A Quick Introduction to Machine Learning (Clustering)

Lecturer: John Guttag

Clustering

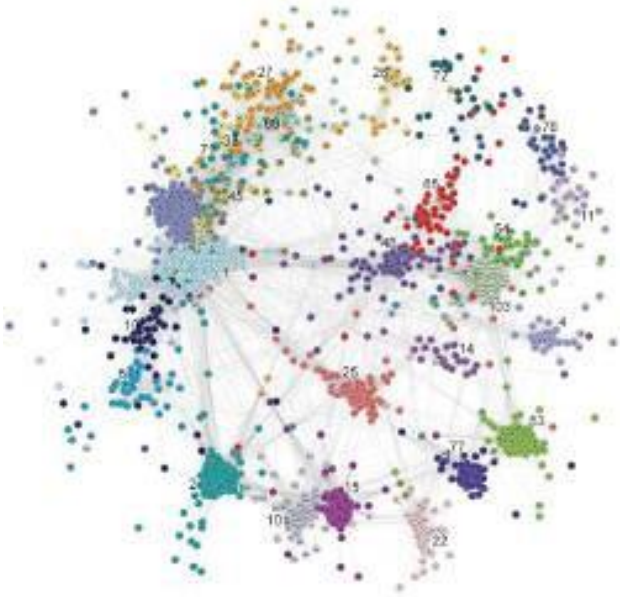
Find an intrinsic grouping in set of unlabeled examples

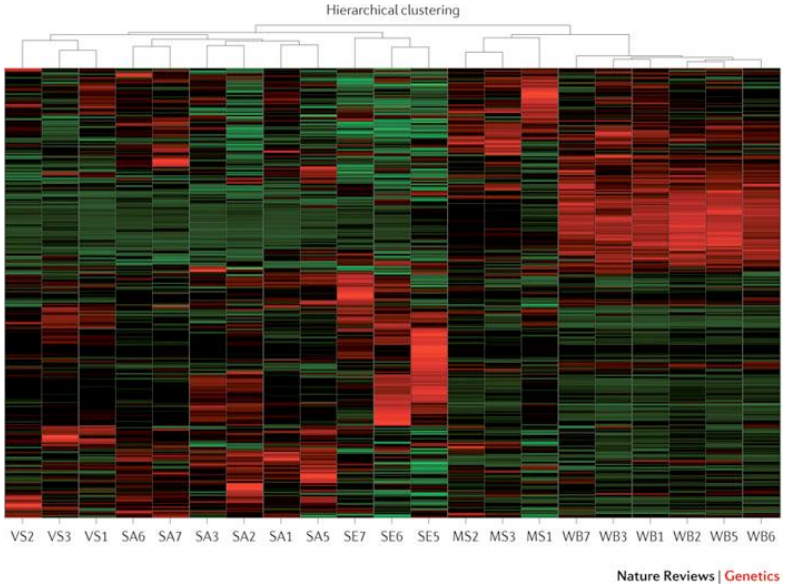
Of great practical utility

- Marketing
- Biology
- Insurance
- Medicine
- …

Marketing

Biology and Medicine



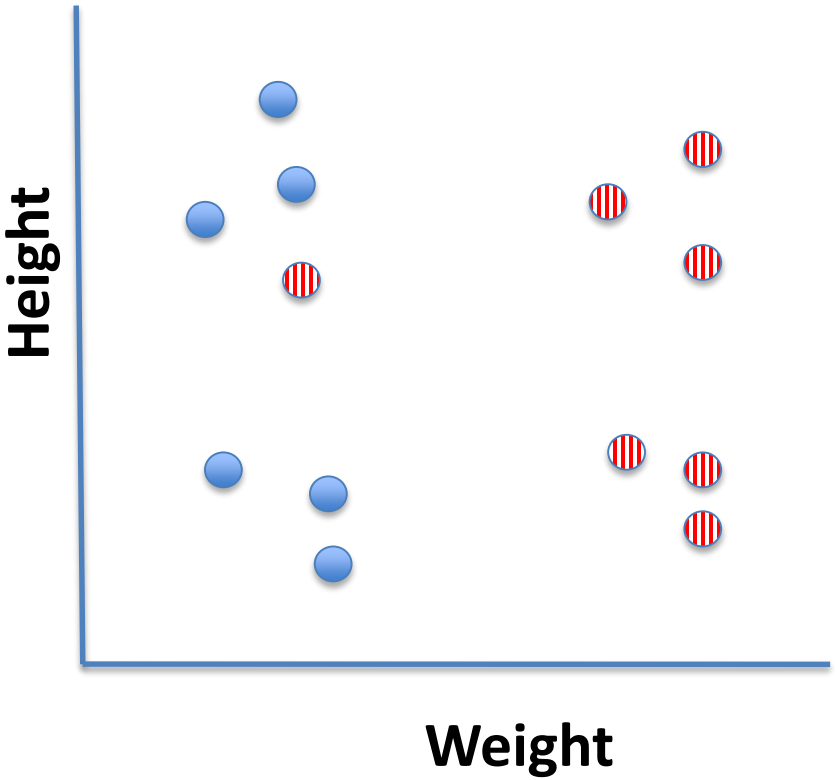


Insurance


Medicare Fraud Bust
91 arrests
8 U.S. cities
$295 million
Medicare
John Doe
000-00-0000
Male
HOSPITAL (PART A) 12
MEDICAL (PART B) 12

What Makes a Clustering Good?

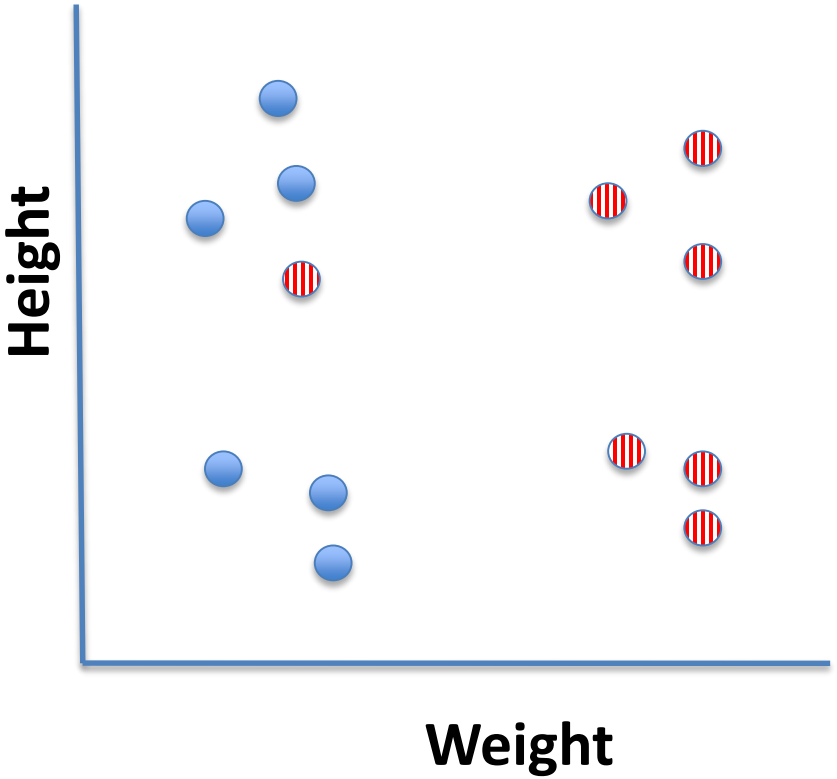
Depends upon the application



What Makes a Clustering Good?

Depends upon the application

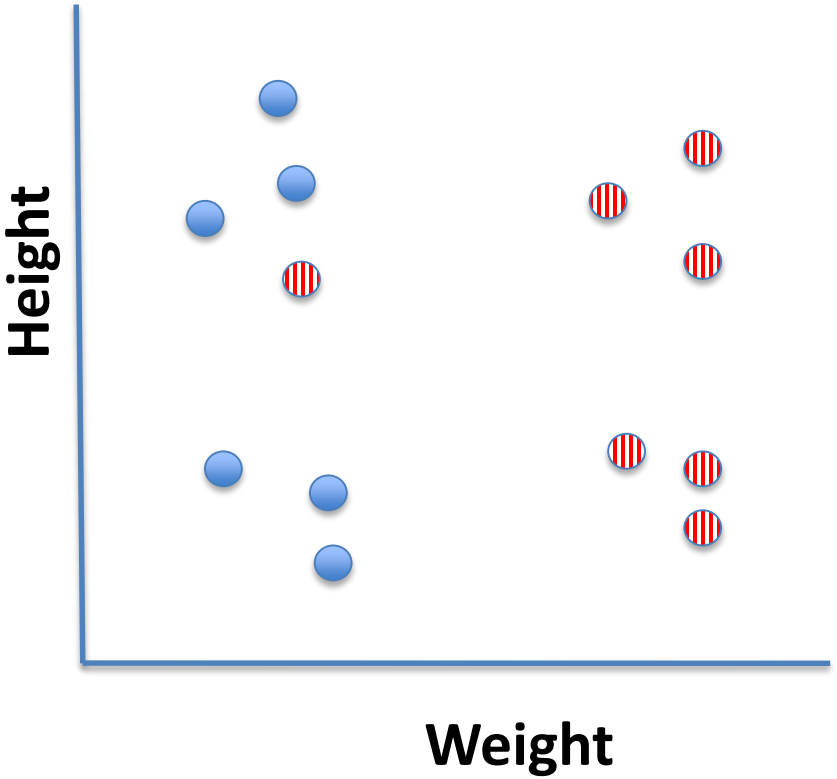
Basketball player



What Makes a Clustering Good?

Depends upon the application

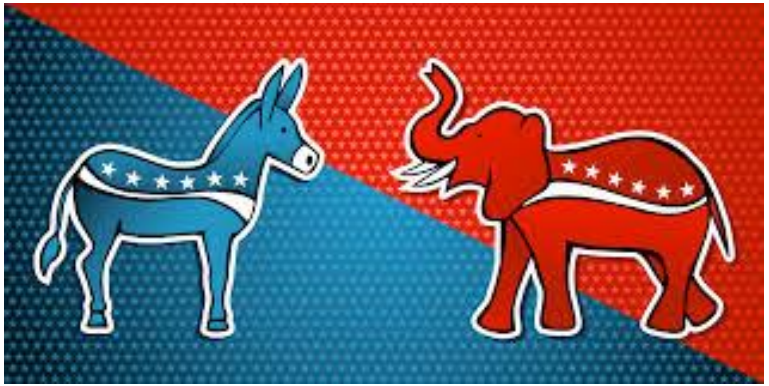
Sumo wrestler

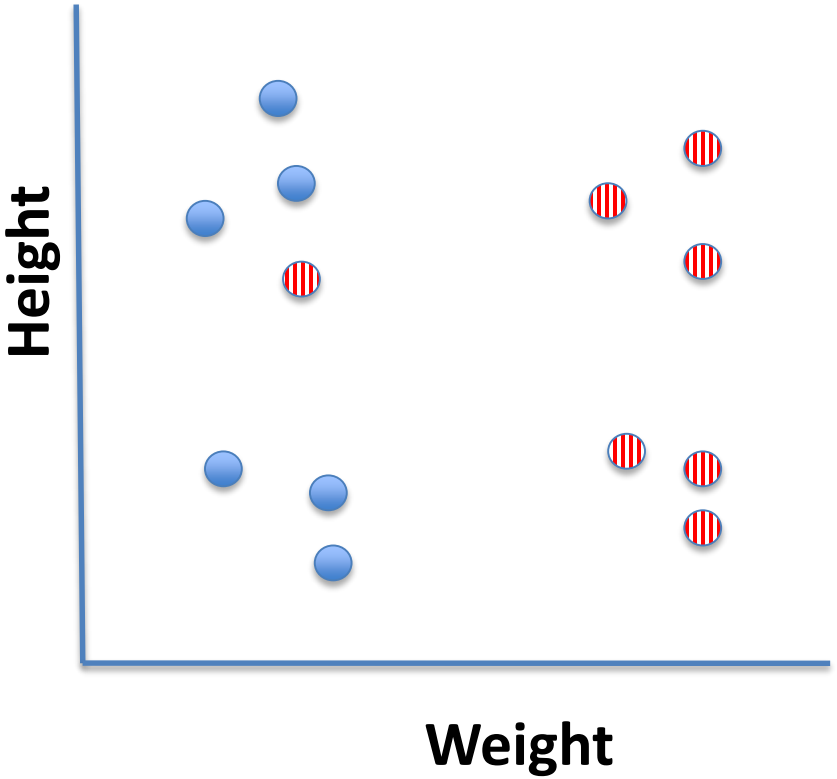


What Makes a Clustering Good?

Depends upon the application

Political candidates





Like All ML, It's and Optimization Problem

Like All ML, It's and Optimization Problem

Need an objective function

- Low intra-cluster dissimilarity
- High inter-cluster dissimilarity

Intra-cluster Dissimilarity

$$V(c) = \sum_{x \in c} (mean(c) - x)^2$$

$$badness(C) = \sum_{c \in C} V(c)$$

Are We Done?

Sufficient to find a set of clusters, C, such that badness(C) is minimized?

Suppose each example is in a cluster of size 1?

badness(C) = ?

What do we need?

Need a Constraint

Maximum distance between clusters is D

The maximum number of clusters is k

A Classic Formulation of Optimization

An objective function and a constraint

Like many optimization problems, computationally nasty

Usually rely on a greedy approximation

- K-means
- Hierarchical