

The Turnout Tracker

Jonathan Tannen



sixty-six wards

- A datascience blog about Philadelphia Politics.
- *“FiveThirtyEight for Philadelphia”*
- www.sixtysixwards.com

For me the challenges are

- rigor with simplification
- storytelling.

Alternatively: Question formation.

Alternatively 2: The killer plot.

sixty-six wards

- A datascience blog about Philadelphia Politics.
- *“FiveThirtyEight for Philadelphia”*
- www.sixtysixwards.com

Here are Andrew Stober and Kristin Combs's maps from 2015.

▼ View code

```
library(sf)
divs <- st_read("../data/gis/2019/Political_Divisions.shp")
rename(warlddiv = DIVISION_N)df_council <- df_council %>%
  group_by(year, election, WARD19, DIV19) %>%
  mutate(pvote = votes/sum(votes))

ggplot(
  divs %>% left_join(
    df_council %>%
      filter(
        year == 2015,
        election=="general",
        CANDIDATE %in% c("ANDREW C STOBER", "KRISTIN COMBS")
      ) %>%
      mutate(warlddiv = paste0(WARD19, DIV19), Candidate = for
    )
  ) +
  geom_sf(aes(fill = 100 *pvote), color=NA) +
  facet_wrap(~Candidate) +
  scale_fill_viridis_c("Percent\nof Votes") +
  theme_map_sixtysix() %>%replace%
  theme(legend.position = "right") +
  ggtitle("Third Party votes come from the Wealthy Progressiv
```

Third Party votes come from the Wealthy Progressive Divisions



Today

Two data visualization stories:

- The Turnout Tracker
- (Maybe) Philadelphia Voting Blocs

The Turnout Tracker

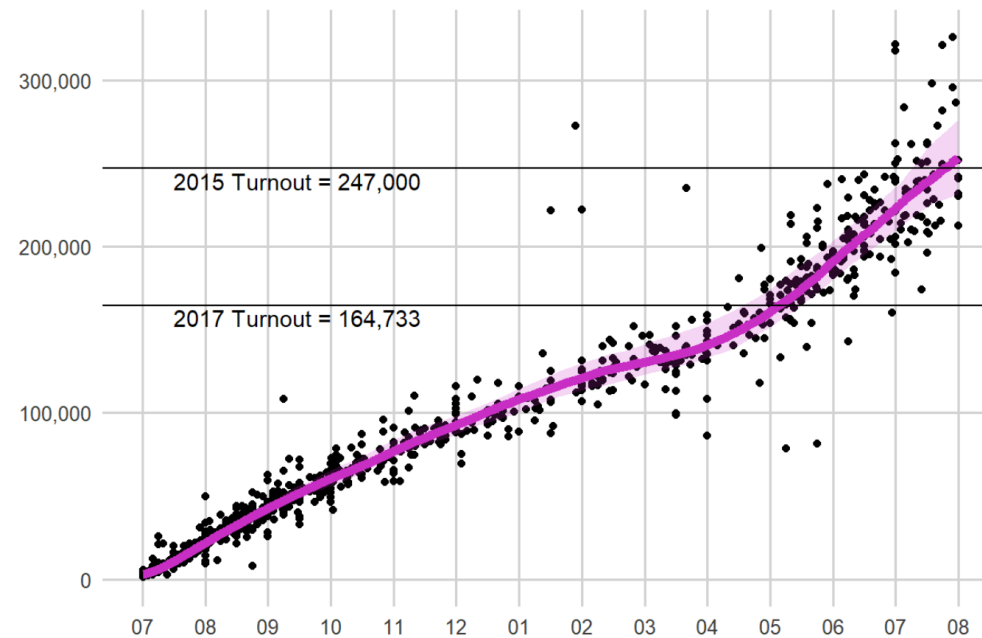


Live 2019 Election Turnout Tracker

NOTE: This is not an example of expected work for this course.

Live Results

Estimated Election Turnout

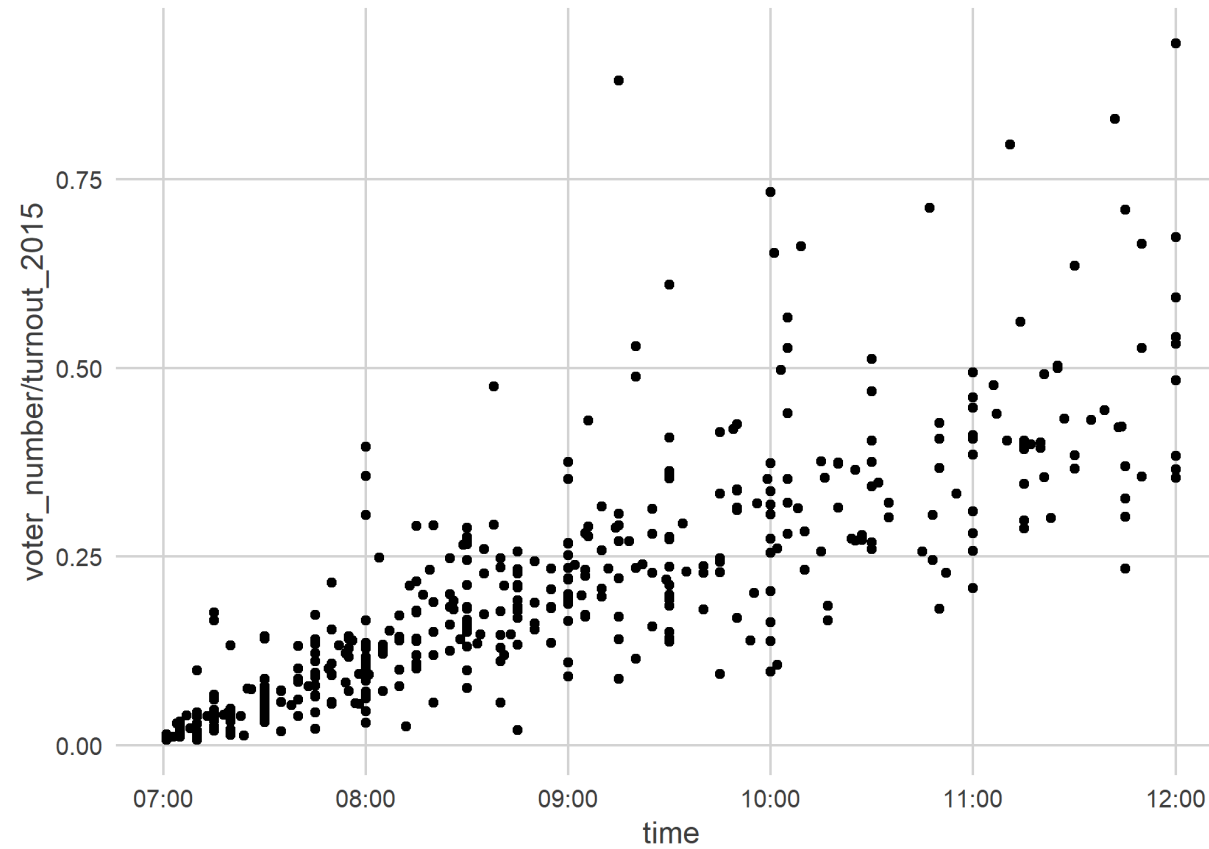


The Turnout Tracker


- If people share where and when they vote in real time, can we predict total turnout?



The Turnout Tracker



The Turnout Tracker

	A	B 	C	D	E	F
1	Timestamp	Ward (1 - 66)	Division (1 - 51)	Time of day	Voter number at	I voted...
2	6/1/2020 8:01:58	27	14	9:00:00 AM	215	by mail
3	6/1/2020 8:47:39	2	5	2:22:00 AM	8	by mail
4	6/1/2020 8:48:25	2	5	3:07:00 AM	567	by mail
5	6/1/2020 9:04:45	18	17	12:00:00 PM	69	by mail
6	6/1/2020 9:16:40	18	17	12:00:00 PM	69	by mail
7	6/1/2020 12:27:00	36	35	12:00:00 AM	0	by mail
8	6/1/2020 12:33:32	18	17	6:00:00 PM	2	by mail
9	6/1/2020 13:30:37	27	11	10:00:00 AM	0	by mail
10	6/2/2020 7:32:18	18	10	7:00:00 AM	1	by mail
11	6/2/2020 8:17:07	39	46	7:00:00 AM	1	by mail
12	6/2/2020 8:19:45	60	08	7:22:00 AM	22	by mail
13	6/2/2020 8:25:46	2	24	3:00:00 PM	22	by mail
14	6/2/2020 8:28:30	22	1	4:20:00 PM	69	by mail
15	6/2/2020 8:32:02	14	4	11:00:00 AM	20	by mail
16	6/2/2020 8:32:08	36	2	12:34:00 PM	55378008	by mail
17	6/2/2020 8:35:04	34	30	8:00:00 AM	20	by mail
18	6/2/2020 9:01:06	36	37	4:20:00 PM	255	by mail
19	6/2/2020 9:07:06	27	6	12:00:00 PM	68	by mail

The Turnout Tracker



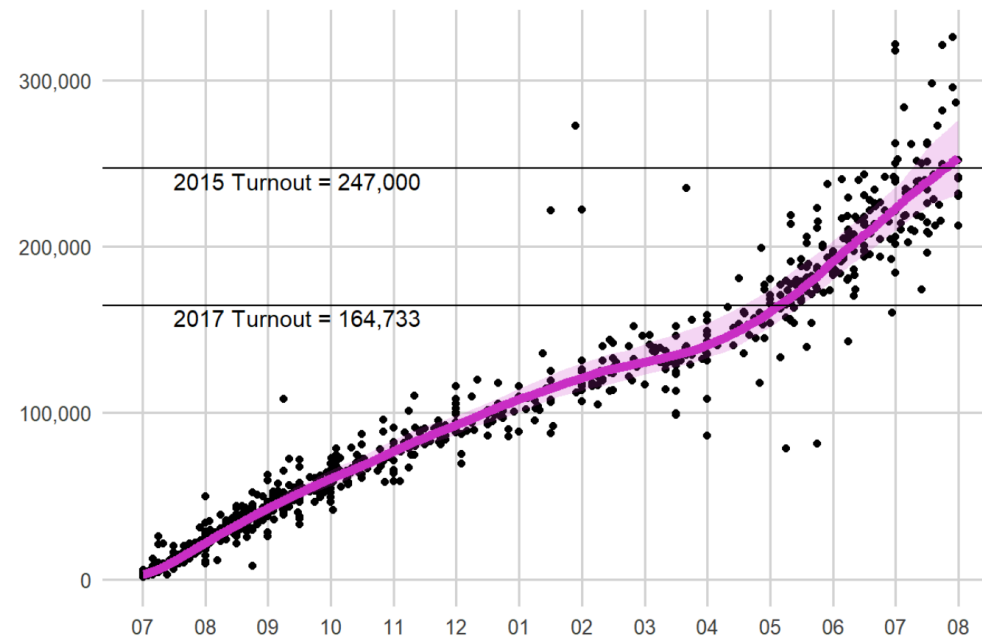
Live 2019 Election Turnout Tracker

Welcome to the [Sixty-Six Wards](#) turnout tracker! Voters across Philadelphia are sharing their turnout to support citizen science.

First, vote! Then, share your division and Voter Number at <http://bit.ly/sixtysixturnout>.

Live Results

Estimated Election Turnout



A meta view: the nature of this work

Table 1 | A schematic for organizing empirical modelling along two dimensions, representing the different levels of emphasis placed on prediction and explanation

	No intervention or distributional changes	Under interventions or distributional changes
Focus on specific features or effects	Quadrant 1: Descriptive modelling Describe situations in the past or present (but neither causal nor predictive)	Quadrant 2: Explanatory modelling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	Quadrant 3: Predictive modelling Forecast outcomes for similar situations in the future (but can break under changes)	Quadrant 4: Integrative modelling Predict outcomes and estimate effects in as yet unseen situations

From Hofman, J.M., Watts, D.J., Athey, S. *et al.* Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).

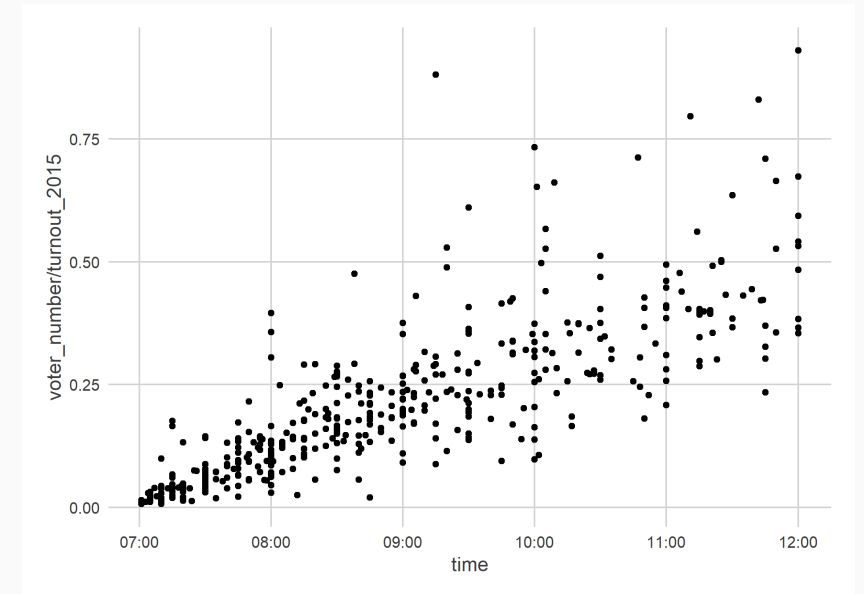
Meta takeaways

- Understanding-focused approach to statistical modeling.
- Practical full-stack problem solving.
- It's crucial that we build systems & playbooks for this style of work.

A naïve approach

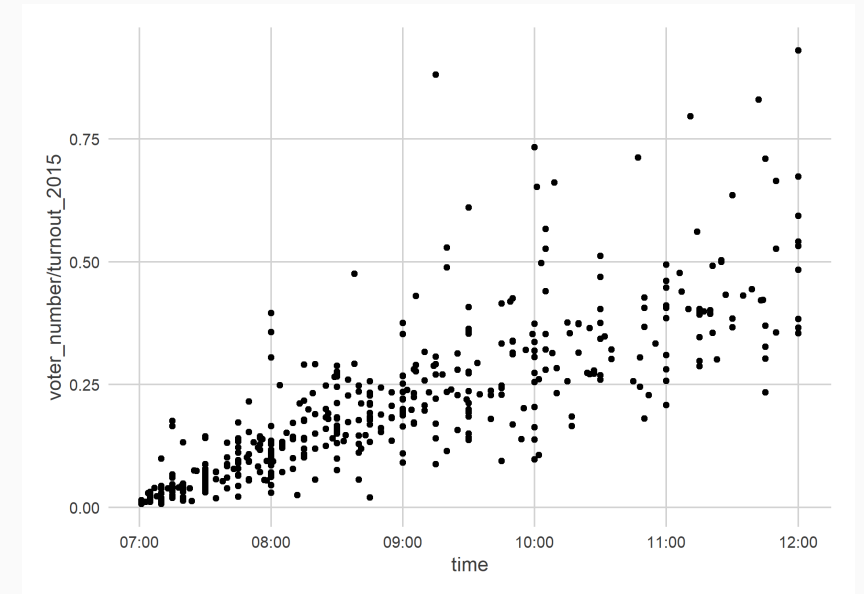
- x_i = response i
- t_i = time of response i (in hours)
- d_i = division of response i
- T_y = City-wide turnout in year y
- $T_{d,y}$ = Final turnout for division d in year y

$$\hat{T}_{2019} = T_{2015} \cdot \text{avg} \left(\frac{x_i}{T_{d_i,2015}} \cdot \frac{13}{t_i} \right)$$



Turnout Tracker: the challenges

- Different divisions have different baseline turnouts.
- Divisions may swing together.
- We don't know the time pattern.
- There's *definitely* selection bias into who shares.
- Knowing uncertainty in the estimate is everything.



$$\log(x_i) = \alpha_{y_i} + \mu_{d_i} + \gamma_{d_i y_i} + f(t_i) + e_i$$

Turnout Tracker: the model

At time t in division d , year y the voter number x_i so far is

$$\log(x_i) = \alpha_{y_i} + \mu_{d_i} + \gamma_{d_i y_i} + f(t_i) + e_i$$

Turnout Tracker: the model

At time t in division d , year y the voter number x_i so far is

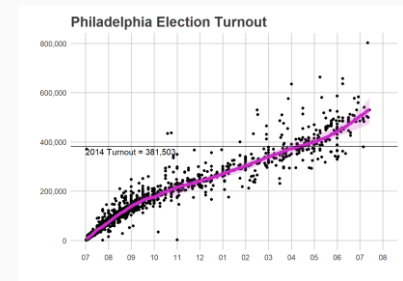
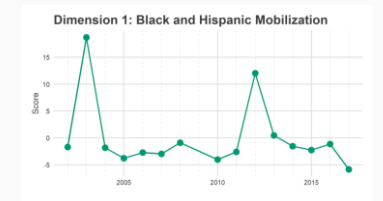
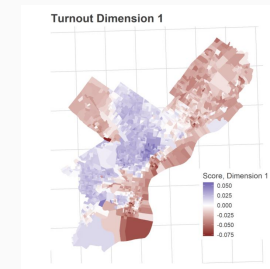
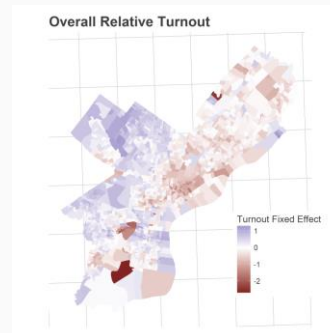
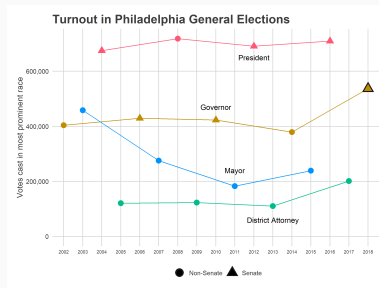
$$\log(x_i) = \alpha_{y_i} + \mu_{d_i} + \gamma_{d_i y_i} + f(t_i) + e_i$$

Overall turnout for
this year

Usual turnout for
this division

This year's "swing"
for the division

Time pattern



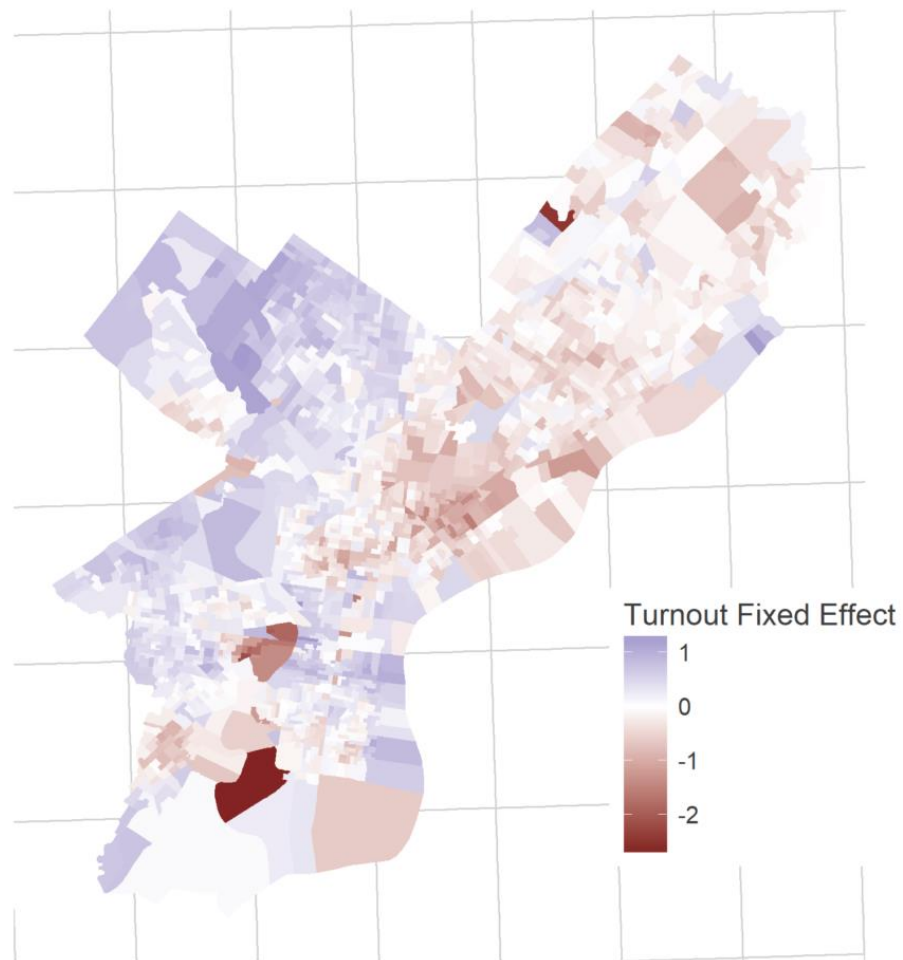
Estimating turnout

$$\log(x_i) = \alpha_{y_i} + \mu_{d_i} + \gamma_{d_i y_i} + f(t_i) + e_i$$

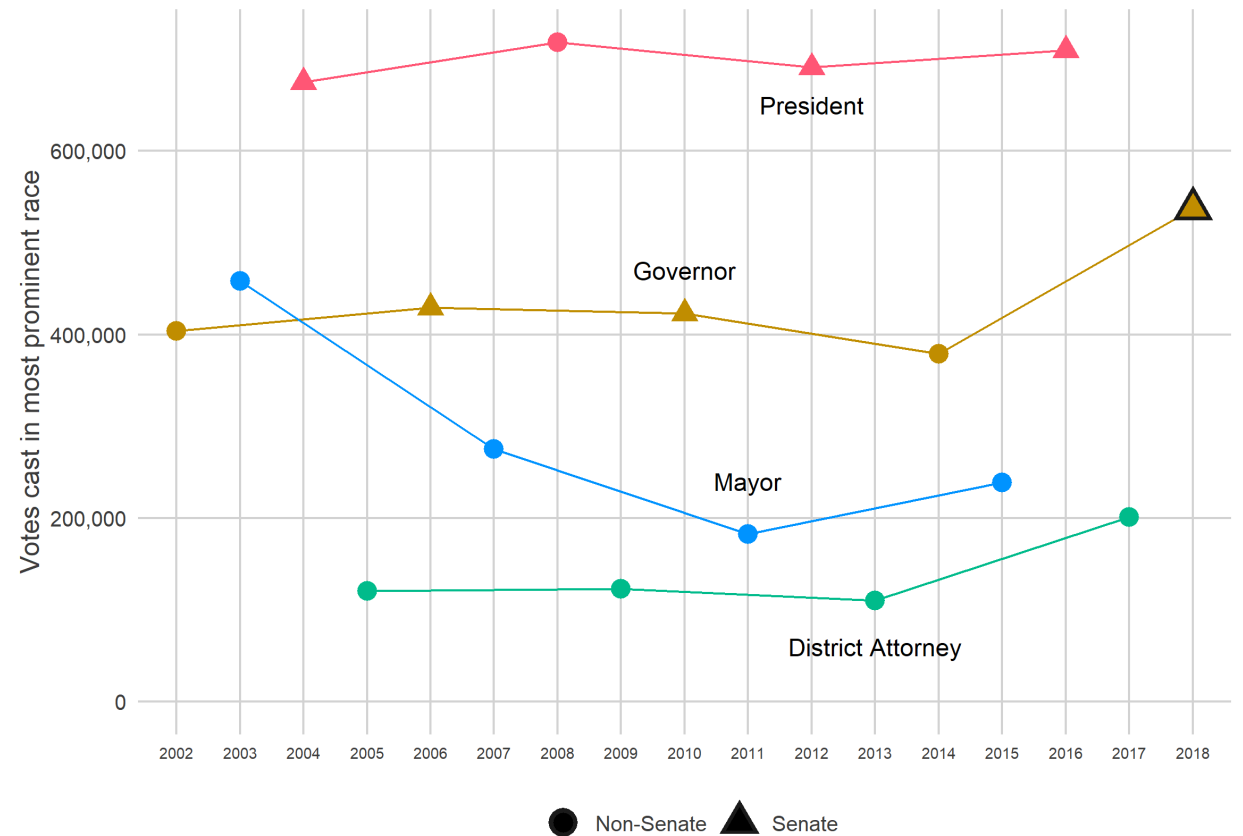
- We can estimate μ_{d_i} from historic data.
- $\gamma_{d_i y_i}$ needs to allow divisions to covary.
 - Model it as $\gamma_{d_i y_i} \sim N(0, \Sigma)$
 - Can estimate Σ from historic data.
- $\alpha_{y_i} + f(t_i)$ need to be estimated in real time.

Estimating Turnout: Baseline Levels μ

Overall Relative Turnout



Turnout in Philadelphia General Elections



Estimating Turnout: Correlated Districts

Singular Value Decomposition

We are interested in a giant $D \times D$ matrix of how all divisions correlate with each other. This would require more than 1,703 elections.

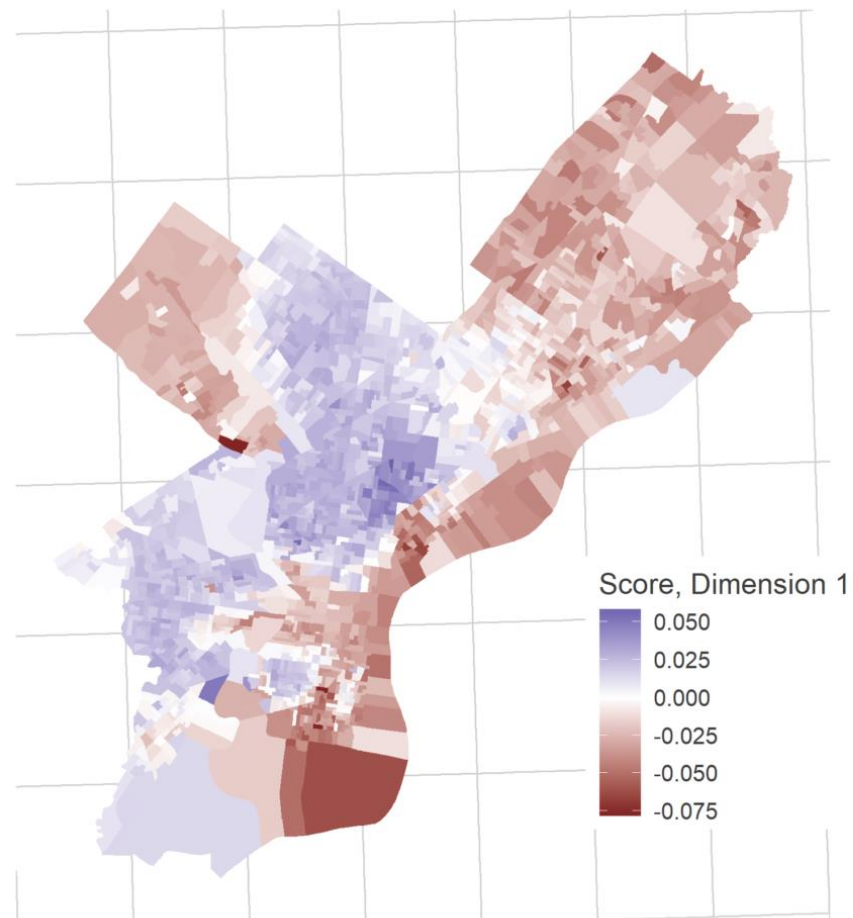
Instead, find an approximation using “dimension reduction”. (Note: I do this for the $D \times E$ matrix of Divisions to Elections).

$$\Sigma = U' D V$$

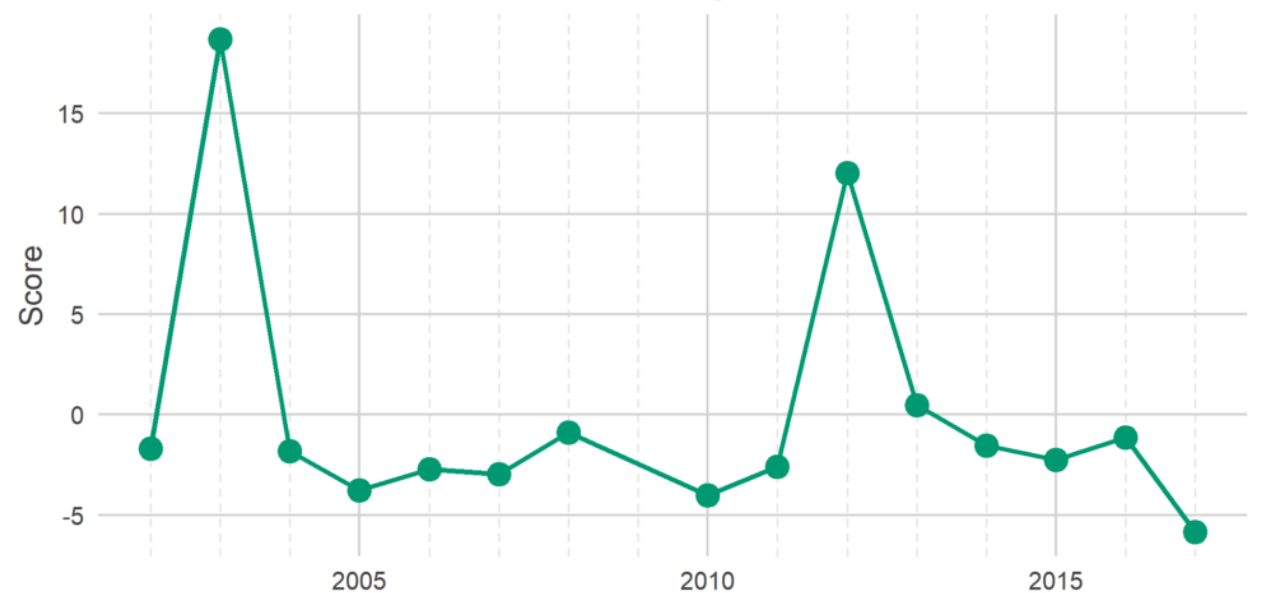
Σ_{11}	Σ_{12}	Σ_{13}	$=$	U_{11}	U_{12}	$\begin{vmatrix} D_1 & 0 \\ 0 & D_2 \end{vmatrix}$	$\begin{vmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \end{vmatrix}$
Σ_{21}	Σ_{22}	Σ_{23}		U_{21}	U_{22}		
Σ_{31}	Σ_{32}	Σ_{33}		U_{31}	U_{32}		
Σ_{41}	Σ_{42}	Σ_{43}		U_{41}	U_{42}		

Estimating Turnout: Correlated Districts

Turnout Dimension 1

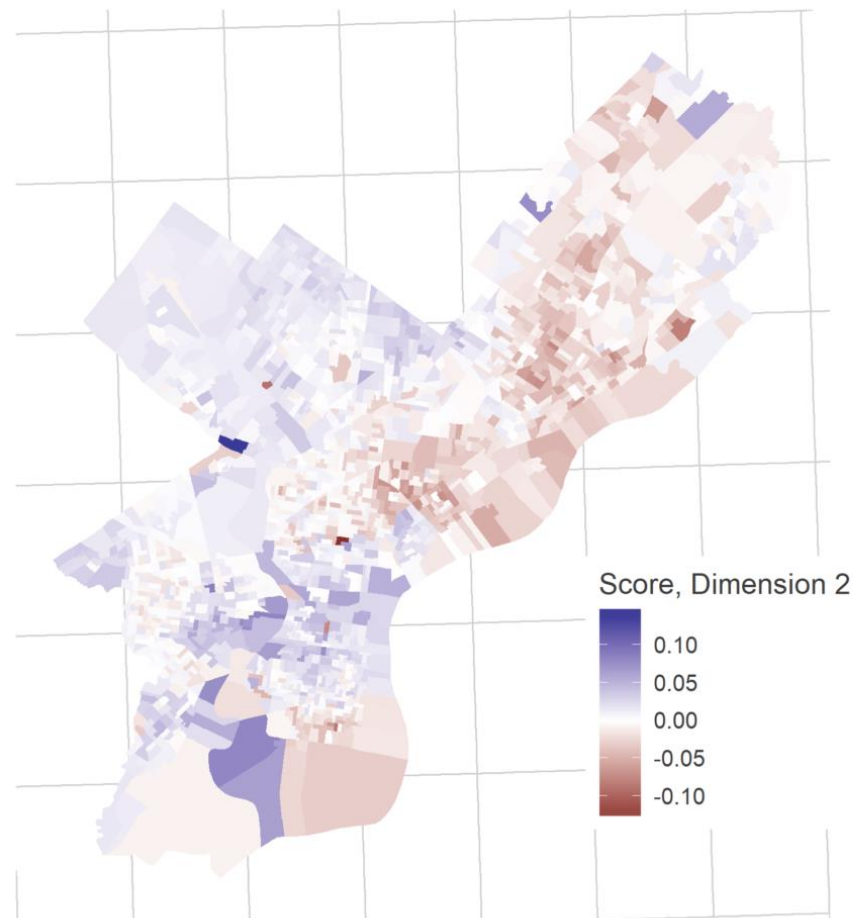


Dimension 1: Black and Hispanic Mobilization

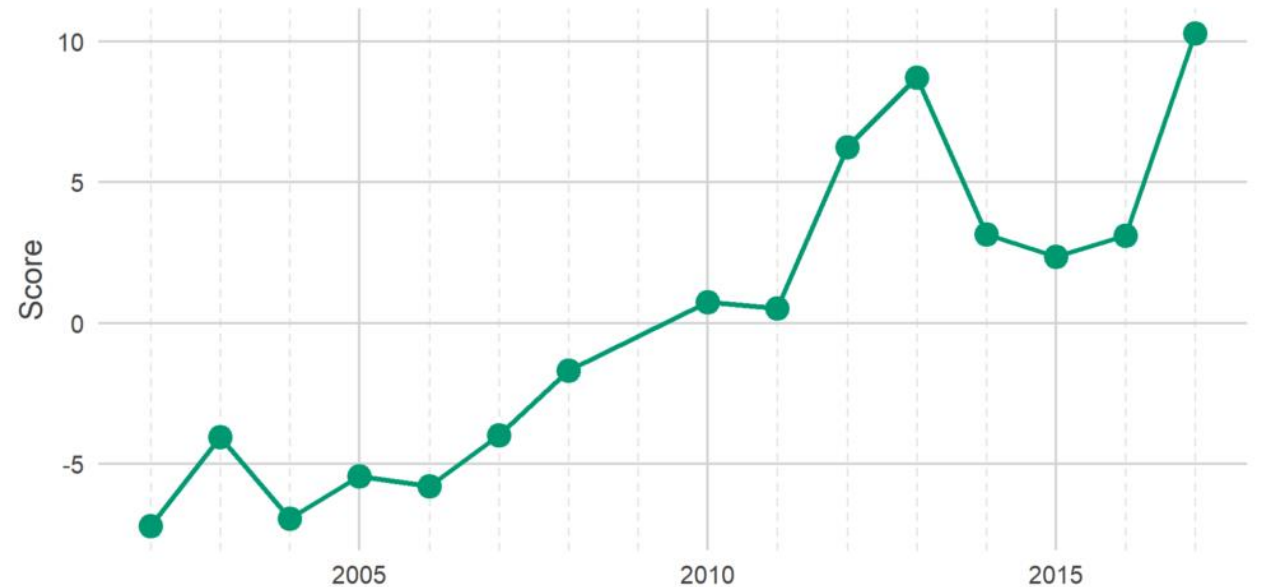


Estimating Turnout: Correlated Districts

Turnout Dimension 2



Dimension 2: Increasing Turnout



Estimating Turnout: Time Pattern

Even after we've adjusted for (a) each division's baseline turnout and (b) covarying swings, how do we compare data at 8am with data from 2pm?

$$\log(x_i) = \alpha_y + \mu_d + \gamma_{dy} + f(t) + e_i$$

Estimating Turnout: Time Pattern

Even after we've adjusted for (a) each division's baseline turnout and (b) covarying swings, how do we compare data at 8am with data from 2pm?

$$\log(x_i) = \alpha_y + \mu_d + \gamma_{dy} + f(t) + e_i$$

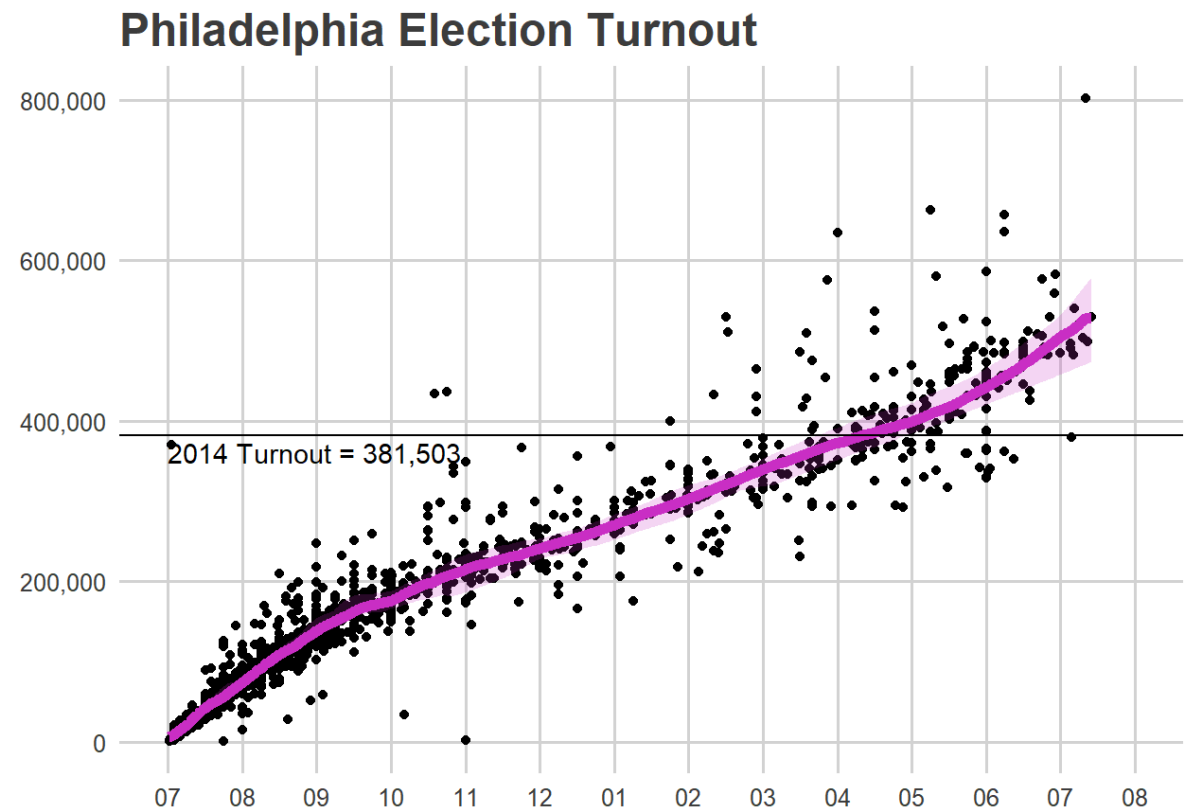
$$\alpha_y + f(t) = E[\log(x_i)] - \mu_d - \gamma_{dy}$$

Estimating Turnout: Time Pattern

Even after we've adjusted for (a) each division's baseline turnout and (b) covarying swings, how do we compare data at 8am with data from 2pm?

$$\log(x_i) = \alpha_y + \mu_d + \gamma_{dy} + f(t) + e_i$$

$$\alpha_y + f(t) = E[\log(x_i)] - \mu_d - \gamma_{dy}$$



Two methods to fit the model

$$\log(x_i) = \alpha_y + \mu_d + \gamma_{dy} + f(t) + e_i$$

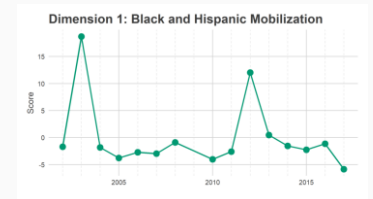
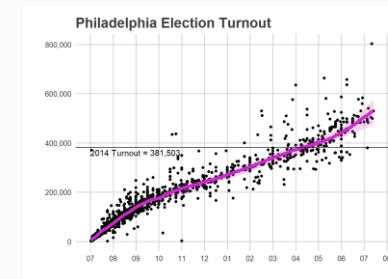
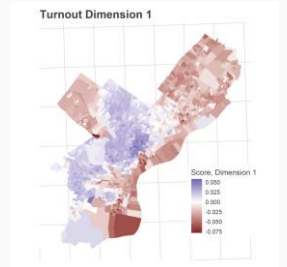
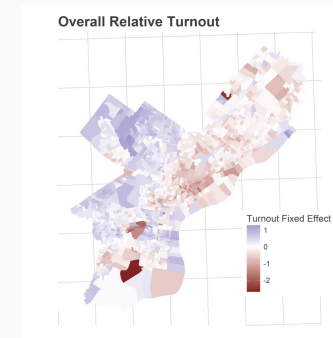
- Iteratively fit random effects γ_{dy} and time effects $\alpha_y + f(t)$.
 - Maximum Likelihood approach. γ_{dy} has closed form solution. Use loess smoother for $\alpha_y + f(t)$.
- Model $f(t)$ as a Gaussian Process, so everything is a conditional normal.

Turnout Tracker: the challenges

- Different divisions have different baseline turnouts.
- Divisions may swing together.
- We don't know the time pattern, and no ground truth.
- There's *definitely* selection bias into who shares.
- Knowing uncertainty in the estimate is everything.

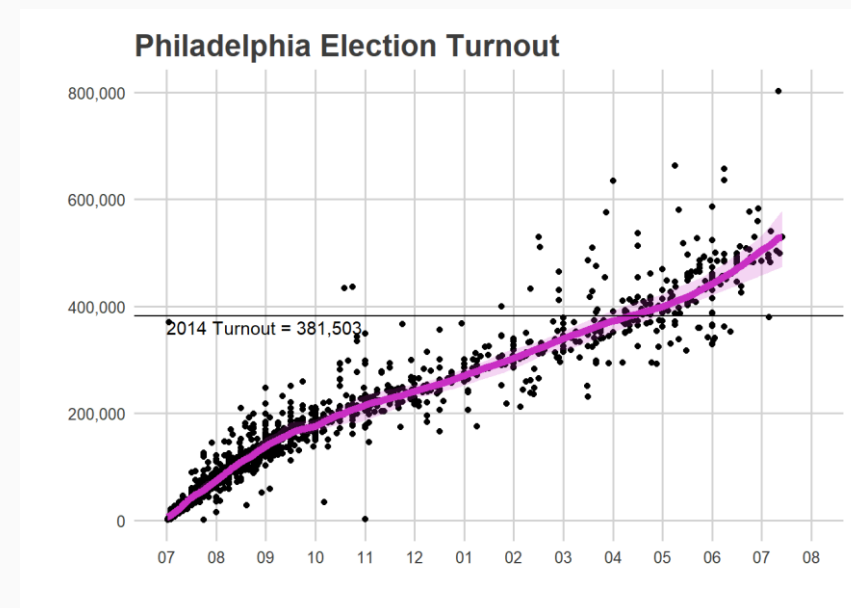
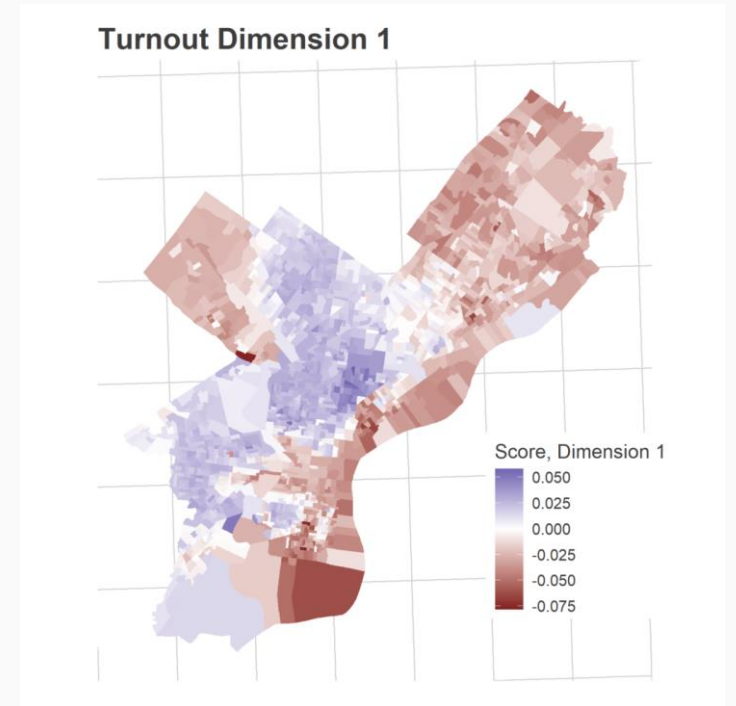
Turnout Tracker: the challenges

- Different divisions have different turnouts.
 - Use each division's baseline rate from past elections.
- Divisions may swing together.
 - Estimate the covariance in divisions' swings from election to election.
- We don't know the time pattern , and no ground truth.
 - Estimate it on the fly. 😏
- There's *definitely* selection bias into who shares.
 - oof.
- Knowing uncertainty in the estimate is everything. ???



Selection bias & Uncertainty

- If selection is independent of γ_d conditional on Σ , we're safe.
- Use bootstrapping to generate uncertainty.
- How do we know if this works?
 - Integration Tests!



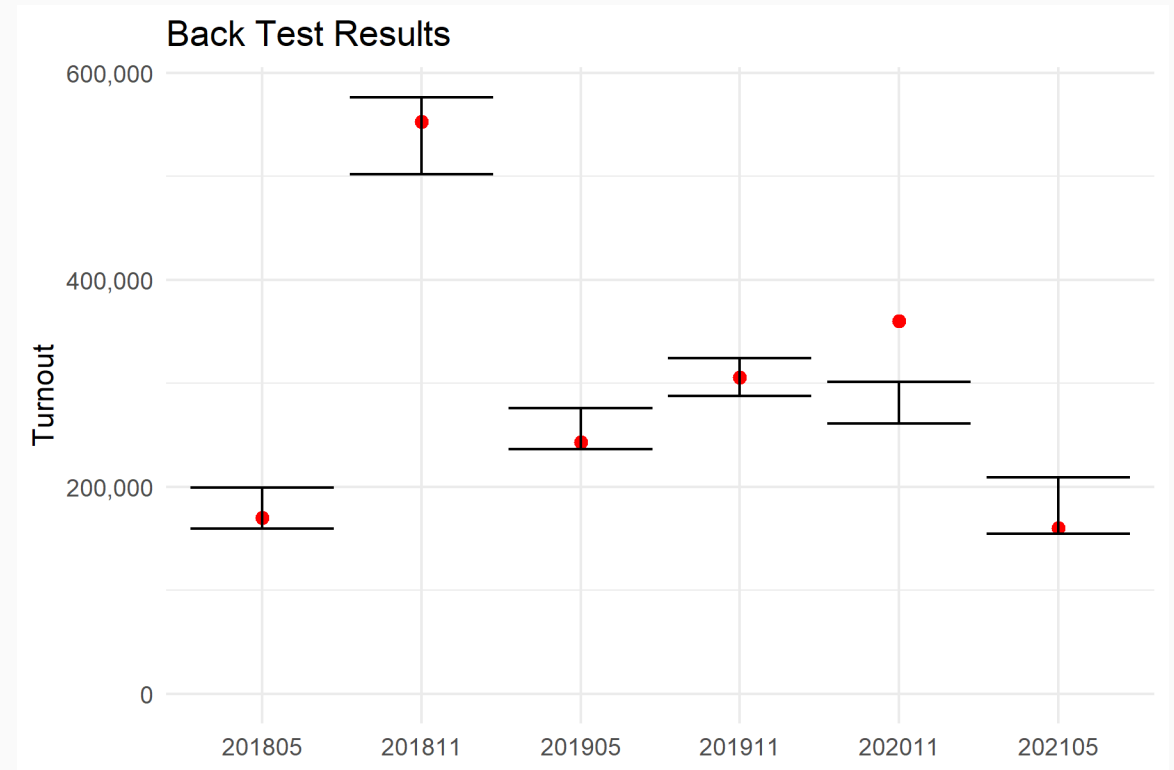
Practical notes for a live tool

- Need aggressive handling of outliers.
 - Winsorizing and sanity checks.
- Provide understanding-based visualizations.
 - Maps, plots of intermediate calculations.
- *Everything you look at to debug should be a unit test or an output.*

Integration/Back Tests

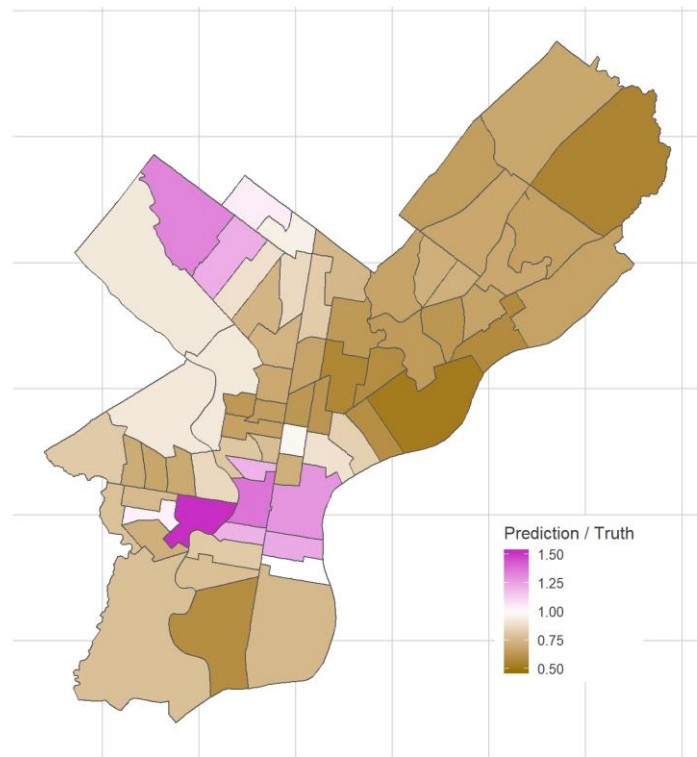
- Test full runs of the package on known results.
- Tests for
 - Composition errors
 - Correctness errors

```
1 library(testthat)
2 library(dplyr)
3
4
5 TRACKER_DIR <- "C:/Users/Jonathan Tannen/Dropbox/sixty_six/posts/turnout_
6 olddir <- setwd(TRACKER_DIR)
7 METHOD <- "loess"
8
9 source("R/load_data.R", chdir=TRUE)
10 source("R/fit_submissions.R", chdir=TRUE)
11 source("R/bootstrap.R", chdir=TRUE)
12 source("R/precalc_params.R", chdir=TRUE)
13
14 setwd(olddir)
15
16 test_elections <- tribble(
17   ~folder, ~turnout,
18   "phila_202105", 160e3,
19   "phila_202011", 360e3,
20   "phila_201911", 306e3,
21   "phila_201905", 243e3,
22   "phila_201811", 553e3,
23   "phila_201805", 170e3
24 )
25
```

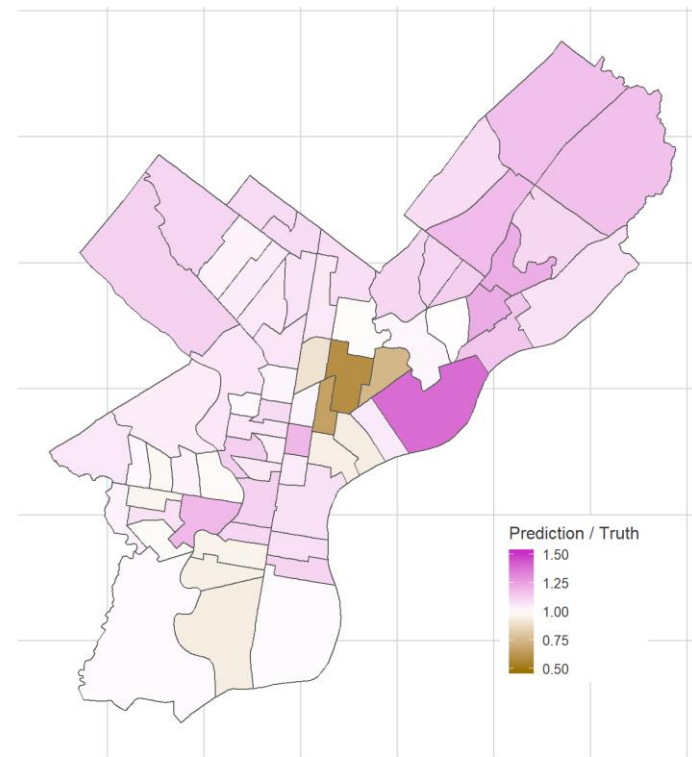


What broke the Tracker in 2020?

Tracker Predictions vs Eventual Results

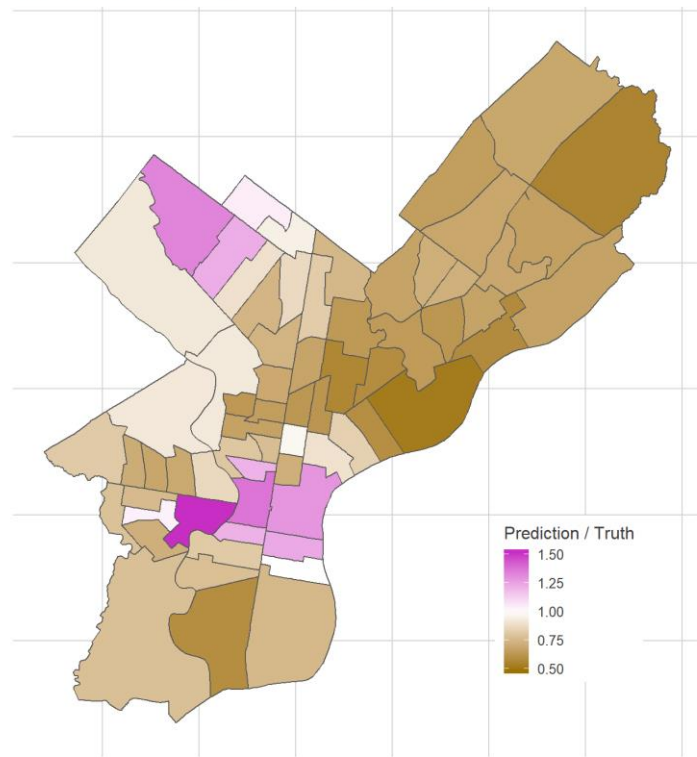


Tracker Predictions vs Eventual Results, 2019 Primary

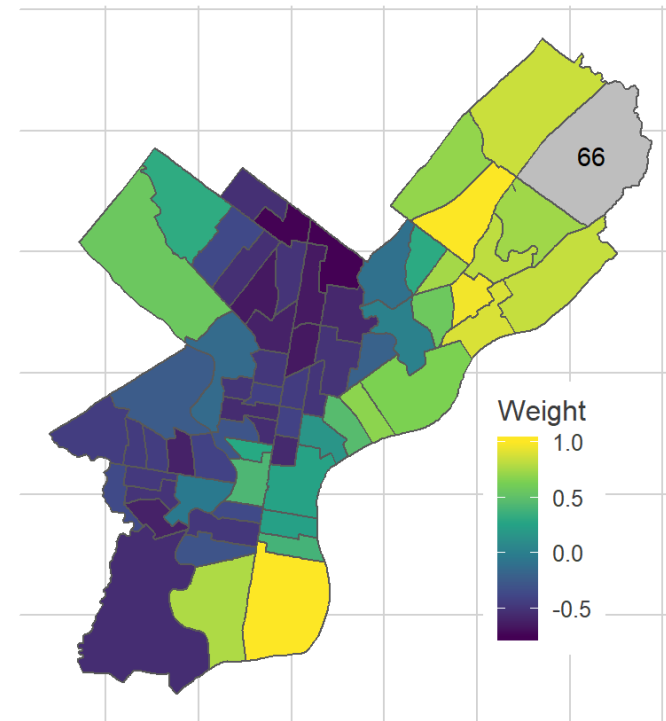


What broke the Tracker in 2020?

Tracker Predictions vs Eventual Results

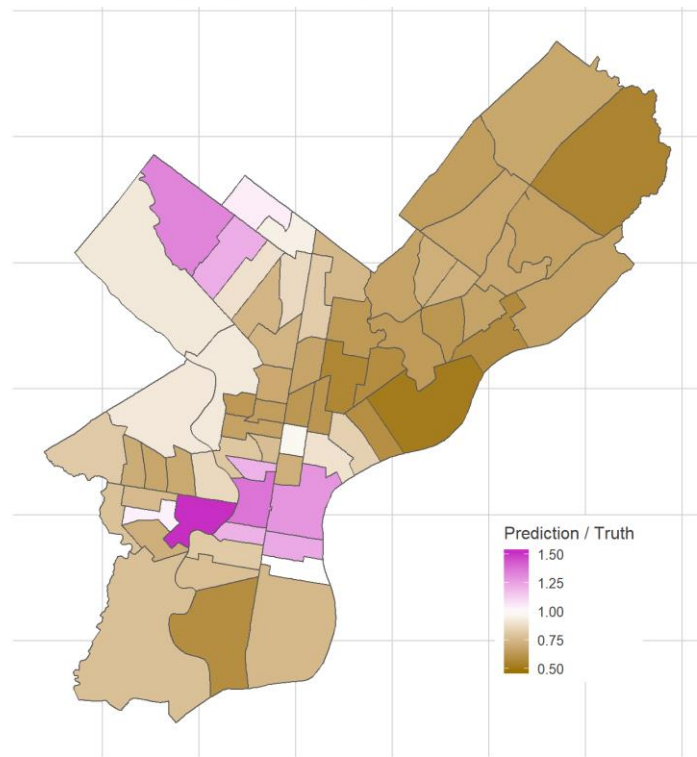


Weights used to predict Ward 66



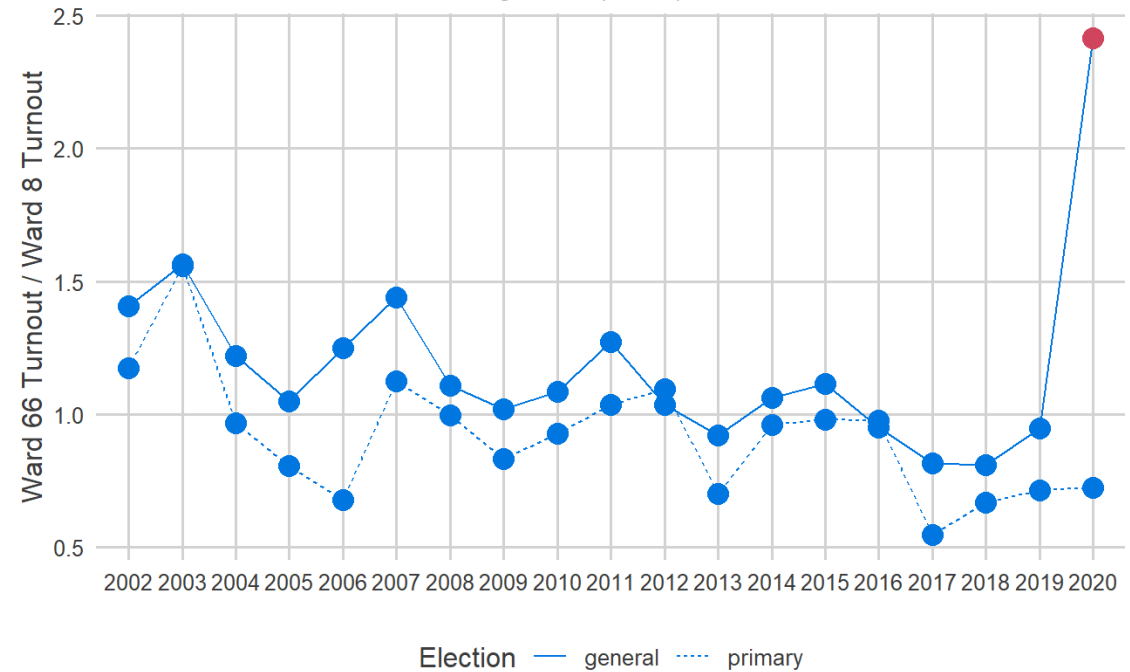
What broke the Tracker in 2020?

Tracker Predictions vs Eventual Results



Distribution of 66th Ward turnout vs 8th

Elections from 2002 to the 2020 general (in red).



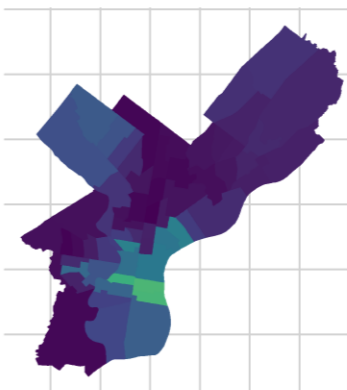
The Turnout Tracker: The Tech

- Google Form to collect responses
- R functions that download google sheet, process data.
- R functions that generate predictions and bootstraps.
- RMarkdown document generates HTML report.
- httr command to push html to website.
- TODO:
 - Docker Container

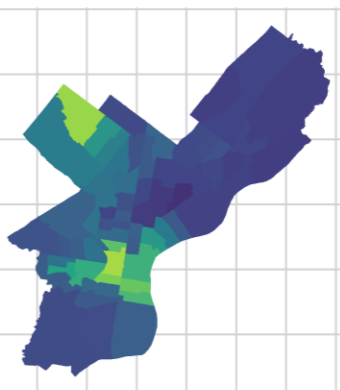
https://github.com/jtannen/turnout_tracker

Philadelphia's Voting Blocs

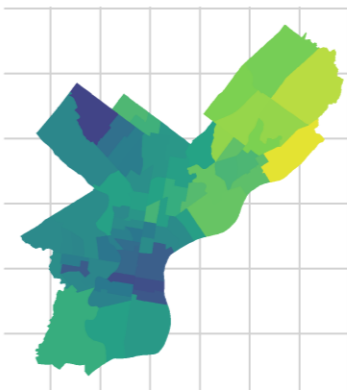
Jen Devor



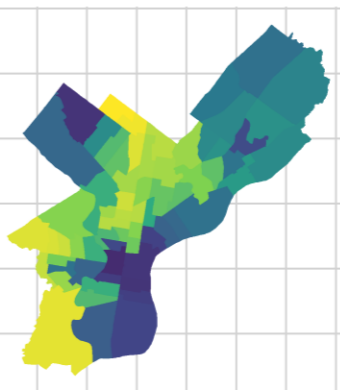
Kahlil Williams



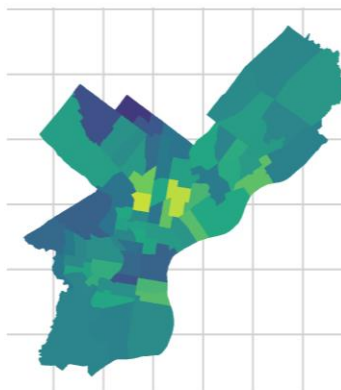
Lisa Deeley



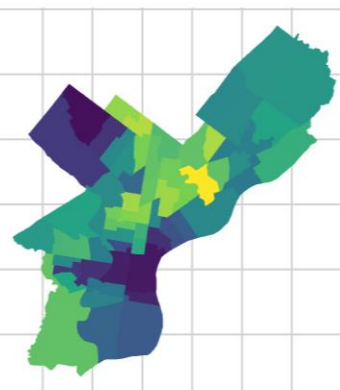
Omar Sabir



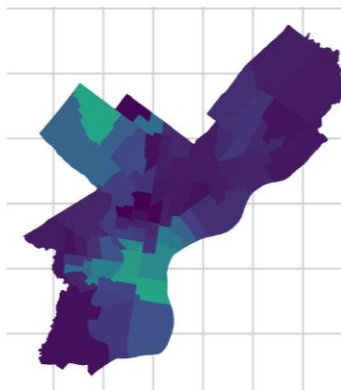
Jennifer Schultz



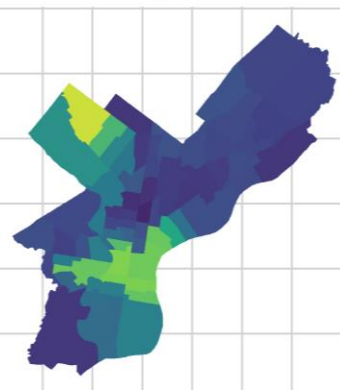
Joshua Roberts



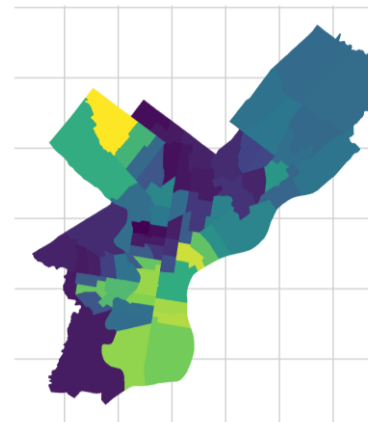
Kay Yu



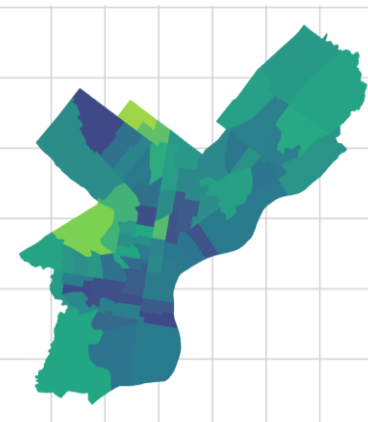
Tiffany Palmer



Justin Diberardinis



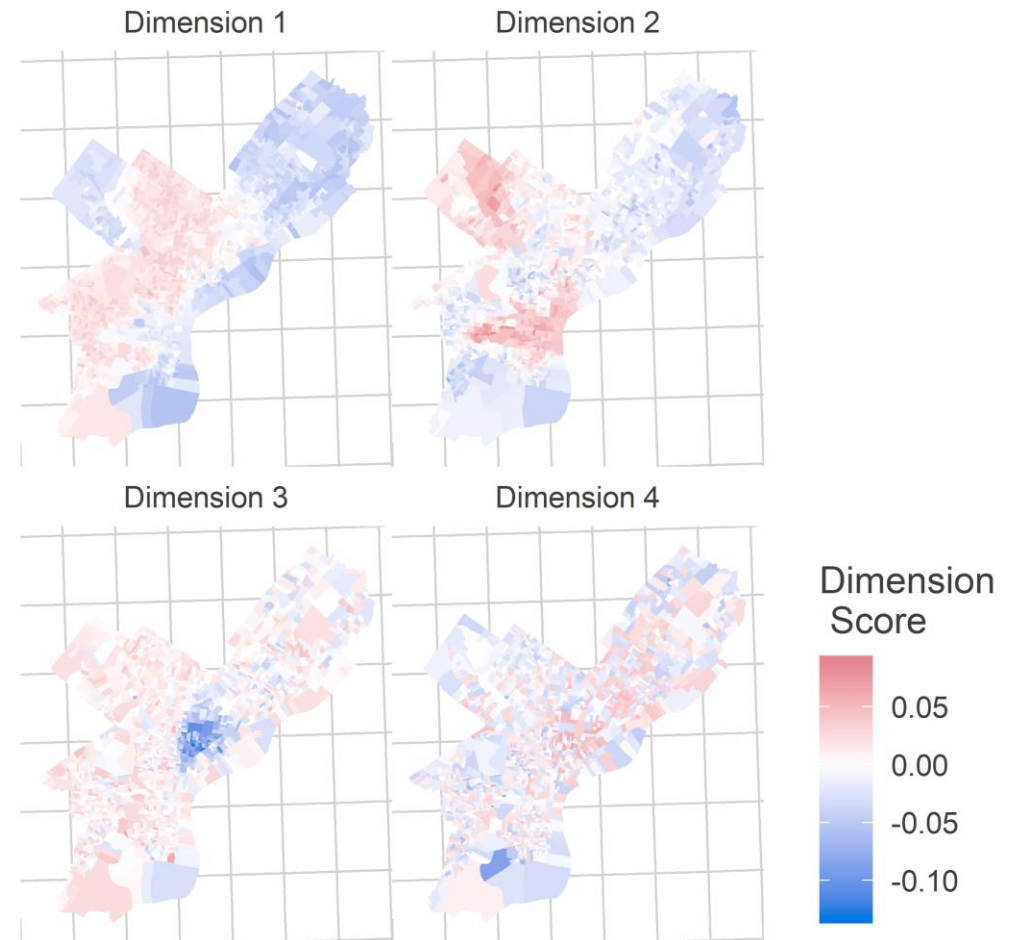
Katherine Gilmore Richardson



Philadelphia's Voting Blocs

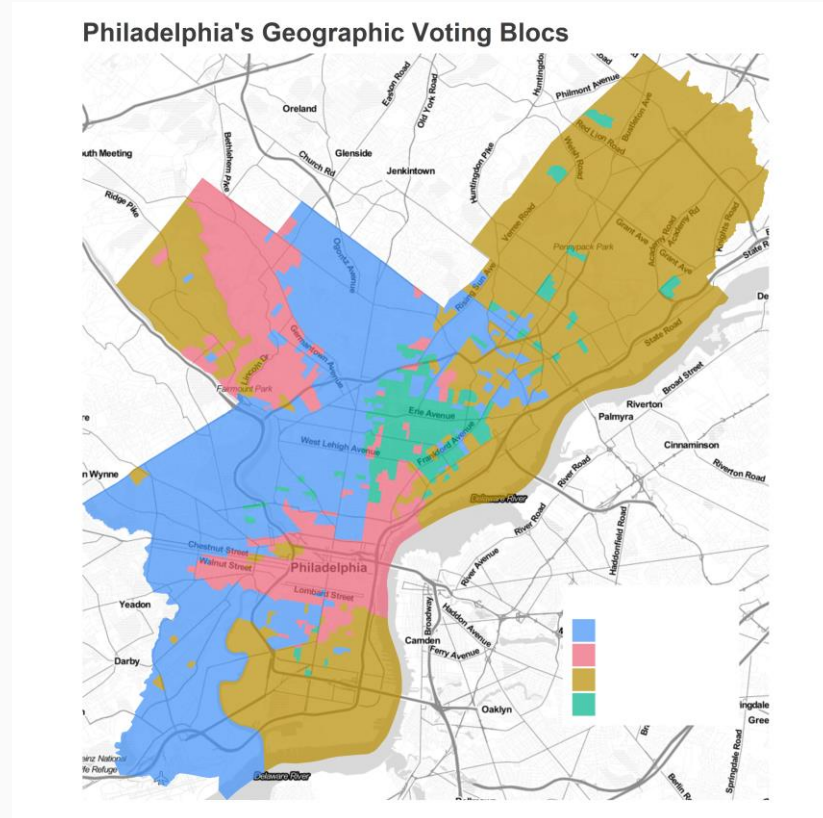
SVD to the rescue!

Dimensions of Philadelphia's Votes

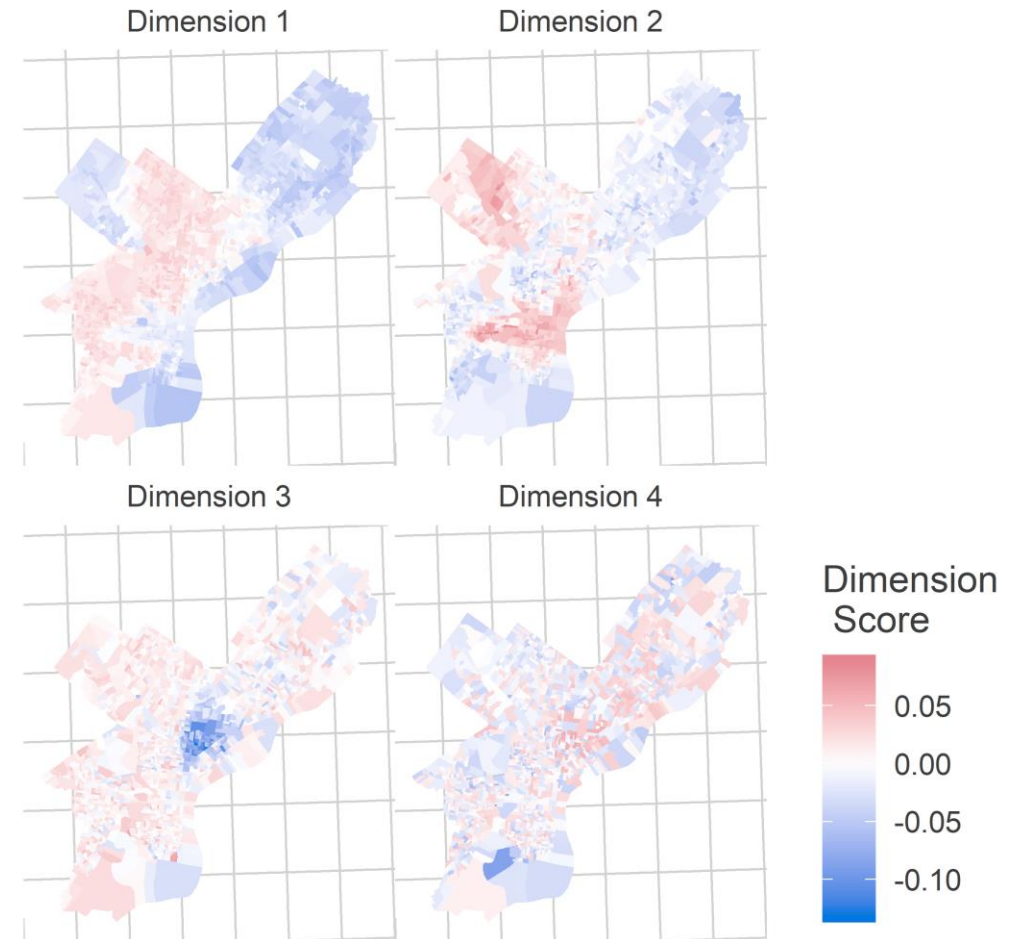


Philadelphia's Voting Blocs

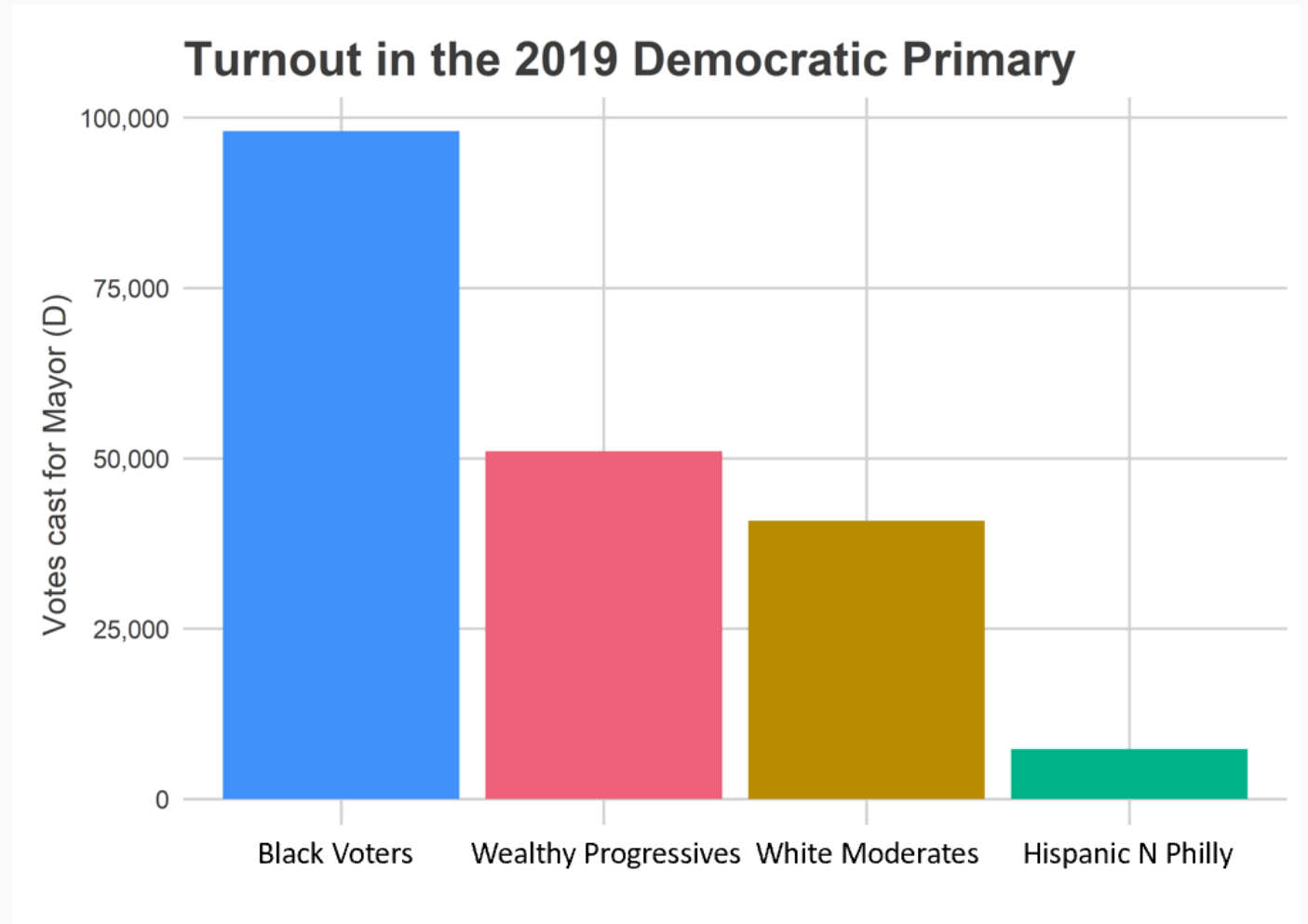
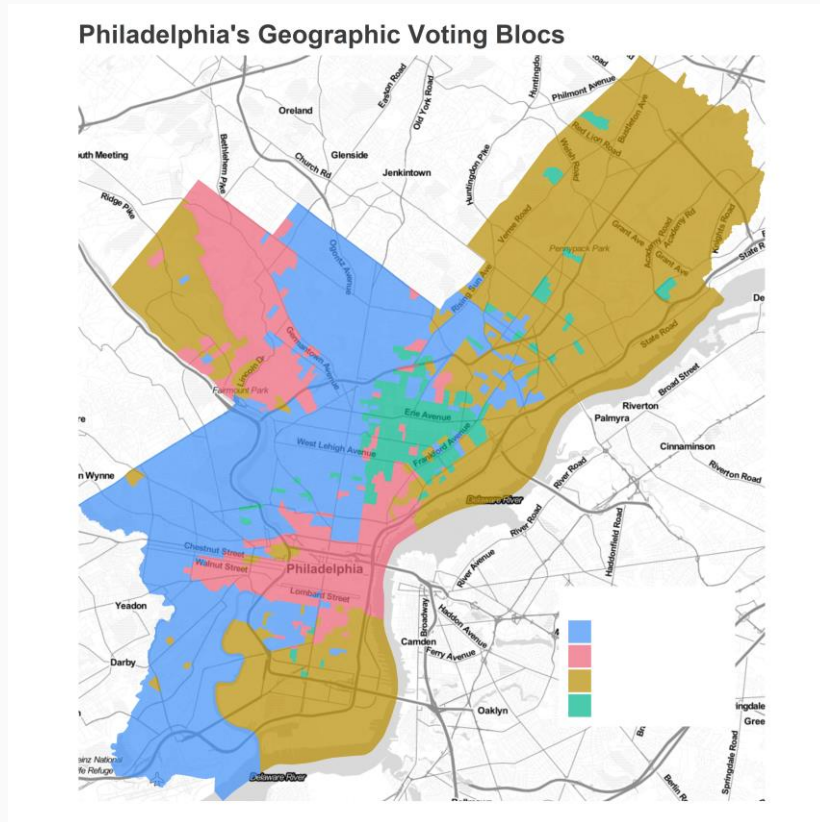
SVD + buckets to the rescue!



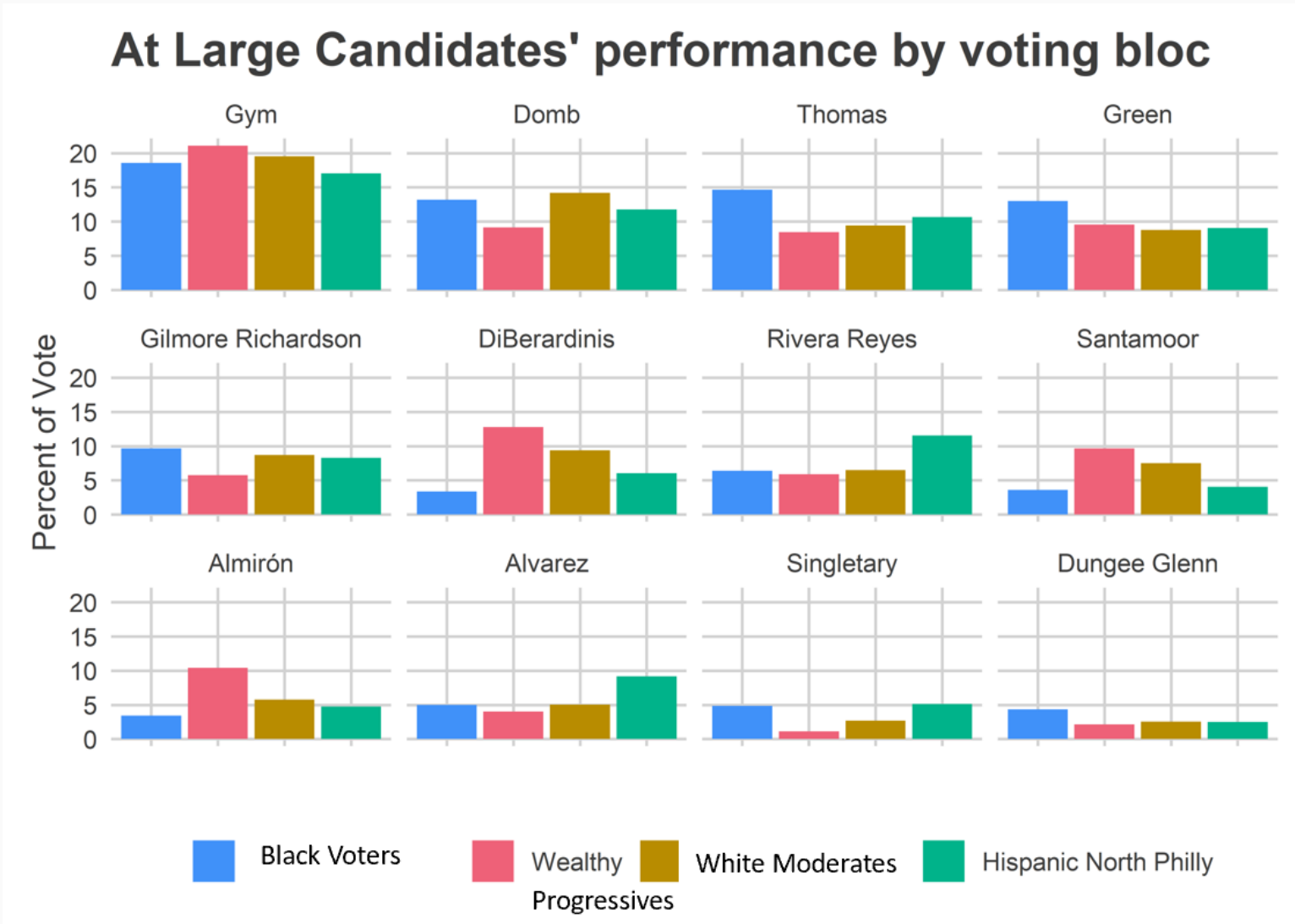
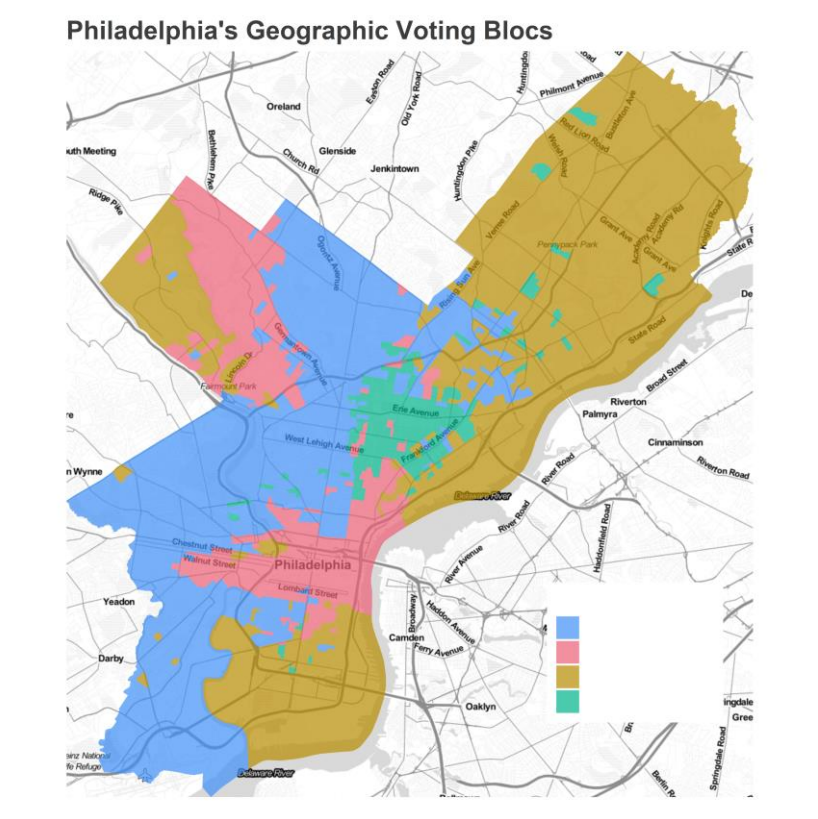
Dimensions of Philadelphia's Votes



Philadelphia's Voting Blocs



Philadelphia's Voting Blocs



The end!

Questions?

Appendix

Estimating Turnout: Time Pattern

$$\log(x_i) = \alpha_y + \mu_d + \gamma_{dy} + f(t)$$

The total city turnout at the end of the day is...

$$\begin{aligned} & \sum_d \exp\{\alpha_y + \mu_d + \gamma_{dy} + f(t)\} \\ &= \exp\{\alpha_y + f(t)\} \sum_d \exp\{\mu_d + \gamma_{dy}\} \end{aligned}$$

