SDC Baseball

Gregory J. Matthews[1]
Pétala Gardênia da Silva Estrela Tuy[1]
Robert Arthur[2]

[1]Department of Mathematics and Statistics
Loyola University Chicago
@statsinthewild
[2]FiveThirtyEight.com
New York, NY
@no_little_plans

Skidmore College - June 23, 2016

Outline

Motivation

- Dissertation topic was statistical disclosure control.
- What happens when someone has a subset of true data?
- Simple (but real) example: 3 students in a class.
- Looking for a more realistic example.....

Hall of Fame Voting

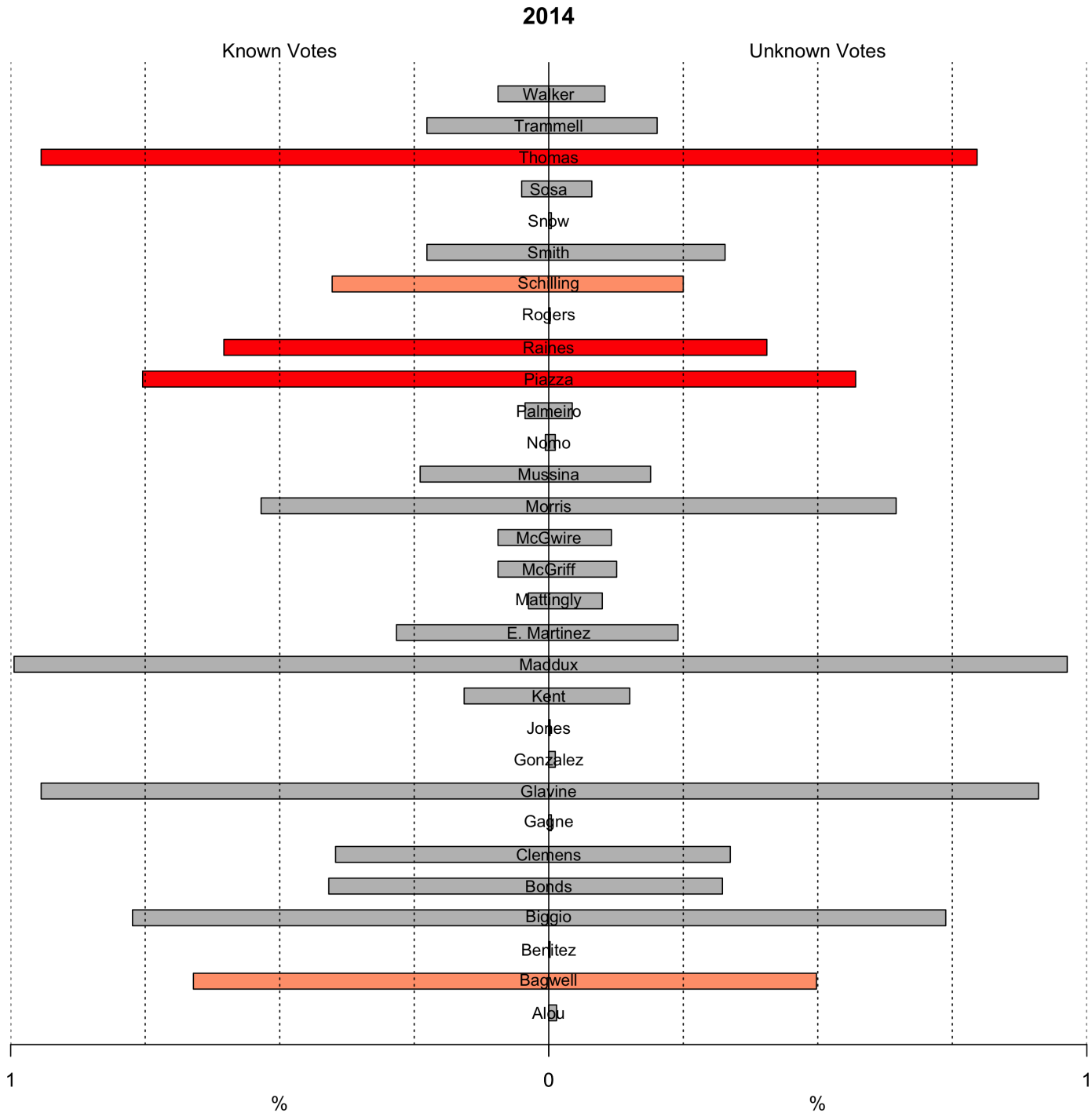
- Baseball Hall of Fame voting!!!
- The Baseball Writers Association of America (BBWAA) elects retired players to the Hall of Fame.
- The BBWAA releases players' vote totals, but does not release individual ballots.
- However, many BBWAA voters are writers and release their ballots publicly.
- So we know vote totals and a subset of the full data!

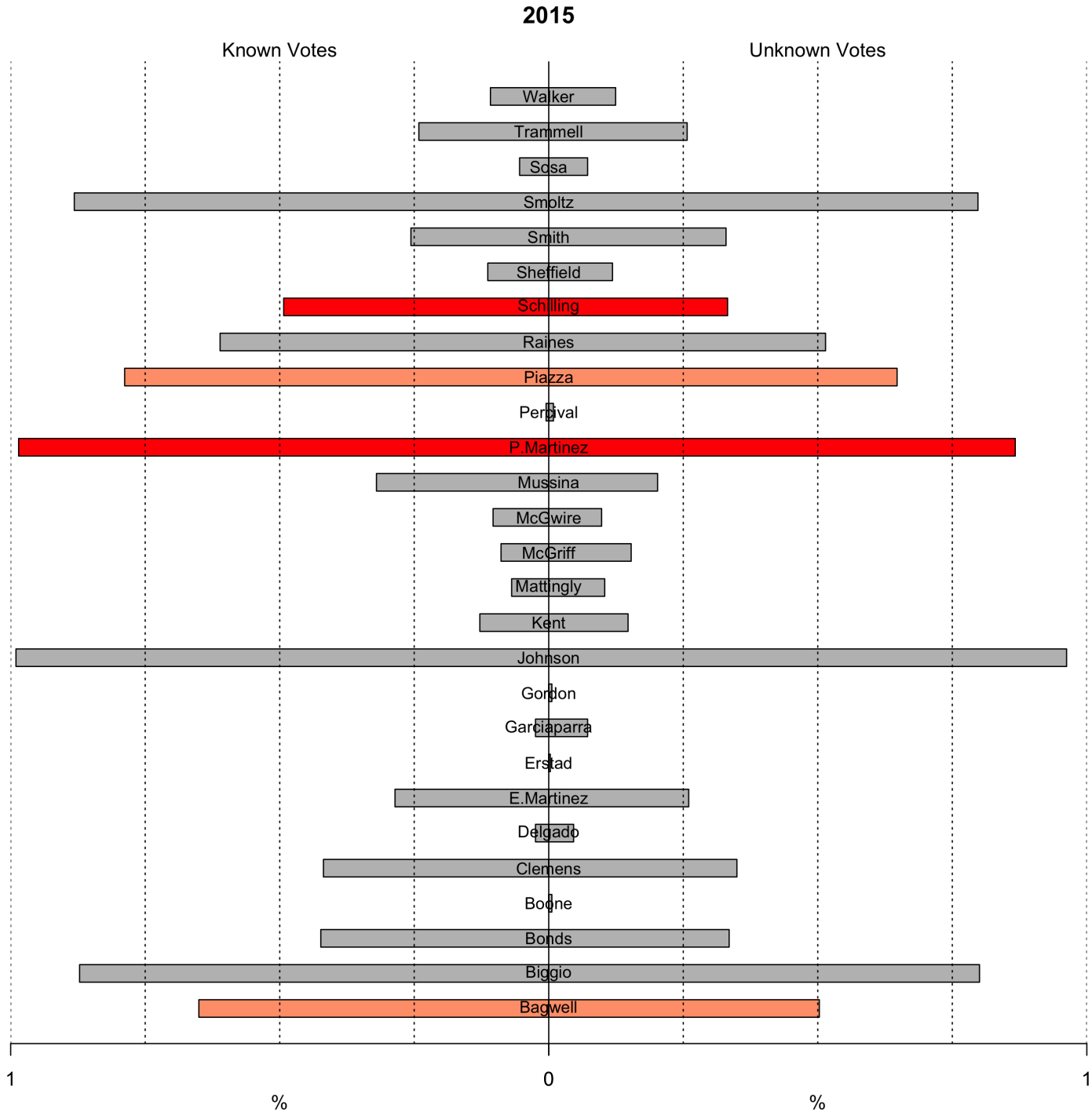
Hall of Fame Voting Rules

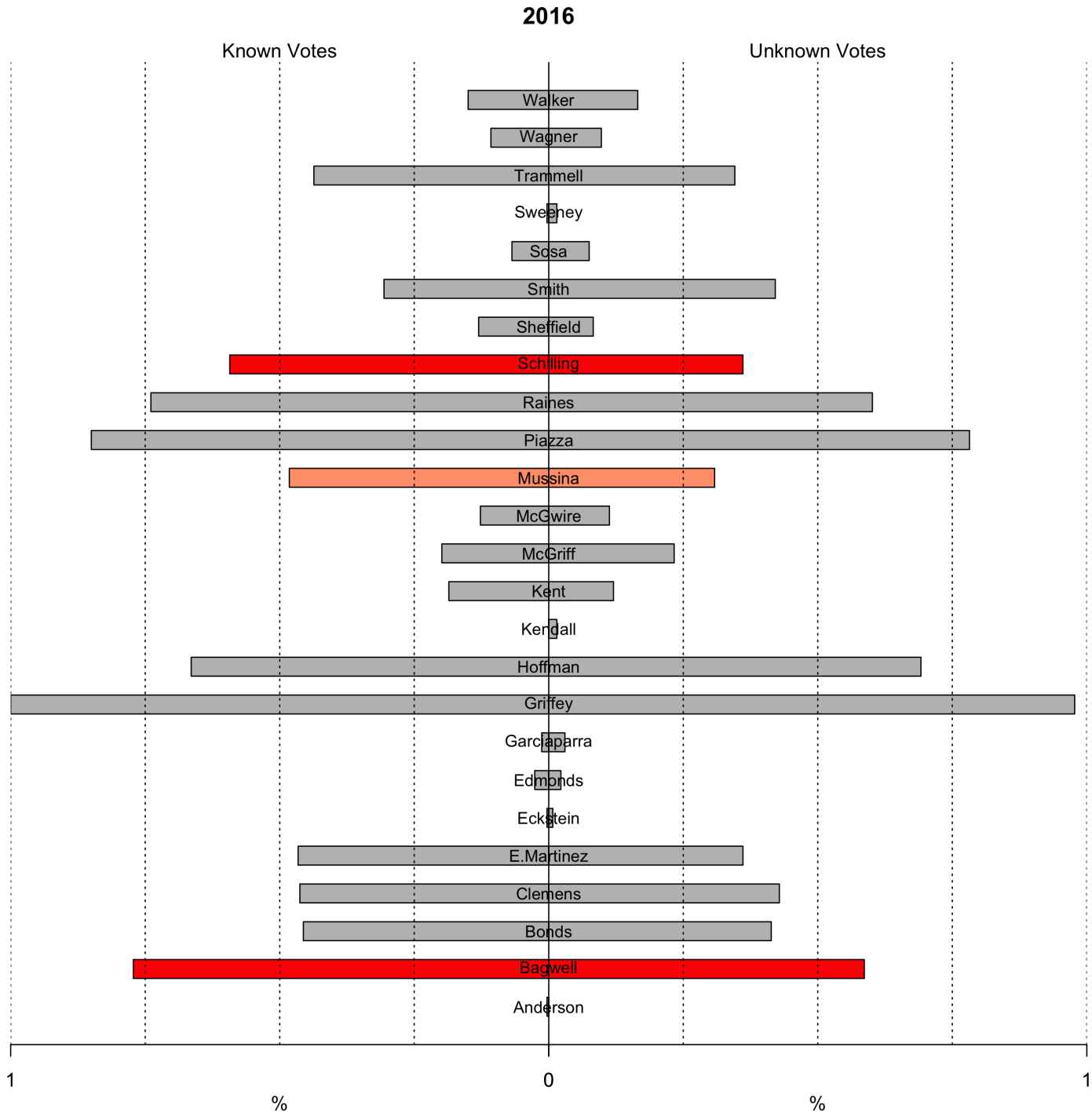
- Players become eligible 5 years after they retire from playing baseball.
- Each member of the BBWAA (approximately 625 eligible members) get to vote for up to 10 players [1]
- A player receiving votes of 75% of cast ballots gains entry into the Hall of Fame.
- Players must receive 5% of the vote to remain on ballot the next year.

Year	Total Ballots	Known Ballots	% Known
2014	571	159	27.85%
2015	549	203	36.98%
2016	440	307	69.80%

Table: Total ballots cast, total ballots known, and percentage of ballots known for the years 2014, 2015 and 2016


2014
Known Votes
Unknown Votes
Walker
Trammell
Thomas
Sosa
Snow
Smith
Schilling
Rogers
Raines
Piazza
Palmeiro
Nomo
Mussina
Morris
McGwire
McGriff
Mattingly
E. Martinez
Maddux
Kent
Jones
Gonzalez
Glavine
Gagne
Clemens
Bonds
Biggio
Benitez
Bagwell
Alou
1
0
1
%
%


2015
Known Votes
Unknown Votes
Walker
Trammell
Sosa
Smoltz
Smith
Sheffield
Schilling
Raines
Piazza
Percival
P.Martinez
Mussina
McGwire
McGriff
Mattingly
Kent
Johnson
Gordon
Garciaparra
Erstad
E.Martinez
Delgado
Clemens
Boone
Bonds
Biggio
Bagwell
1
0
1
%
%


2016
Known Votes
Unknown Votes
Walker
Wagner
Trammell
Sweeney
Sosa
Smith
Sheffield
Schilling
Raines
Piazza
Mussina
McGwire
McGriff
Kent
Kendall
Hoffman
Griffey
Garciaparra
Edmonds
Eckstein
E.Martinez
Clemens
Bonds
Bagwell
Anderson
1
0
1
%
%

- p-values calculated using Fisher's exact test
- Odds in the unknown group as the reference group

Players	OR 2014	P-values 2014	OR 2015	P-values 2015	OR 2016	P-values 2016
Bagwell	1.96	<0.001*	1.83	<0.001*	2.39	<0.001*
Piazza†	2.32	<0.001*	2.02	<0.001*	1.58	0.11
Raines	2.24	<0.001*	1.48	0.030	1.88	0.01
Schilling	2.02	<0.001*	1.95	<0.001*	2.58	<0.001*
Mussina	1.34	0.23	1.85	0.0028	2.09	<0.001*
Thomas†	4.27	<0.001*				
P.Martinez†			10.20	<0.001*		

*Significant after using the Holm correction.
†Elected to Baseball Hall of Fame

Table: Significant odds ratios and unadjusted p-values for each player for the years 2014, 2015 and 2016.

2 steps:

- Multiple Imputation [2]
- Latent Class Analysis [3]

Imputations were generated using fully conditional specification (FCS) implemented by the 'mice' package in R using predictive mean matching (PMM).

M=10 imputations were used.

Imputations were subject to to the restrictions that:

1. The total votes on a ballot had to be 10 or fewer
2. The total votes per player had to match up with released totals.

Target distribution for imputation:

$$P(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{Y}^{mis}\mathbf{1}_J \leq 10\mathbf{1}_J, \mathbf{1}'_{n_{mis}}\mathbf{Y}^{mis} = V - \mathbf{1}'_{n_{obs}}\mathbf{Y}^{obs})$$

where V is a vector of the player vote totals, $\mathbf{Y}^{mis}$ are the unknown ballots, and $\mathbf{Y}^{obs}$ are the observed ballots.

Latent Class Analysis

- After imputation LCA was performed on each of the completed data sets.
- We are interested in LCA with a covariate, namely, an indicator for is the ballot known or unknown.
- Formally, we are interested in estimating

$$P(L = c \mid X = x) = \frac{\exp^{\beta_{0c}+\beta_{1c}x}}{1 + \sum_{c=1}^{C-1} \exp^{\beta_{0c}+\beta_{1c}x}}$$

where $X = 1$ if the ballot is known and 0 otherwise, L is the latent class, and C is the number of latent classes.

Combining Rules

- After estimating the regression coefficients, they can be combined using Rubin's combining rules.

$$\bar{Q}_M = \sum_{m=1}^{M} \frac{\beta_1^{(m)}}{M}$$

$$B_M = \sum_{m=1}^{M} \frac{(\beta_1^{(m)} - \bar{Q}_M)^2}{M-1}$$

$$\bar{U}_M = \sum_{m=1}^{M} \frac{var(\beta_1^{(m)})}{M}$$

- Note: Here we set $\bar{U}_M = 0$ as we view the full collection of ballots as a population, rather than a sample from all possible ballots.
- This gives us an estimate of the variance $\bar{Q}_M$ as $T_M = (1 + \frac{1}{M})B_M$.

Questions of interest:

- How many latent classes are there?
- What players are in which latent classes?
- Does the probability of a voter belonging to a latent class differ of the known and unknown groups?

Questions of interest:

- How many latent classes are there? 2.
- What players are in which latent classes?
- Does the probability of a voter belonging to a latent class differ of the known and unknown groups?

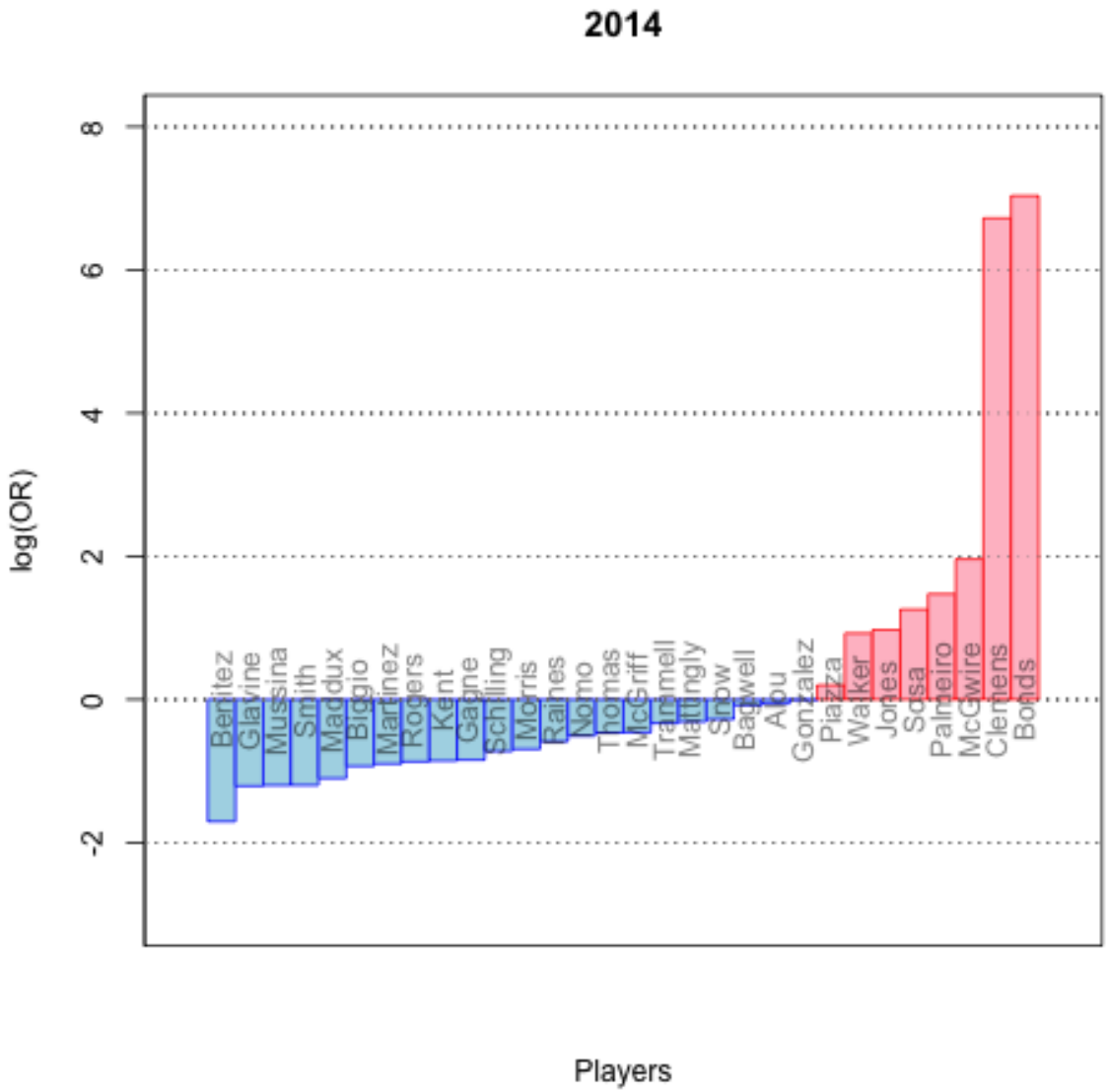


Figure: Log odds ratio comparing the likelihood of a voter from class 1 or class 2 voting for a particular player in 2014

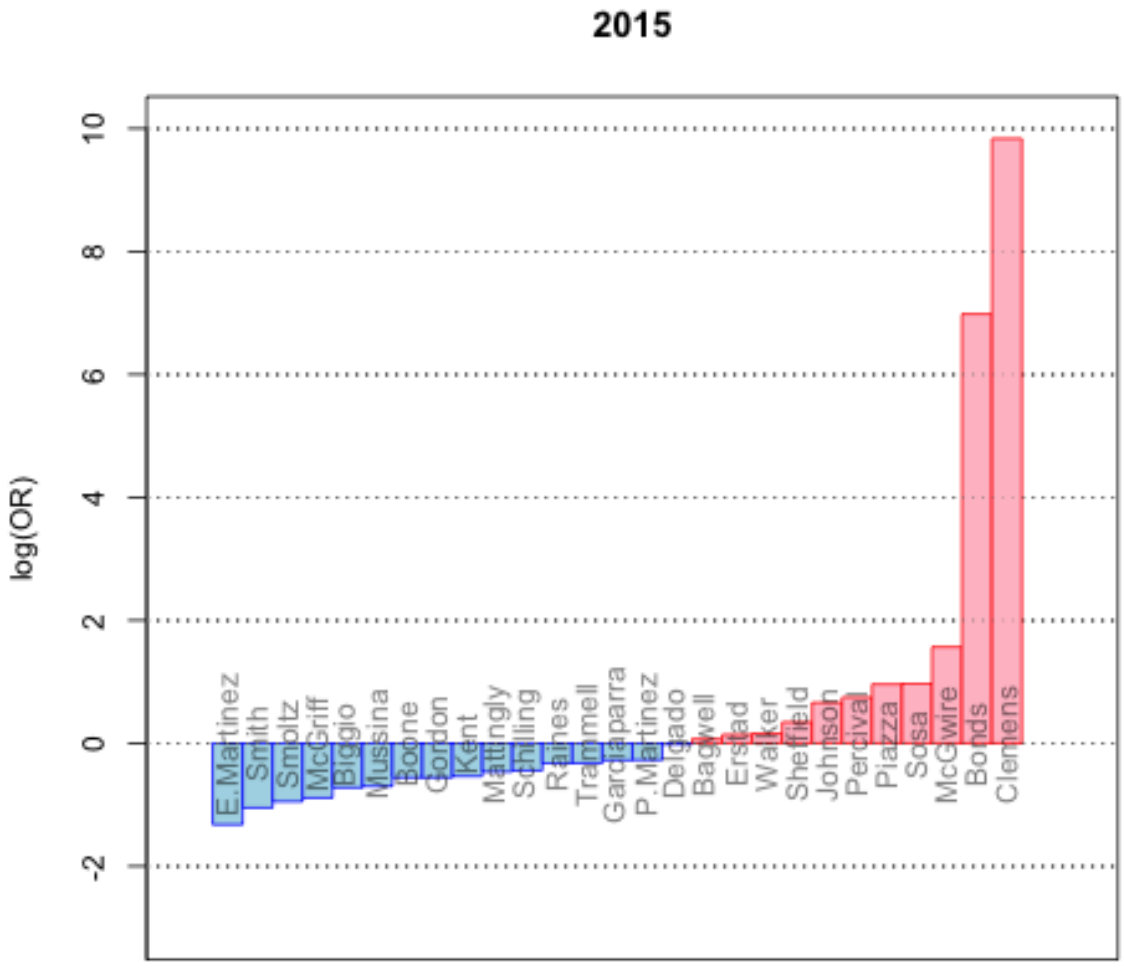


Figure: Log odds ratio comparing the likelihood of a voter from class 1 or class 2 voting for a particular player in 2015

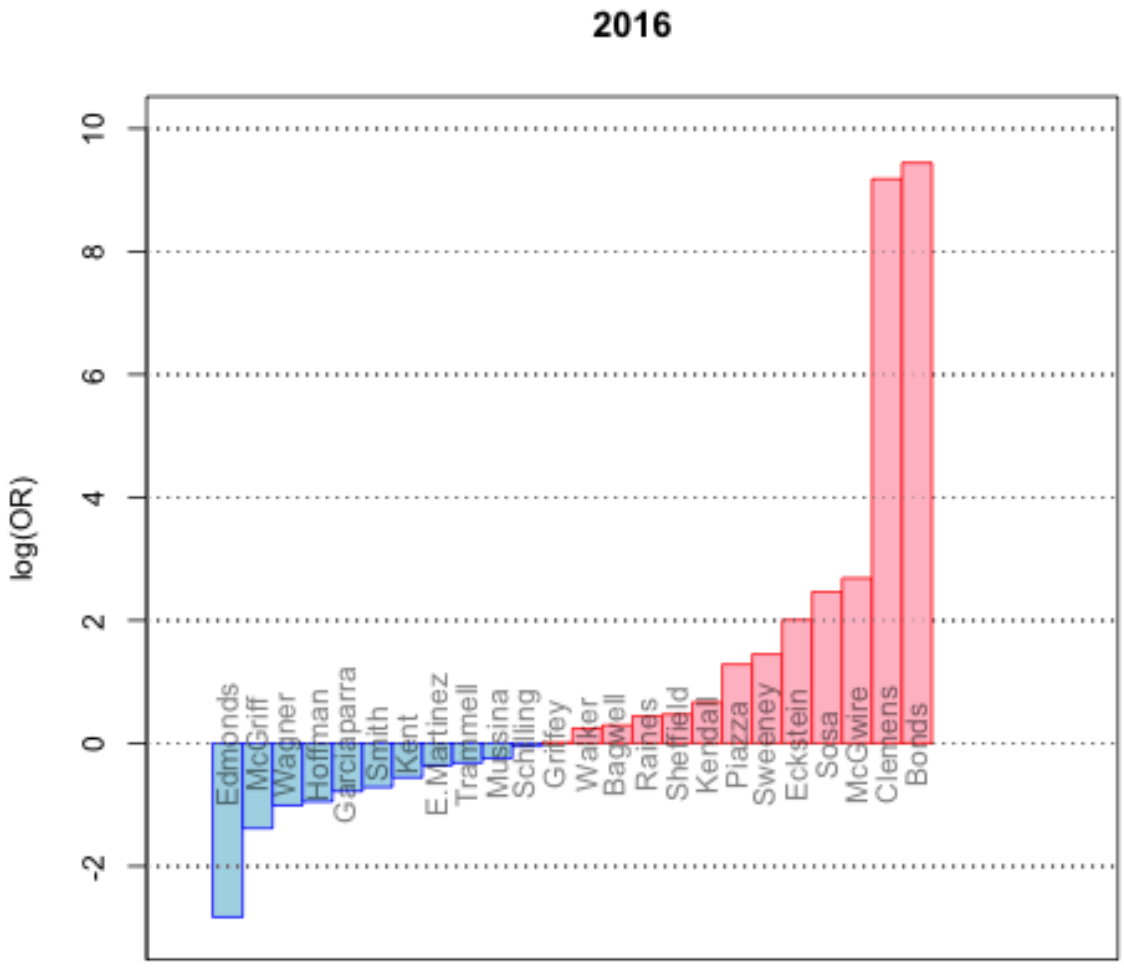


Figure: Log odds ratio comparing the likelihood of a voter from class 1 or class 2 voting for a particular player in 2016

Latent Classes

- Class 1: Pro-PED's
- Class 2: Anti-PED's

Questions of interest:

- How many latent classes are there? 2.
- What players are in which latent classes? (Bonds, Clemens, etc. vs Glavine, Mussina, McGriff, Smith, etc.
- Does the probability of a voter belonging to a latent class differ of the known and unknown groups?

	Known	Unknown
Class 1 (pro-PED's)	0.4063	0.3461
Class 2 (anti-PED's)	0.5937	0.6539

Table: 2014

	Known	Unknown
Class 1 (pro-PED's)	0.417	0.3401
Class 2 (anti-PED's)	0.583	0.6599

Table: 2015

	Known	Unknown
Class 1 (pro-PED's)	0.4469	0.4089
Class 2 (anti-PED's)	0.5531	0.5911

Table: 2016

Probability that a voter belongs to class 1 or class 2 given that a voter is in the known or unknown group.

Are these significant differences? Yes.

Year	Odds Ratio (Confidence Interval)
2014	1.293 (1.073, 1.557)
2015	1.387 (1.234, 1.560)
2016	1.168 (1.077, 1.278)

Table: Odds ratios for belonging to latent class 1 (pro-PED's) vs latent class 2 (anti-PED's) comparing the group or known and unknown ballots. An odds ratio of 1 here indicates that an individual voter in the known group is more likely to be in latent class 1 than latent class 2.

Questions of interest:

- How many latent classes are there? 2.
- What players are in which latent classes? (Bonds, Clemens, etc. vs Glavine, Mussina, McGriff, Smith, etc.
- Does the probability of a voter belonging to a latent class differ of the known and unknown groups? Yup.

- Here we have released statistics (vote totals) and a subset of the full data (released ballots) and we seek to learn about the group of voters who did not release their ballots.
- There are several significant differences in the voting habits of known and unknown group of voters (e.g. known group more likely to vote for Thomas in 2014 OR: 4.27)
- Voters fall into two latent classes: pro-PED's and anti-PED's.
- Voters in the known group are significantly more likely to belong to latent class 1 than voters in the unknown group.

- The difference in support for these two classes depending on whether a voter released their ballot or not is consistent with older voters taking more conservative attitudes toward PED users, which is also noted in Pollis [4].
- Here we have not learned anything specifically about an *individual*, but we have learned something about a *group*
- From a privacy perspective we need to ask what a database participant owes to other participants in terms of privacy.

Citations

1. Baseball Writers Association of America (2016), "BBWAA election rules", http://baseballhall.org/hall-of-famers/bbwaa-rules-for-election.

2. Little, RJA Rubin, DB (1987), Statistical Analysis with Missing Data, John Wiley Sons.

3. Collins, LM Lanza, ST (2010), Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences., Wiley, New York.

4. Pollis, L (2015), "Ninety percent mental: Are secret ballots ruining cooperstown?" Baseball Prospec- tus.