The Harris Poll

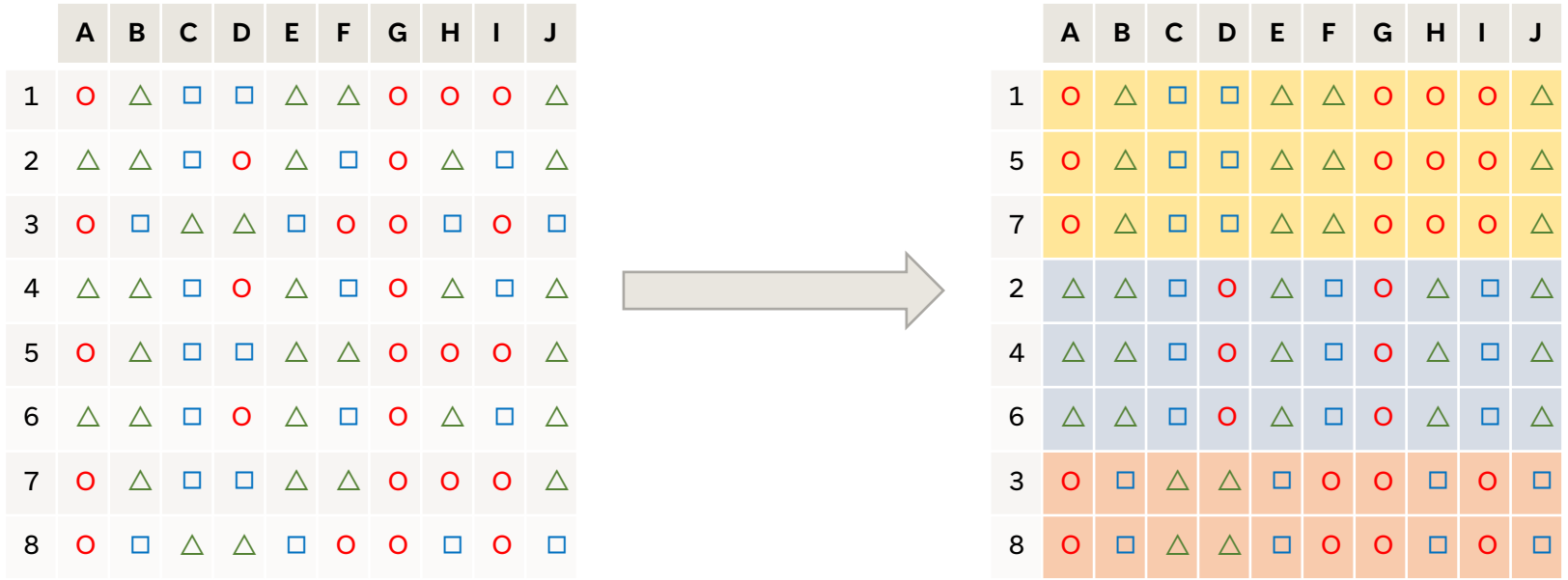
ARCHETYPAL ANALYSIS

MAINTAINING CONTRASTIVE GROUPS IN CLUSTER ANALYSIS

Jacob Nelson / Senior Data Scientist / Harris Poll

REVIEW: WHAT IS CLUSTER ANALYSIS

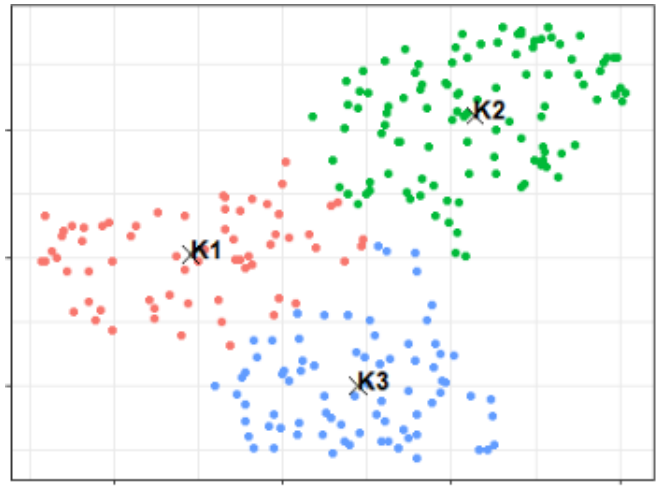
- Divides data observations into meaningful / useful groups
- Helps us to understand, organize, and discover patterns in seemingly unstructured data



LET'S TALK ABOUT K-MEANS

K-MEANS

Finds the optimal "group averages" or "centroids" that minimize the sum of squared distances between observations and their closest centroid



ARCHETYPAL ANALYSIS

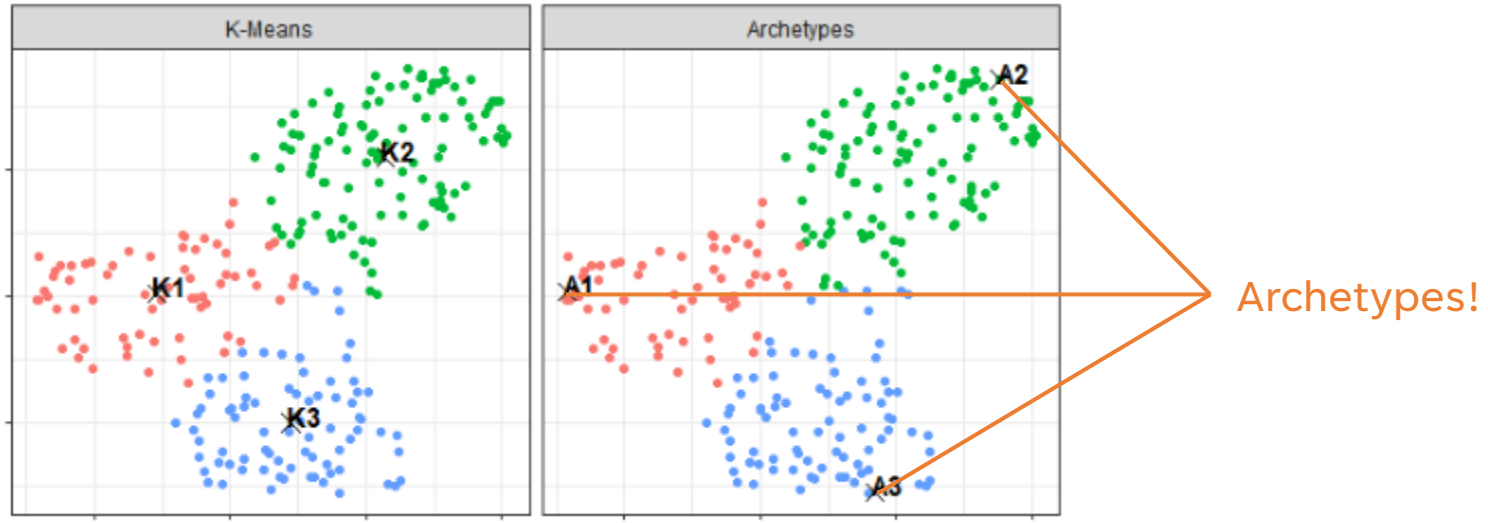
K-Means Seeks “Typical” Observations

Centroids exemplify cluster membership

Archetypal Analysis Seeks Extreme Observations!

Describes the data as convex combinations of extreme / periphery points

Clutler & Breiman (1994)



SO, WHAT IS AN ARCHETYPE?

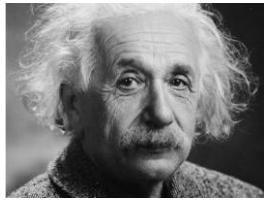
An archetype represents a pure form of a class or type...

The **champions** of a class

WE USE “ARCHETYPES” TO CLASSIFY A LOT OF THINGS!

ATHLETIC

SMART



MANY MORE...

Conservatives/Liberals

Activists

Value Shoppers

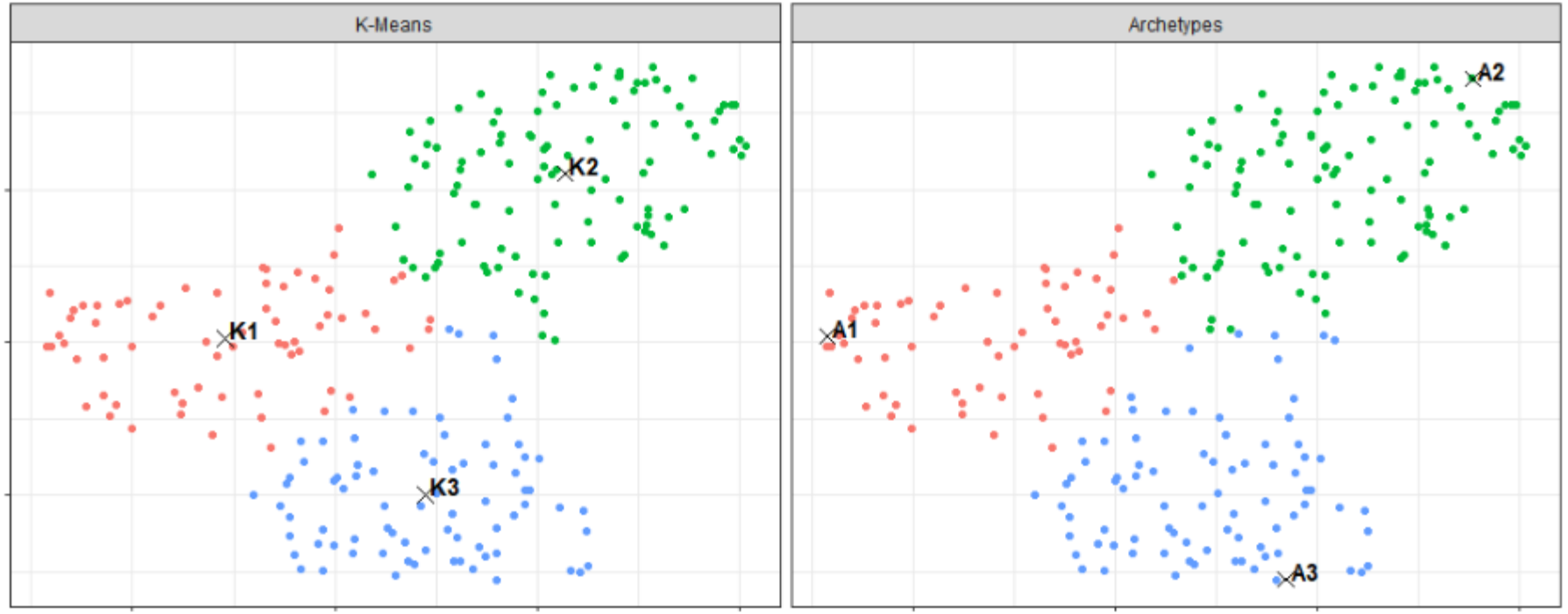
Superheroes

Planets

Life

WHY DOES IT MATTER IN CLASSIFICATION?

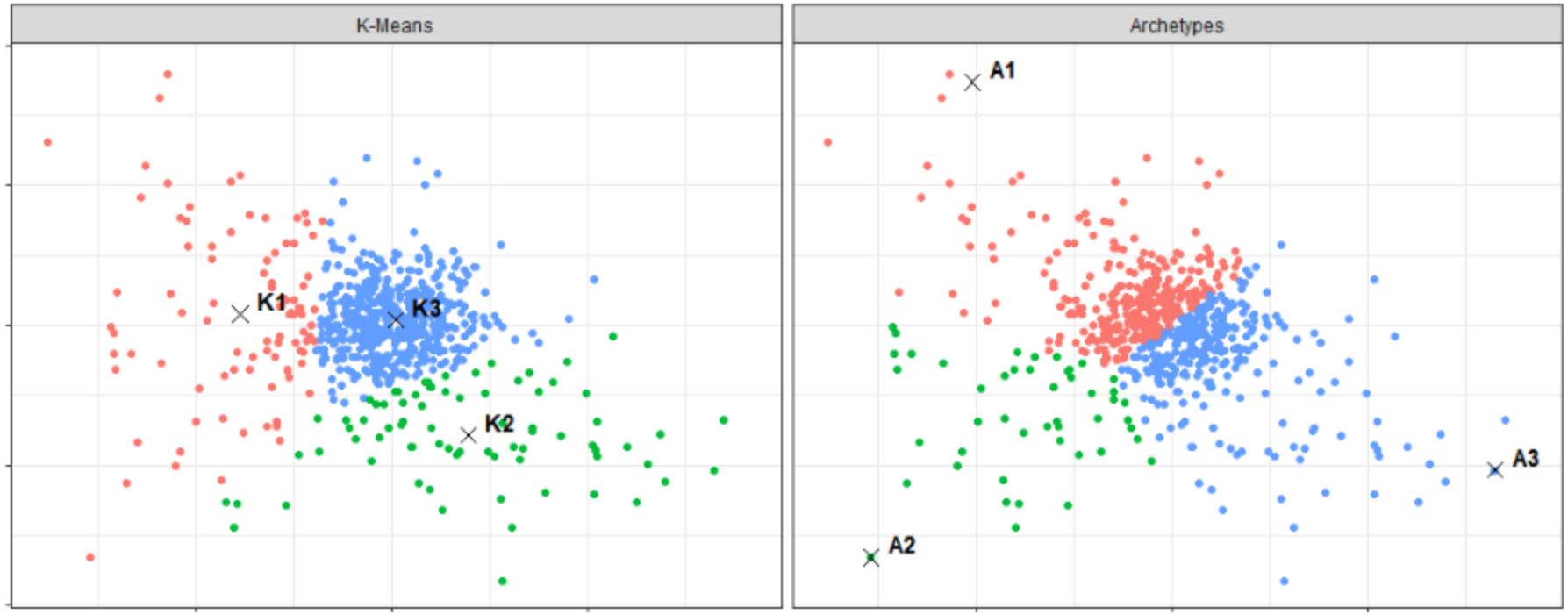
Despite very different classification rules, nearly identical classification



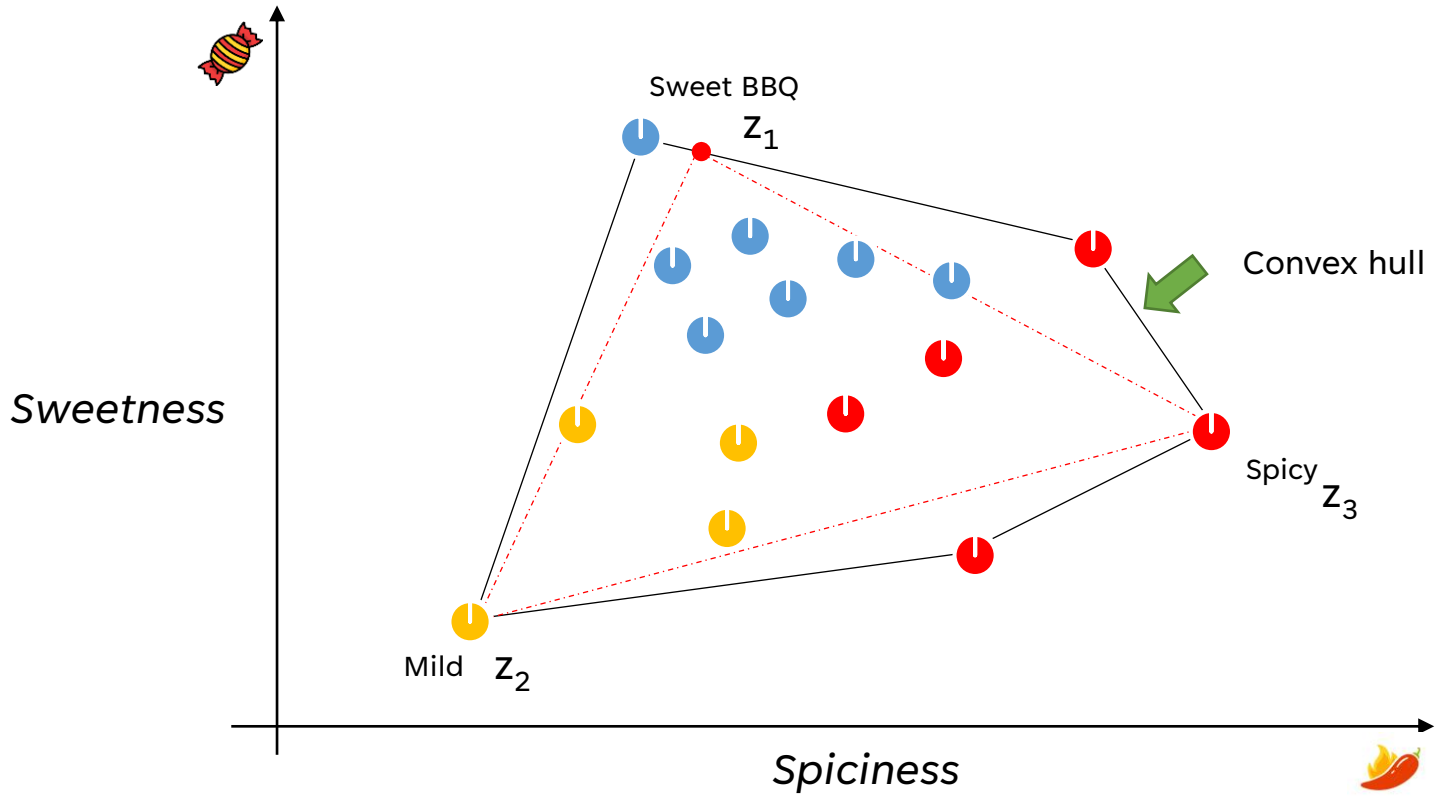
And yet...

WHY DOES IT MATTER IN CLASSIFICATION?

Contrastive categories at the expense of homogenous clusters!



HOW IT WORKS



Archetypal analysis approximates the convex hull

Observations are convex combinations of the archetypes

$$\alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3$$

(non-negative and sum to 1)

ARCHETYPAL ANALYSIS MATHEMATICAL SUMMARY

THE DATA PROBLEM

Represent each individual observation as a convex combination of the archetypes.

Archetypes are a convex combination of the observations.

Find the $n * k$ matrices α and β, and matrix Z (the archetypes) that minimizes the residuals:

$$RSS = ||X - \alpha Z^T||_2 \; ; Z = X^T \beta$$

Constraints...

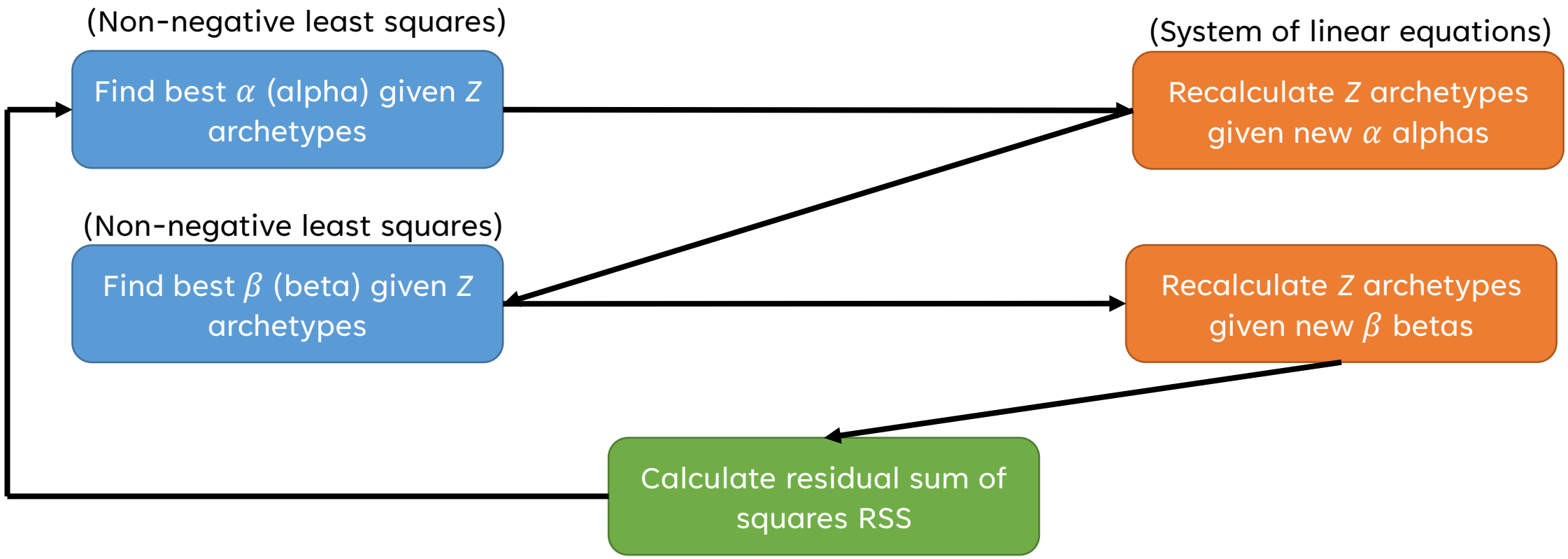
$$\sum_{j=1}^{k} \alpha_{ij} = 1 \; with \; \alpha \geq 0 \; and \; i = 1, \dots, n,$$

$$\sum_{i=1}^{n} \beta_{ji} = 1 \; with \; \beta \geq 0 \; and \; i = 1, \dots, n,$$

$RSS = ||X - \alpha Z^T||_2 ; Z = X^T\beta$

OPTIMIZATION

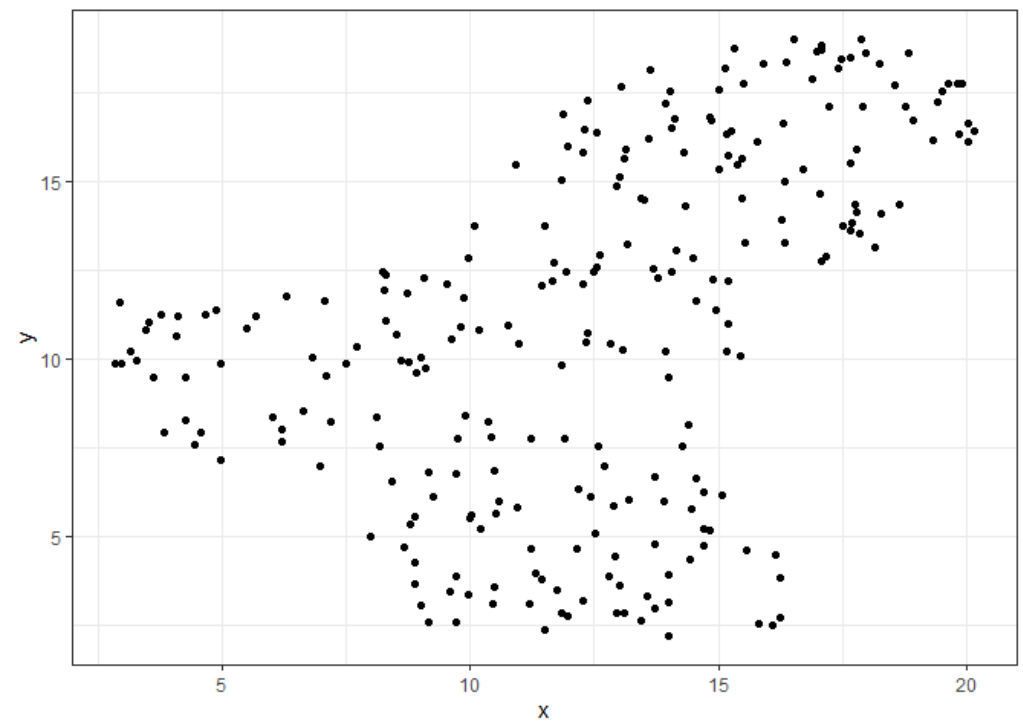
(The alternating least squares approach!)



If RSS meets threshold, or maximum iterations reached, exit loop

SETUP

```
library(archetypes)
#> Loading required package: modeltools
#> Loading required package: stats4
#> Loading required package: nnls
library(ggplot2)
data(toy)
head(toy)
#>              x         y
#> [1,]  8.749423  9.922297
#> [2,]  4.965050  7.163594
#> [3,] 10.092057 13.750449
#> [4,] 17.068966 12.776089
#> [5,] 13.591363 16.201112
#> [6,]  8.282601 11.925104
dim(toy)
#> [1] 250   2
ggplot(as.data.frame(toy), aes(x = x, y = y)) +
  geom_point() +
  theme_bw()
```



RUNNING ARCHETYPAL ANALYSIS

```
set.seed(1)
archetypes_model  <- archetypes(
  data = toy,
  k = 3,
  verbose = TRUE
)
#> 1: rss = 0.02084210, improvement = 0.05663049
#> 2: rss = 0.01232307, improvement = 0.00851903
#> 3: rss = 0.00931637, improvement = 0.00300669
#> 4: rss = 0.00798909, improvement = 0.00132728
#> 5: rss = 0.00757768, improvement = 0.00041141
#> 6: rss = 0.00738986, improvement = 0.00018783
#> 7: rss = 0.00729980, improvement = 0.00009006
#> 8: rss = 0.00728891, improvement = 0.00001089
#> 9: rss = 0.00728196, improvement = 0.00000695
#> 10: rss = 0.00727594, improvement = 0.00000602
#> 11: rss = 0.00727163, improvement = 0.00000431
#> 12: rss = 0.00726895, improvement = 0.00000268
#> 13: rss = 0.00726761, improvement = 0.00000133
#> 14: rss = 0.00727484, improvement = -0.00000723
```

ACCESSING THE COEFFICIENTS

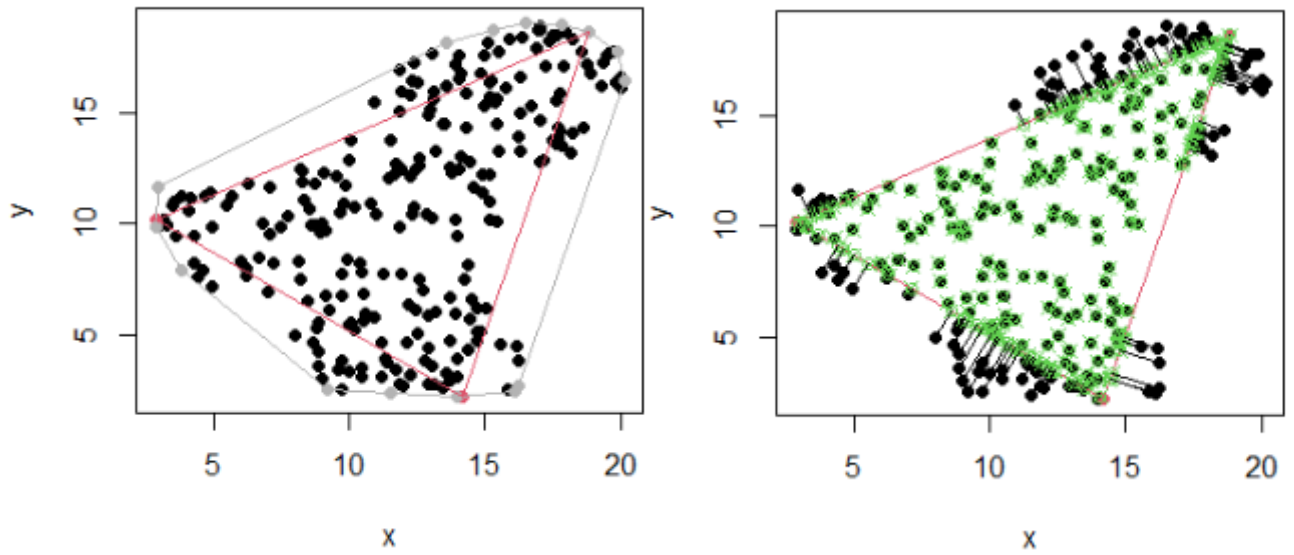
```
alphas <- coef(archetypes_model)
head(alphas)
#>             [,1]      [,2]       [,3]
#> [1,] 0.561725919 0.1964401 0.24180620
#> [2,] 0.764299891 0.0000000 0.23567946
#> [3,] 0.542378800 0.4395223 0.01807073
#> [4,] 0.008318144 0.6390827 0.35257598
#> [5,] 0.321406919 0.6785679 0.00000000
#> [6,] 0.637714583 0.2818481 0.08040853

at <- archetypes_model$archetypes
at
#>              x         y
#> [1,]  2.872939 10.186436
#> [2,] 18.831127 18.621895
#> [3,] 14.210083  2.241151
```

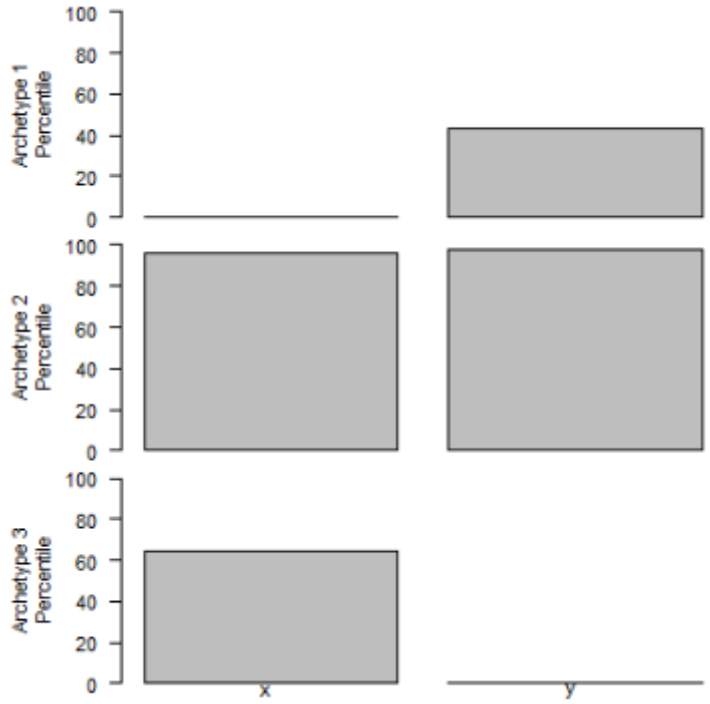
PLOTTING THE RESULTS

```
xyplot(archetypes_model, toy, chull = chull(toy))
```

```
xyplot(archetypes_model, toy, adata.show = TRUE)
```

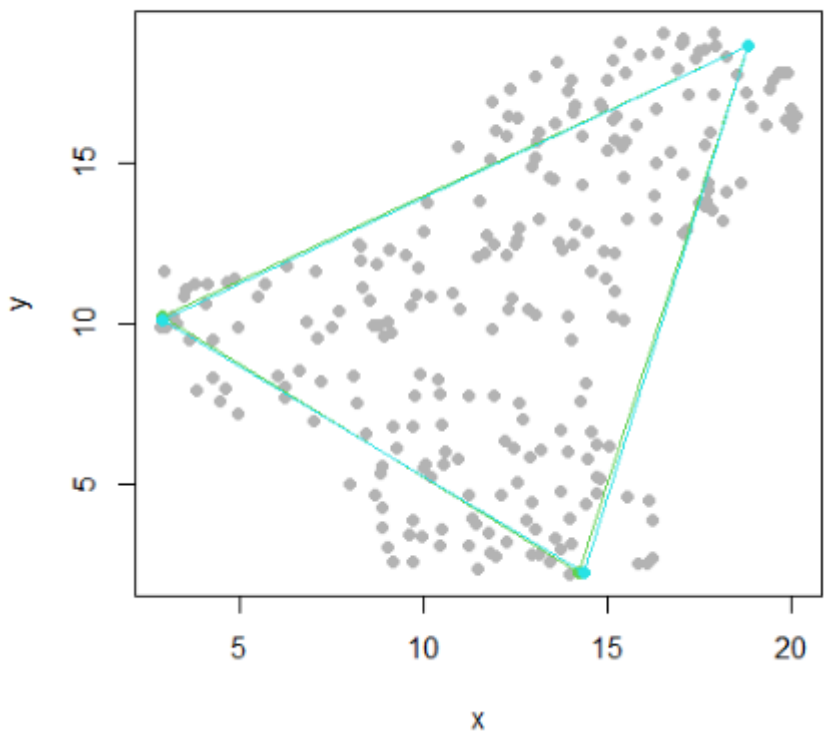
```
barplot(archetypes_model, toy, percentiles = TRUE)
```





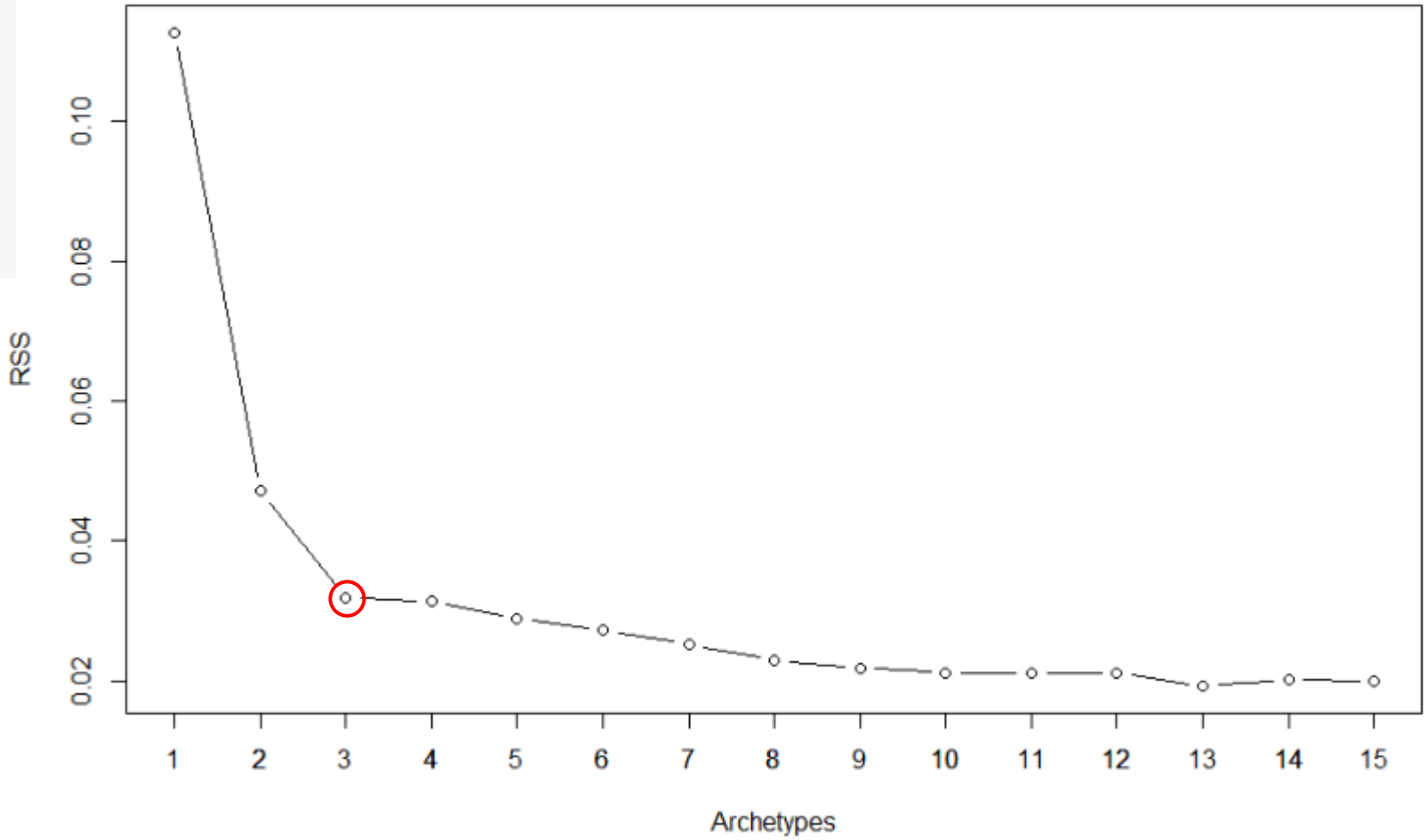
AVOID LOCAL MINIMA

```r
set.seed(1)
archetypal_steps2 <- stepArchetypes(
  data = toy,
  k = 3,
  nrep = 4,
  verbose = FALSE
)

xyplot(archetypal_steps2, toy)

final_model <- bestModel(archetypal_steps2)
```



CHOOSING THE RIGHT “K”

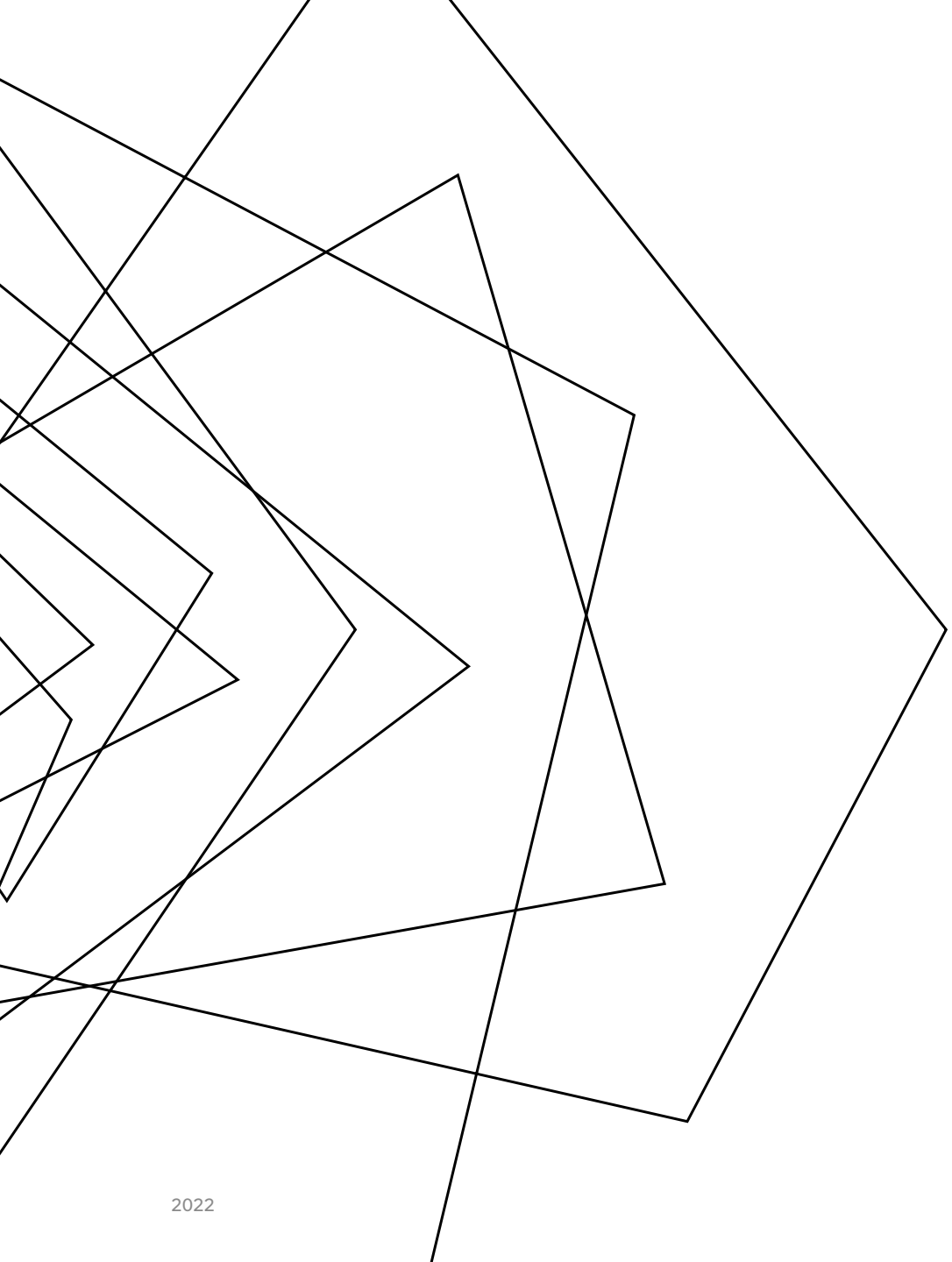
```r
set.seed(1)
archetypal_steps <- stepArchetypes(
  data2,
  k = 1:15,
  verbose = FALSE,
  nrep = 3

screeplot(archetypal_steps)
```



A LOT MORE OPTIONS

CHECK OUT THE VIGNETTE

```
vignette("archetypes")
```



SELF REPORTED VOTER MOTIVATIONS

HARRIS POLL'S ANNUAL CUE RESEARCH

Annual research serving a variety of miscellaneous needs

15 MOTIVATIONS

Part of the research we tacked on to Harris Poll's annual Cue Research was measuring 15 Voter Motivations

TRADE OFF EXERCISE

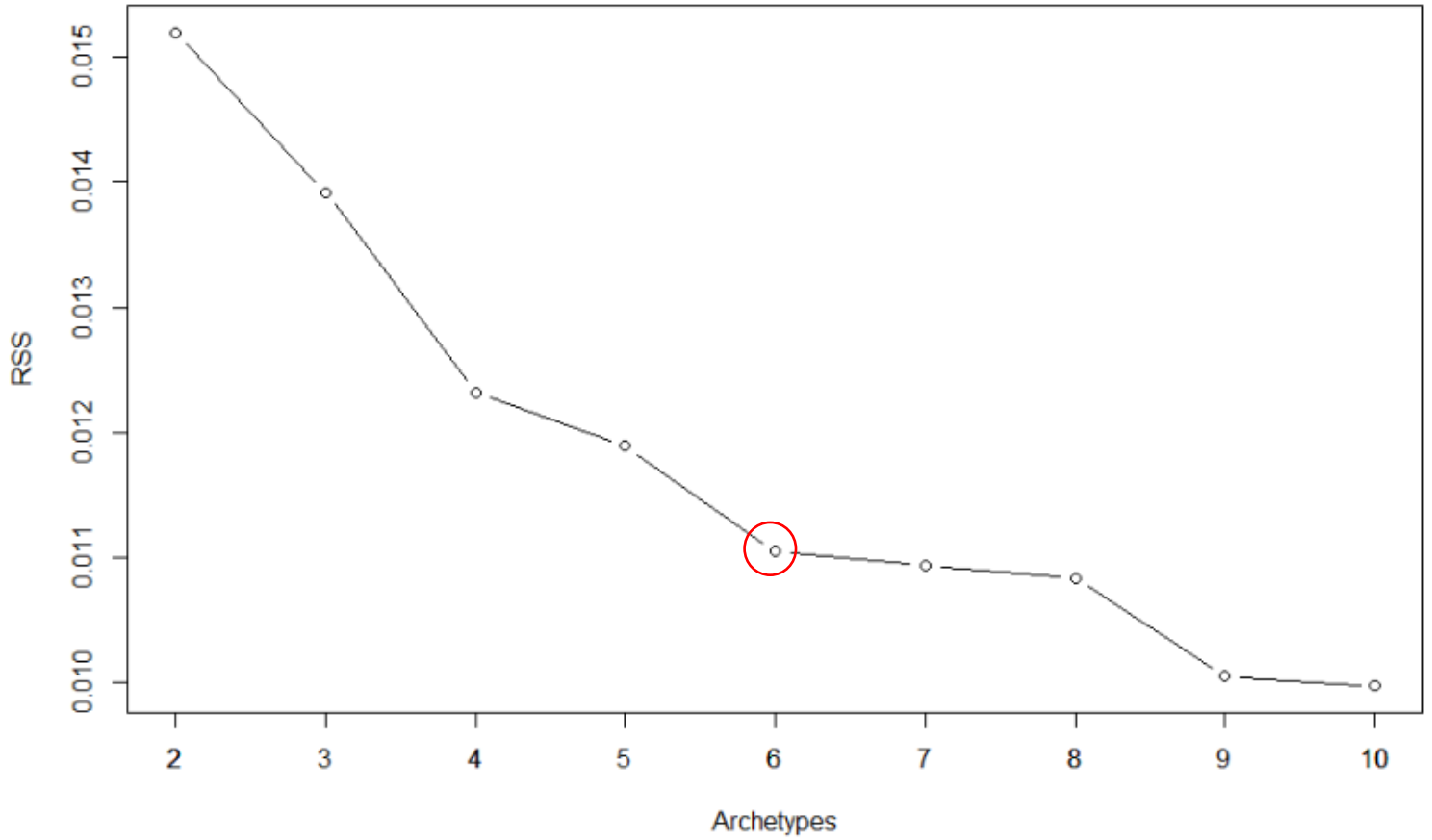
Each respondent in our survey evaluated sets of motivations and evaluated which was most and least motivating for them to vote.

We score them based on choice based modelling

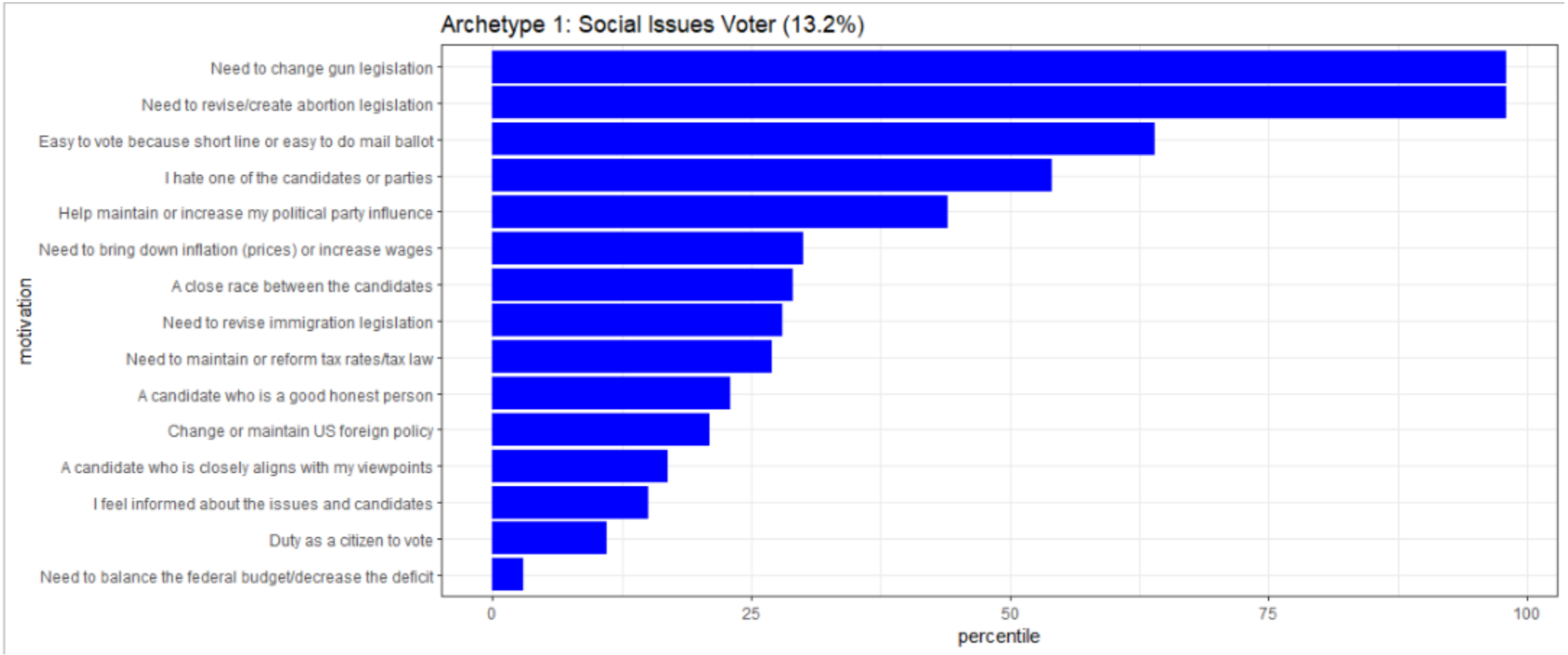
15 MOTIVATIONS

1. A close race between the candidates
2. Duty as a citizen to vote
3. Need to revise/create abortion legislation
4. A candidate who is a good honest person
5. A candidate who is closely aligns with my viewpoints
6. Help maintain or increase my political party influence
7. Need to revise immigration legislation
8. Need to bring down inflation (prices) or increase wages
9. Need to balance the federal budget/decrease the deficit
10. Change or maintain US foreign policy
11. Need to change gun legislation
12. I feel informed about the issues and candidates
13. Easy to vote because short line or easy to do mail ballot
14. I hate one of the candidates or parties
15. Need to maintain or reform tax rates/tax law

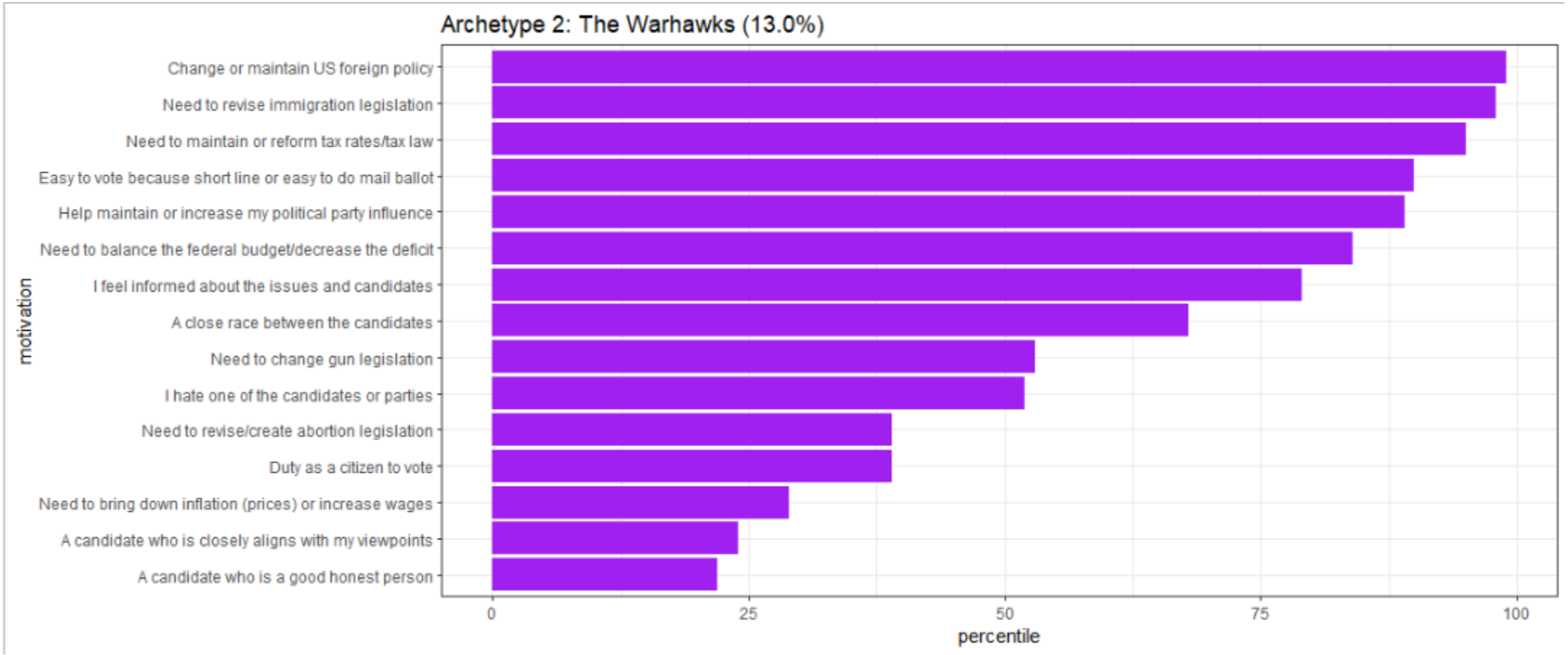
CHOOSING THE RIGHT MODEL



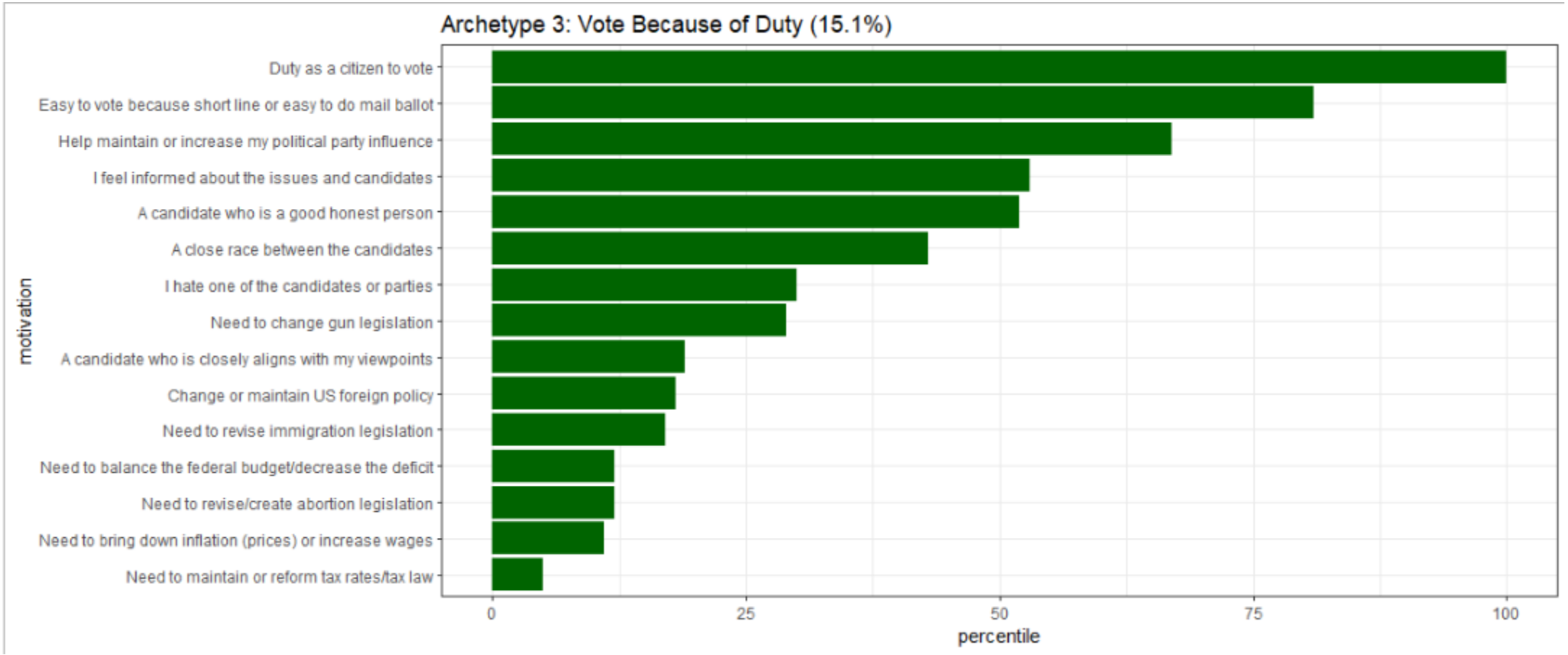
ARCHETYPE 1 – SOCIAL ISSUES VOTERS



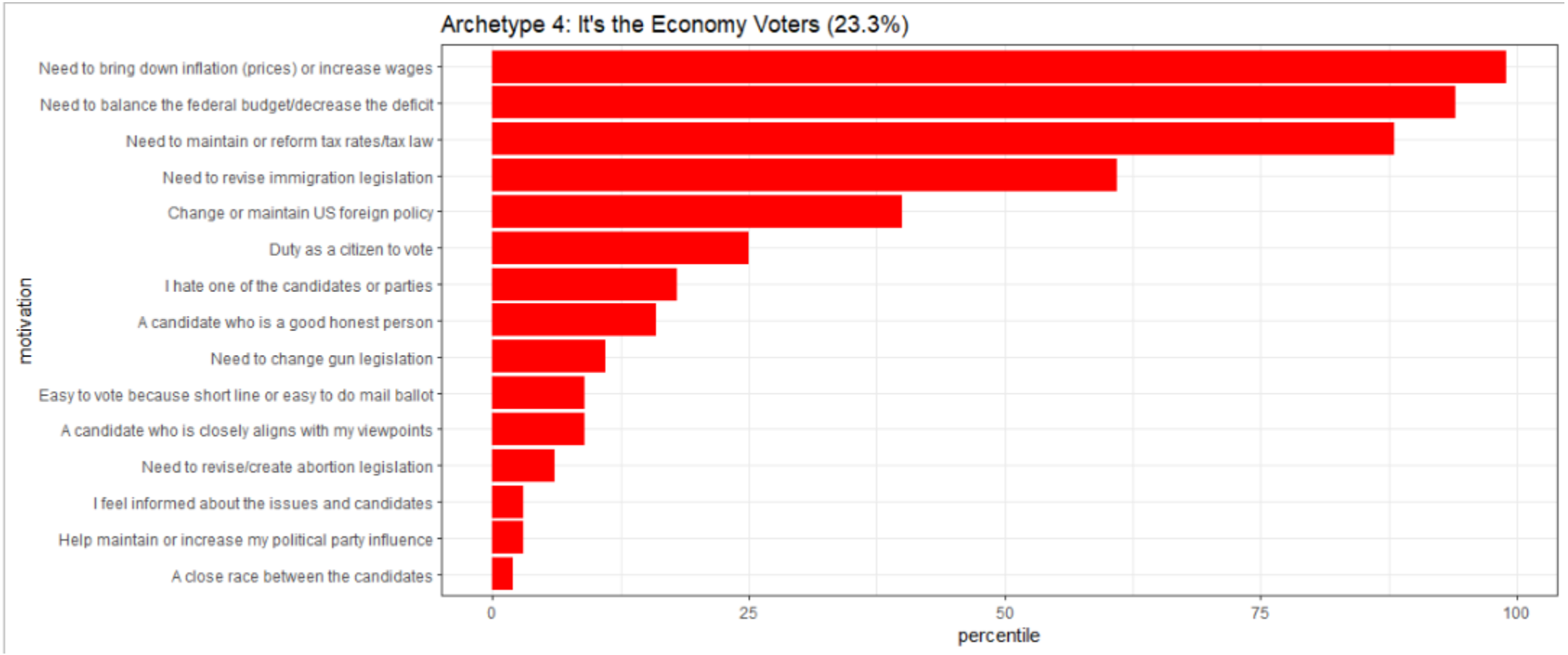
ARCHETYPE 2 – THE WARHAWKS



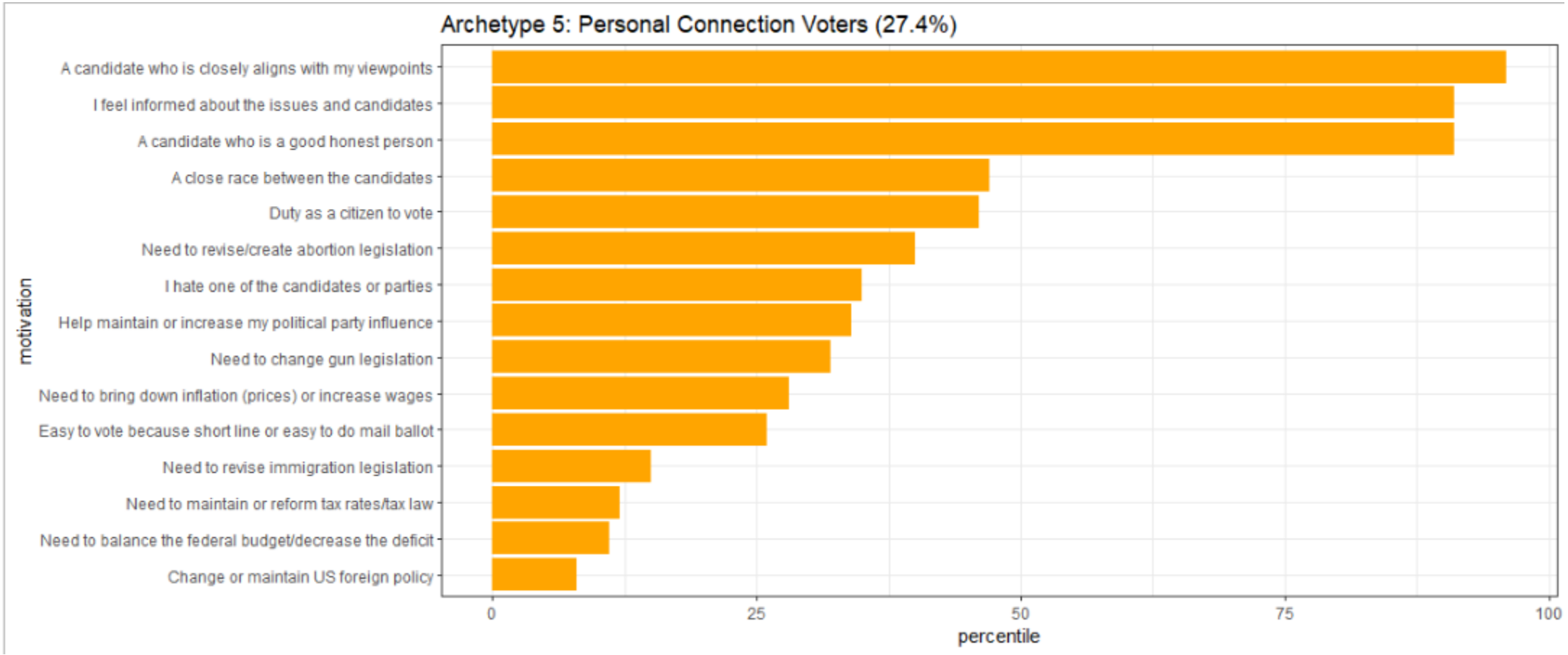
ARCHETYPE 3 – DUTIFUL CITIZENS



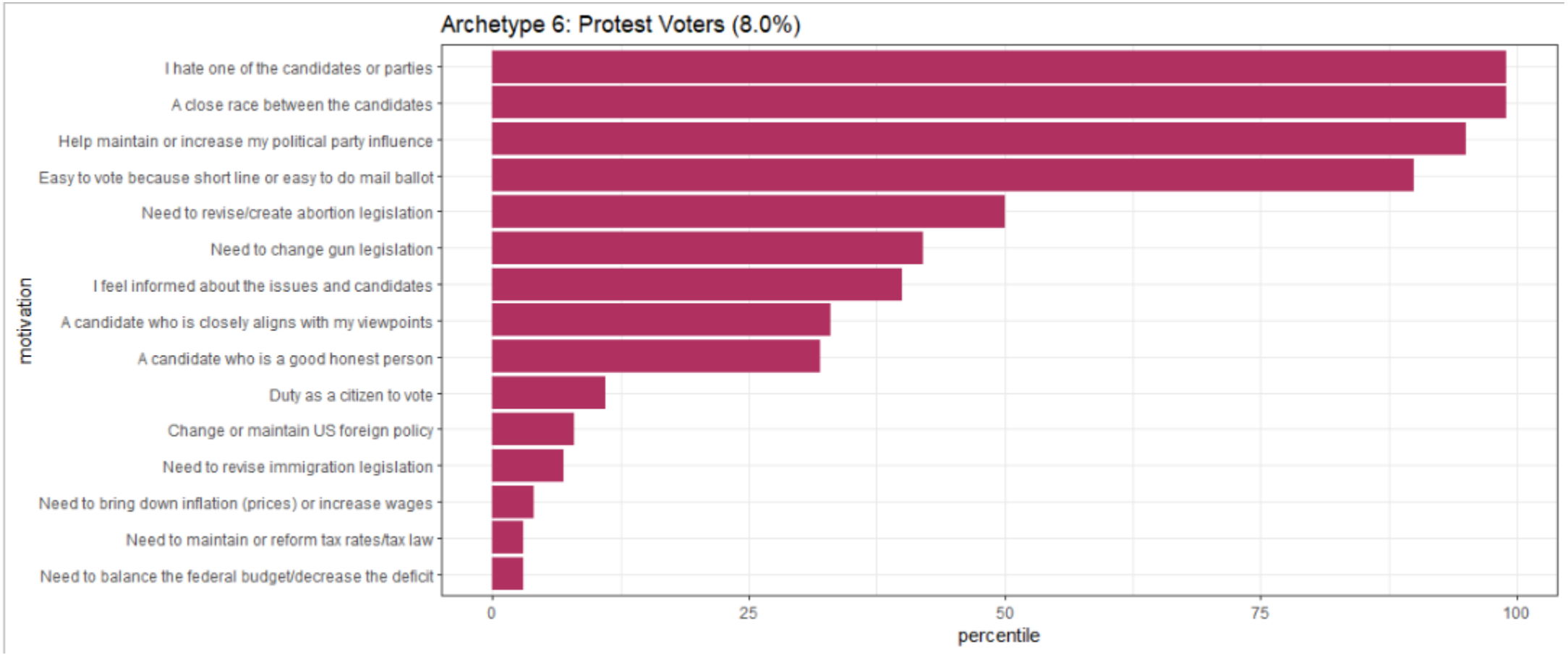
ARCHETYPE 4 – IT'S THE ECONOMY

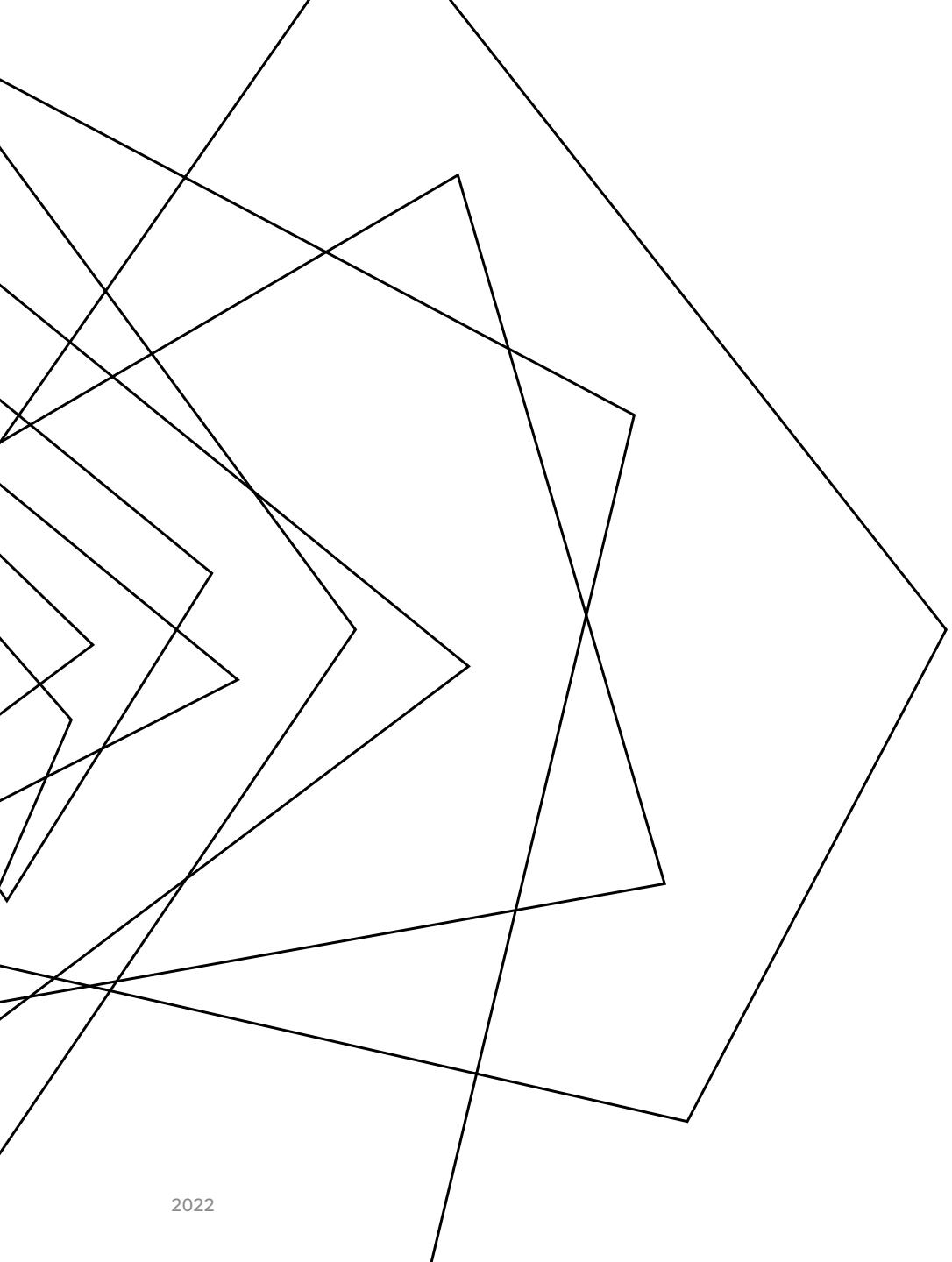


ARCHETYPE 5 – PERSONAL CONNECTION VOTERS



ARCHETYPE 6 – PROTEST / CLOSE RACE VOTERS





SUMMARY

ARCHETYPAL ANALYSIS IS...

A way of finding and organizing data into contrastive categories and groups

AND IS USEFUL BECAUSE...

Finding the CHAMPIONS of a class, and exemplifying class membership based on contrastive features.

BEST WAY TO IMPLEMENT IS...

Via the archetypes package in R. Check out the vignette!

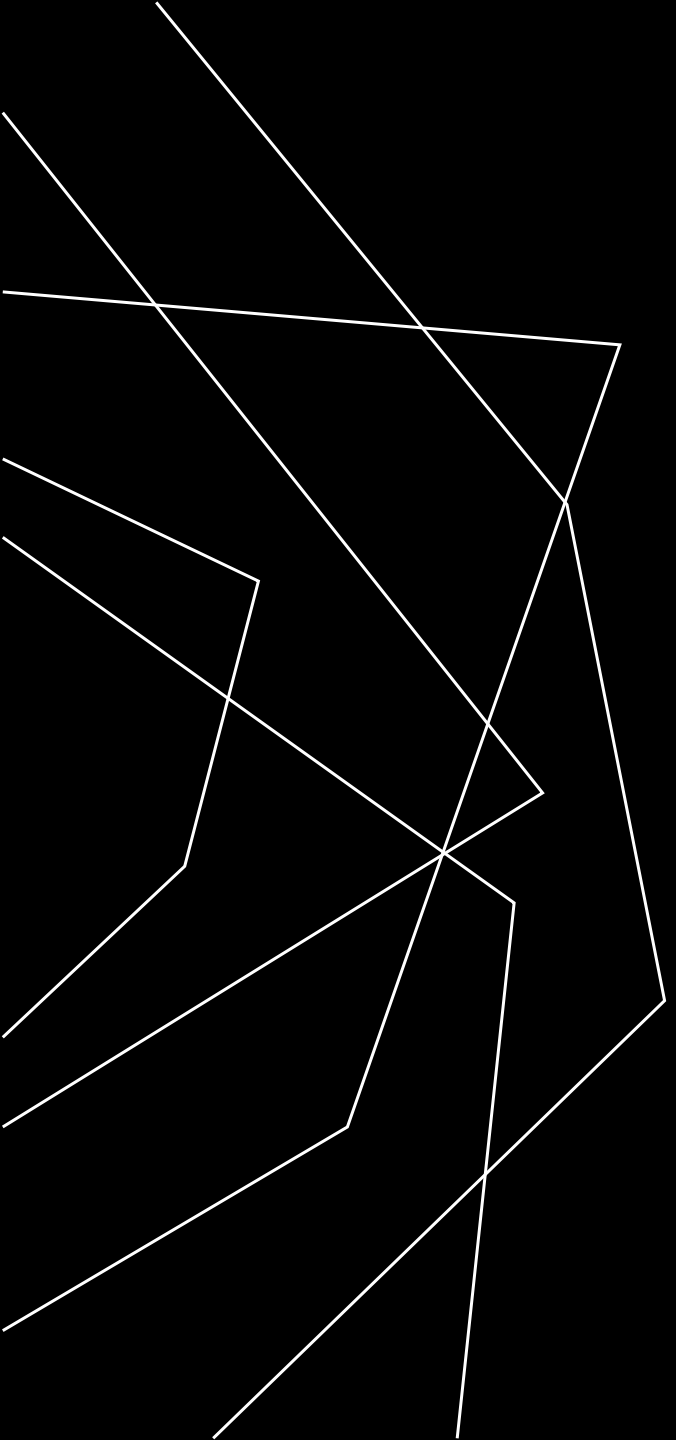
MORE INFORMATION

ORIGINAL PAPER ON ARCHETYPAL ANALYSIS

Adele Cutler and Leo Breiman. Archetypal analysis. Technometrics, 36(4):338–347, November 1994.

R PACKAGE + VIGNETTE

Manuel J. A. Eugster and Friedrich Leisch. From Spider-Man to Hero -- Archetypal Analysis in R. Journal of Statistical Software, 30(8), 1-23, 2009. http://www.jstatsoft.org/v30/i08/



THANK YOU!!