Week 1, Video 5

Classifiers, Part 3

Classification

- There is something you want to predict (“the label”)
- The thing you want to predict is categorical
  - The answer is one of a set of categories, not a number

In a Previous Class

- Step Regression
- Logistic Regression
- J48/C4.5 Decision Trees

Today

- More Classifiers

Decision Rules

- Sets of if-then rules which you check in order

Decision Rules Example

- **IF** time < 4 and knowledge > 0.55 then **CORRECT**
- **ELSE IF** time < 9 and knowledge > 0.82 then **CORRECT**
- **ELSE IF** numattempts > 4 and knowledge < 0.33 then **INCORRECT**
- **OTHERWISE CORRECT**

Many Algorithms

- Differences are in terms of how rules are generated and selected

- Most popular subcategory (including JRip and PART) repeatedly creates decision trees and distills best rules

Generating Rules from Decision Tree

1. Create Decision Tree
2. If there is at least one path that is worth keeping, go to 3 else go to 6
3. **Take the "Best" single path from root to leaf and make that path a rule**
4. **Remove all data points classified by that rule from data set**
5. **Go to step 1**
6. **Take all remaining data points**
7. **Find the most common value for those data points**
8. **Make an "otherwise" rule using that**

Relatively conservative

- Leads to simpler models than most decision trees

Very interpretable models

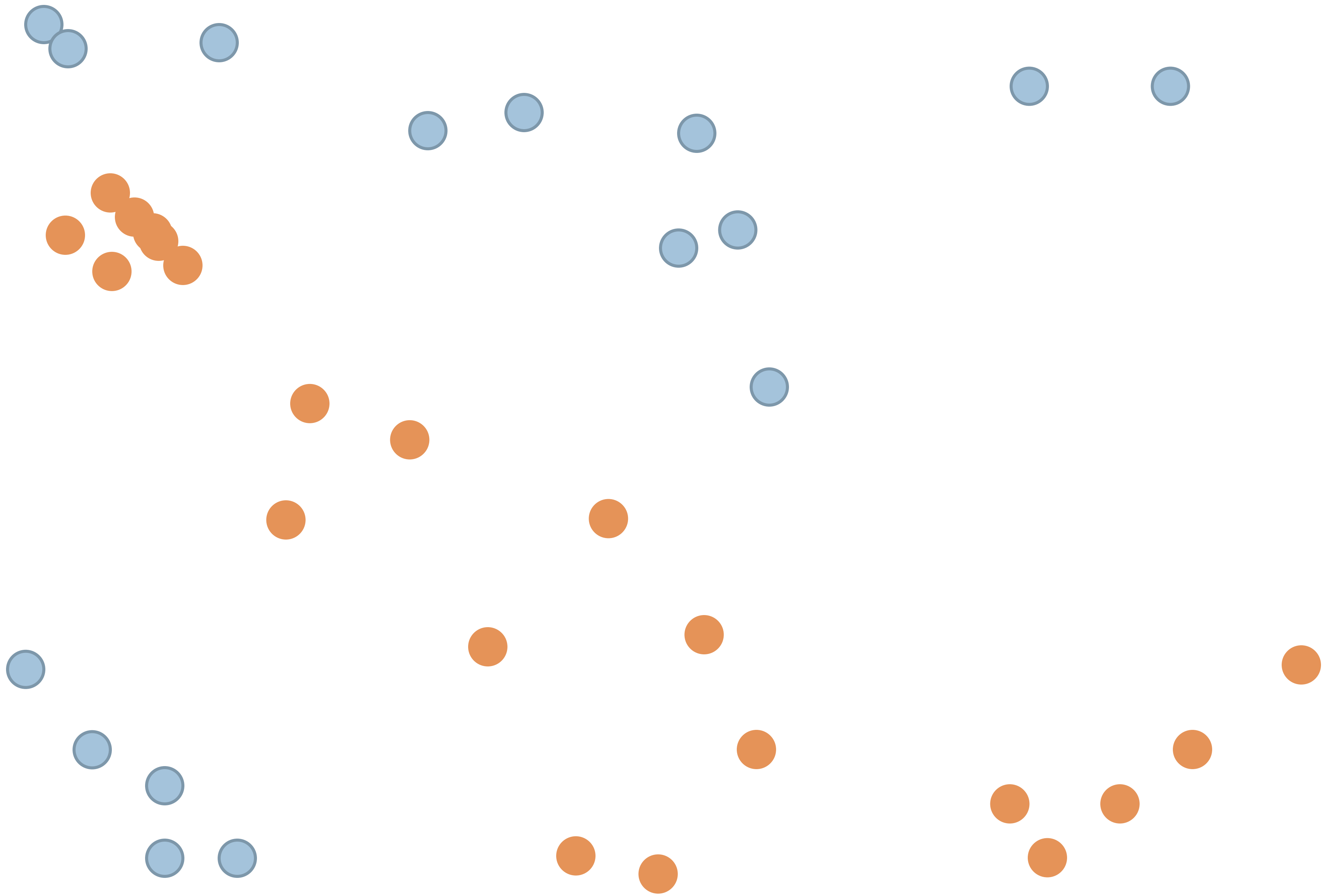
- Unlike most other approaches

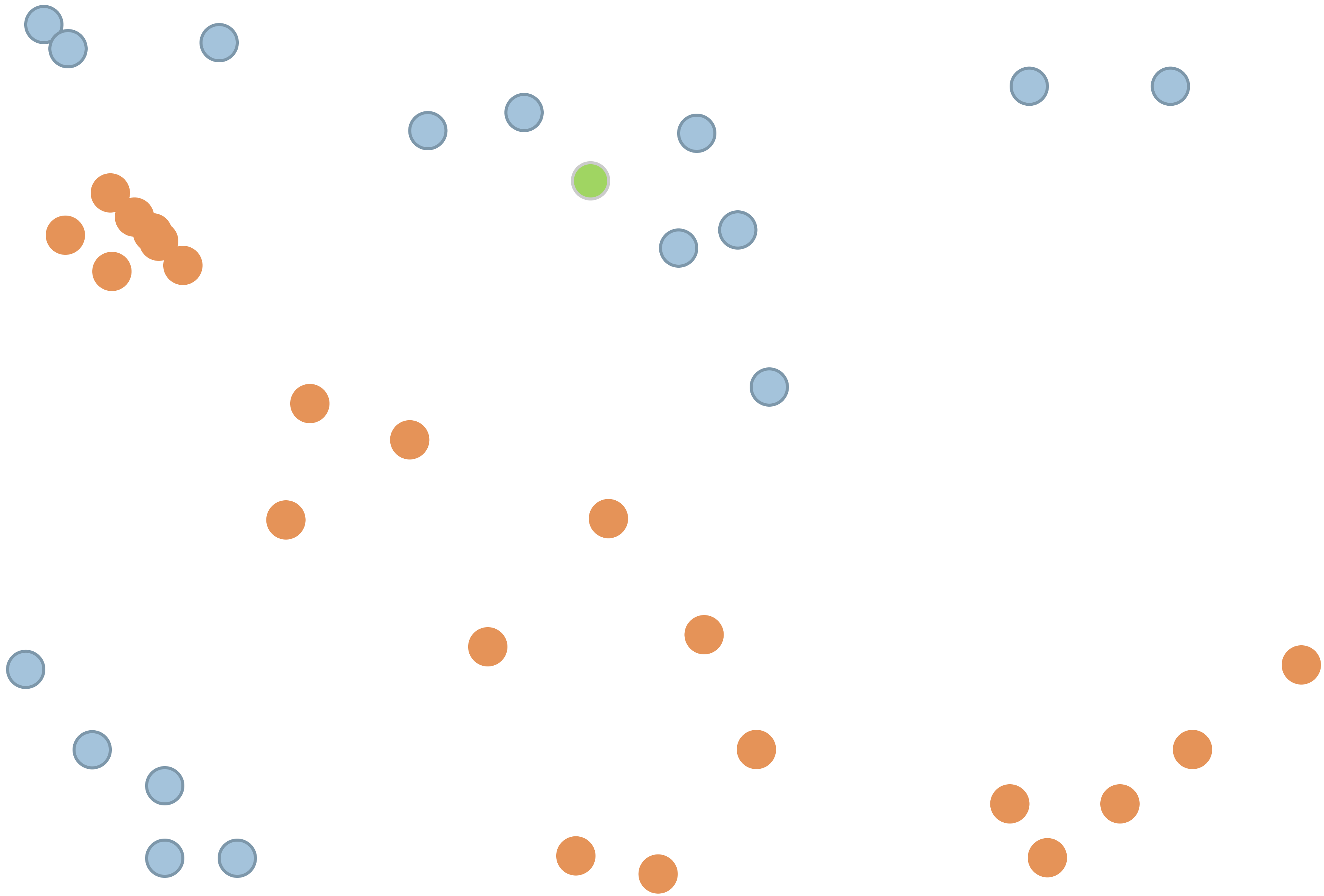
Good when multi-level interactions are common

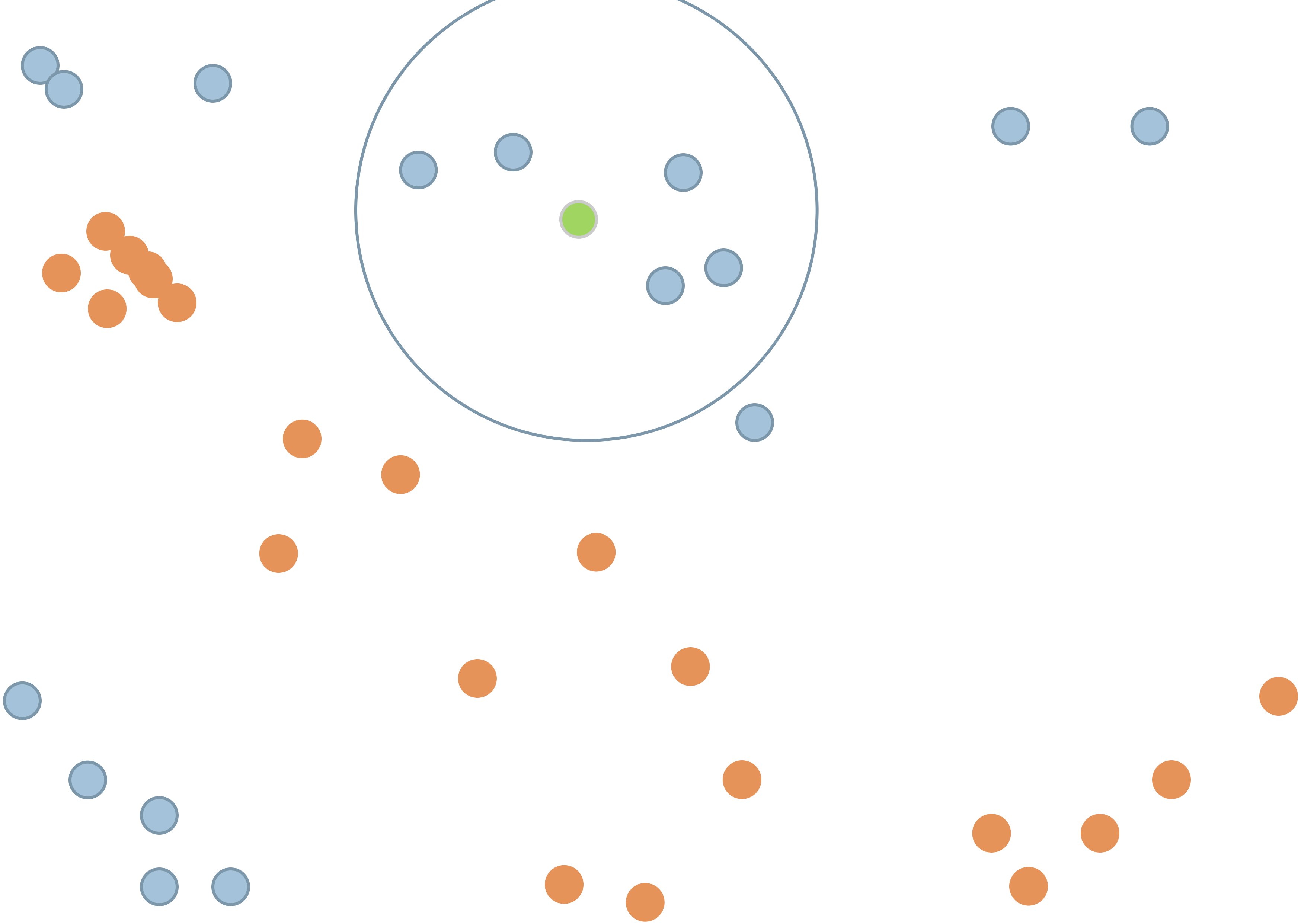
- Just like decision trees

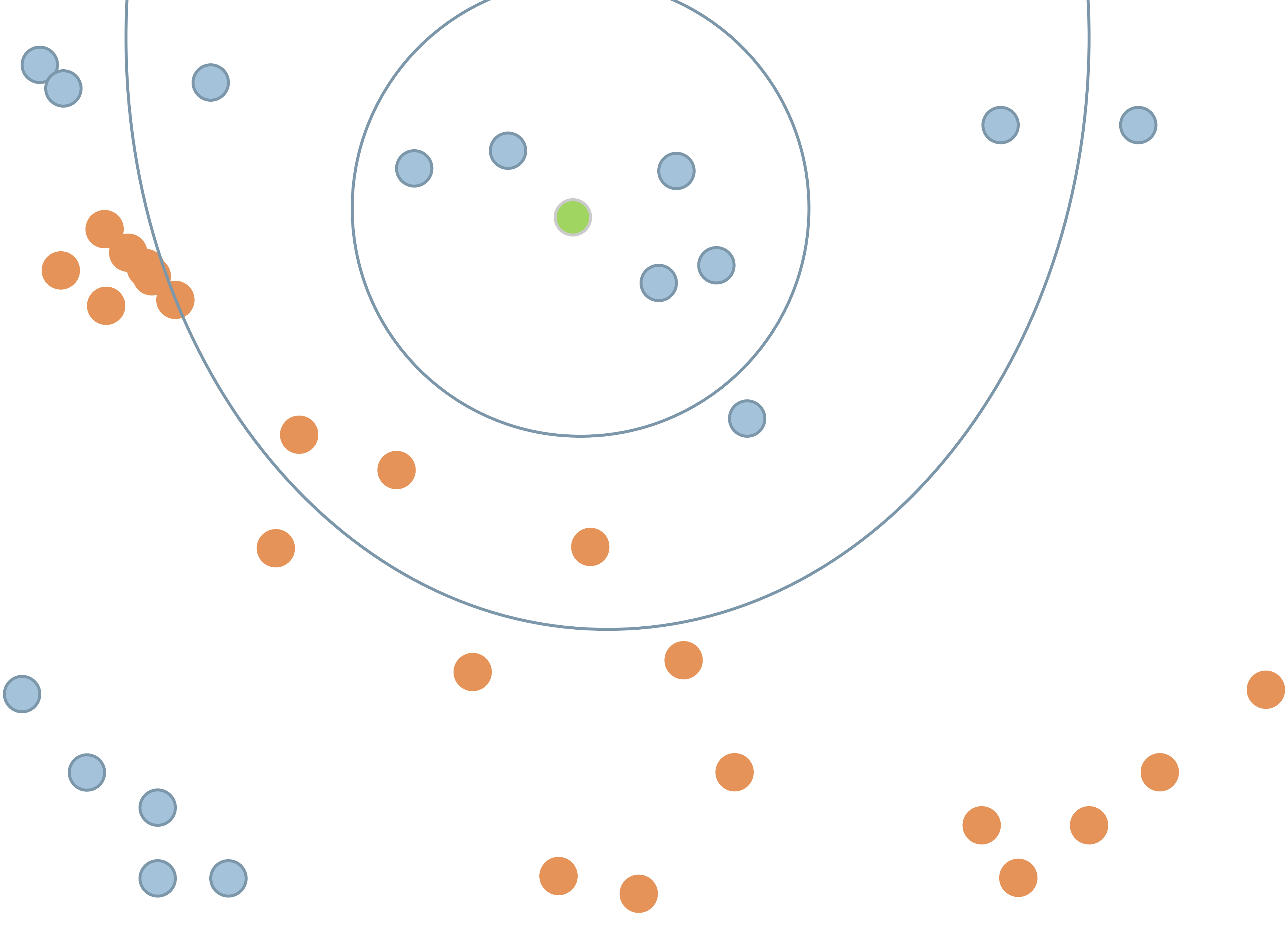
K*

- Predicts a data point from neighboring data points
  - Weights points more strongly if they are nearby









Good when data is ***very*** divergent

- Lots of different processes can lead to the same result

- Intractable to find general rules

- But data points that are similar tend to be from the same group

Big Advantage

- Sometimes works when nothing else works
- Has been useful for my group in detecting emotion from log files (Baker et al., 2012)

Big Drawback

- To use the model, you need to have the whole data set

Bagged Stumps

- Related to decision trees
- Lots of trees with only the first feature
- Relatively conservative

- A close variant is Random Forest

Common Thread

- So far, all the classifiers I've discussed are conservative
  - Find simple models
  - Don't over-fit

- These algorithms appear to do better for most educational data mining than less conservative algorithms
  - In brief, educational data has lots of systematic noise

Some less conservative algorithms

Support Vector Machines

- Conducts dimensionality reduction on data space and then fits hyperplane which splits classes
- Creates very sophisticated models
- Great for text mining
- Great for sensor data
- Not optimal for most other educational data
  - Logs, grades, interactions with software

Genetic Algorithms

- Uses mutation, combination, and natural selection to search space of possible models
- Can produce inconsistent answers

Neural Networks

- Composes extremely complex relationships through combining "perceptrons"
- Finds *very* complicated models

Soller & Stevens (2007)

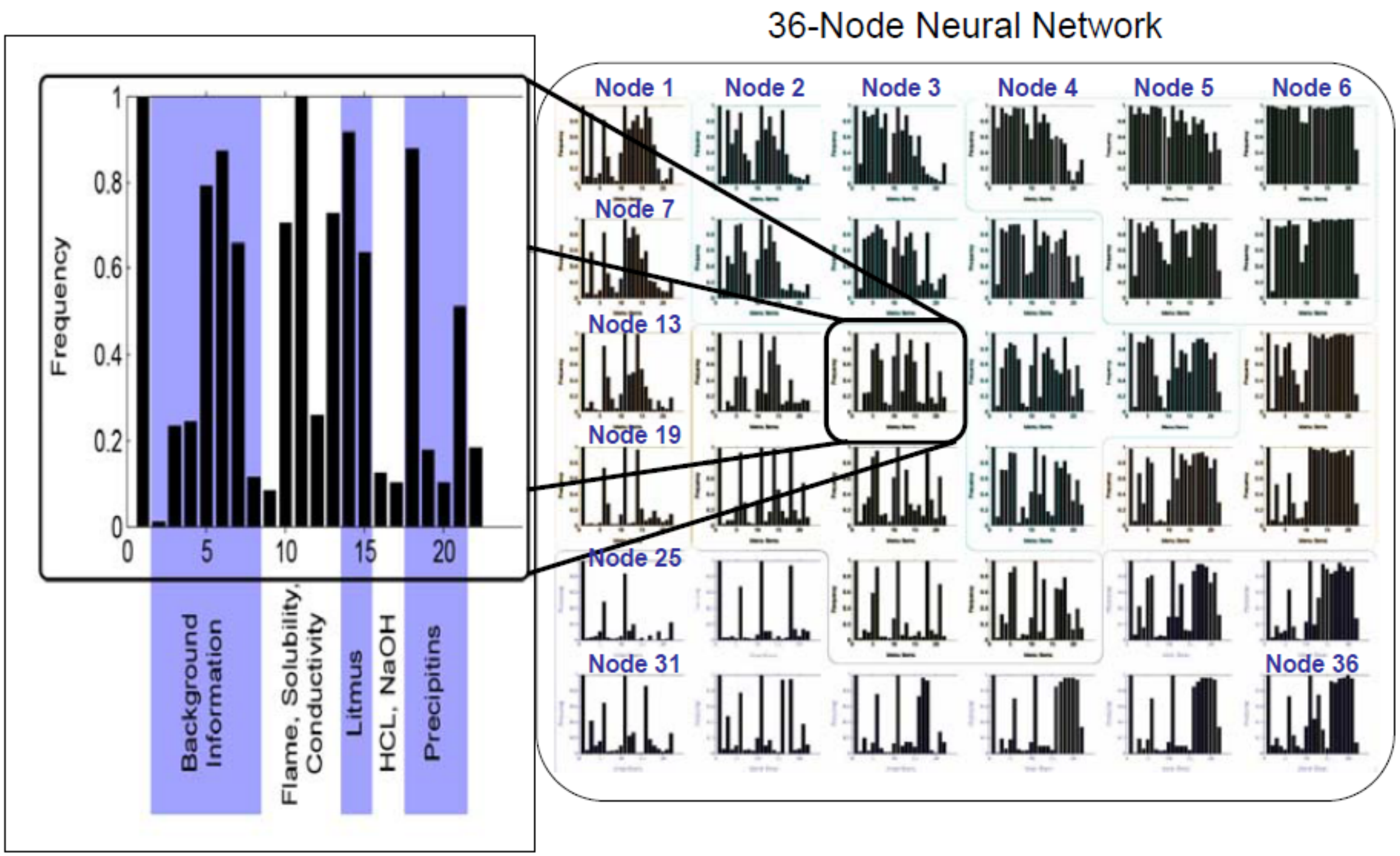
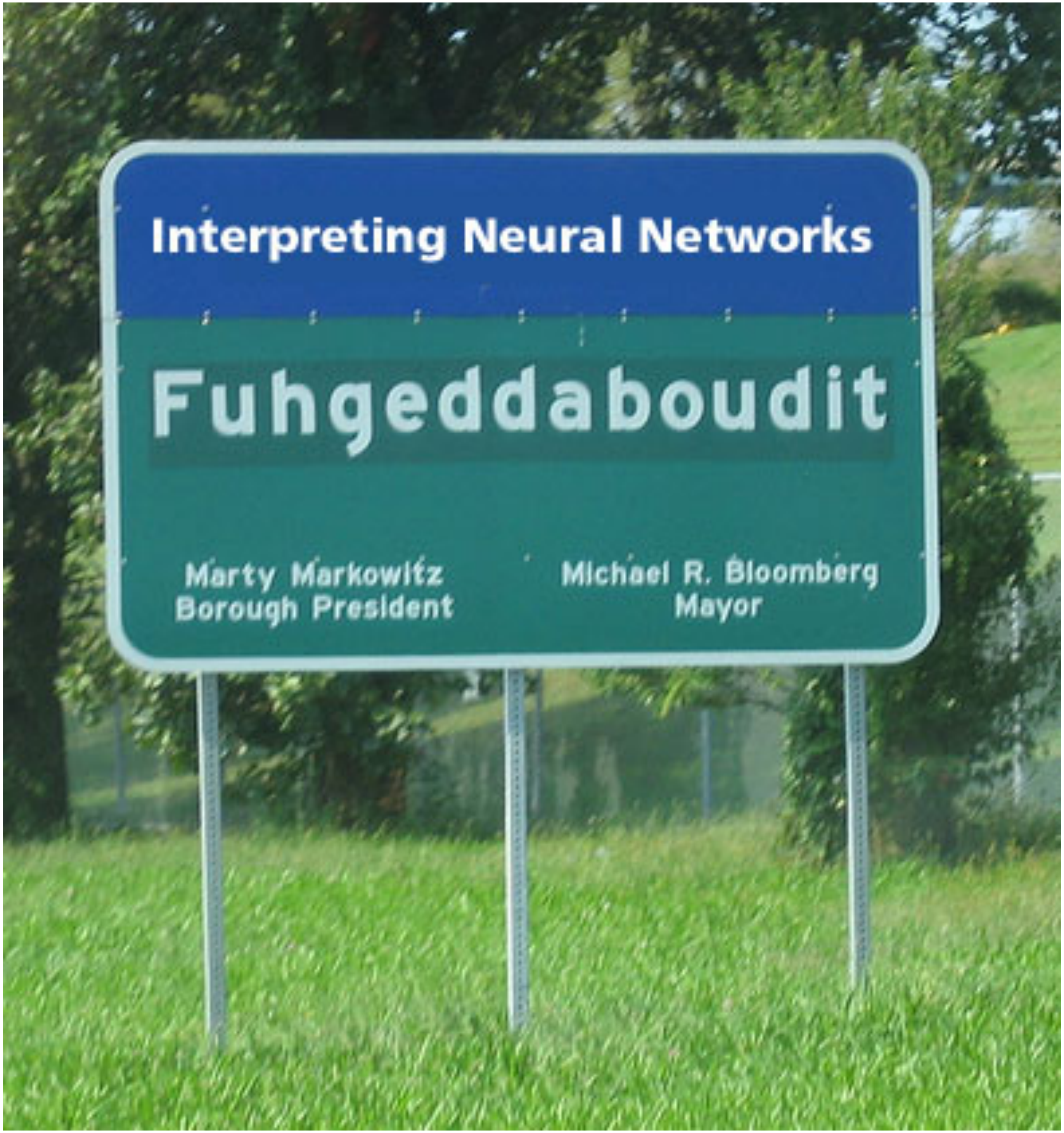


Figure 11. A Neural Network Showing the 36 Nodes, Each Describing a Different Subset of the Population

In fact

- The difficulty of interpreting non-linear models is so well known, that they put up a sign about it in New York City


Interpreting Neural Networks
Fuhgeddaboudit
Marty Markowitz
Borough President
Michael R. Bloomberg
Mayor

Note

- Support Vector Machines, Genetic Algorithms, and Neural Networks are ***great*** for some problems
- For most types of educational data, they have not historically produced the best solutions

Later Lectures

- Goodness metrics for comparing classifiers
- Validating classifiers
- Classifier conservatism and over-fitting

Next Lecture

- A case study in classification