Statistical Overview and Visualization of Election Data

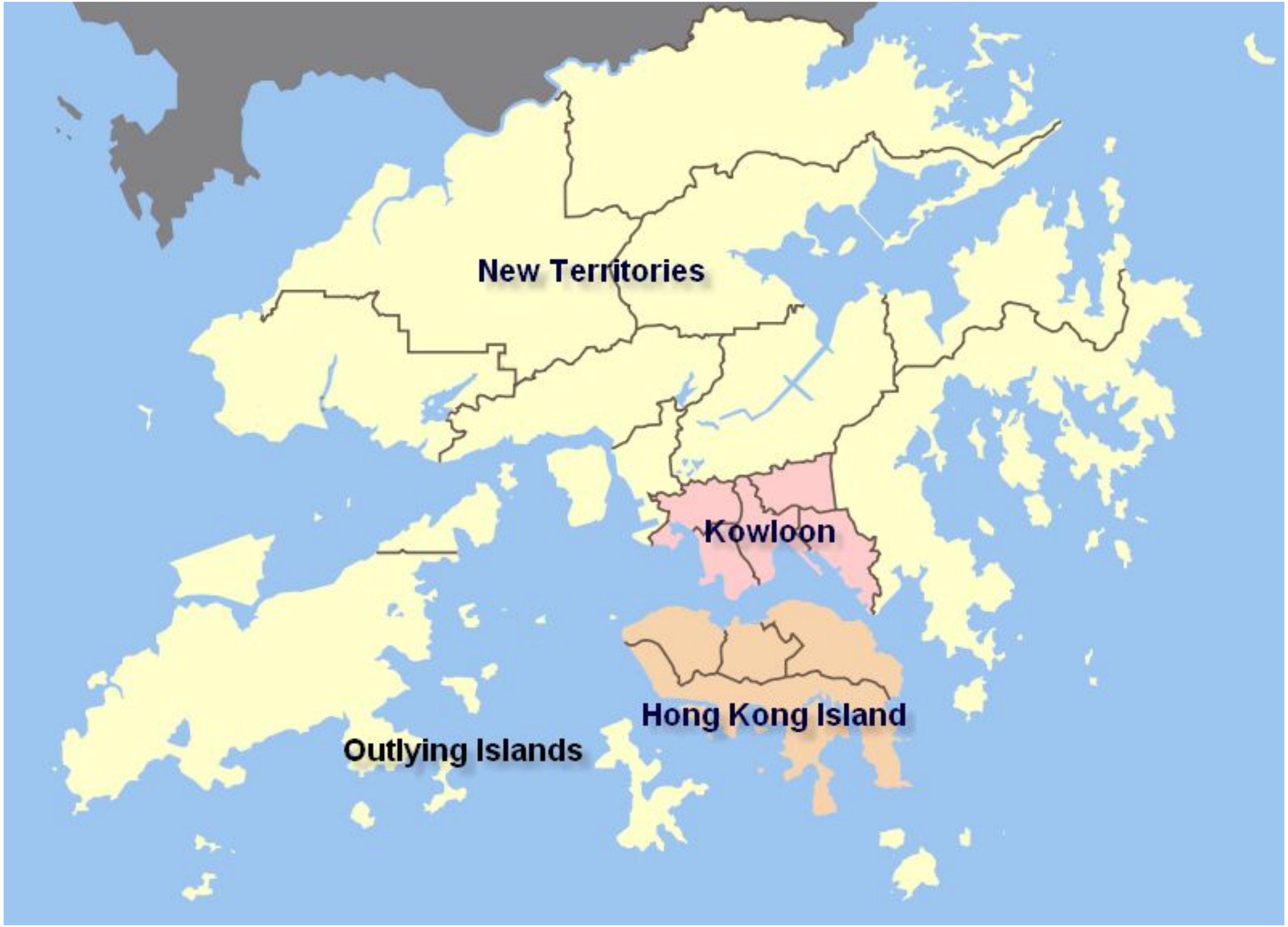
Mentor: Dr. Kwai Wong
Student: Ng Ching Tao (CityU), Jorge Garcia (NMSU), Frank Betancourt (UTK)

Objectives of the project

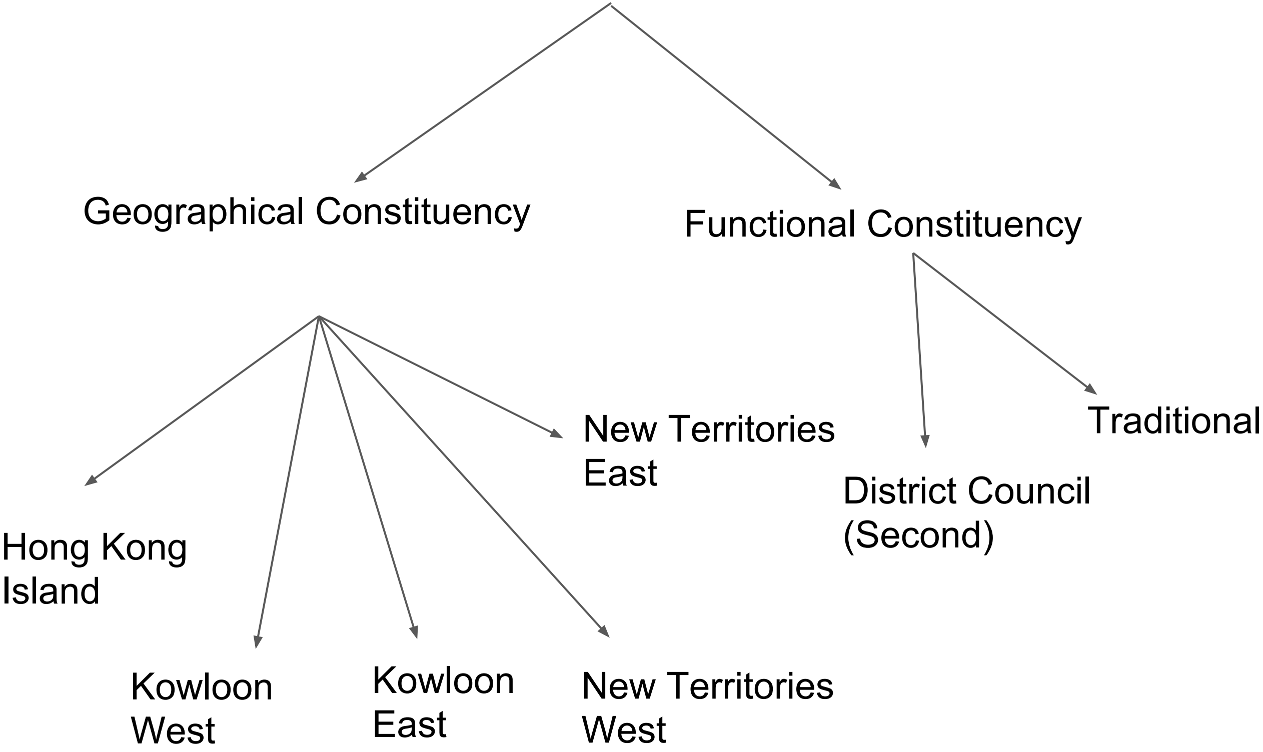
- Using Pearson Product Moment Correlation to quantify how certain a factor affects electoral results
- Experiment with feasibility of predicting electoral results with deep neural networks
- Visualize election data

Background of information

	Hong Kong (HK)	United States (U.S.)
Source of data	1. Public Opinion Program, the University of Hong Kong (HKUPOP) (Year: 2008, 2012, 2016)	1. ANES Time Series Study (ANES) (Year: 1992-2016) 2. American Community Survey (US Census Bureau) (Year: 2015)
Political Battle	Pro- government vs Pro- choice	Republican vs Democrats
Election	Hong Kong Legislative Council Election	House of Representatives


New Territories
Kowloon
Hong Kong Island
Outlying Islands

Hong Kong Legislative Council Election



Geographical Constituency

Functional Constituency

New Territories East

Hong Kong Island

Kowloon West

Kowloon East

New Territories West

District Council (Second)

Traditional

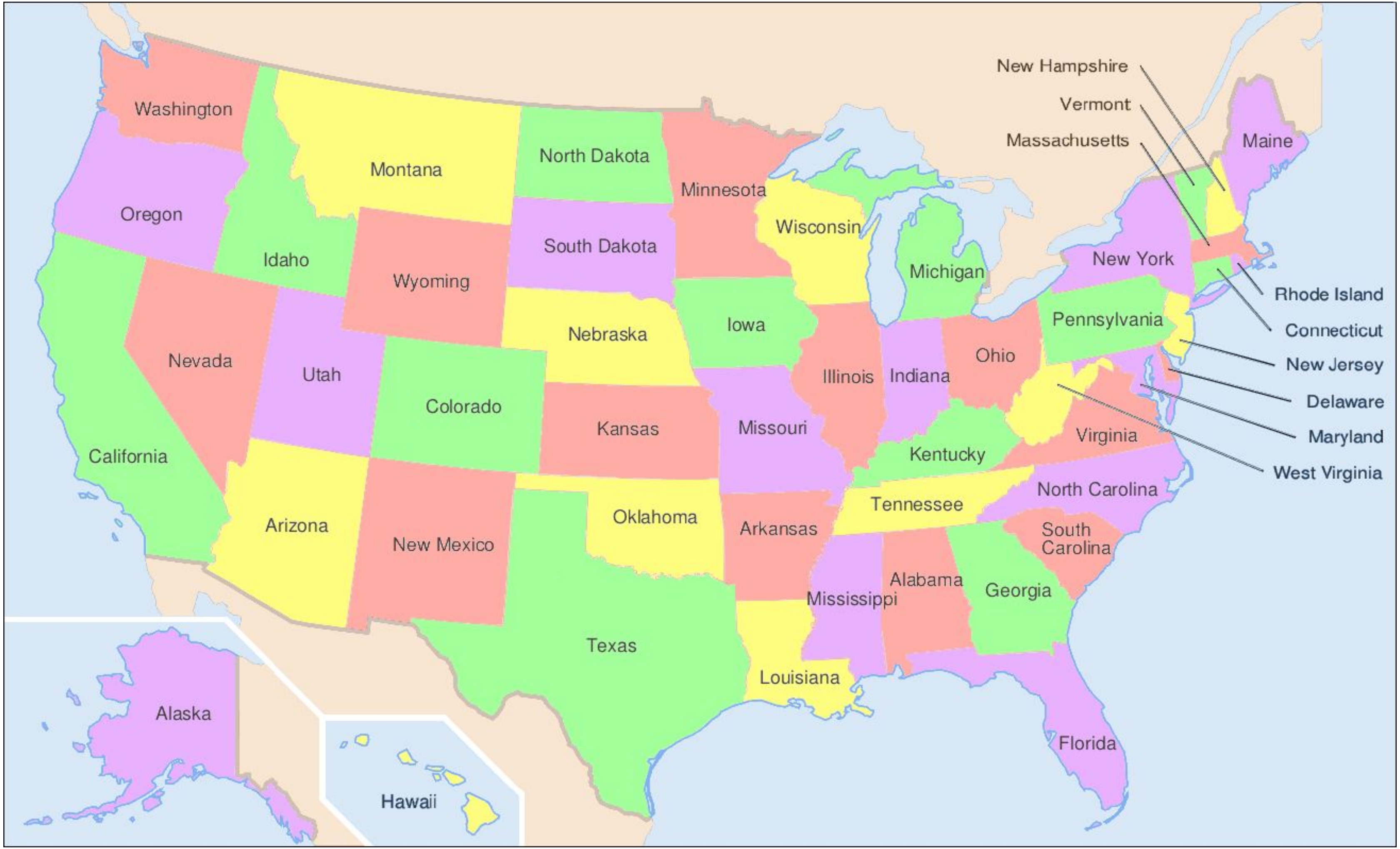
*Traditional functional constituency is not included

Dominant factors across election period

	Hong Kong Island	Kowloon West	Kowloon East	New Territories West	New Territories East	District Council (Second)
2008	Political inclination	Political inclination	Political inclination	Political inclination	Political inclination	/
2012	Emphasis on relationship with central government raised by candidate	Follow strategic plan raised by candidate	Emphasis on relationship with central government raised by candidate	Voting decision	Education level	Political Inclination
2016	Voting decision	Voting decision	Education level	Voting decision	Voting decision	Voting decision

Dominant factors on election day

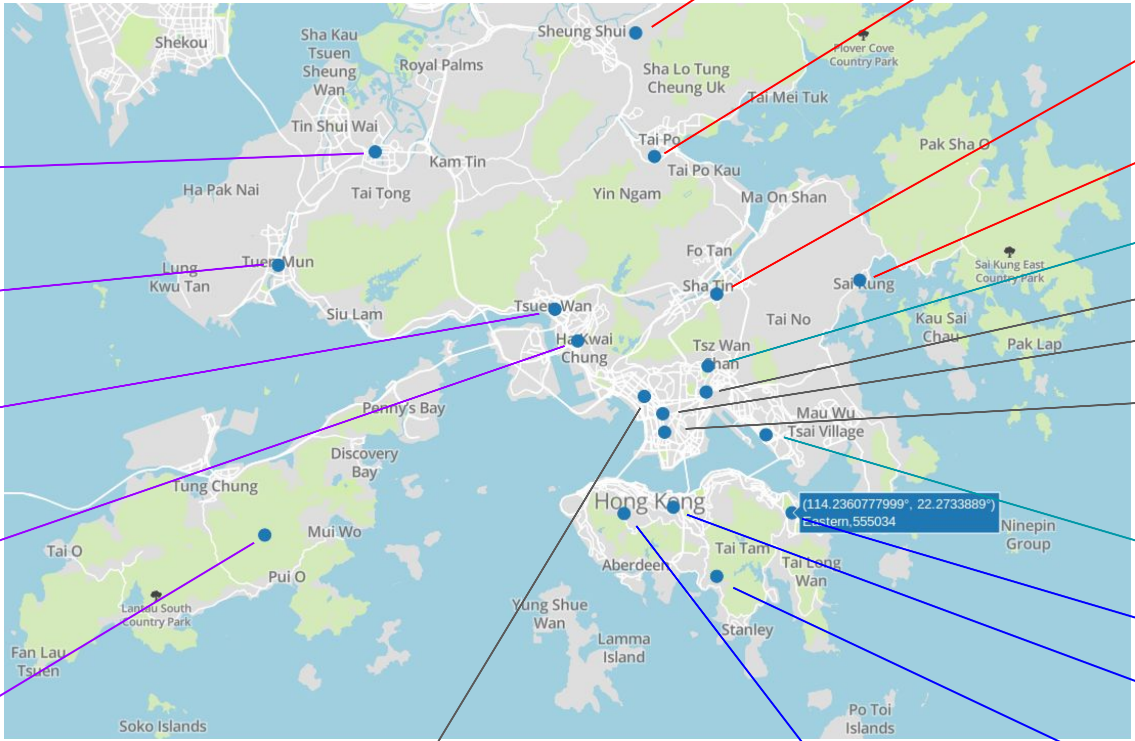
	Hong Kong Island	Kowloon West	Kowloon East	New Territories West	New Territories East	District Council (Second)
2008	Join July first demonstration	Occupation	Preference of candidates	Duration of being voter	Education level	/
2012	Channels of knowing candidates	Preference of candidates	Voting decision	Reasons of voting	Voting decision	Age


New Hampshire
Vermont
Massachusetts
Maine
Washington
Montana
North Dakota
Minnesota
Oregon
Wisconsin
Idaho
South Dakota
Michigan
New York
Wyoming
Rhode Island
Connecticut
Iowa
Pennsylvania
Nebraska
Nevada
Ohio
New Jersey
Utah
Illinois
Indiana
Colorado
Delaware
Kansas
Missouri
Maryland
Virginia
California
Kentucky
West Virginia
North Carolina
Tennessee
Oklahoma
Arizona
Arkansas
South Carolina
New Mexico
Alabama
Georgia
Mississippi
Texas
Louisiana
Alaska
Florida
Hawaii

U.S. Data Correlation

1. Which party does respondent vote, and under straight ticket or split ticket
2. Ticket splitting presidential vs congressional vote
3. Vote in National Elections
4. Vote on Election day or before
5. Intended Presidential Vote vs Actual Presidential Vote
6. Vote for a candidate for congress
7. Straight ticket vs split ticket
8. Respondent registers and votes
9. Congressional votes for House of Representatives
10. Vote for winner in House of Representatives

Visualization-Hong Kong Map



North, 315270

Tai Po, 303926

Sha Tin, 659194

Sai Kung, 461864

Wong Tai Sin, 425235

Kowloon City, 418732

Mong Kok, 130000

Yau Tsim Mong, 212970

Kwun Tong, 648541

Eastern, 555034

Wan Chai, 180123

Southern, 274994

Central and Western, 243266

Sham Shui Po, 405869

Yuen Long, 614178

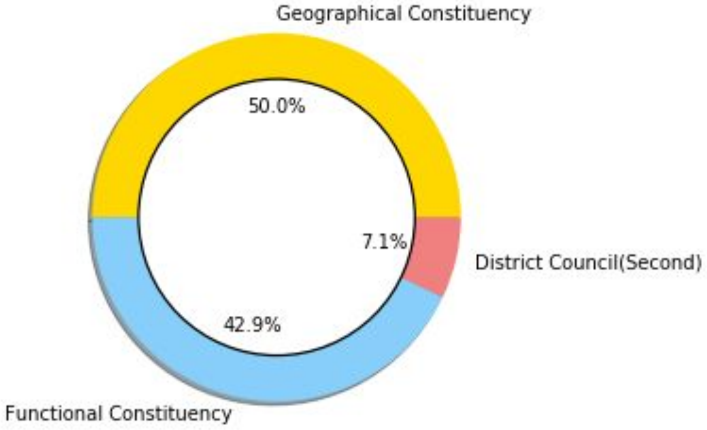
Tuen Mun, 489299

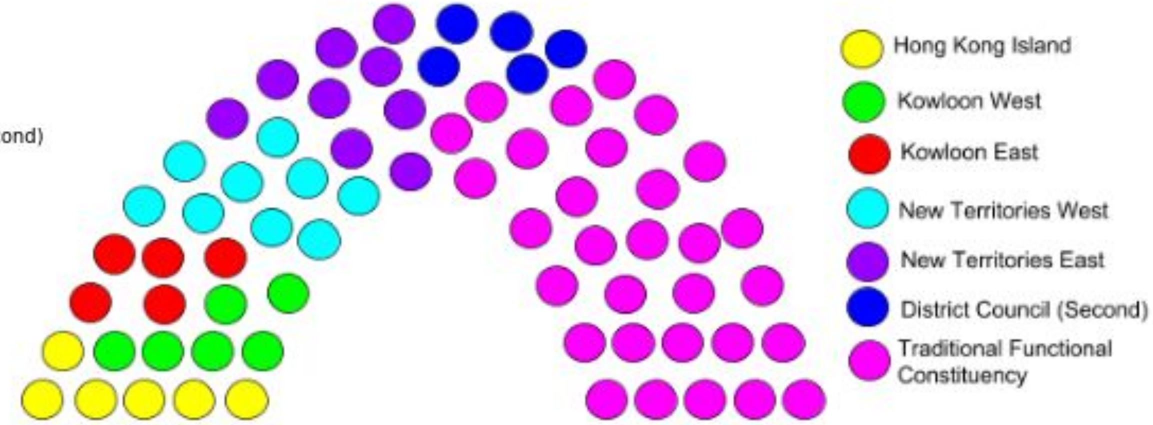
Tsuen Wan, 318916

Kwai Tsing, 520572

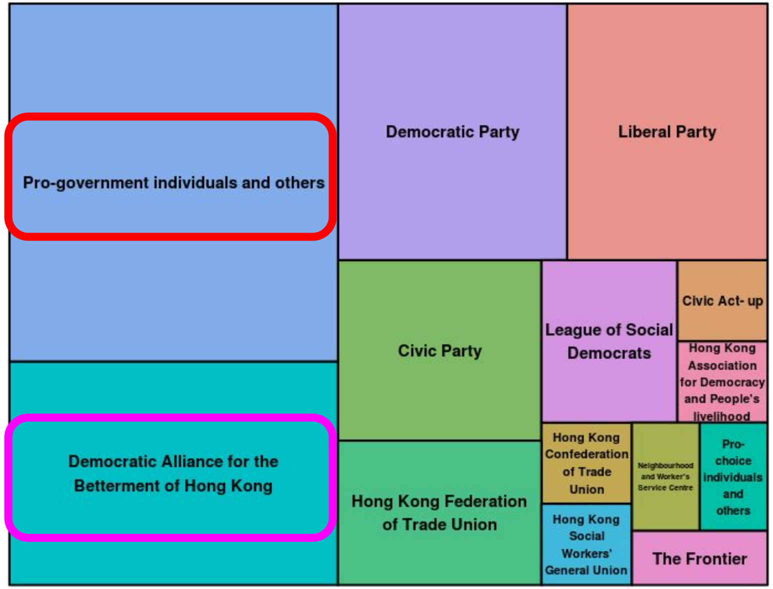
Islands, 156801

Visualization of Hong Kong election data

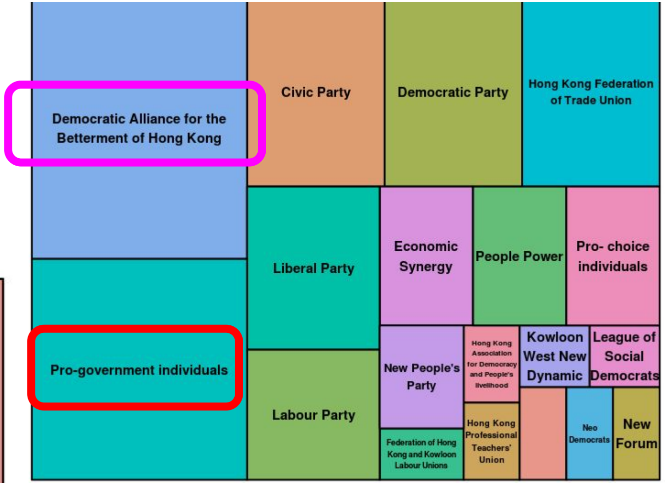




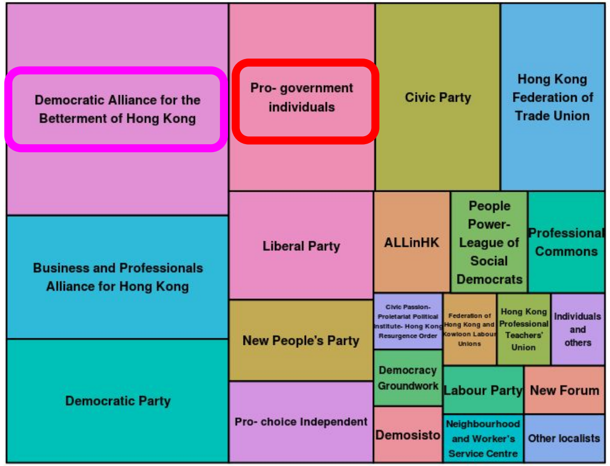
Election Results



2008

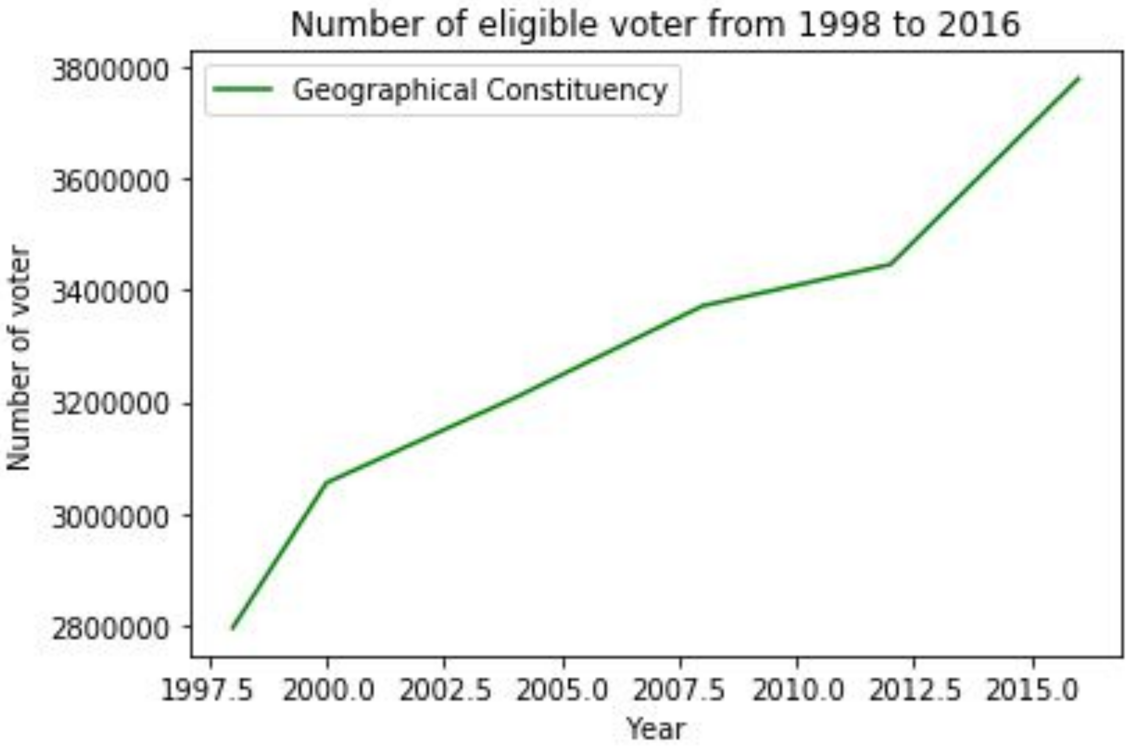


2012

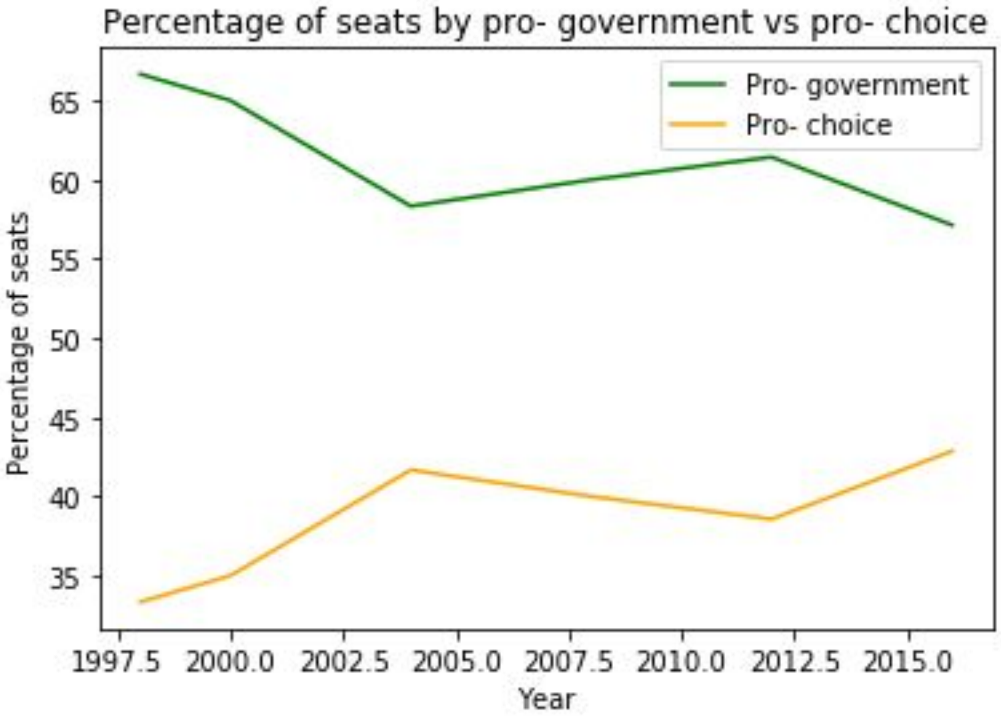


2016

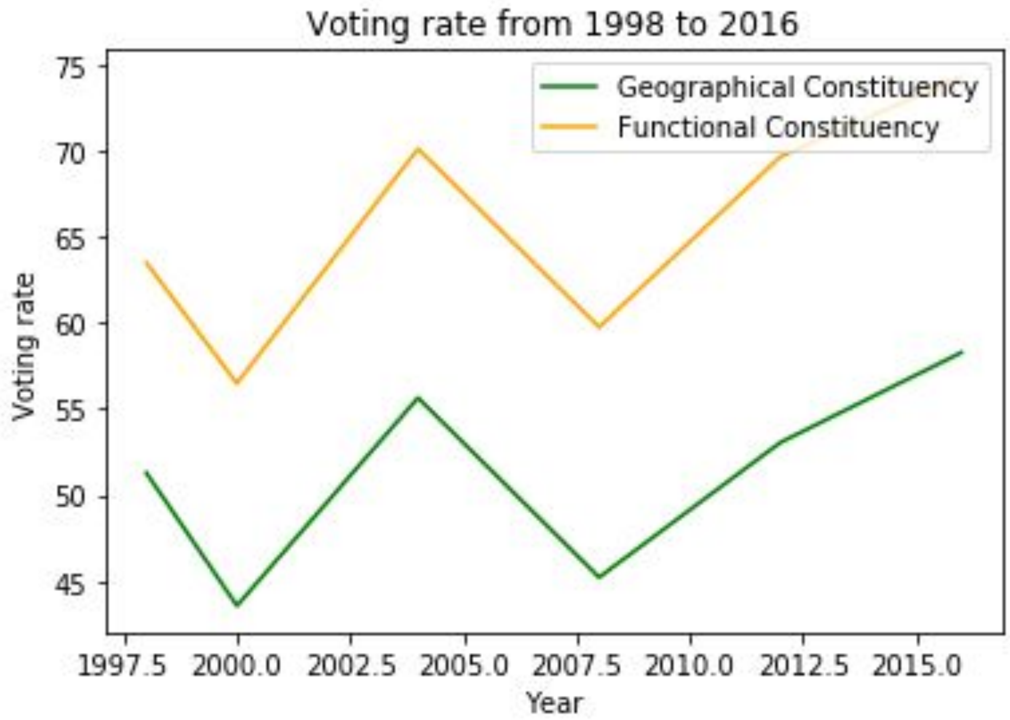
Number of eligible voters



Percentage of seats gained by pro- government and pro- choice



Voting rate

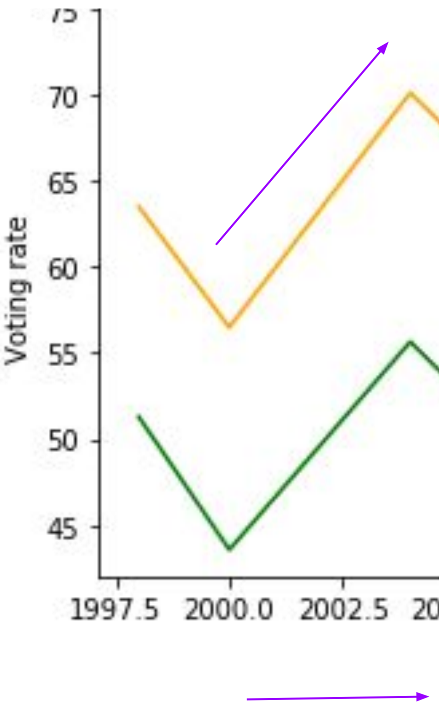

Voting rate from 1998 to 2016
Geographical Constituency
Functional Constituency
Voting rate
75
70
65
60
55
50
45
1997.5
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
Year

Year 2000-2004

Ask Tung to step down

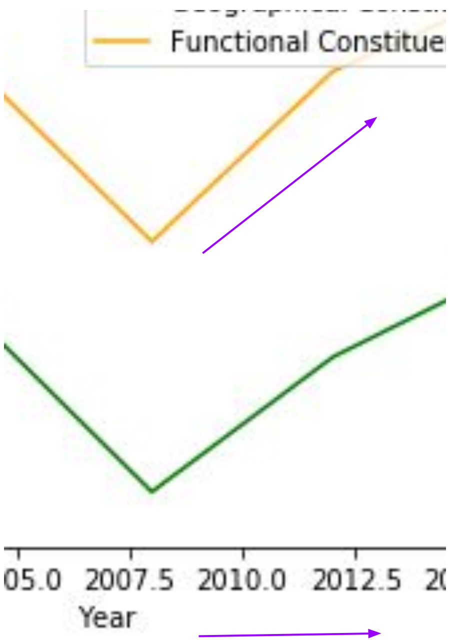
2003 Demonstration:
Against political incompetence &
maladministration of Tung Chee Hwa

2004 Demonstration:
Striving For Universal Suffrage in 2007
& 2008 for the chief executive and
Legislature respectively



Year 2008-2012

July 2012:
Moral and National Education Controversy



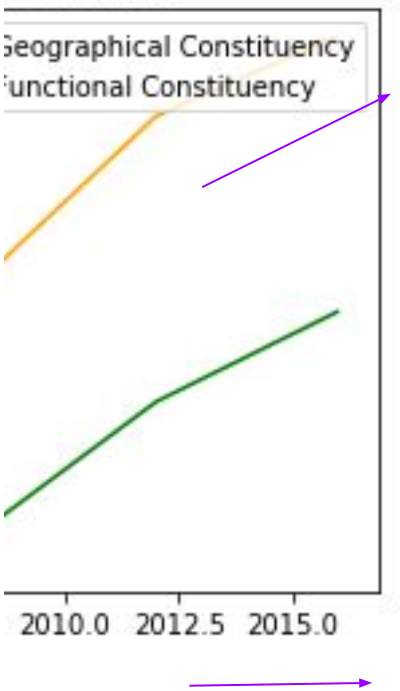
Year 2012-2016

October 2013:

Free-to-air TV license controversy

28 September 2014 -
15 December 2014:

Occupy Central

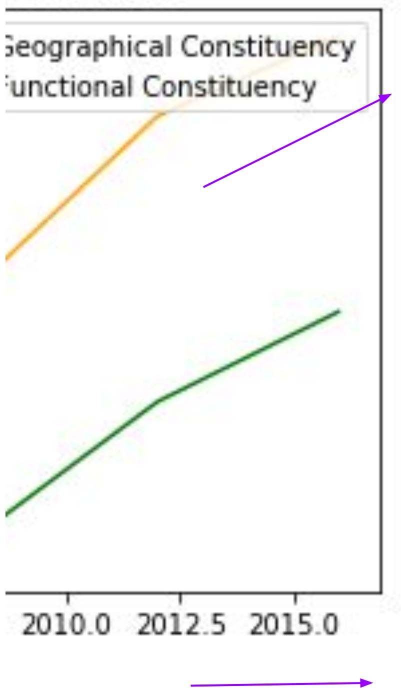


Year 2012-2016

Protect One country, Two system
Emancipate five booksellers immediately

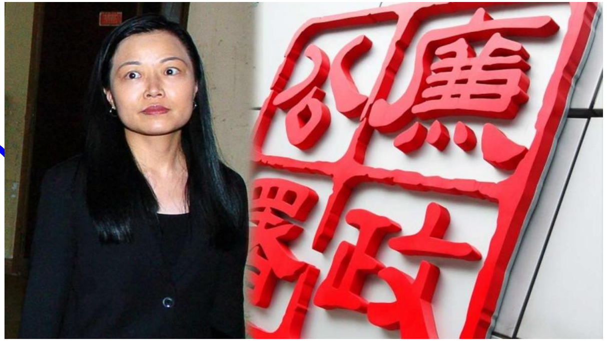
October 2015 - June 2016:

Causeway Bay bookseller disappearance



July 2016:

Resignations of ICAC heads controversy



Interactive Map Visualization

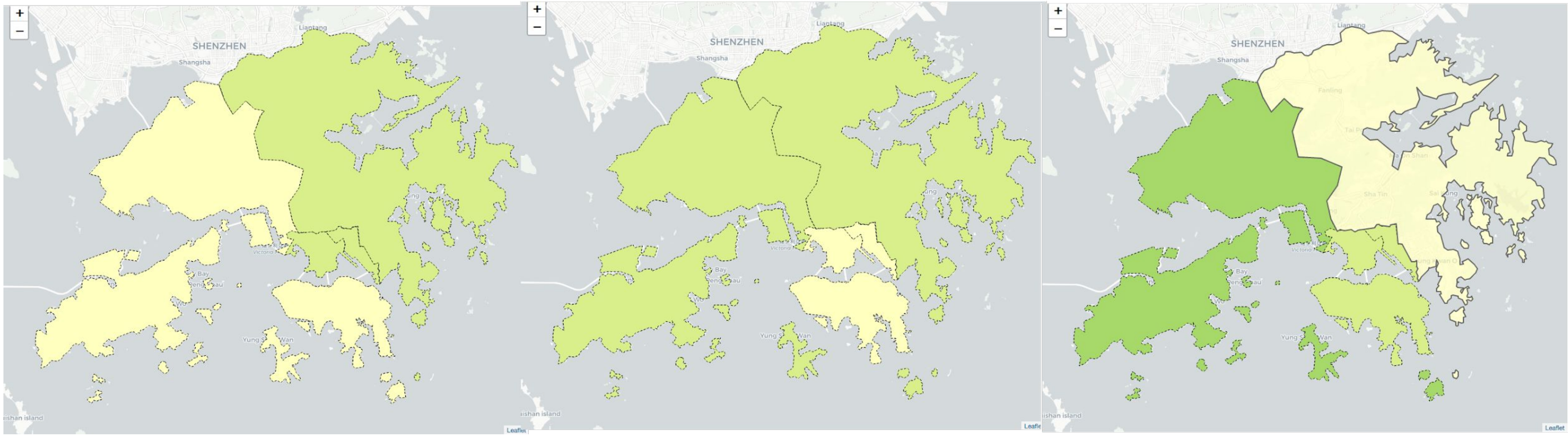


Figure: Chloropleth map showing correlations for 2008, 2012, and 2016

-A choropleth map was created using the Python library Folium, Leaflet.js, and geojson to specify the shape of regions

-Colors change depending on what factor, year pair are being displayed, greener being more positive correlation, and redder more negative

-Interactive visualization can be viewed at: http://web.eecs.utk.edu/~fbetanco/visualization/chloropleth.html

Election Data Deep Learning

Regions

Hong Kong:

- Hong Kong Island
- Kowloon West
- Kowloon East
- New Territories West
- New Territories East

Training: HKUPOP Survey (2008, 2012)

Prediction: HKUPOP Survey (2016)

Parameters: 8

United States:

- California
- Texas
- Alabama
- Minnesota
- Florida

Training: ANES Time Series (1992 - 2014)

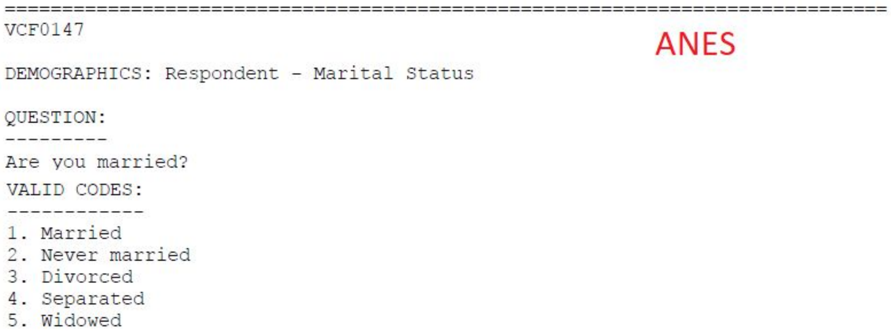
Prediction: US Census Bureau ACS (2015)

Parameters: 9

Preprocessing: Homogenize

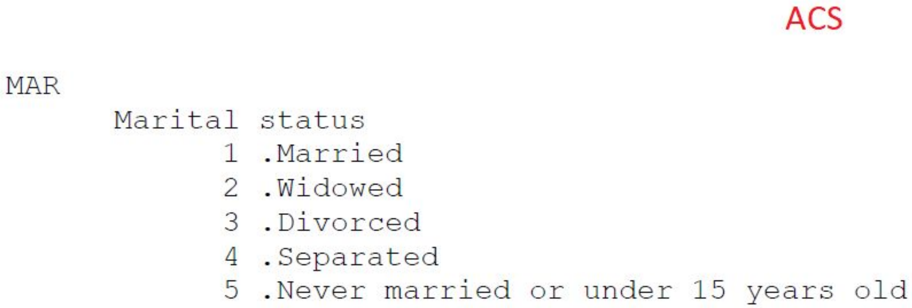
Surveys have different questions and answer keys:

1. Identify shared parameters
2. Modify values to have same answer key



ANES

```
==============================================================================
VCF0147

DEMOGRAPHICS: Respondent - Marital Status

QUESTION:
---------
Are you married?
VALID CODES:
------------
1. Married
2. Never married
3. Divorced
4. Separated
5. Widowed
```



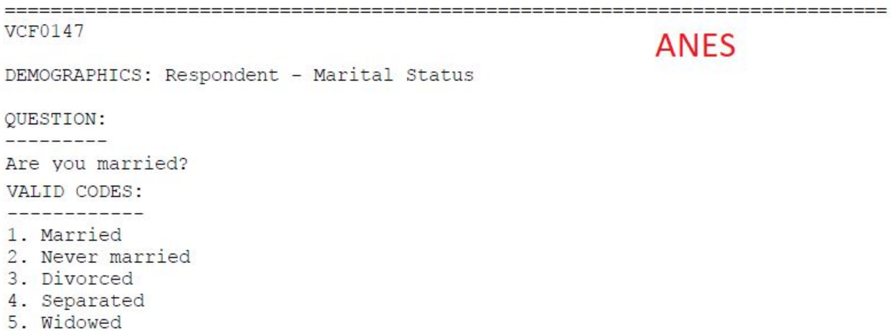
ACS

```
MAR
      Marital status
            1 .Married
            2 .Widowed
            3 .Divorced
            4 .Separated
            5 .Never married or under 15 years old
```

Preprocessing: Homogenize

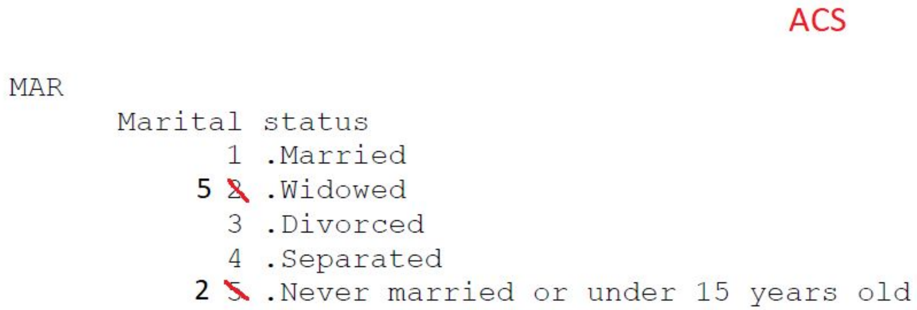
Surveys have different questions and answer keys:

1. Identify shared parameters
2. Modify values to have same answer key



ANES

```
==============================================================================
VCF0147

DEMOGRAPHICS: Respondent - Marital Status

QUESTION:
---------
Are you married?
VALID CODES:
------------
1. Married
2. Never married
3. Divorced
4. Separated
5. Widowed
```



ACS

MAR

Marital status

1 .Married

5 ~~2~~ .Widowed

3 .Divorced

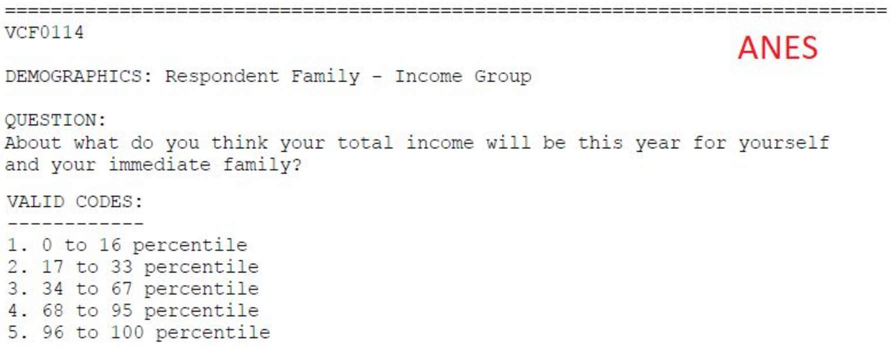
4 .Separated

2 ~~5~~ .Never married or under 15 years old

Preprocessing: Homogenize

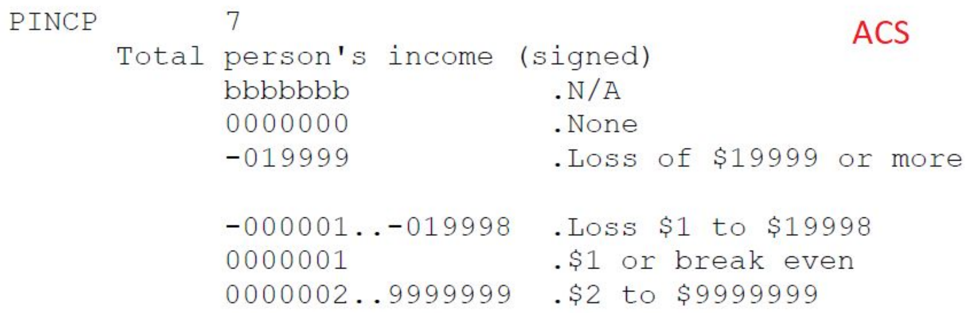
Tricky parameters:

- Same information, but hidden
- Want to keep as many parameters as possible



ANES

```
==============================================================================
VCF0114

DEMOGRAPHICS: Respondent Family - Income Group

QUESTION:
About what do you think your total income will be this year for yourself
and your immediate family?

VALID CODES:
------------
1. 0 to 16 percentile
2. 17 to 33 percentile
3. 34 to 67 percentile
4. 68 to 95 percentile
5. 96 to 100 percentile
```



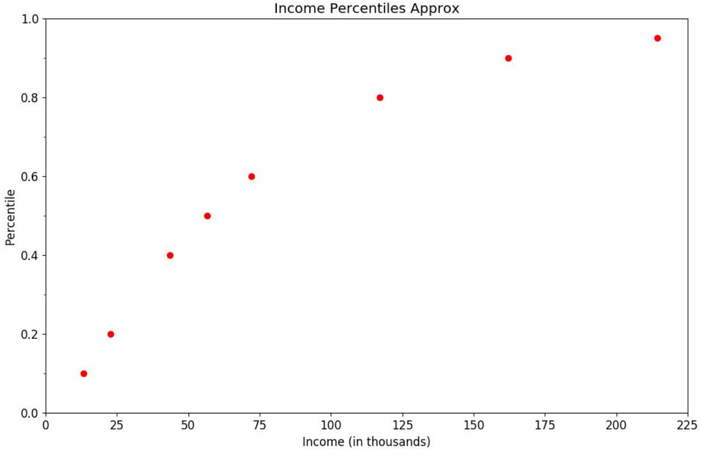
ACS

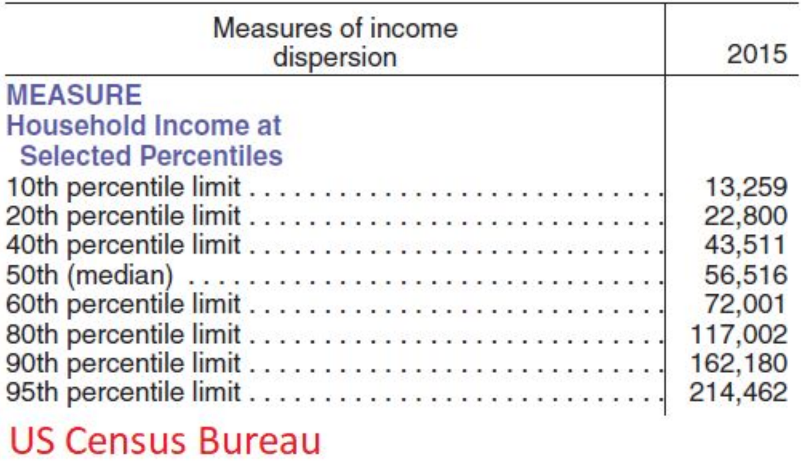
```
PINCP        7
     Total person's income (signed)
           bbbbbbb             .N/A
           0000000             .None
           -019999             .Loss of $19999 or more

           -000001..-019998    .Loss $1 to $19998
           0000001             .$1 or break even
           0000002..9999999    .$2 to $9999999
```

Preprocessing: Homogenize

Solution:

1. Identify percentile brackets
2. Interpolate to find needed percentiles
   - ○ Lagrange Interp: Approx is good





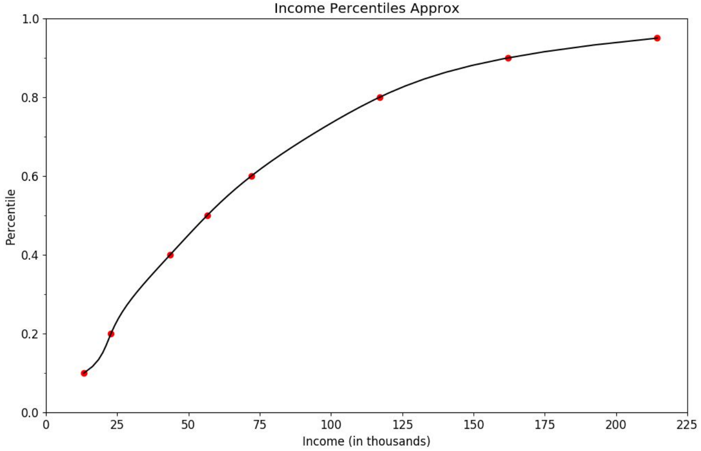
Measures of income dispersion	2015
MEASURE	
Household Income at Selected Percentiles	
10th percentile limit	13,259
20th percentile limit	22,800
40th percentile limit	43,511
50th (median)	56,516
60th percentile limit	72,001
80th percentile limit	117,002
90th percentile limit	162,180
95th percentile limit	214,462

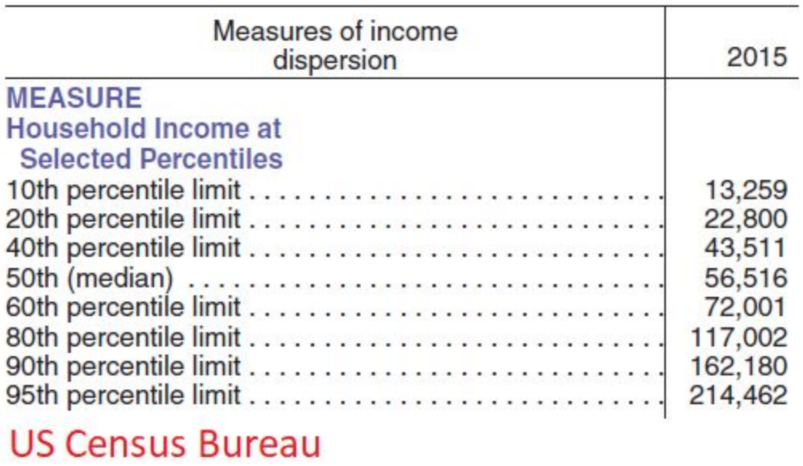
US Census Bureau

Preprocessing: Homogenize

Solution:

1. Identify percentile brackets
2. Interpolate to find needed percentiles
   - ○ Lagrange Interp: Approx is good





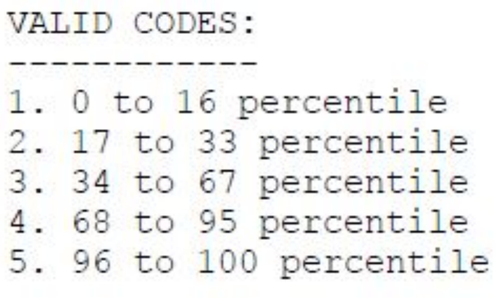
Measures of income dispersion	2015
MEASURE	
Household Income at Selected Percentiles	
10th percentile limit	13,259
20th percentile limit	22,800
40th percentile limit	43,511
50th (median)	56,516
60th percentile limit	72,001
80th percentile limit	117,002
90th percentile limit	162,180
95th percentile limit	214,462

US Census Bureau

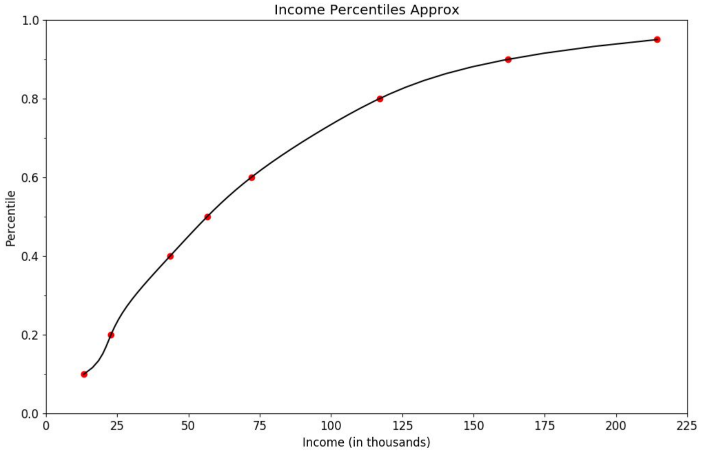
Preprocessing: Homogenize

Solution:

1. Identify percentile brackets
2. Interpolate to find needed percentiles
   - ○ Lagrange Interp: Approx is good



```
VALID CODES:
------------
1. 0 to 16 percentile
2. 17 to 33 percentile
3. 34 to 67 percentile
4. 68 to 95 percentile
5. 96 to 100 percentile
```

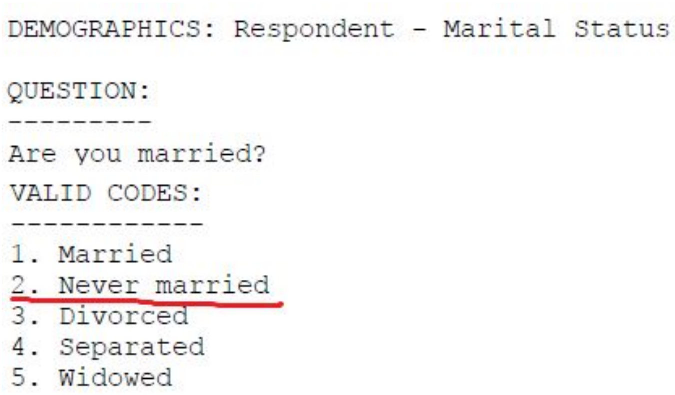


1. 0 to $20,787
2. $20,787 to $35,345
3. $35,345 to $86,701
4. $86,701 to $214,462
5. + $214,462

Preprocessing: One Hot Encoding

Survey answers is nominal data:

- Numbers have no real value - Represent an idea
- Model only needs to know True/False (Mutually exclusive)

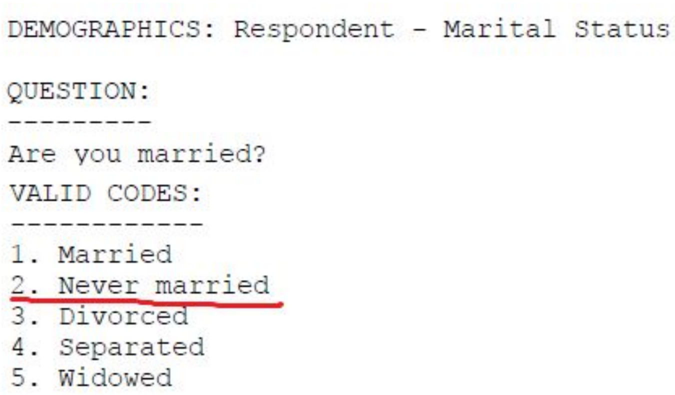


```
DEMOGRAPHICS: Respondent - Marital Status

QUESTION:
----------
Are you married?
VALID CODES:
------------
1. Married
2. Never married
3. Divorced
4. Separated
5. Widowed
```

Preprocessing: One Hot Encoding

Survey answers is nominal data:

- Numbers have no real value - Represent an idea
- Model only needs to know True/False (Mutually exclusive)



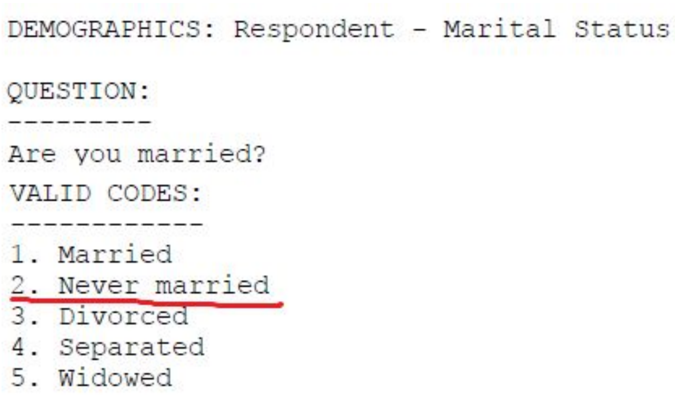
```
DEMOGRAPHICS: Respondent - Marital Status

QUESTION:
----------
Are you married?
VALID CODES:
------------
1. Married
2. Never married
3. Divorced
4. Separated
5. Widowed
```

$$2 \rightarrow \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \\ [0 & 0 & 1 & 0 & 0 & 0] \end{array}$$

Preprocessing: One Hot Encoding

Survey answers is nominal data:

- Numbers have no real value - Represent an idea
- Model only needs to know True/False (Mutually exclusive)



```
DEMOGRAPHICS: Respondent - Marital Status

QUESTION:
----------
Are you married?
VALID CODES:
------------
1. Married
2. Never married
3. Divorced
4. Separated
5. Widowed
```

$$2 \rightarrow \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \\ \end{array}$$

$$\left[\begin{array}{cccccc} 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right]$$

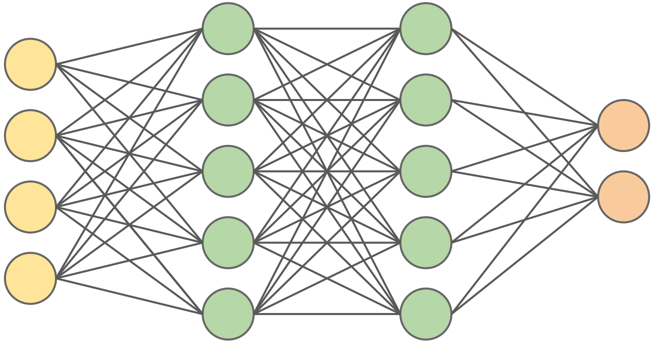
HK Parameters: 8 ---> 40

US Parameters: 9 ---> 58

Prediction: Tools

- Keras for Python 3.x
  - Easy to implement (Math is taken care of)
  - Have to determine optimal architecture and hyperparameters
- Quickly develop a prototype to experiment with the idea

Deep Neural Network:



Prediction: Classification

Two possible approaches (Bipartisan elections):

- Only consider voters
  - Binary classification

$$f(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid

- Consider both voters & nonvoters
  - Multiclass classification

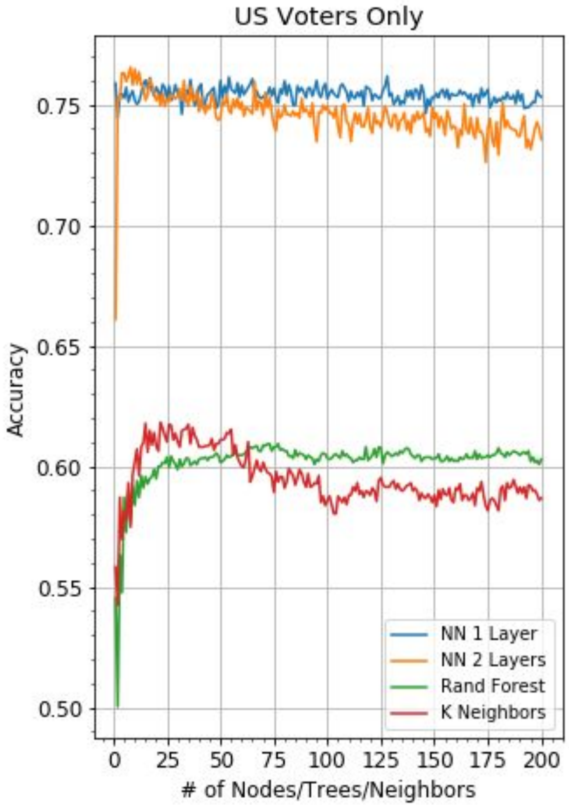
$$f(z_j)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

Softmax

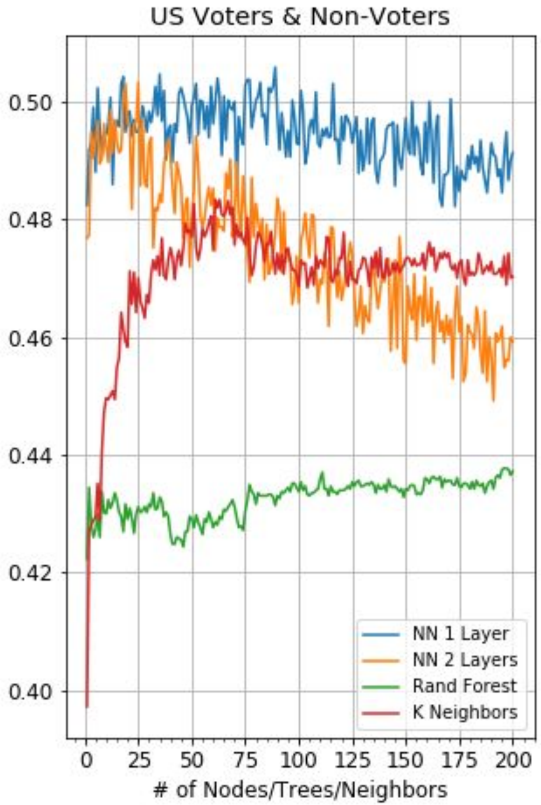
Prediction: Hyperparameters

Comparison between 4 possible ML algorithms:

- Predictions most stable with binary classification
- Best accuracy at ~10 nodes with a DNN



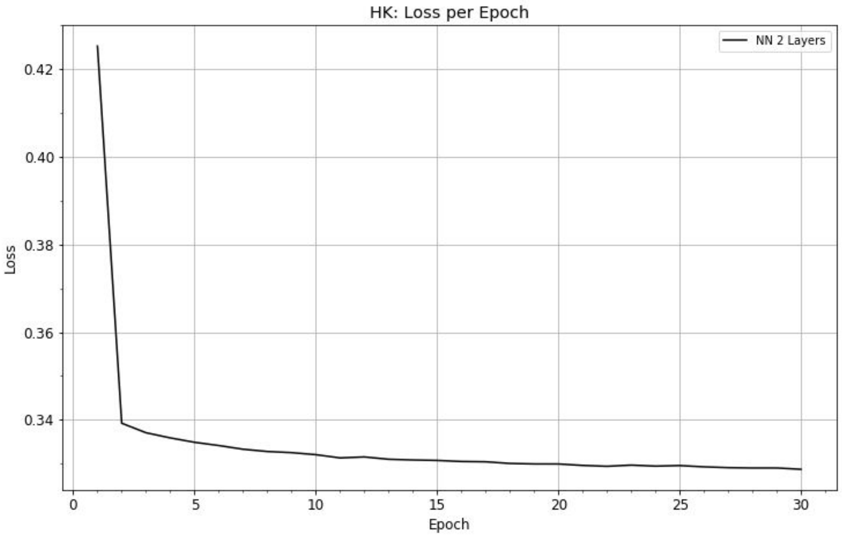
US Voters Only

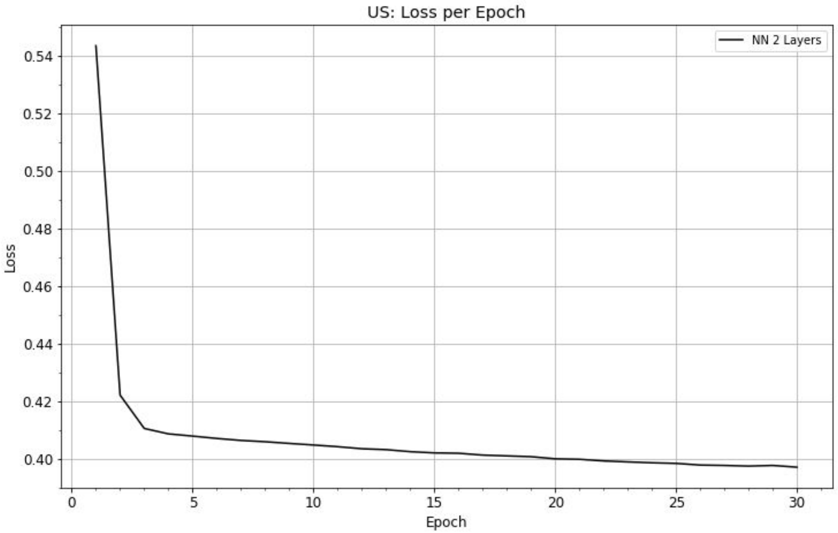


US Voters & Non-Voters

Set fixed seed for reproducibility

Prediction: Hyperparameters


HK: Loss per Epoch


US: Loss per Epoch

Prediction loss per epoch:

- Diminishing returns at ~15 epochs for both predictions

Prediction: Parameters

Hong Kong:

- Political Inclination
- Gender
- Education
- Previous Voter
- Age Group
- Occupation
- Planning to vote
- Area of Residence

United States:

- Age Group
- Gender
- Race
- Census Region
- Income Group
- Occupation
- Employment Status
- Education
- Marital Status

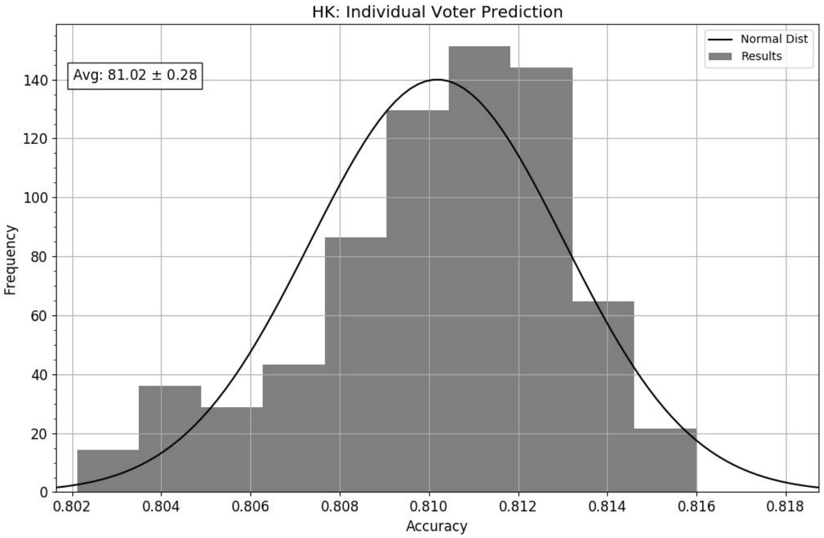
Prediction: Uncertainty

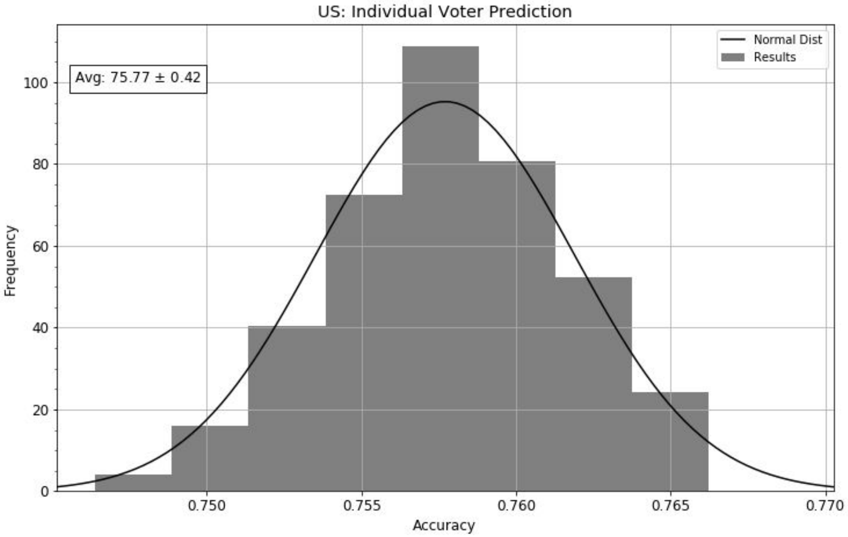
- Model’s weights are randomly initialized
    - Leads to different results every time
    - How good is the model then?

Use a prediction as a “measurement”:

- Sample of N = 100 predictions
- Find distribution that best describes the sample
- Calculate appropriate moments

Prediction: Uncertainty





Best fit is with a normal distribution ---> Can quantify the performance of the DNN

Prediction: Counting Votes

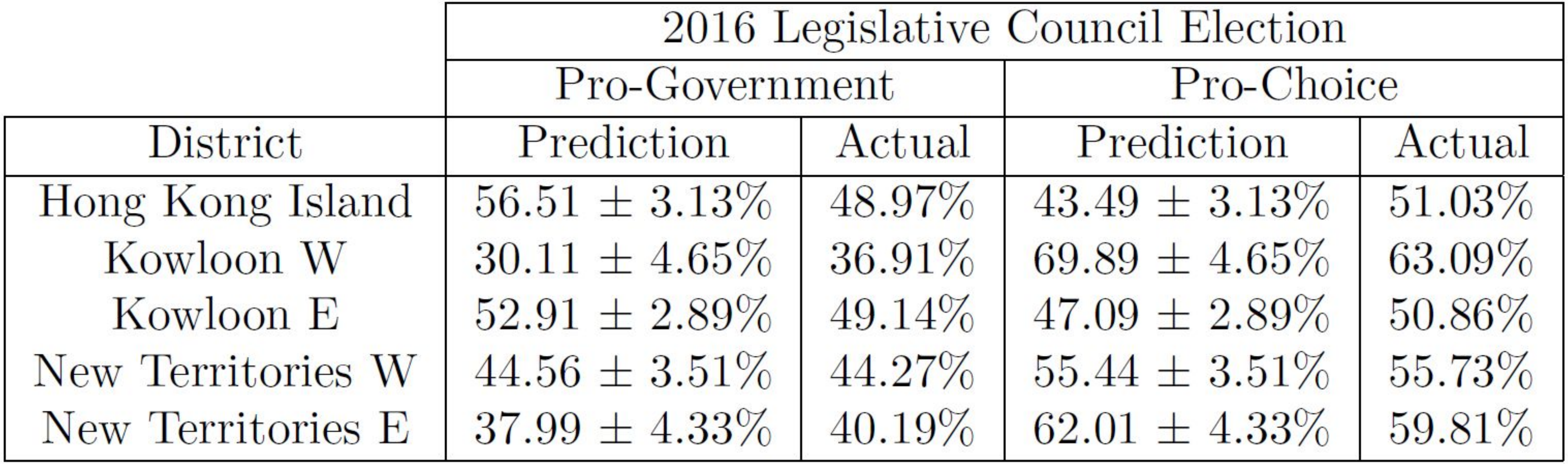
- Each individual has a corresponding statistical weight == Number of votes
  - How many people does this individual represent
- Sum of weights ---> Total number of votes

Voter Turnout:

- US Census Bureau provides voter turnout dependent on various factors
  - Most impactful is racial turnout. Use this as a modifier of the statistical weights.

$$w_i := w_i \cdot turnout_{race}$$

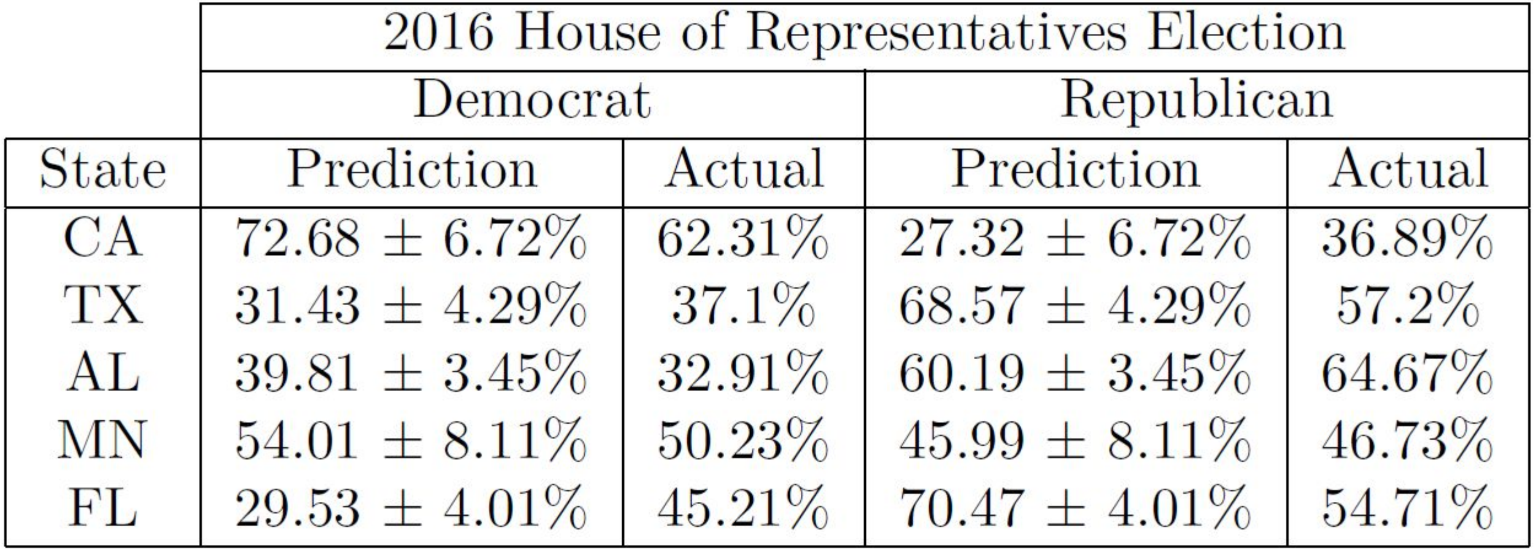
Prediction: Hong Kong Results



	2016 Legislative Council Election			
	Pro-Government		Pro-Choice	
District	Prediction	Actual	Prediction	Actual
Hong Kong Island	56.51 ± 3.13%	48.97%	43.49 ± 3.13%	51.03%
Kowloon W	30.11 ± 4.65%	36.91%	69.89 ± 4.65%	63.09%
Kowloon E	52.91 ± 2.89%	49.14%	47.09 ± 2.89%	50.86%
New Territories W	44.56 ± 3.51%	44.27%	55.44 ± 3.51%	55.73%
New Territories E	37.99 ± 4.33%	40.19%	62.01 ± 4.33%	59.81%

- More parameters could increase precision and accuracy
- Model cannot take into account sudden political shifts and anomalies

Prediction: United States Results



	2016 House of Representatives Election			
	Democrat		Republican	
State	Prediction	Actual	Prediction	Actual
CA	72.68 ± 6.72%	62.31%	27.32 ± 6.72%	36.89%
TX	31.43 ± 4.29%	37.1%	68.57 ± 4.29%	57.2%
AL	39.81 ± 3.45%	32.91%	60.19 ± 3.45%	64.67%
MN	54.01 ± 8.11%	50.23%	45.99 ± 8.11%	46.73%
FL	29.53 ± 4.01%	45.21%	70.47 ± 4.01%	54.71%

- Model can determine dominant party in each state
    - Greatly exaggerates the vote sway at times
- Improve with: Parameters, Turnout percentage