Paranormal Distribution

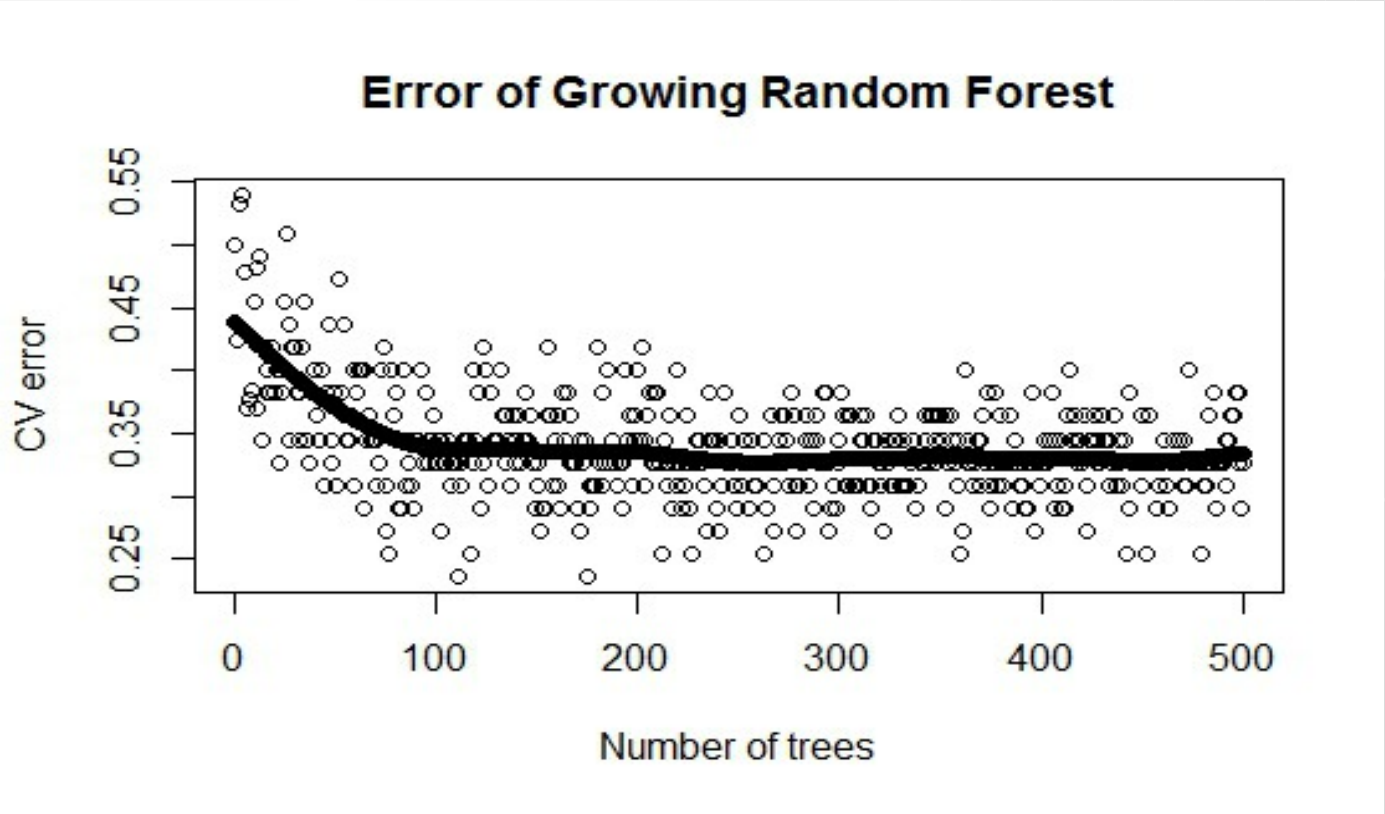
Andy Chung
Kevin Gilbert
Jason Sun

First Step: Random Forest

One decision tree: One classification

500 decision trees: One classification

We bag these trees into a “random forest”

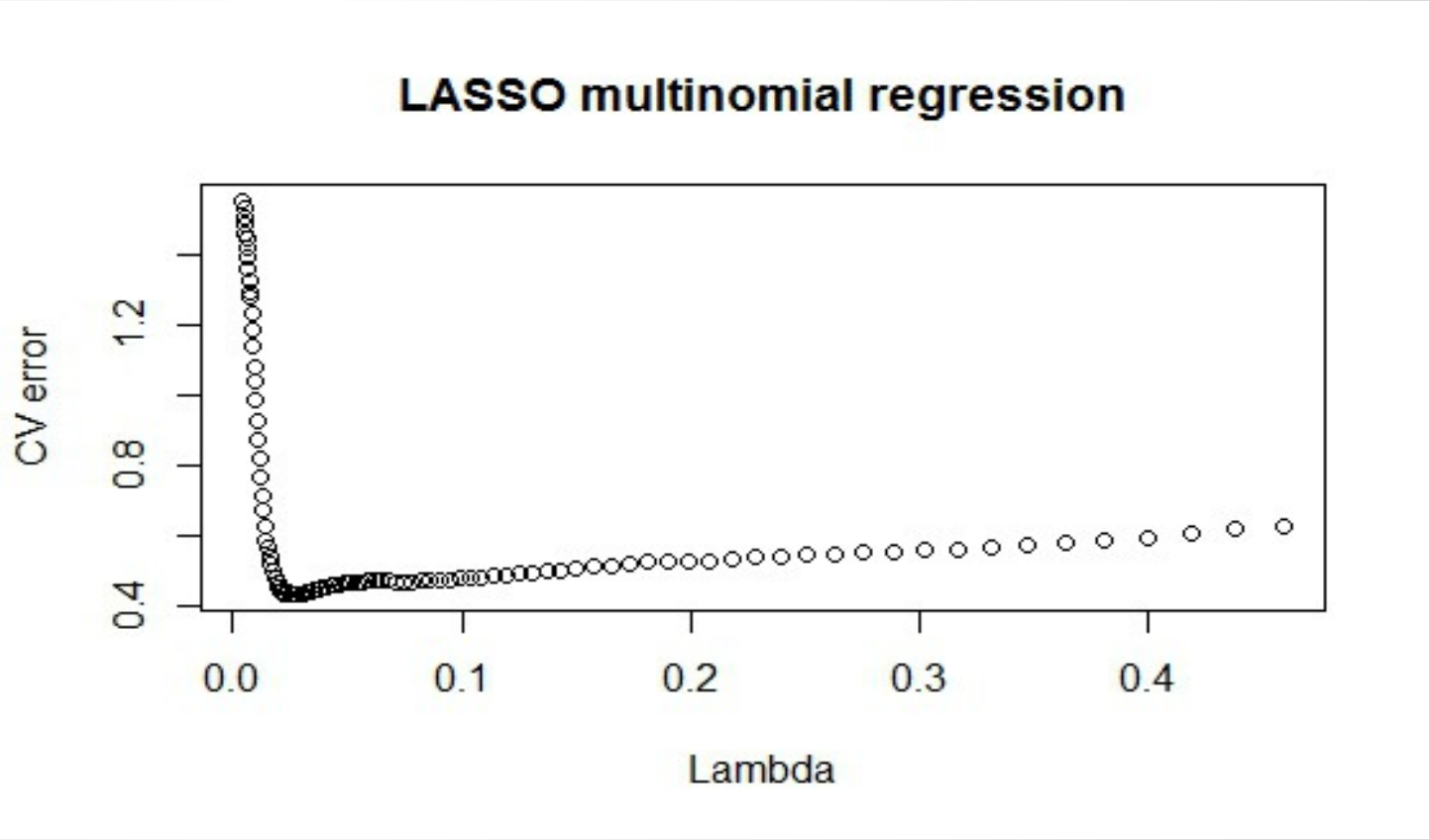


Second Step: LASSO (glmnet)

What if Random Forest isn't sure?

Recalculate unsure nodes with LASSO, ridge

Replacing decisions reduced our CV error



Results on training data

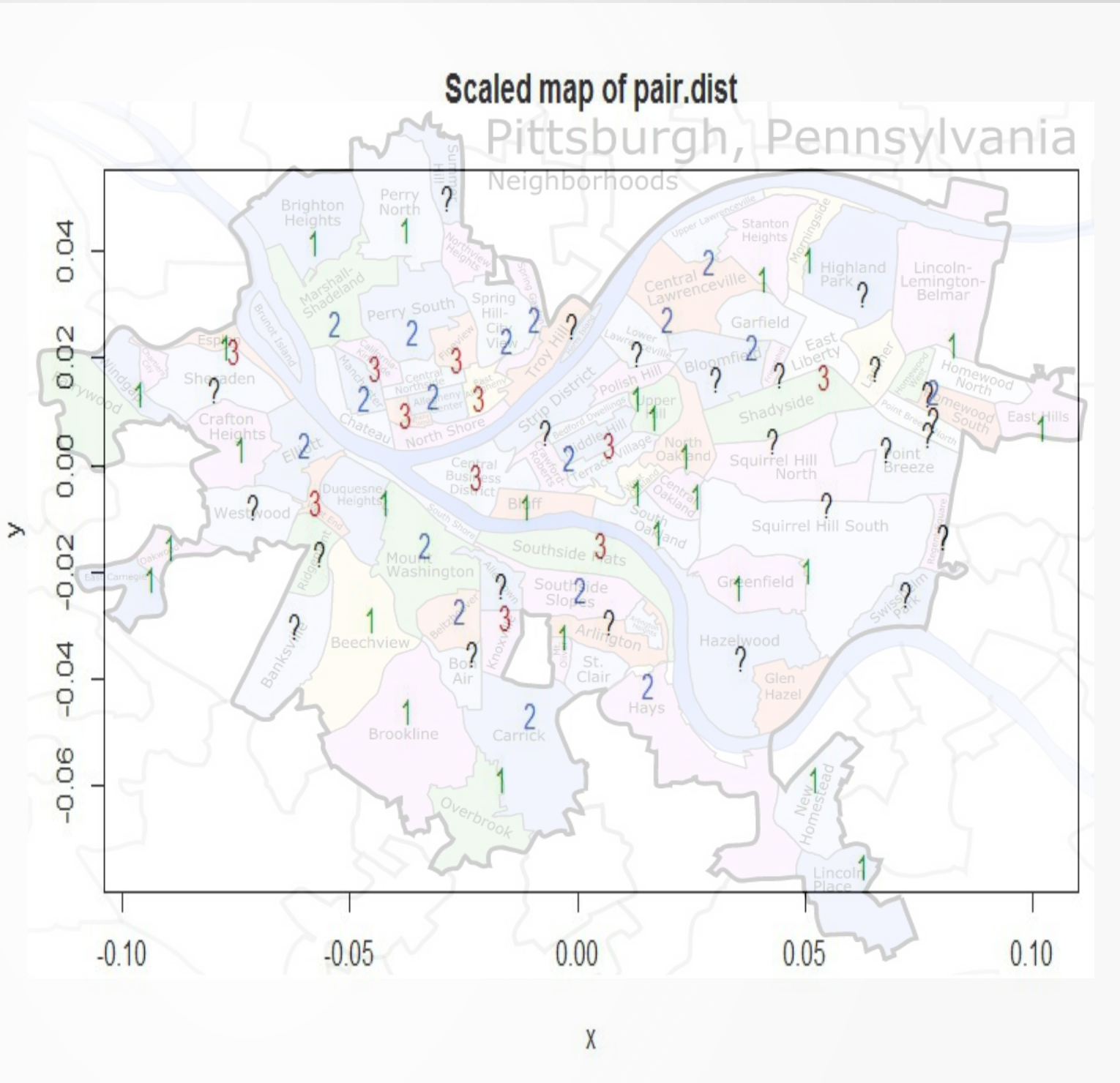
Random Forest prone to overfit

- Individual forests vary between 25% and 40% error

Two-stage algorithm has only 30% error

- Tested on 4-folds cross-validation

We're pretty sure map of neighborhoods is... Pittsburgh!


Scaled map of pair.dist
Pittsburgh, Pennsylvania
Neighborhoods
Brighton Heights
Perry North
Marshall-Shadeland
Perry South
Spring Hill-City View
Troy Hill
Central Lawrenceville
Stanton Heights
Morningside
Highland Park
Lincoln-Lemington-Belmar
Garfield
Bloomfield
East Liberty
Homewood North
Homewood South
East Hills
Shadyside
Point Breeze
Squirrel Hill North
Squirrel Hill South
Strip District
Polish Hill
Upper Hill
Middle Hill
North Oakland
Central Oakland
South Oakland
Central Business District
Bluff
Southside Flats
Southside Slopes
Mount Washington
Chateau
North Shore
Duquesne Heights
Crafton Heights
Elliott
Westwood
Greenfield
Hazelwood
Glen Hazel
Arlington
St. Clair
Hays
Beechview
Brookline
Carrick
Bon Air
Overbrook
New Homestead
Lincoln Place
Swisshelm Park
Banksville
x
y
-0.10
-0.05
0.00
0.05
0.10
0.04
0.02
0.00
-0.02
-0.04
-0.06