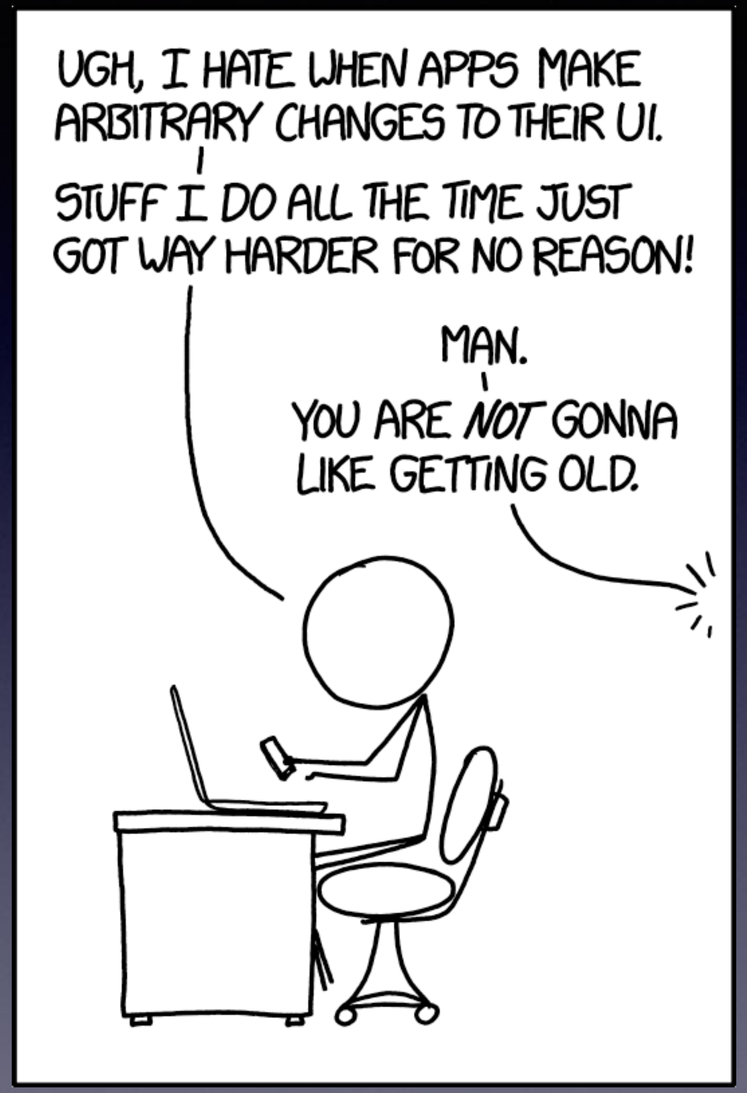
Hierarchical Clustering

Brendan Herger

Slides: https://goo.gl/HhdbmP

Goals

- **References:** Know where to look to dive deeper
- **Use cases:** Know when Hierarchical Clustering is a good fit
- **Proficiency:** Know how to utilize Hierarchical Clustering, with good defaults



https://xkcd.com/1770/

Hierarchical clustering creates groups of similar observations

Intro
Algorithm
Gotchas
Recap

Intro

Why clustering

- We want to understand subgroups, or prescribe different actions for subgroups
  - Unsure of groups (e.g. movie preferences)
  - Labels exist, are not available (e.g. gender isn’t available)

Common use cases

- **Segmentation:** Tailor marketing to each cluster
- **Proxy:** Treat clusters as difficult to observe variables (e.g. gender)
- **Ensemble models:** Train one model per cluster

Algorithm

Algorithm

Merge the 2 most similar clusters,
until 1 cluster remains

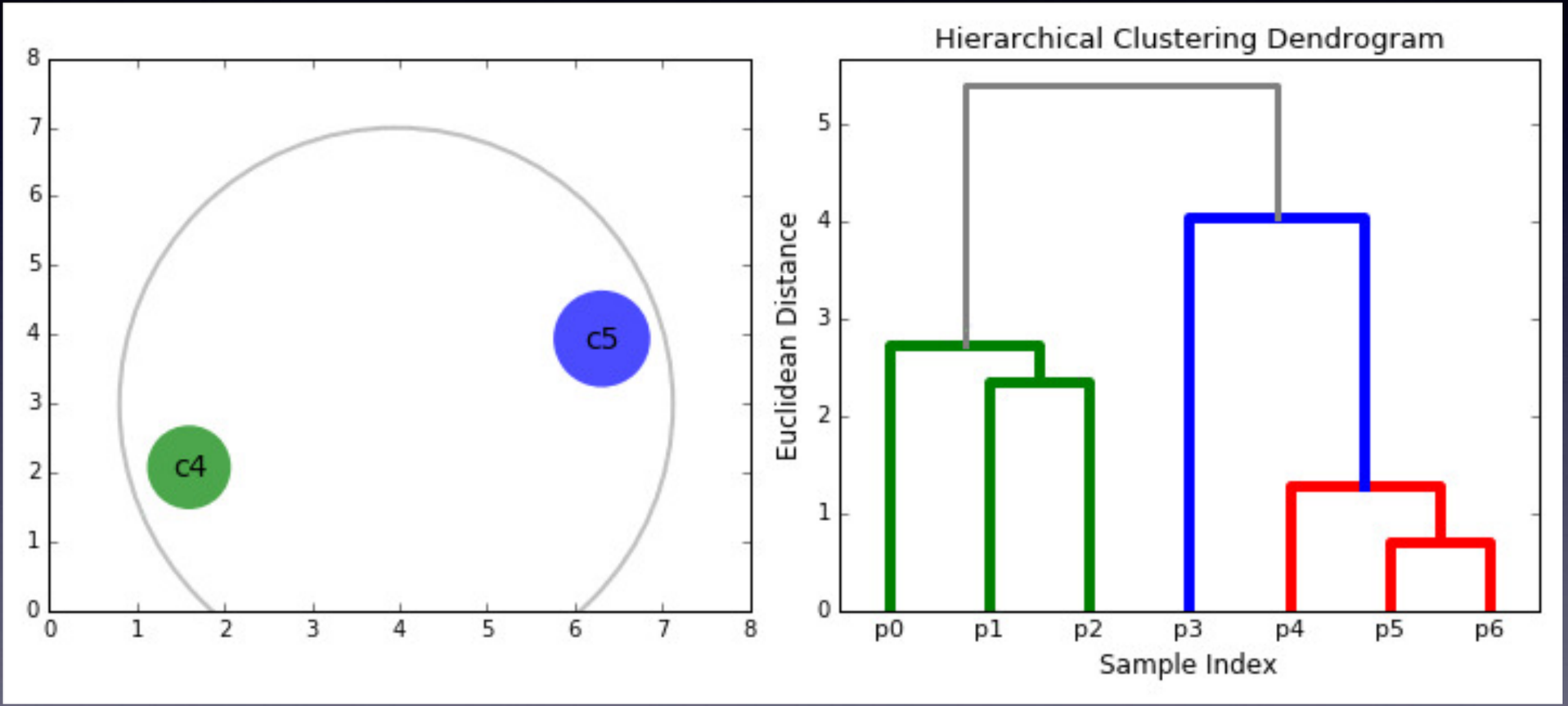
Algorithm

Merge the 2 most similar clusters, until 1 cluster remains

1. Treat every observation as a cluster of size 1
2. Until there is only one cluster:
    a) Measure distance between all current clusters
    b) Merge clusters with smallest distance

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf, page 395

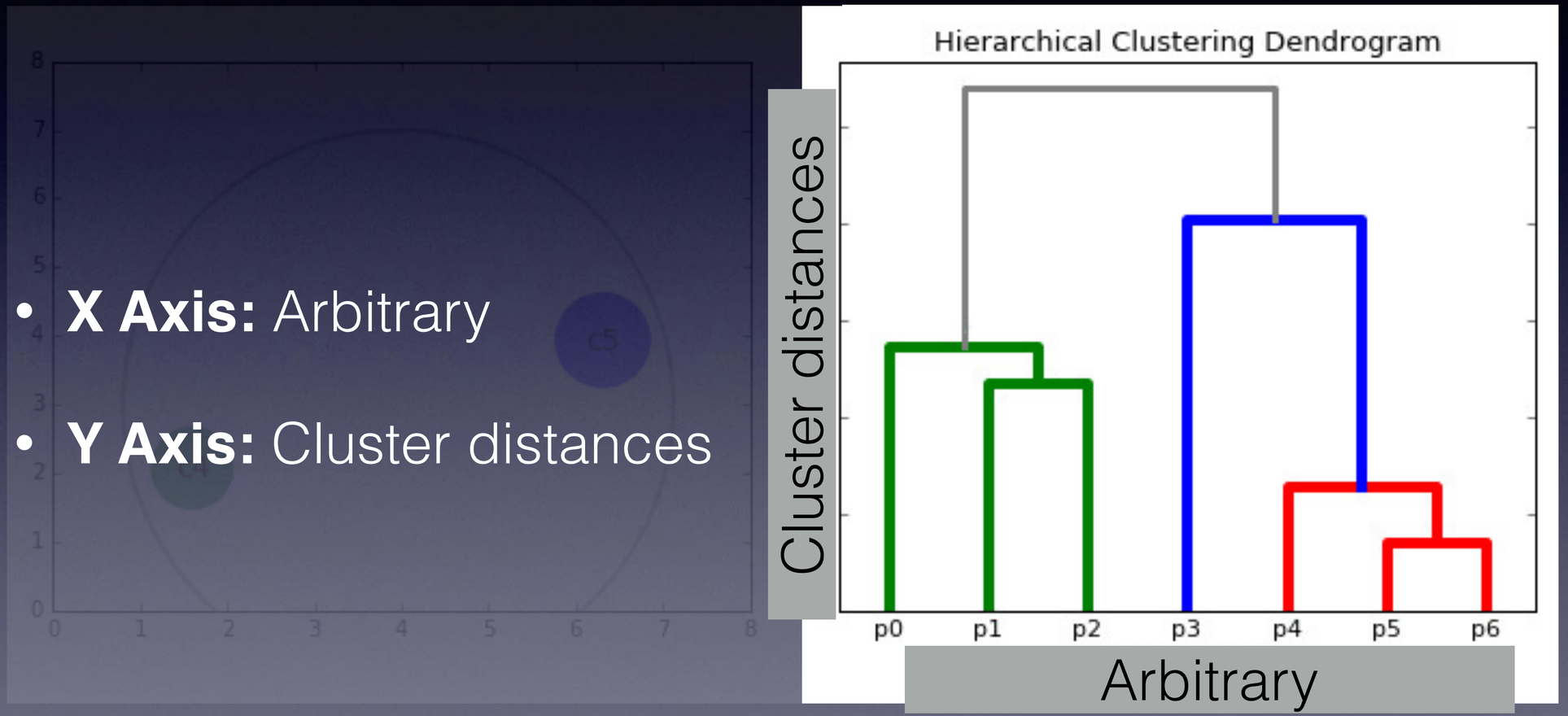
Dendrogram



https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Dendrogram

- **X Axis:** Arbitrary
- **Y Axis:** Cluster distances



https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Gotchas

Linkage (Cluster Distances)

Linkage	Comparator	Description
Complete	Max	Max distance between observations in clusters
Average	Mean	Mean distance between observations in clusters
Single	Min	Minimum distance between observations in clusters
Centroid	N/A	Compute centroid for each cluster, compare centroids
Ward	Sum sq. diff	Merge two clusters which produce the smallest increase of within-cluster variance

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf, page 395
http://homes.mpimf-heidelberg.mpg.de/~mhelmsta/pdf/1963%20Ward%20JASA.pdf

Similarity metrics (Observation Distances)

- **Euclidean:** Standard distance metric
- **Haversine distance:** Geographic data (latitude & longitude)
- **Many, many others**

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf, page 396
http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html

Data types

(All should be normalized)

- **Real valued:** Works well
- **Boolean:** Non-ideal
- **Categorical:** Can be converted to boolean or converted to embedding (real valued vector)
- **Datetime:** Can compare time deltas, by converting to epoch

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf, page 397
https://arxiv.org/abs/1604.06737

Number of clusters

- Highly heuristic
- Generally balances:
  - Number of clusters
  - Distance between merged clusters (Dendrogram y axis)
  - Qualities of clusters

Known issues

- **Robustness:** Model has no default for unseen observations
- **Stability:** Minor changes in the data set can cause large shifts in clusters

Recap

Recap

- Hierarchical clustering creates groups of similar observations
- It is helpful if we want to understand subgroups, or prescribe different actions for subgroups
- Hierarchical clustering creates groups by merging the 2 most similar clusters, until 1 cluster remains
- Common parameters include linkage, distance metric, and number of clusters

Thanks!

Brendan Herger

Slides: https://goo.gl/HhdbmP