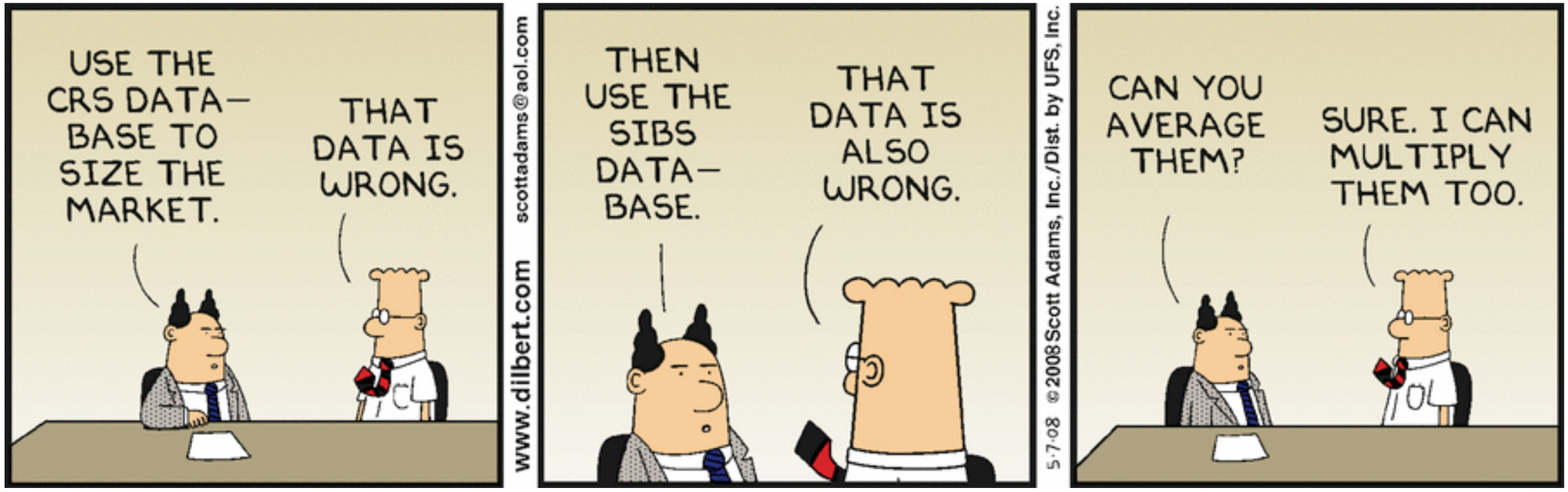
Clustering and Classification

Matt Dickenson
Nov. 13, 2015



Goals

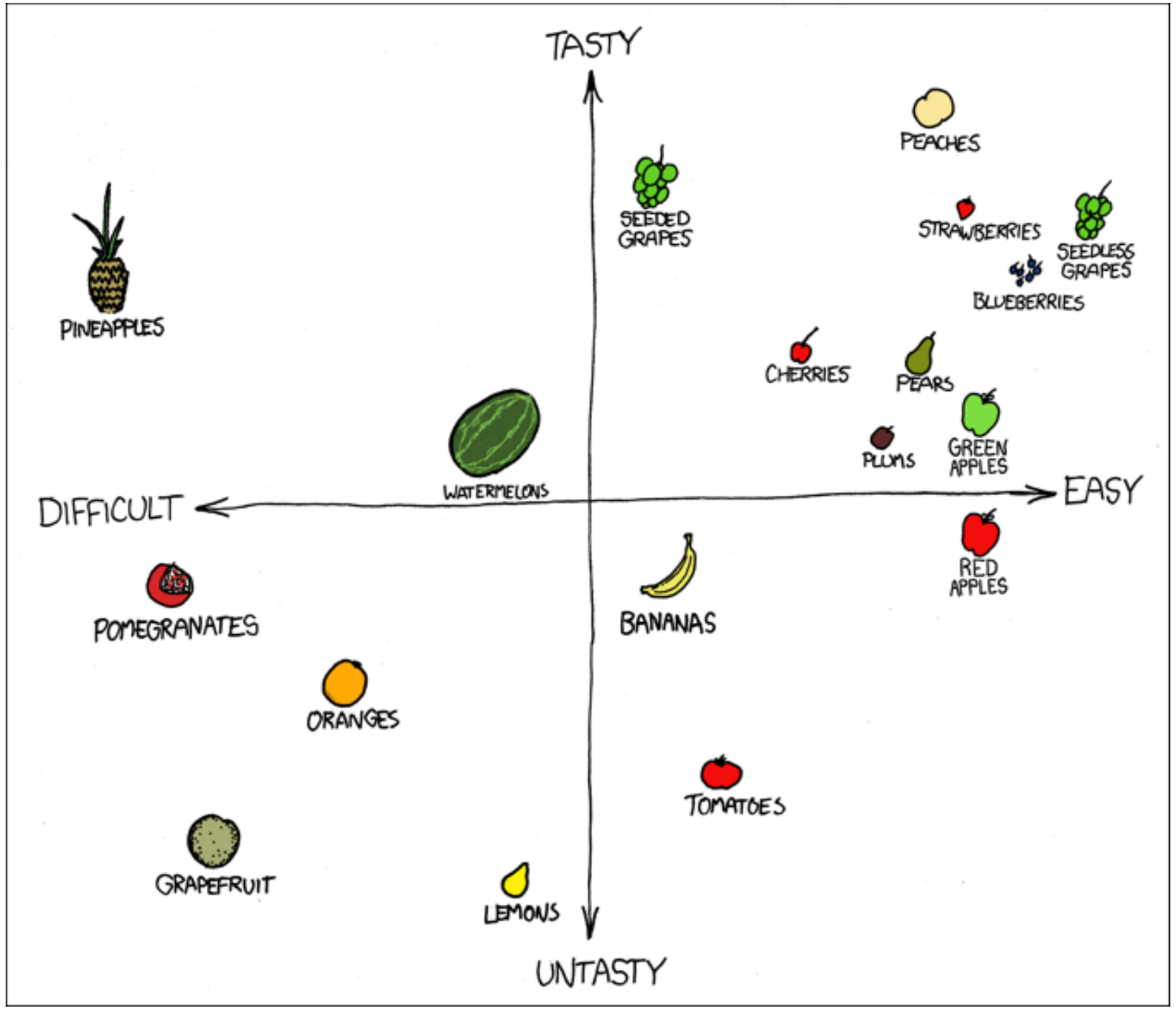
- Understand key terms related to classification
- Recognize options for classification algorithms
- Apply classification to real-world problems

Image: http://dilbert.com/strip/2008-05-07

Motivation: Customer Segmentation

"Given what we know about this customer, will she buy Tide®?"

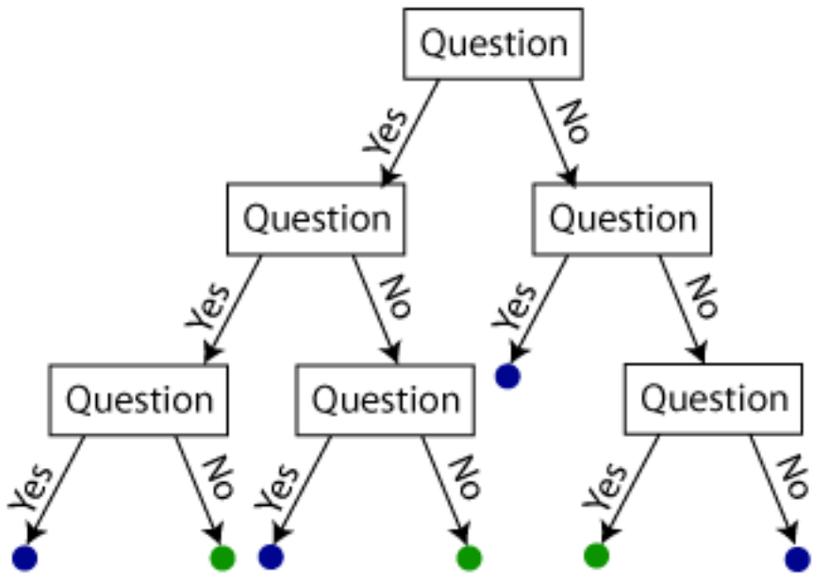
Image: http://savewater.com.au



Background

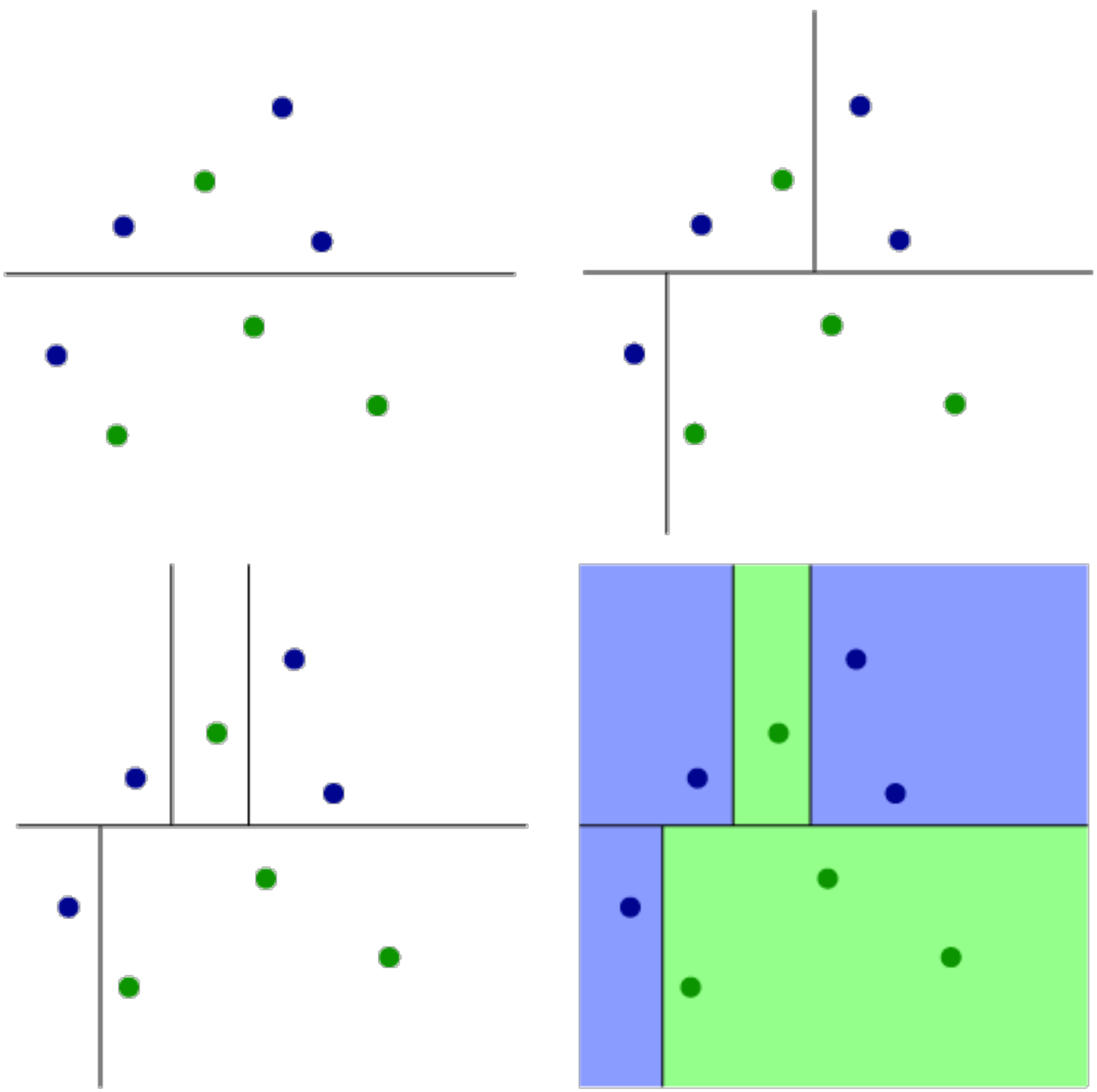
- Discrete & Continuous
- Predictive & Descriptive
- Train & Test

Image: https://xkcd.com/388/



Decision Tree

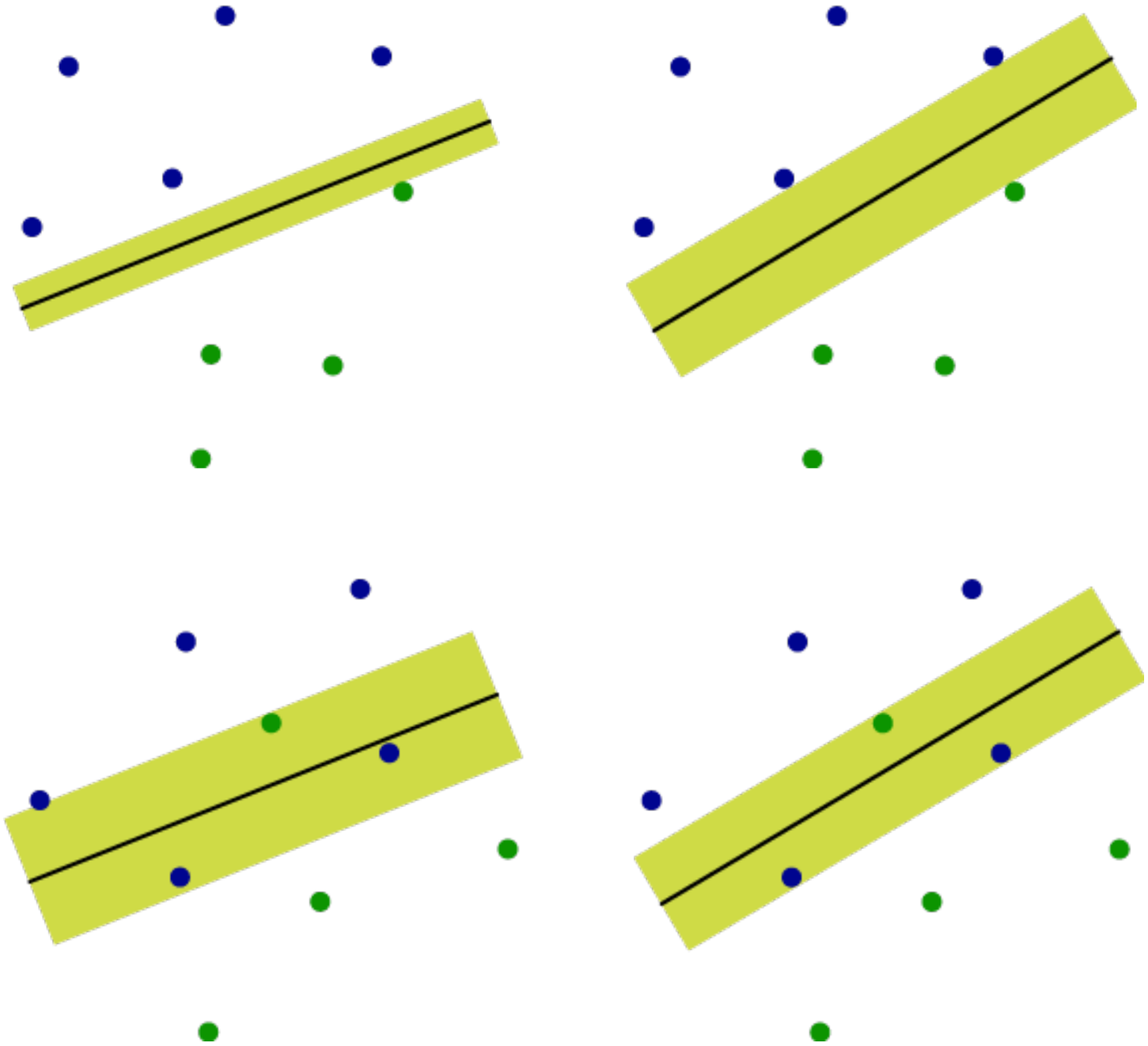
Image: http://shapeofdata.wordpress.com



Decision Tree

- Find a feature with high variance
- Select a cut-off that accurately classifies
- Repeat with subsequent dimensions

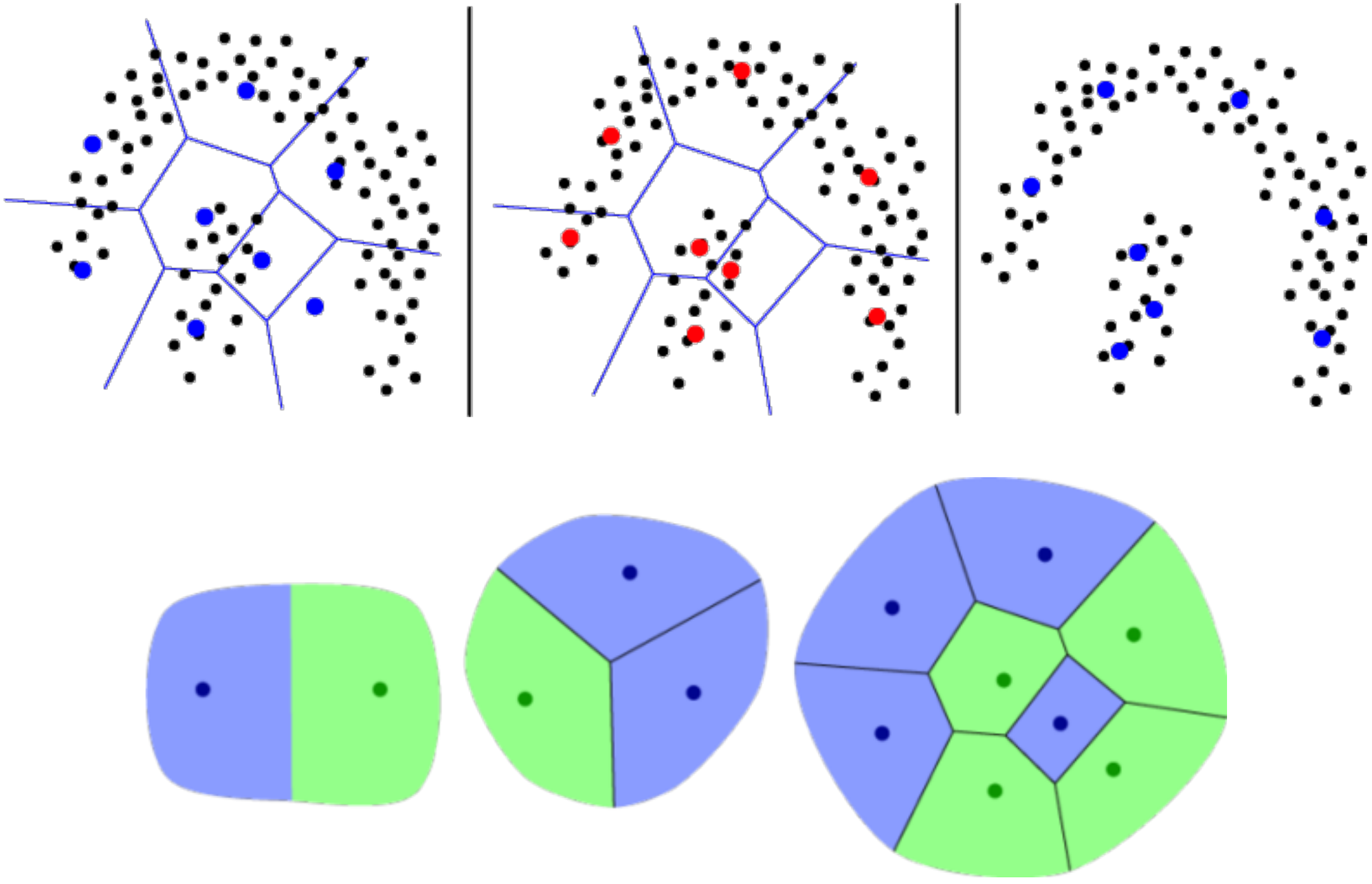
Image: http://shapeofdata.wordpress.com



Support Vector Machine

- Find a line that divides the data points
- Maximize the margin between groups
- Minimize error

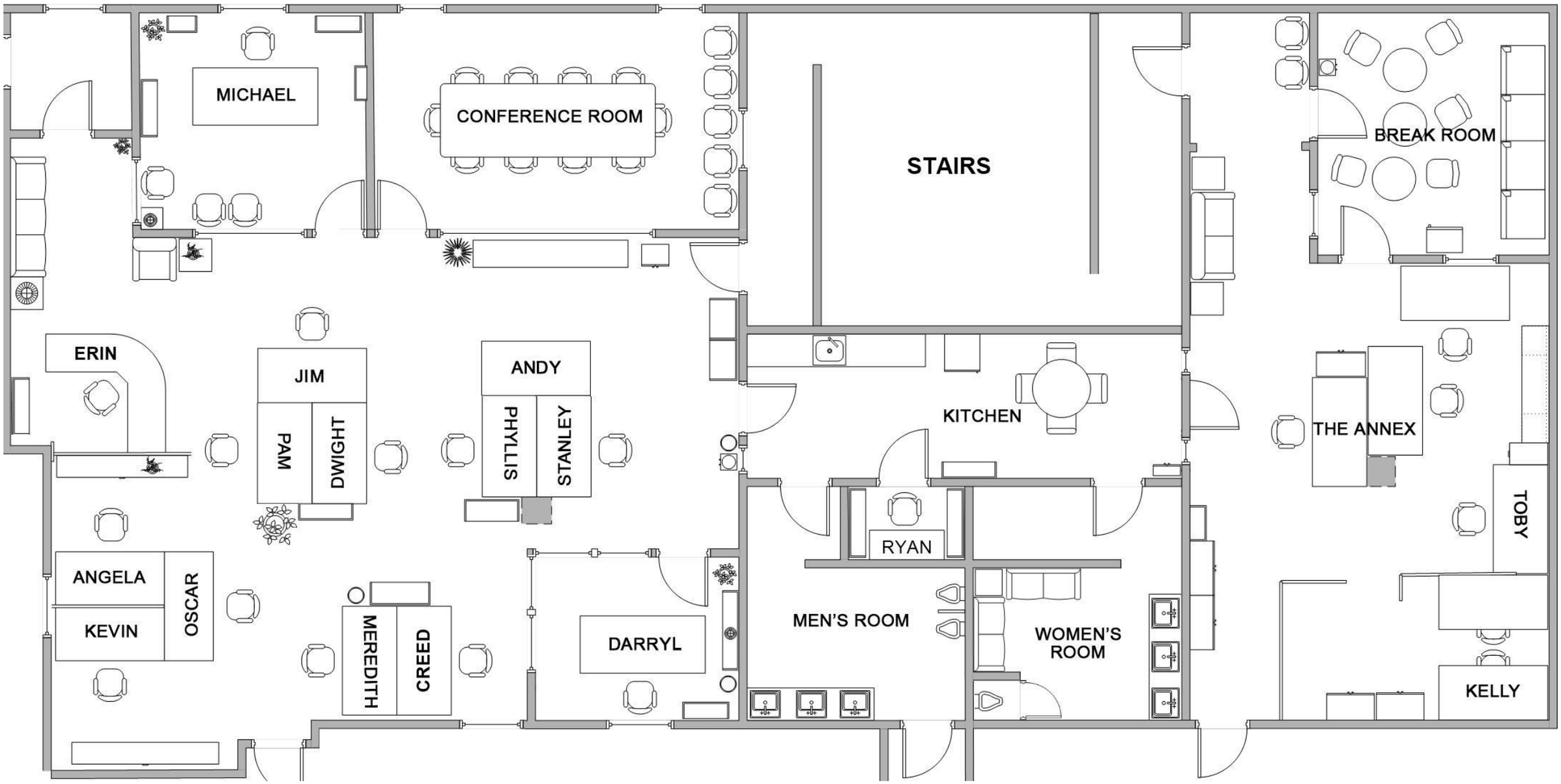
Image: http://shapeofdata.wordpress.com



K-Means

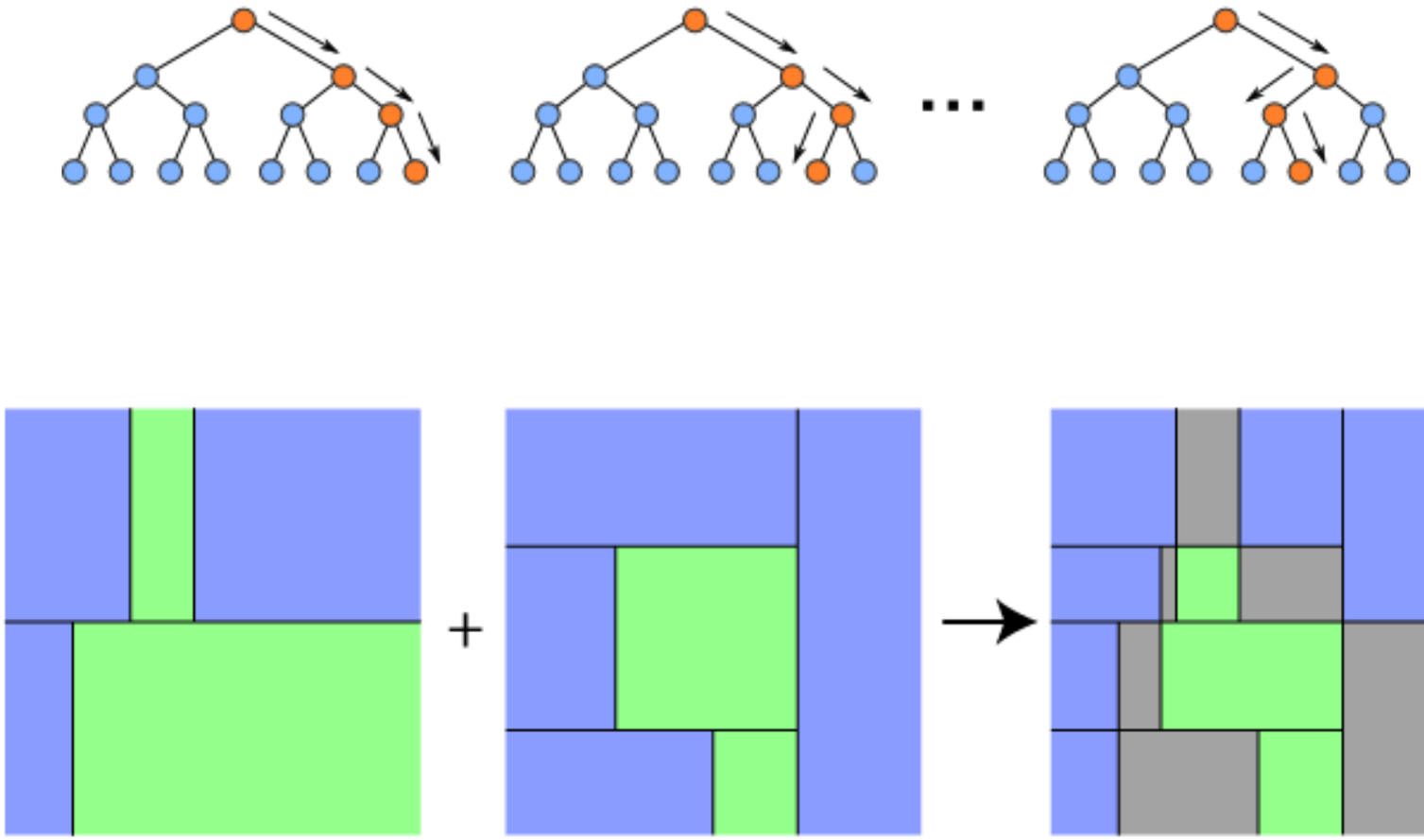
- Labelled centroids
- Classify by closest centroid
- Average new centroids

Image: http://shapeofdata.wordpress.com



K-Means

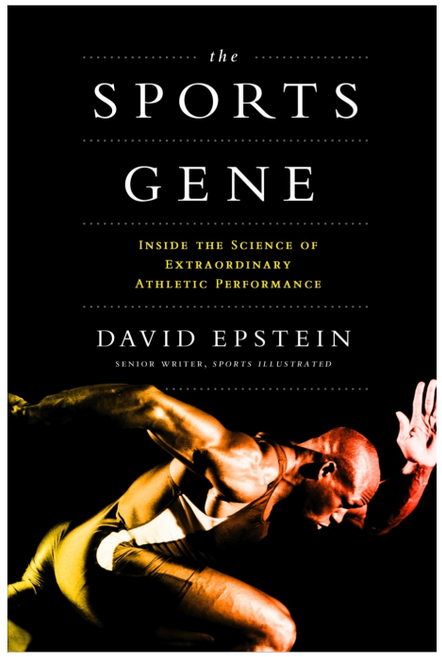
Image: http://theoffice.wikia.com



Random Forest

- Randomly remove features
- Generate many trees
- Average over them

Image: http://shapeofdata.wordpress.com

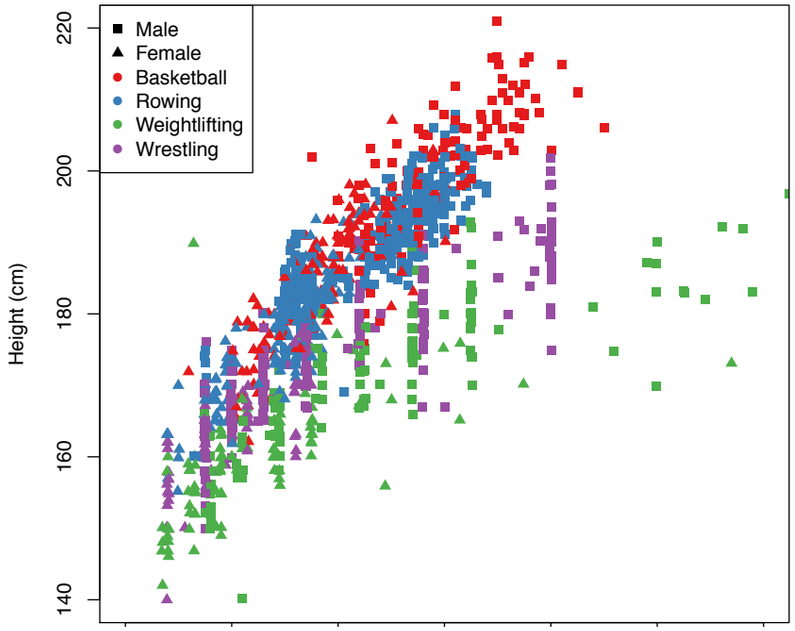


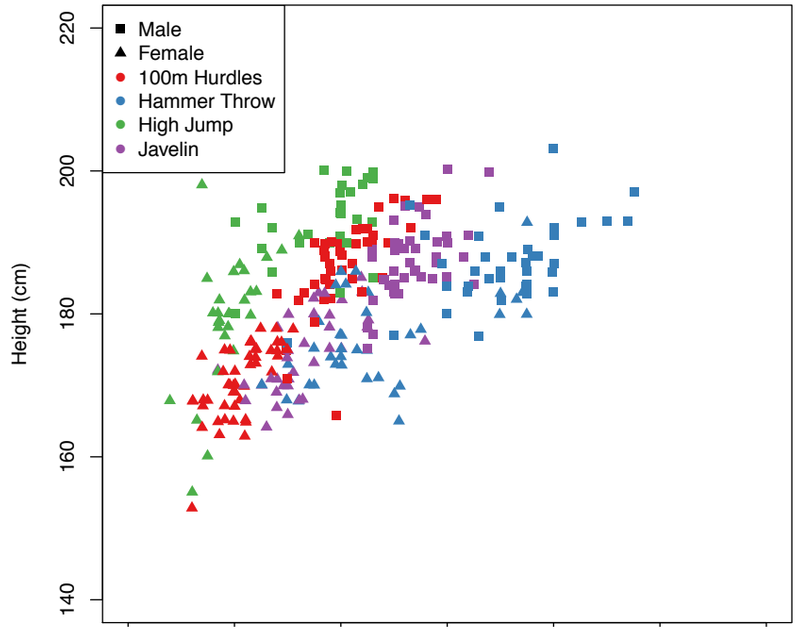
Roberts: [You argue that] if you simply had the height and weight of an Olympic roster, you could do a pretty good job of guessing what their events are. Is that correct?

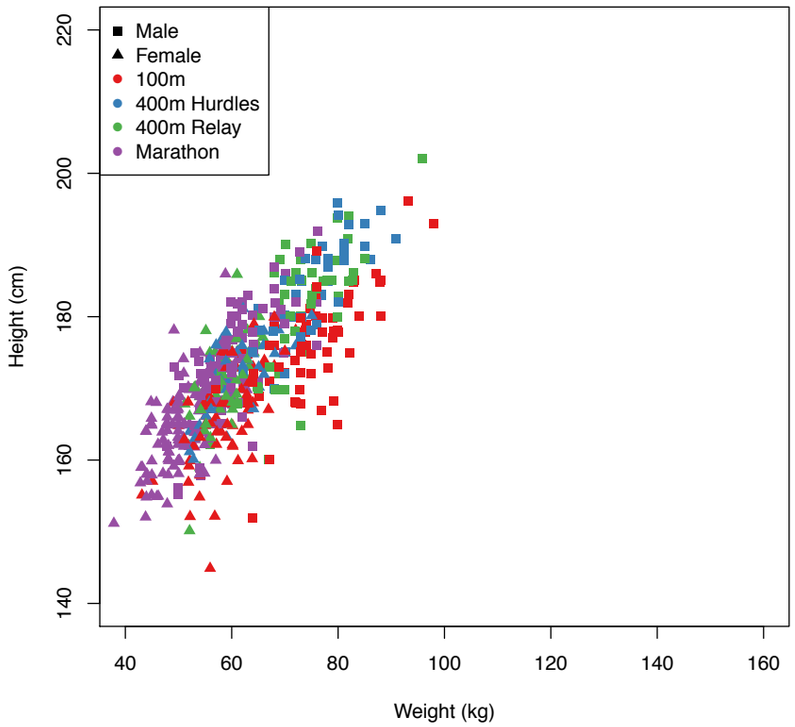
Epstein: *That's definitely correct.* I don't think you would get every person accurately, but… *I think you would get the vast majority of them correctly.* And frankly, you could definitely do it easily if you had them charted on a height-and-weight graph, and I think you could do it for most positions in something like football as well.

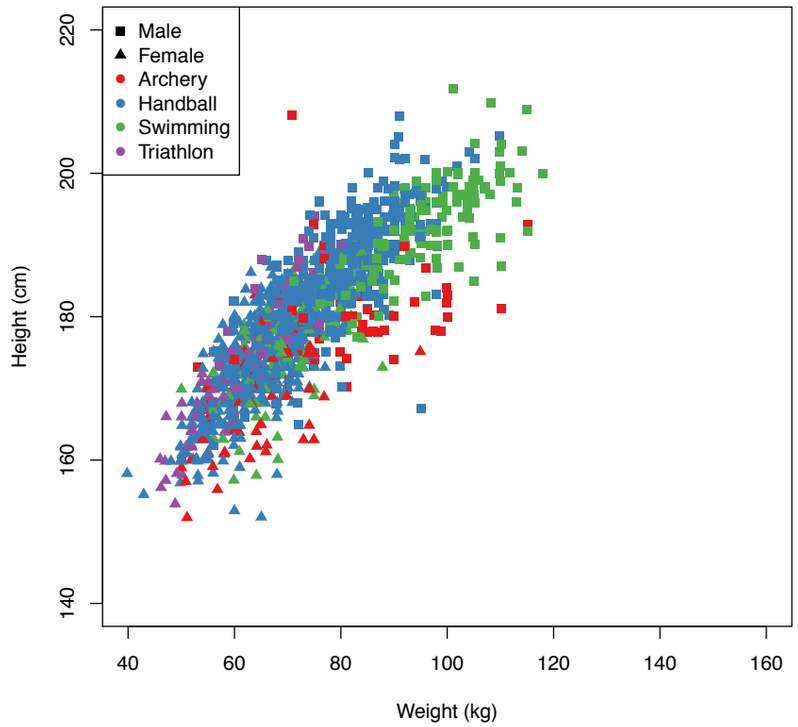
Classifying Olympic Athletes

Image: http://amazon.com
Quote: http://econtalk.org



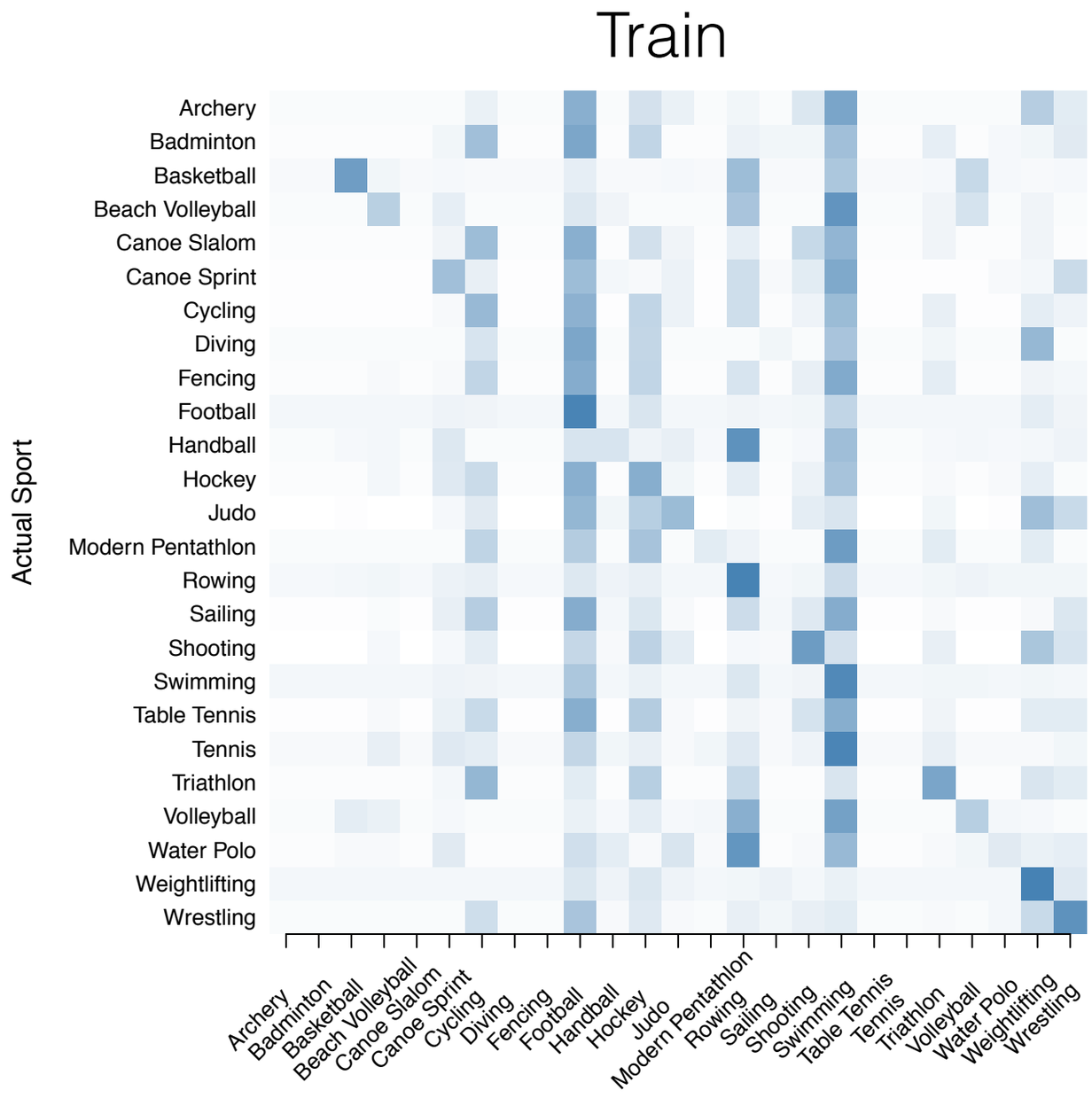


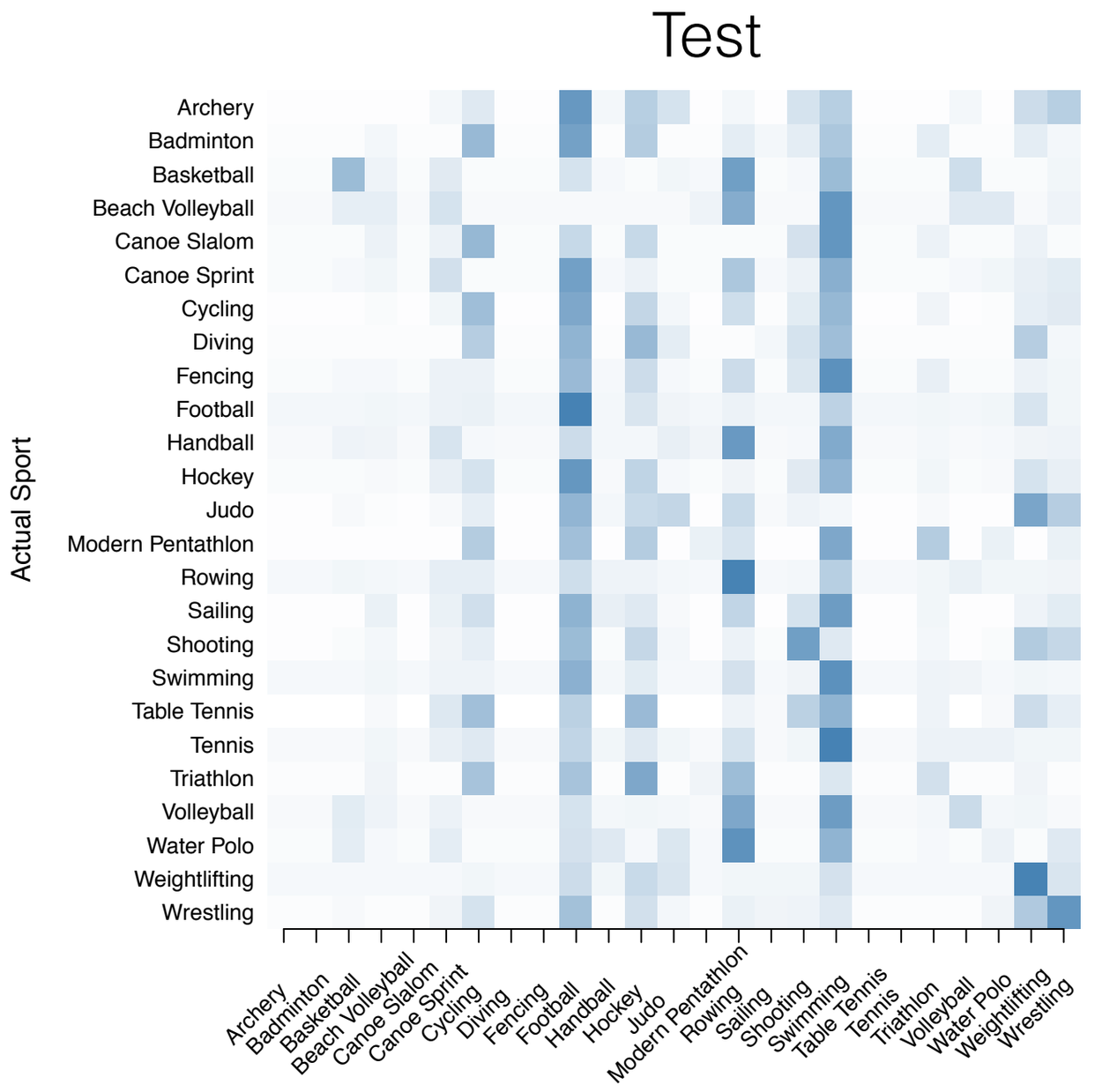




Description

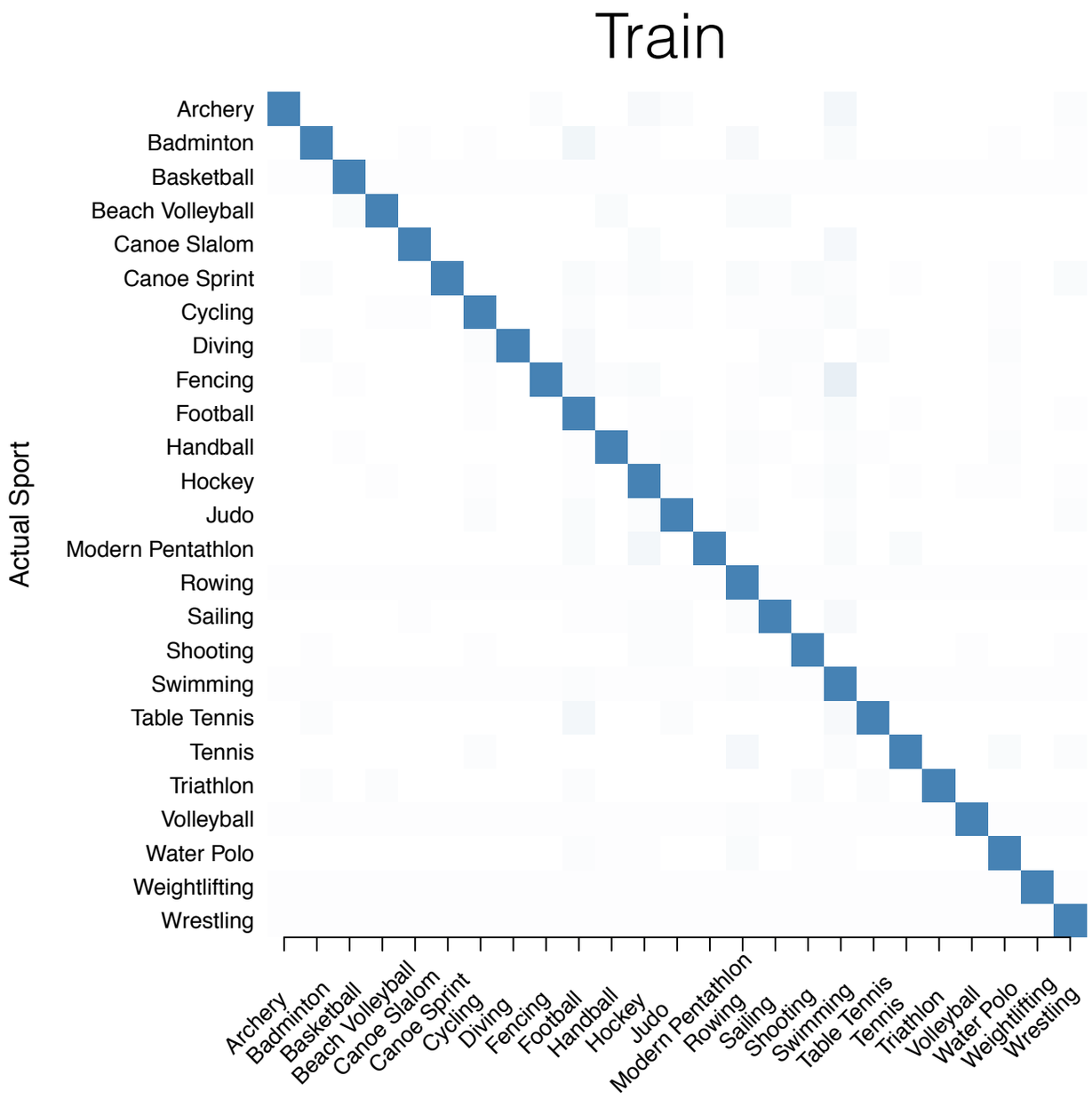
Data: http://theguardian.com

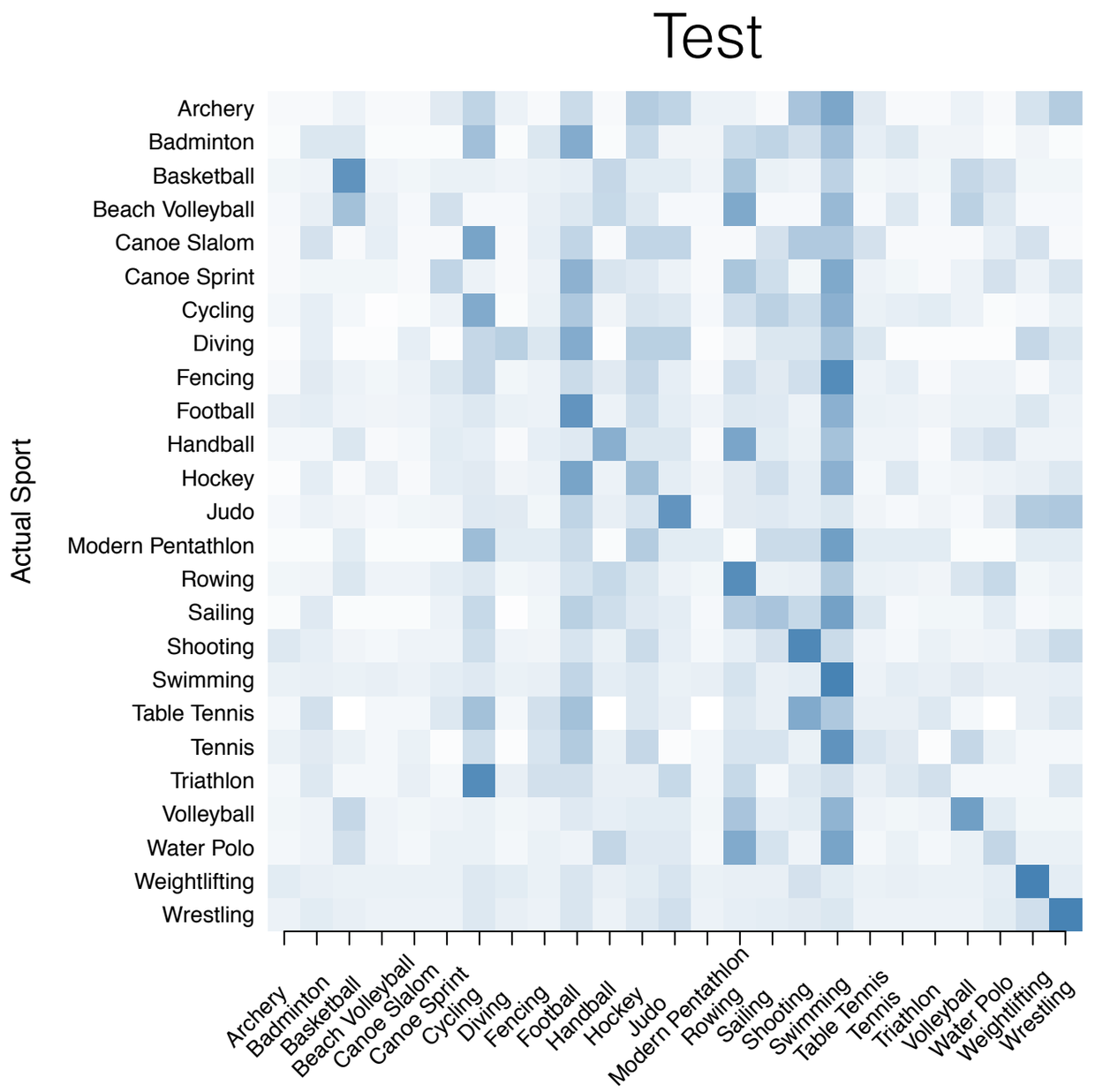




Conditional Inference Tree

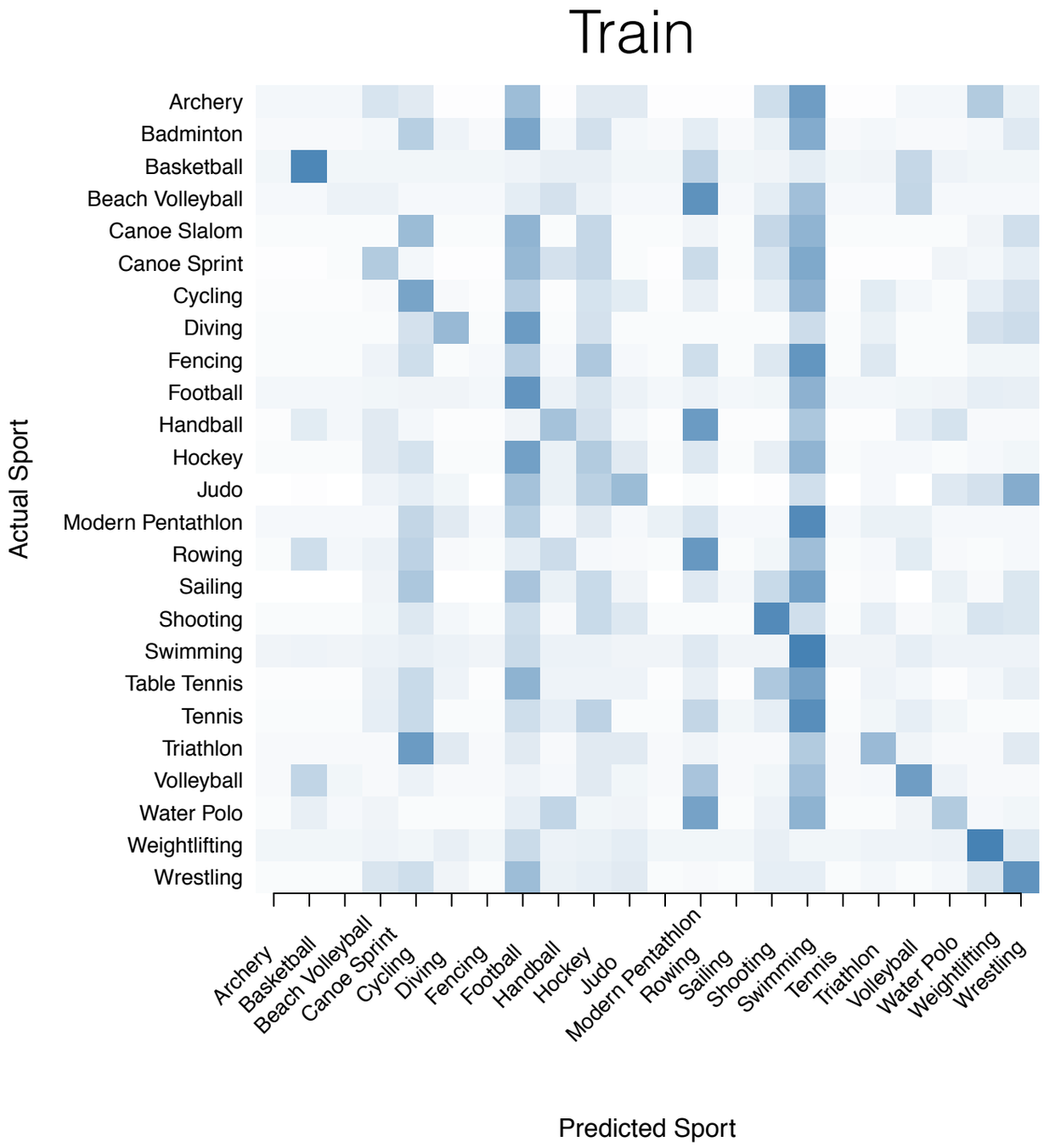
Training set accuracy: .279
Test set accuracy: .219

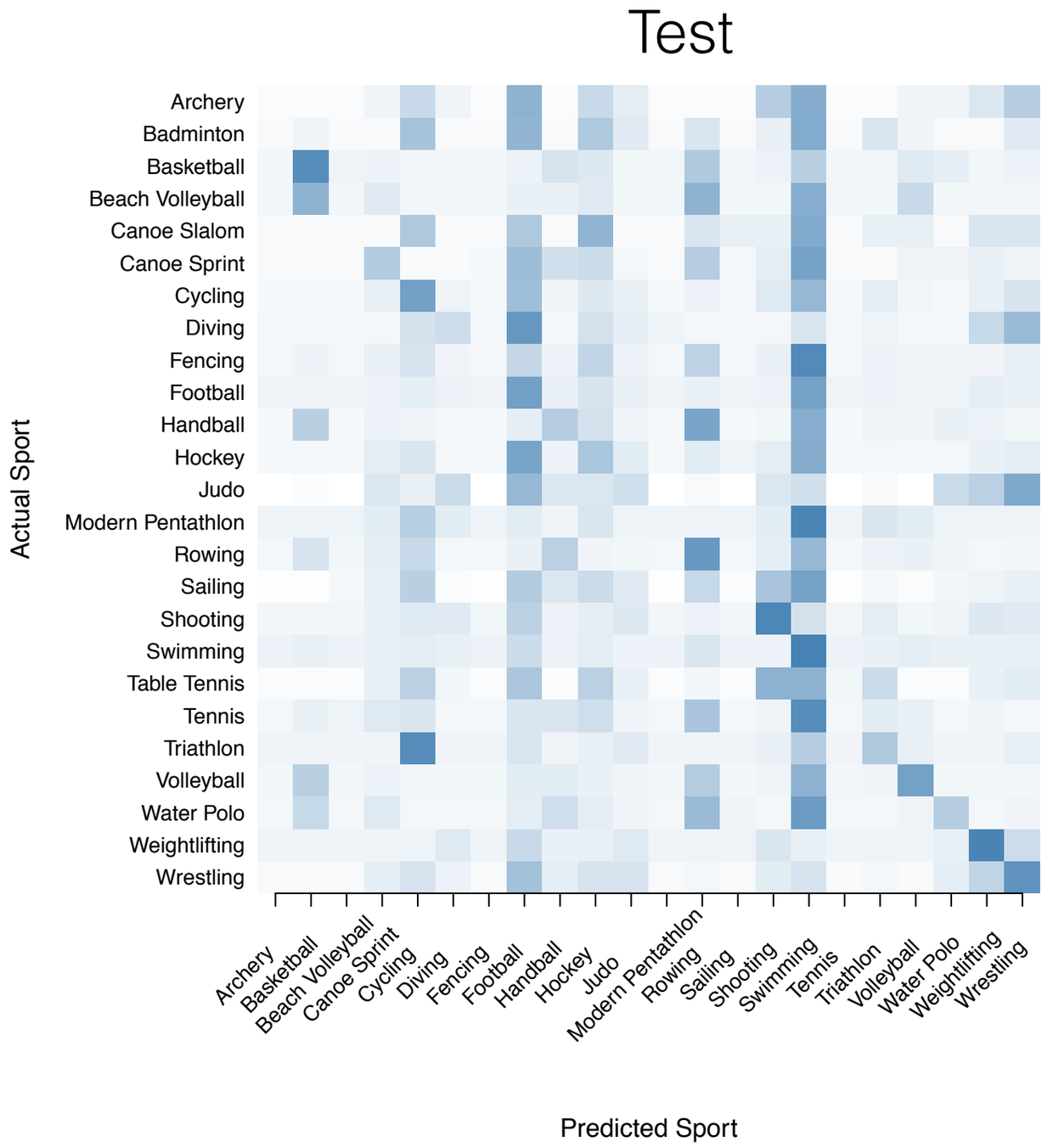




Random Forest

Training set accuracy: .923
Test set accuracy: .244





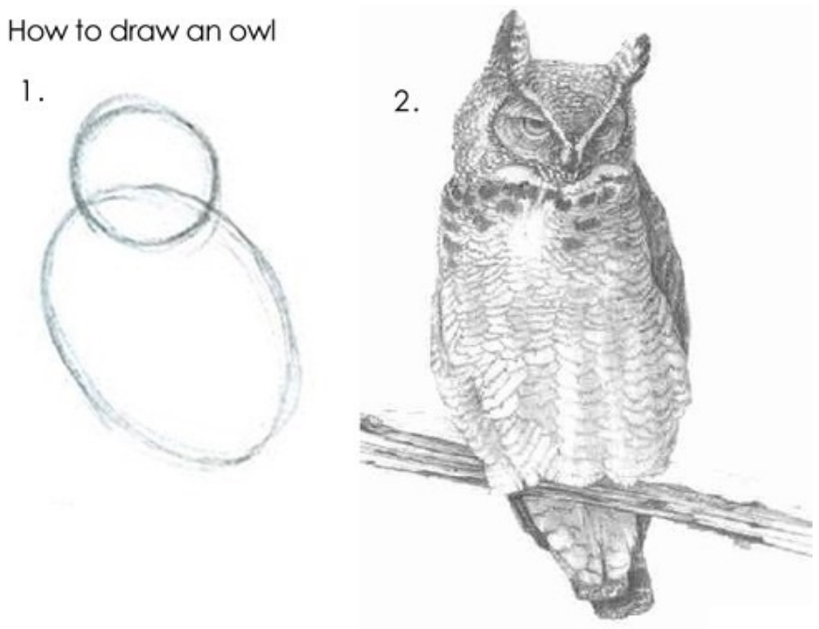
Neural Network

Training set accuracy: .280
Test set accuracy: .265

- Websites
  - shapeofdata.wordpress.com
  - datatau.com
  - research.facebook.com

- Books
  - *Probability & Statistics* (DeGroot & Schervish)
  - *Machine Learning: A Probabilistic Perspective* (Kevin Murphy)

- Podcasts
  - Data Skeptic
  - Not So Standard Deviations
  - Partially Derivative
  - Talking Machines
  - What's the Point?



Resources