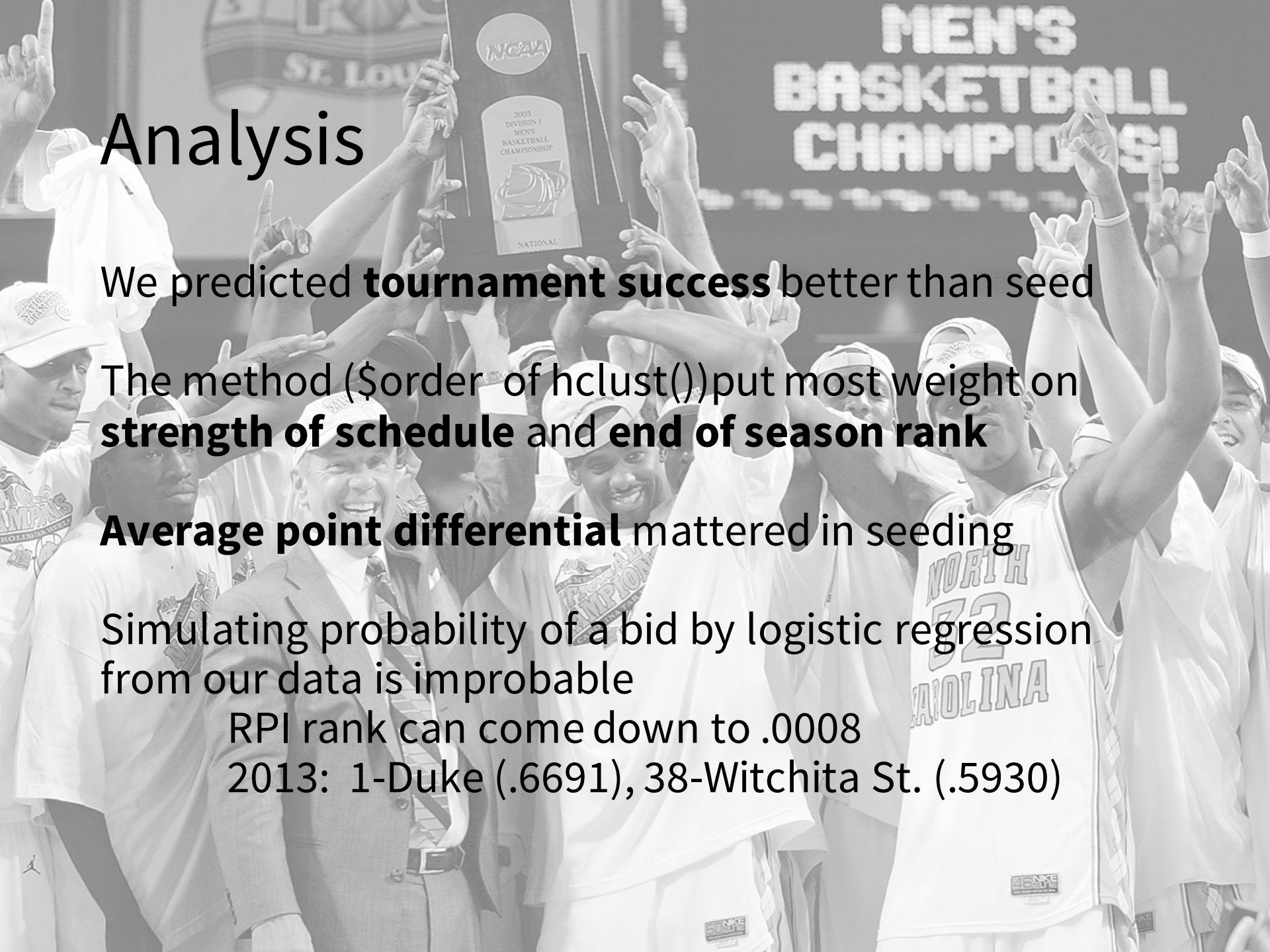
top cluster

Overall:

Predicted **62** out of **68 teams** in the tournament **(91.2%)**

Top teams:

Our Seeding	Actual Tournament Rank	Elite Eight Teams
1) Indiana	3 (1-seed)	Wichita St.
2) Syracuse	16 (4-seed)	Syracuse
3) Ohio State	8 (2-seed)	Ohio St.
4) Florida	10 (3-seed)	Florida
5) Duke	6 (2-seed)	Duke
6) Michigan	13 (4-seed)	Michigan
7) Louisville	1 (1-seed)	Louisville
8) Kansas	2 (1-seed)	Marquette



Analysis

We predicted **tournament success** better than seed

The method ($order of hclust())put most weight on **strength of schedule** and **end of season rank**

Average point differential mattered in seeding

Simulating probability of a bid by logistic regression from our data is improbable

RPI rank can come down to .0008

2013: 1-Duke (.6691), 38-Witchita St. (.5930)

Room for Improvement

Technical flaws:

- Opponents' opponents' record is not included in strength of schedule value
- Missing marquee wins
- Hot streaks

Qualitative flaws:

- Team reputation
- Name recognition probably matters

Bottom line: we're trying to model a small committee of humans with a computer

Potential Future Projects

Predict **seed** better

- Consider what happens when you win the regular season conference but lose in the conference tournament

Improving through **miles to tournament site**

Adding qualitative variables

- Team revenue
- Historical performance

Questions?