‘It’s the Economy, Stupid!’

Joshua Rogin
github:
github.com/WhereisWald0

QUESTION:

Can we look at economic data and election outcomes and apply Machine Learning models to account for a relationship between them?

Pre-methodolgicial principles

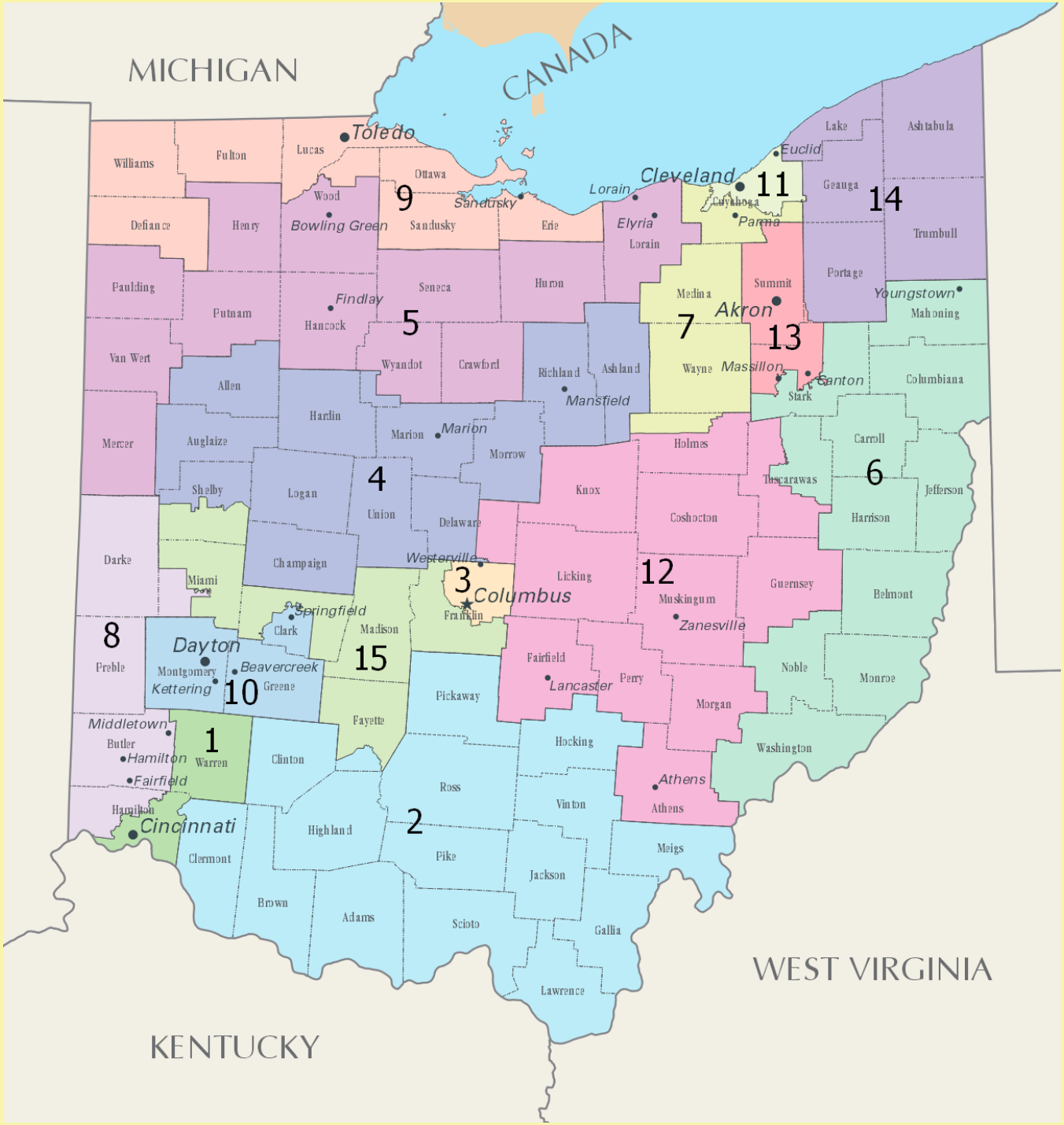
Independence - use primary sources as much as possible

Modularity - formatted data should be easily swapped or connected to new sources

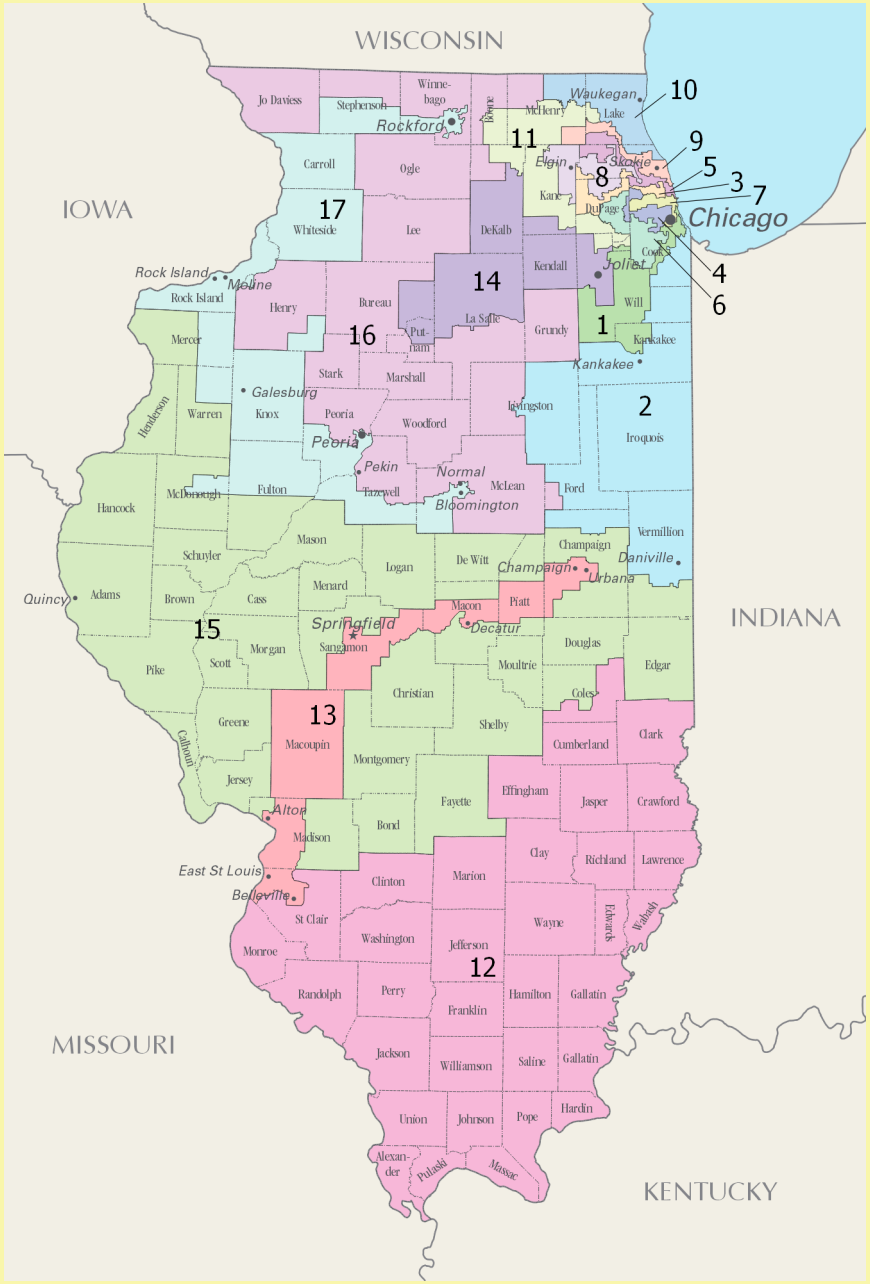
Preservation - avoidance of destructive processes

Granularity - use boundaries of appropriate size.

Future-Proofing -all processes should allow for future data


MICHIGAN
CANADA
WEST VIRGINIA
KENTUCKY
Toledo
Cleveland
Akron
Canton
Youngstown
Euclid
Lorain
Elyria
Parma
Massillon
Mansfield
Marion
Findlay
Bowling Green
Sandusky
Westerville
Columbus
Springfield
Dayton
Beavercreek
Kettering
Middletown
Hamilton
Fairfield
Cincinnati
Zanesville
Lancaster
Athens
Williams
Fulton
Lucas
Ottawa
Wood
Defiance
Henry
Paulding
Putnam
Hancock
Seneca
Huron
Erie
Lorain
Cuyahoga
Lake
Ashtabula
Geauga
Trumbull
Portage
Summit
Medina
Wayne
Stark
Mahoning
Columbiana
Carroll
Jefferson
Harrison
Belmont
Monroe
Noble
Washington
Morgan
Perry
Fairfield
Pickaway
Fayette
Clinton
Ross
Hocking
Vinton
Meigs
Pike
Jackson
Gallia
Lawrence
Scioto
Adams
Brown
Highland
Clermont
Warren
Butler
Preble
Darke
Miami
Champaign
Clark
Madison
Franklin
Delaware
Union
Logan
Shelby
Auglaize
Mercer
Van Wert
Allen
Hardin
Marion
Morrow
Wyandot
Crawford
Richland
Ashland
Knox
Holmes
Tuscarawas
Coshocton
Licking
Muskingum
Guernsey
9
5
4
3
15
10
1
2
7
11
13
14
6
12
8

United States House Districts



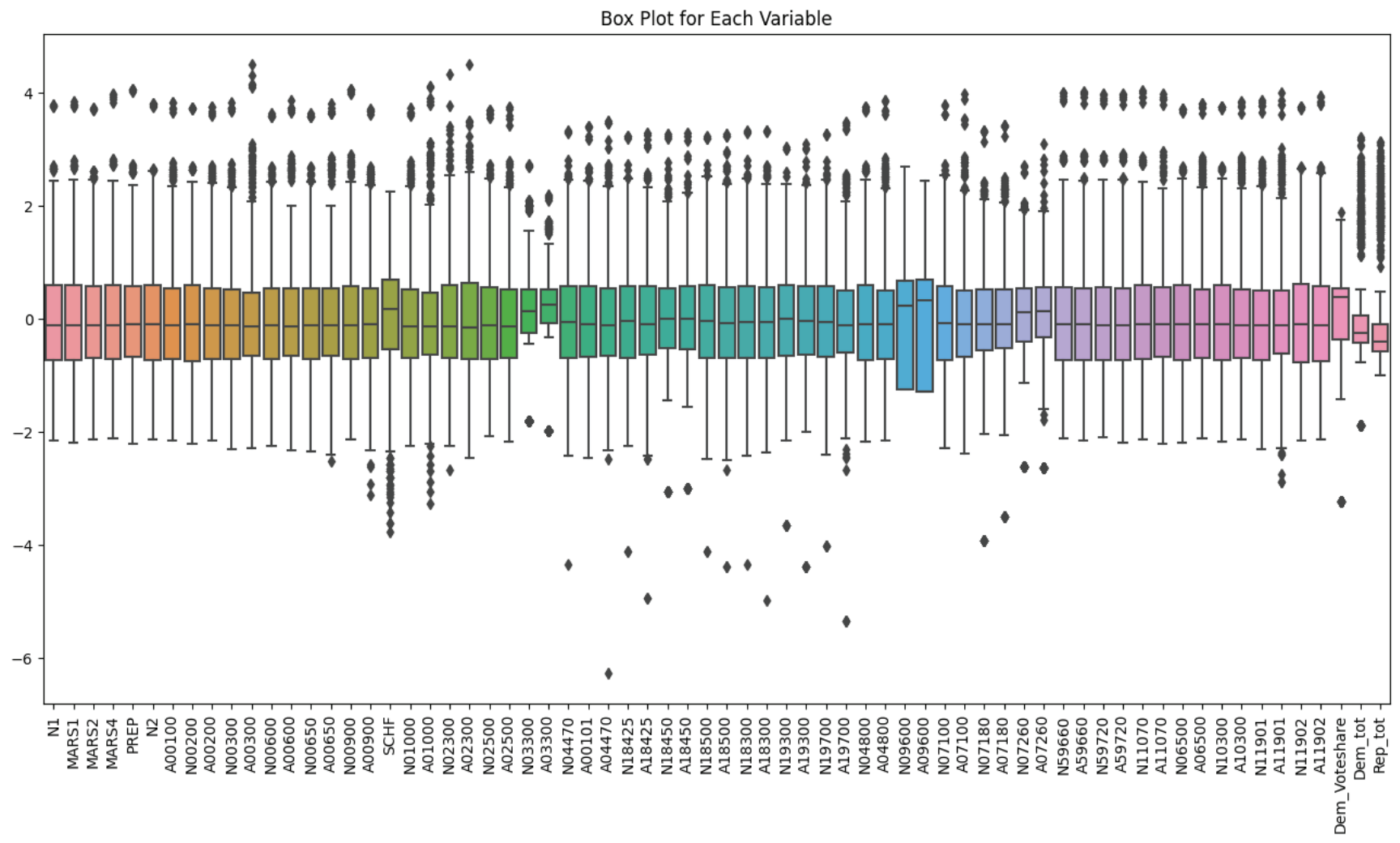
-435 seats

-Elections every two years

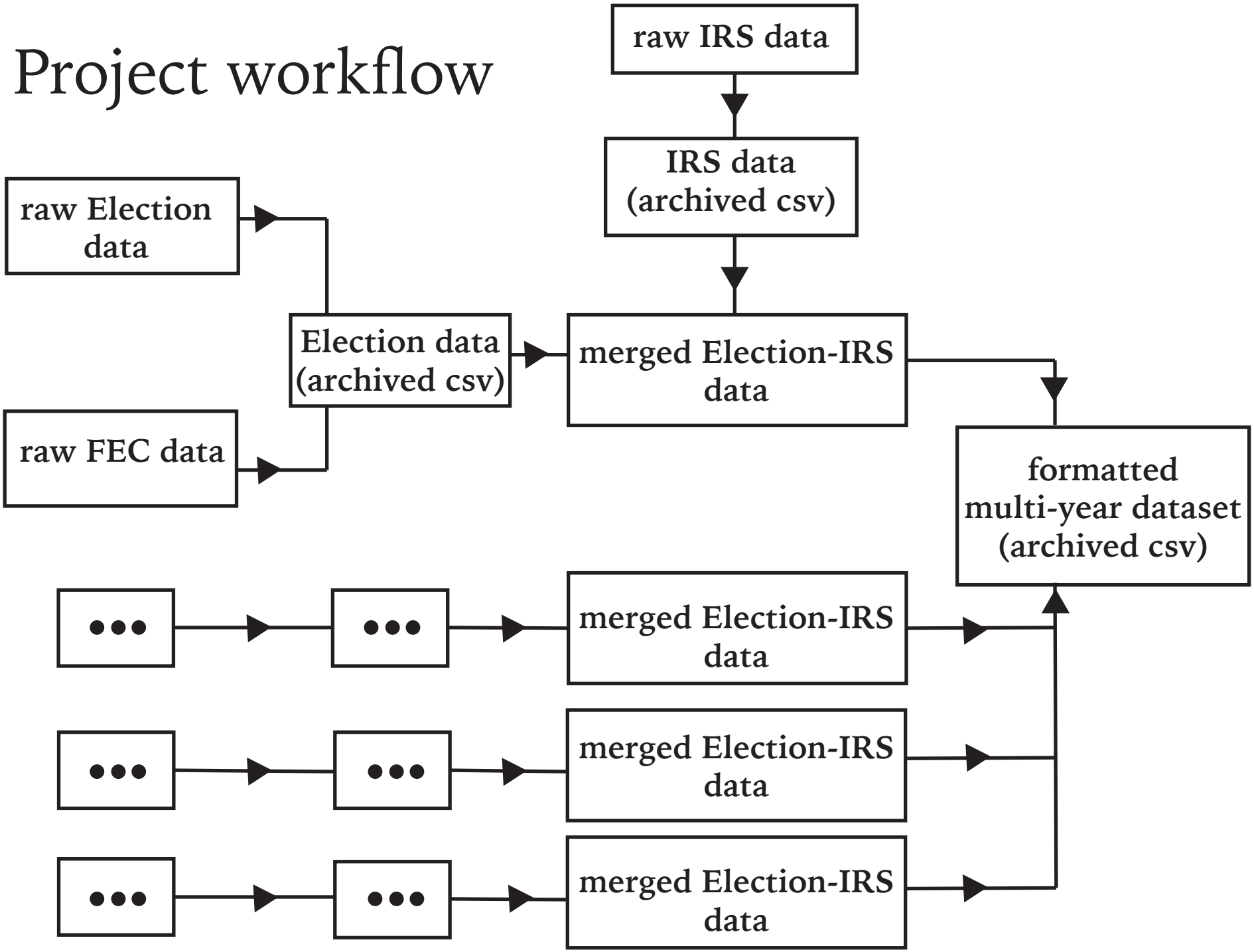
-Numerous candidates

IRS data

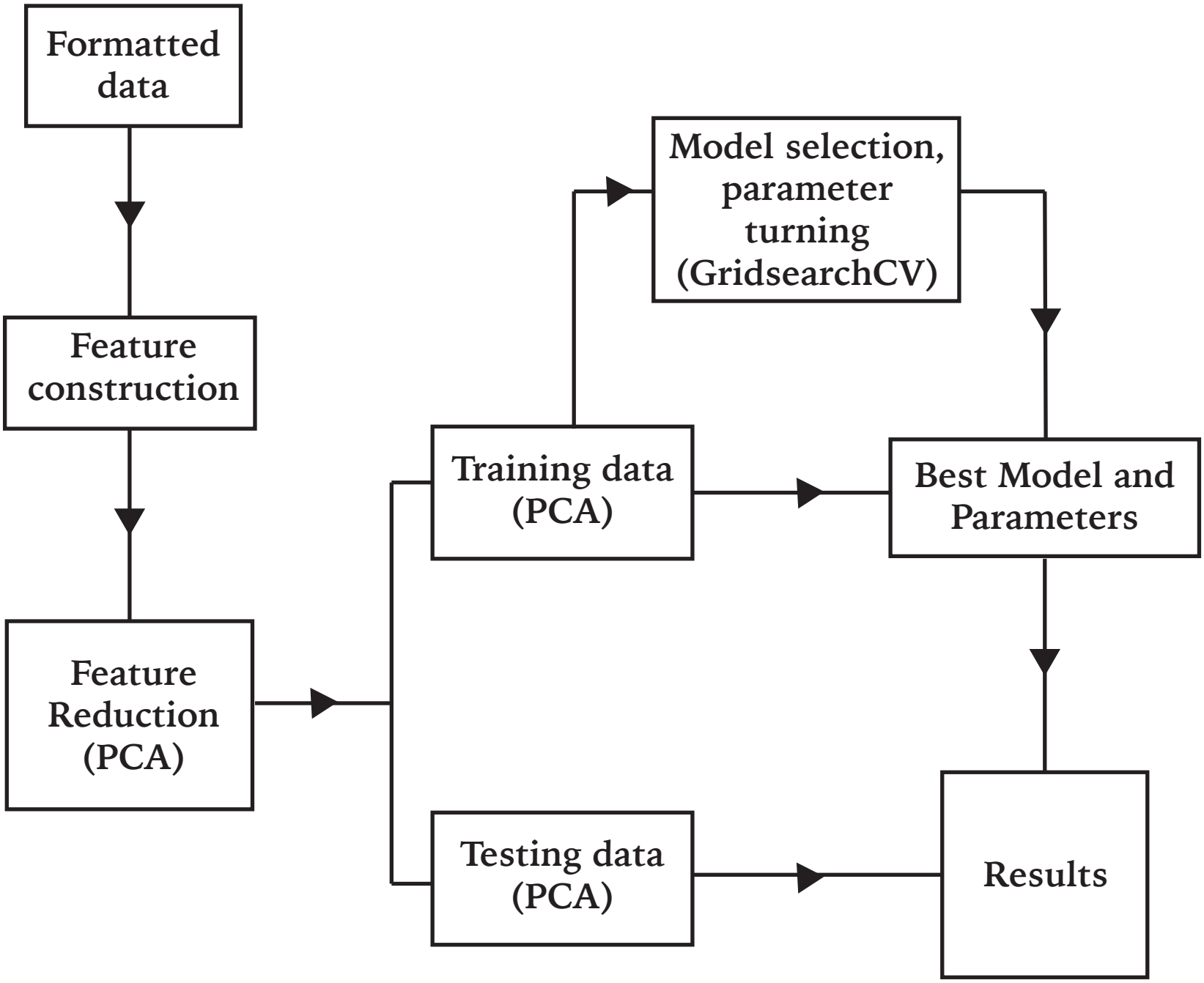
-data shape - 3155 rows, 150 columns (dense)
-lots of outliers (reflects economic reality w/in US)



Project workflow



Machine Learning Pipeline

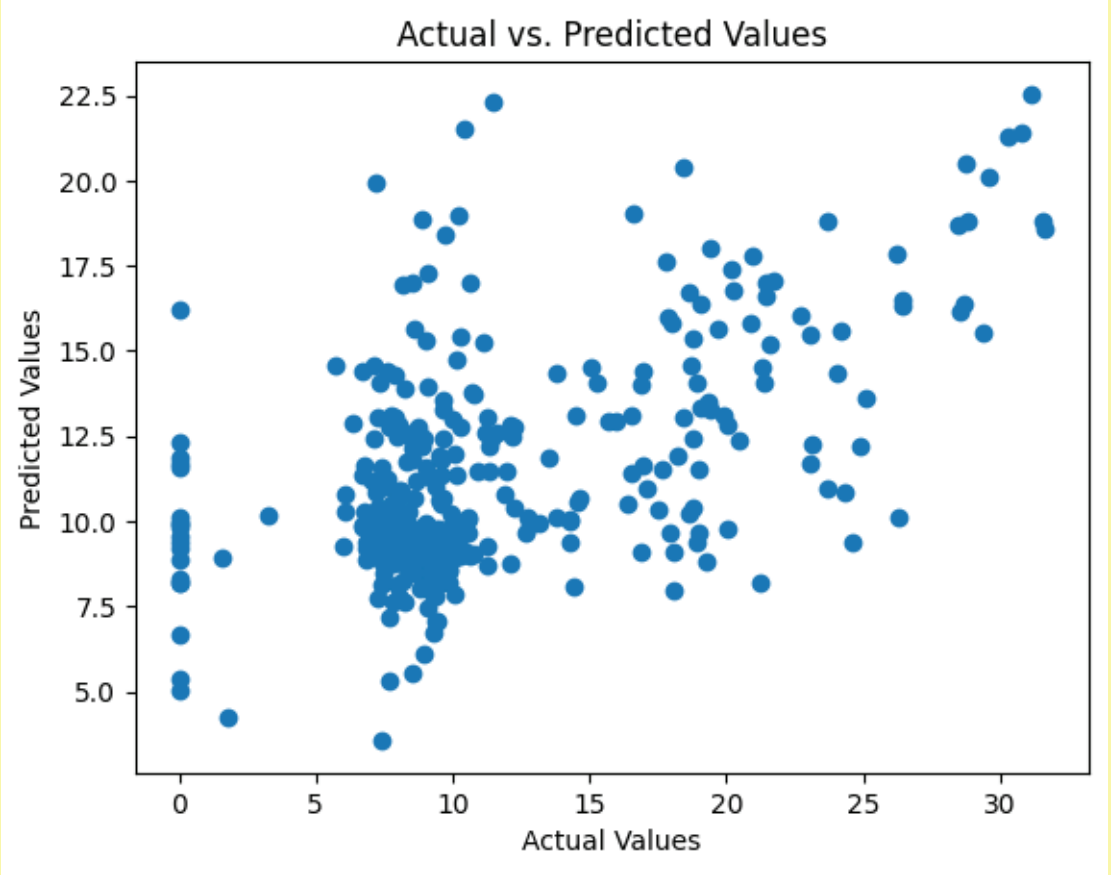


Final Results

After Grid-searchCV:

best model - RandomForest

Gives Us:
Mean Squared Error: 26.77
R-squared: 0.319



concl./ thank you

Next Steps:

- Extract and process additional States data

- Divide IRS data by income tax brackets (could potentially add 7x as many features)