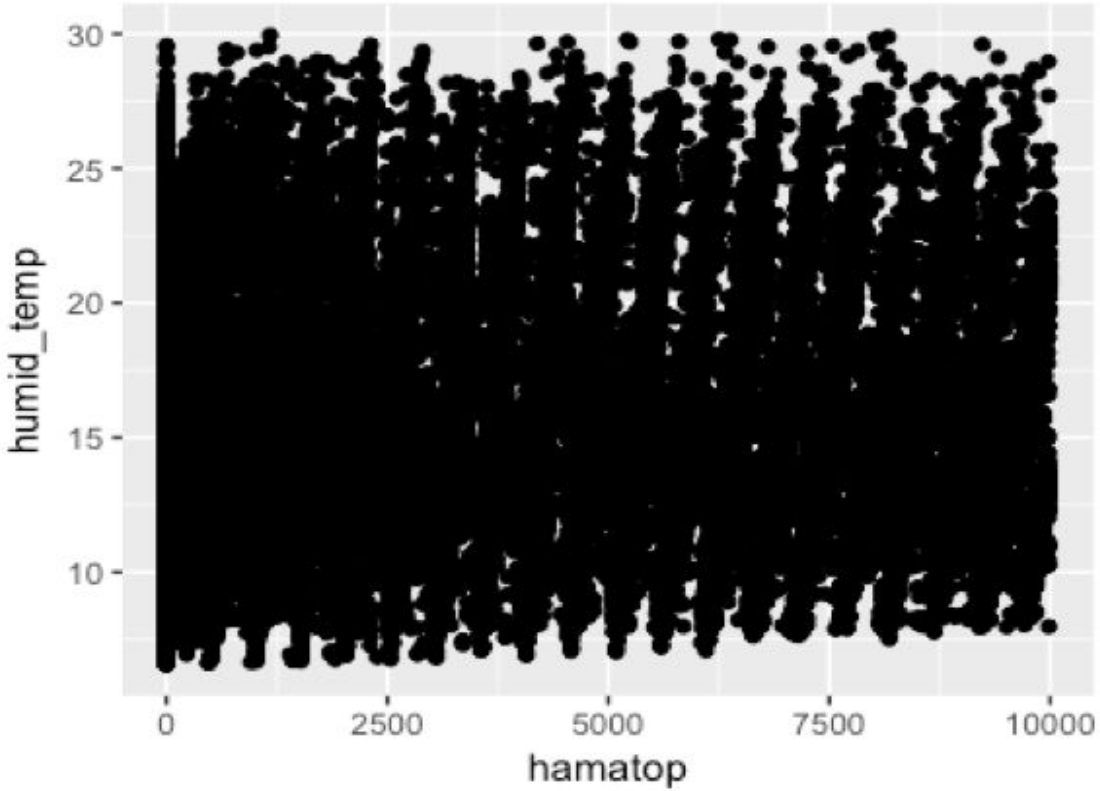
STAT 215A Fall 2017
Week 5

Rebecca Barter
09/22/2017

Peer review

Image source: http://bit.ly/x2pms8

Overplotting



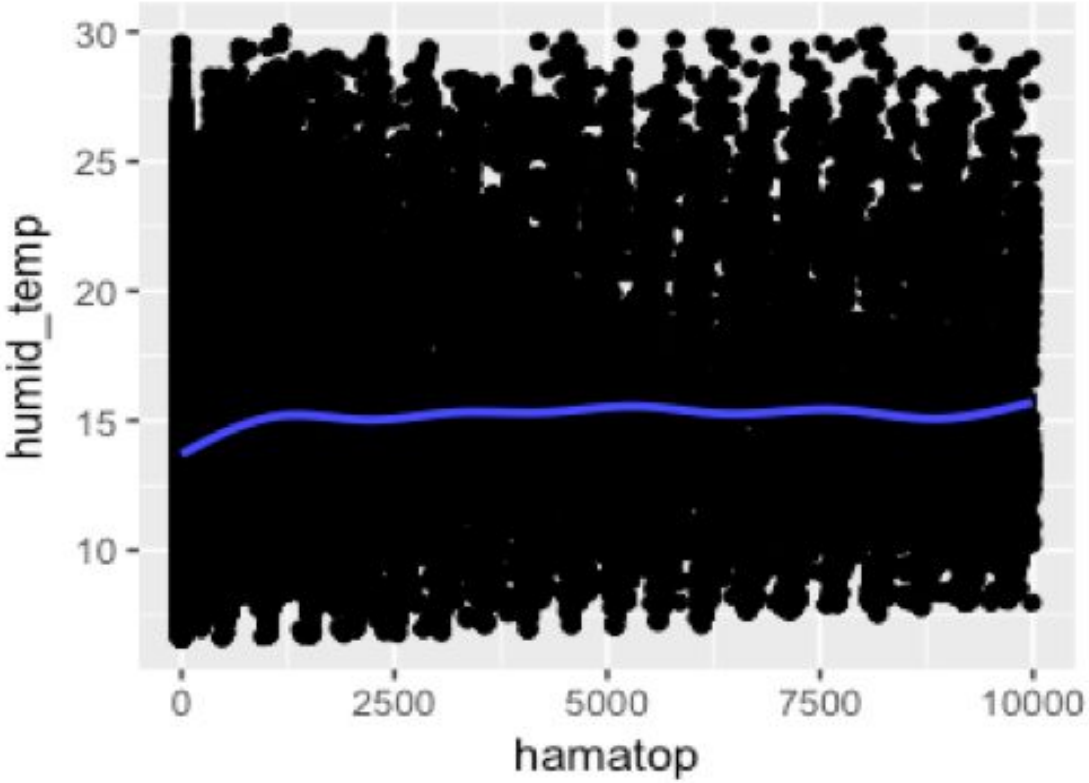
Overplotting

Some report file sizes were so large that I couldn't open them!

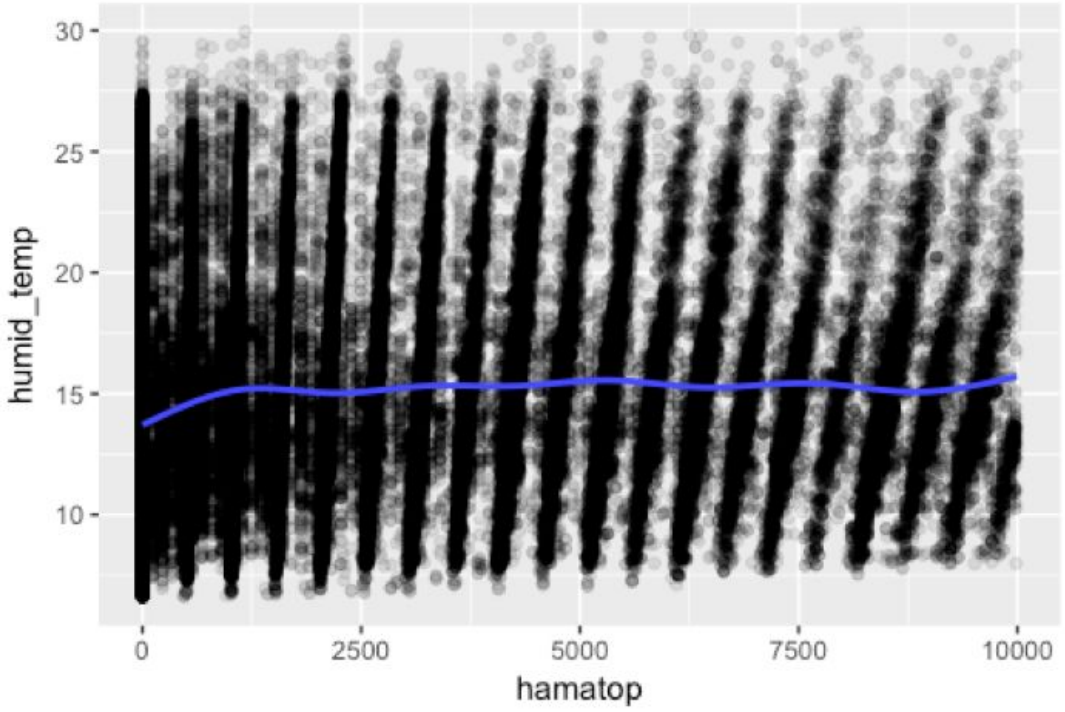
You need to make sure that your figures are rendered as png rather than pdf...

```
<<dev = "png", dpi = 300>>=
ggplot(big_df) + …
@
```

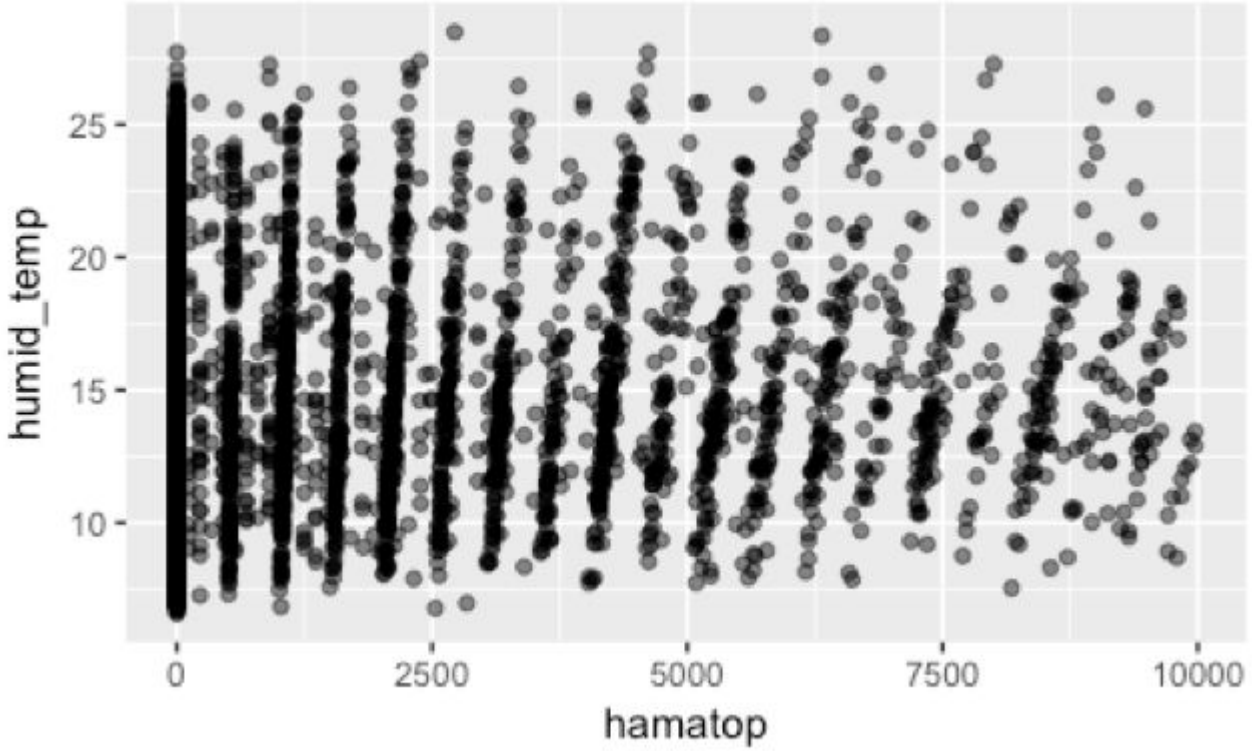
Overplotting: **add a trendline?**



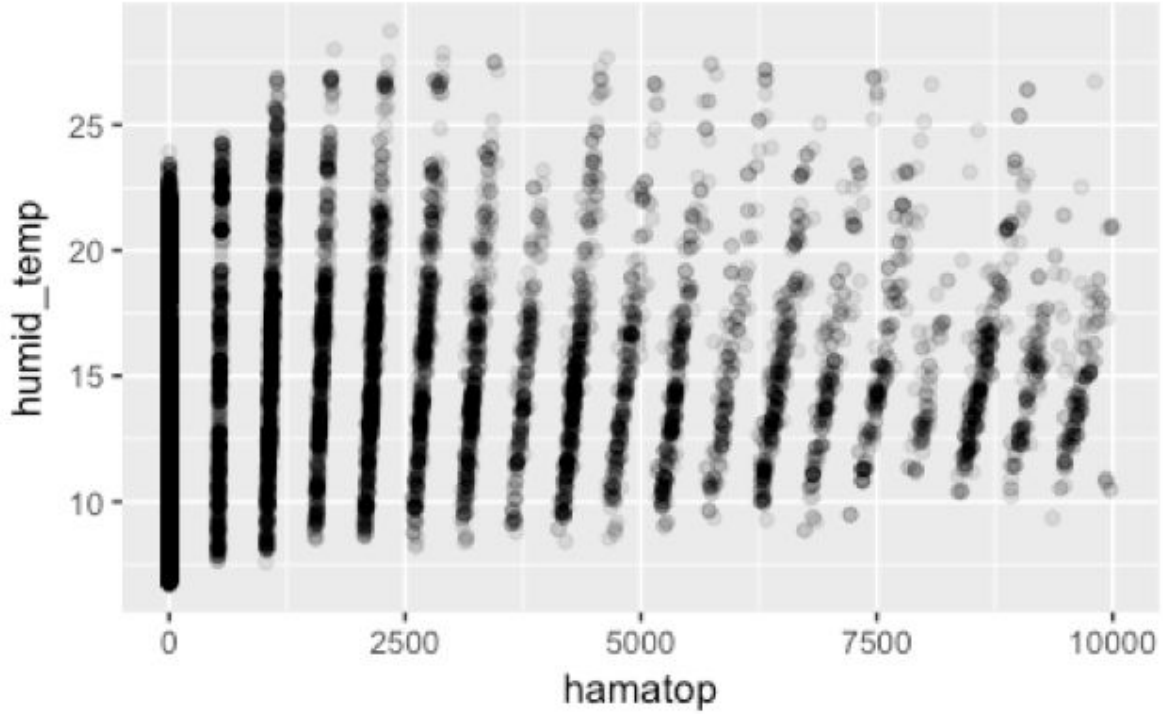
Overplotting: **add transparency?**



Overplotting: **subsampling 10,000 points?**

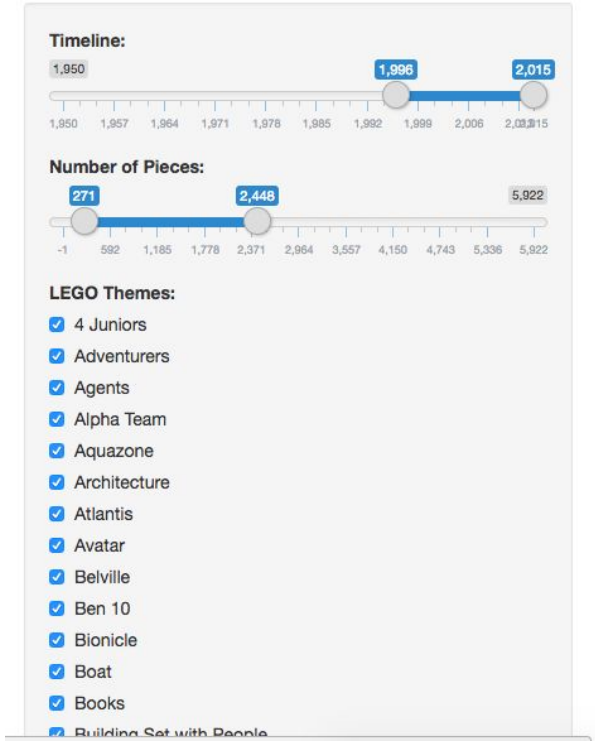


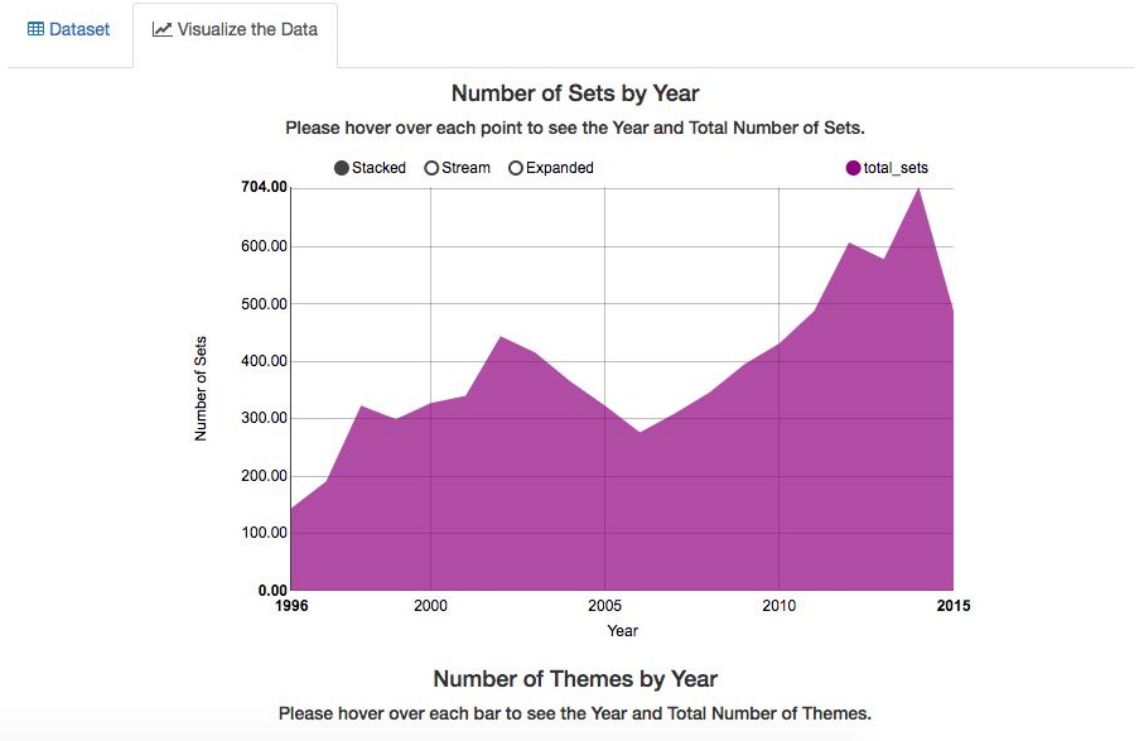
Overplotting: **meaningful subsampling** (plotting only a single node)?



Interactive plotting: shiny apps

https://shiny.rstudio.com/gallery/lego-set.html

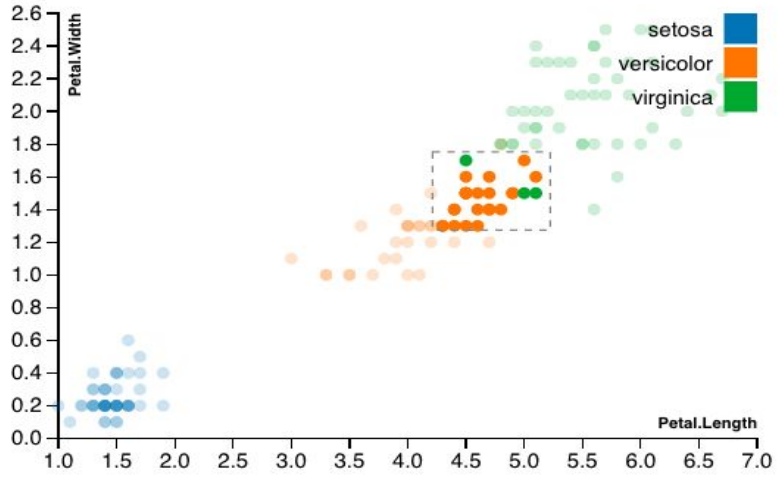


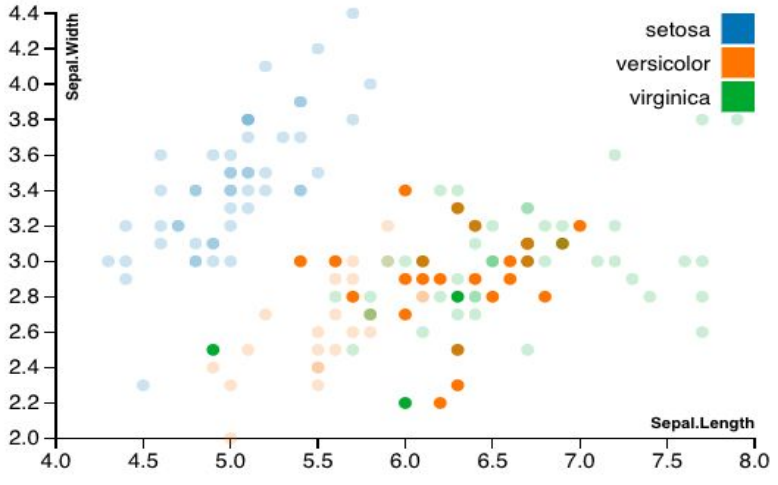


Interactive plotting: linked brushing via crosstalk

https://rstudio.github.io/crosstalk/using.html

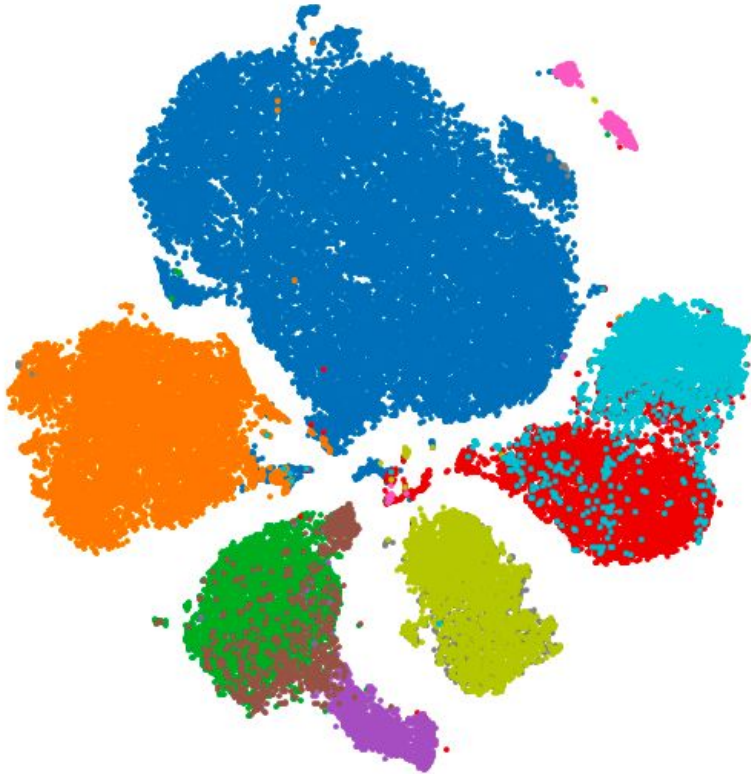
```
library(crosstalk)

shared_iris <- SharedData$new(iris)
bscols(
  d3scatter(shared_iris, ~Petal.Length, ~Petal.Width, ~Species, width="100%", height=300),
  d3scatter(shared_iris, ~Sepal.Length, ~Sepal.Width, ~Species, width="100%", height=300)
)
```





Clustering

K-means
Hierarchical clustering
Spectral clustering
...



Silhouette plots

A measure of the separation between clusters

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

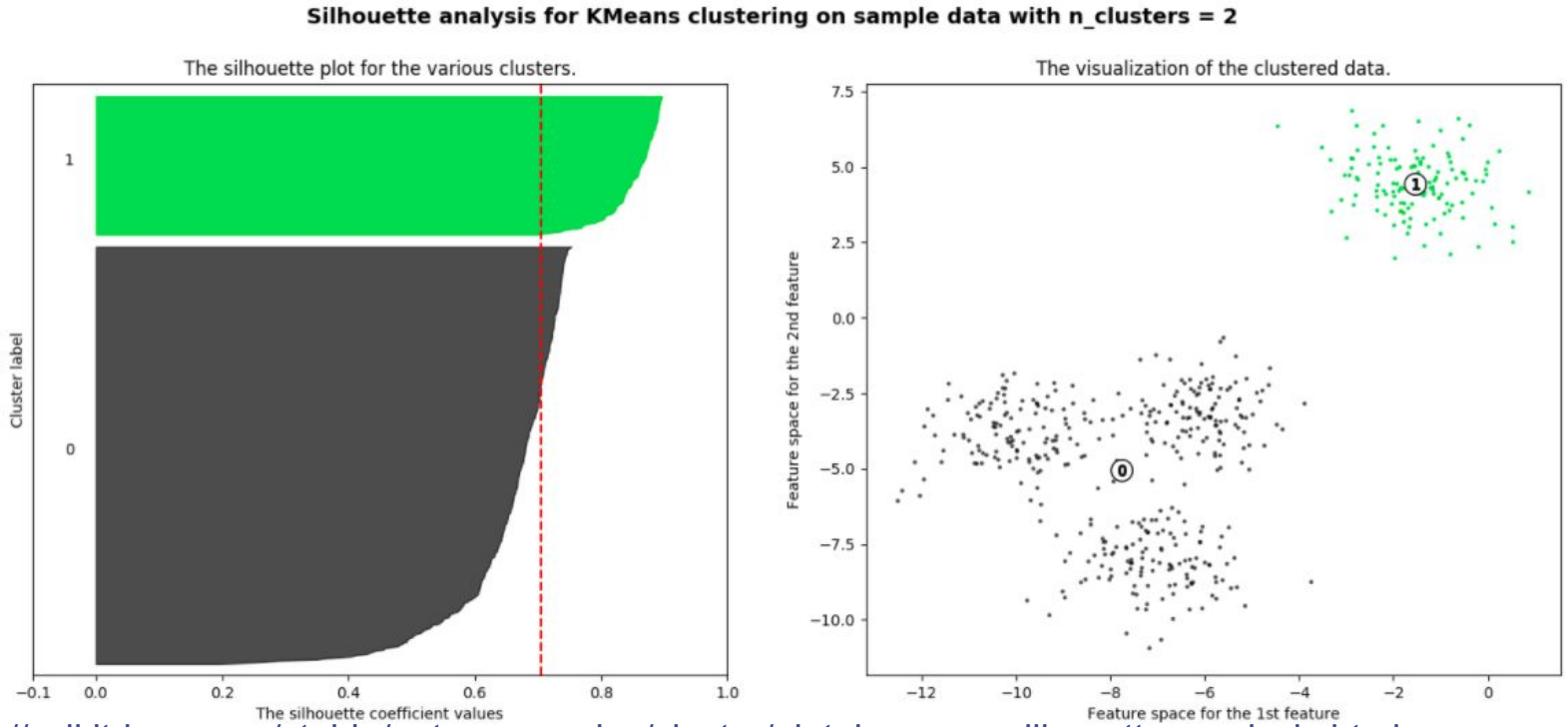
average dissimilarity of data point *i* with all other data within the **same cluster**

lowest average dissimilarity of data point *i* to any **other cluster**

https://en.wikipedia.org/wiki/Silhouette_(clustering)

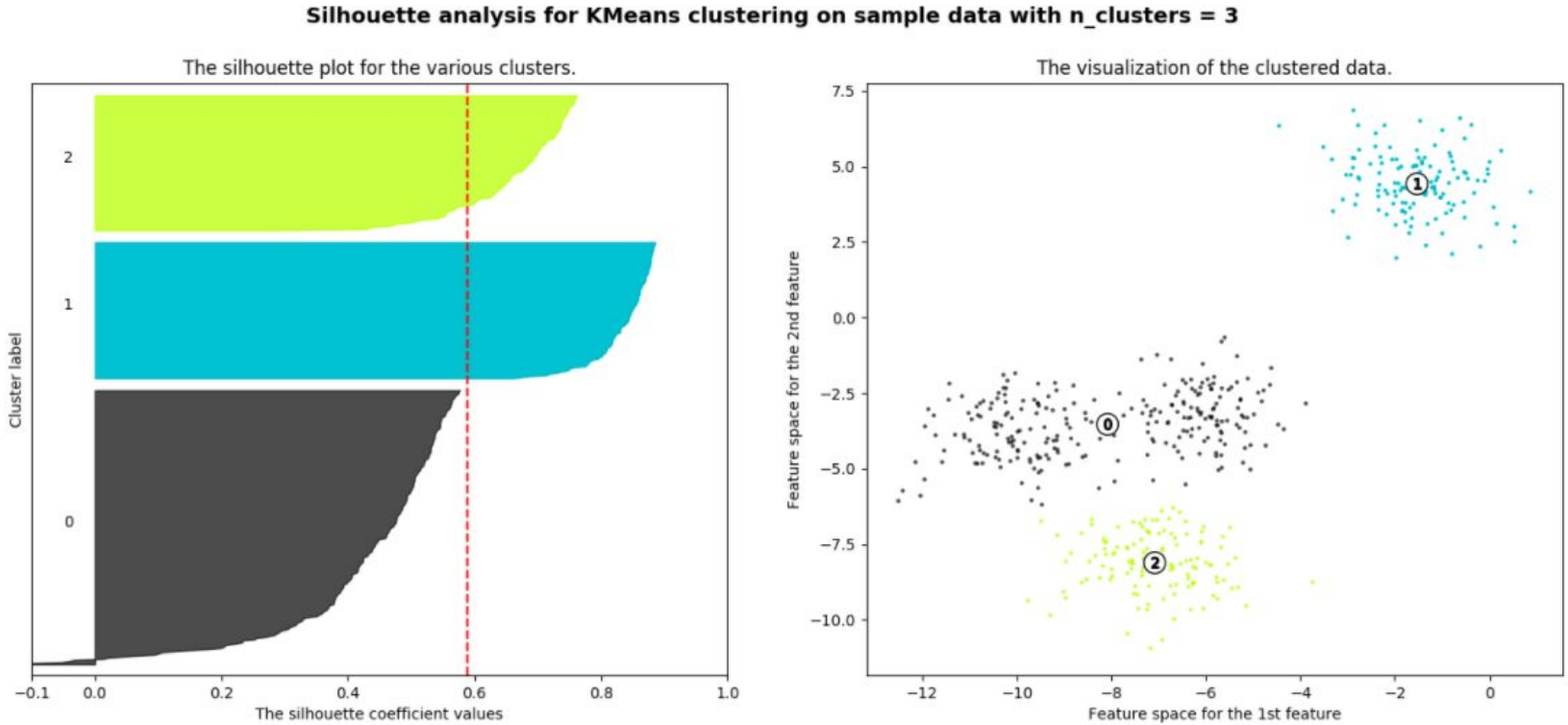
Silhouette plots (**average sil = 0.70**)

Plot silhouette widths in decreasing order, grouped by cluster



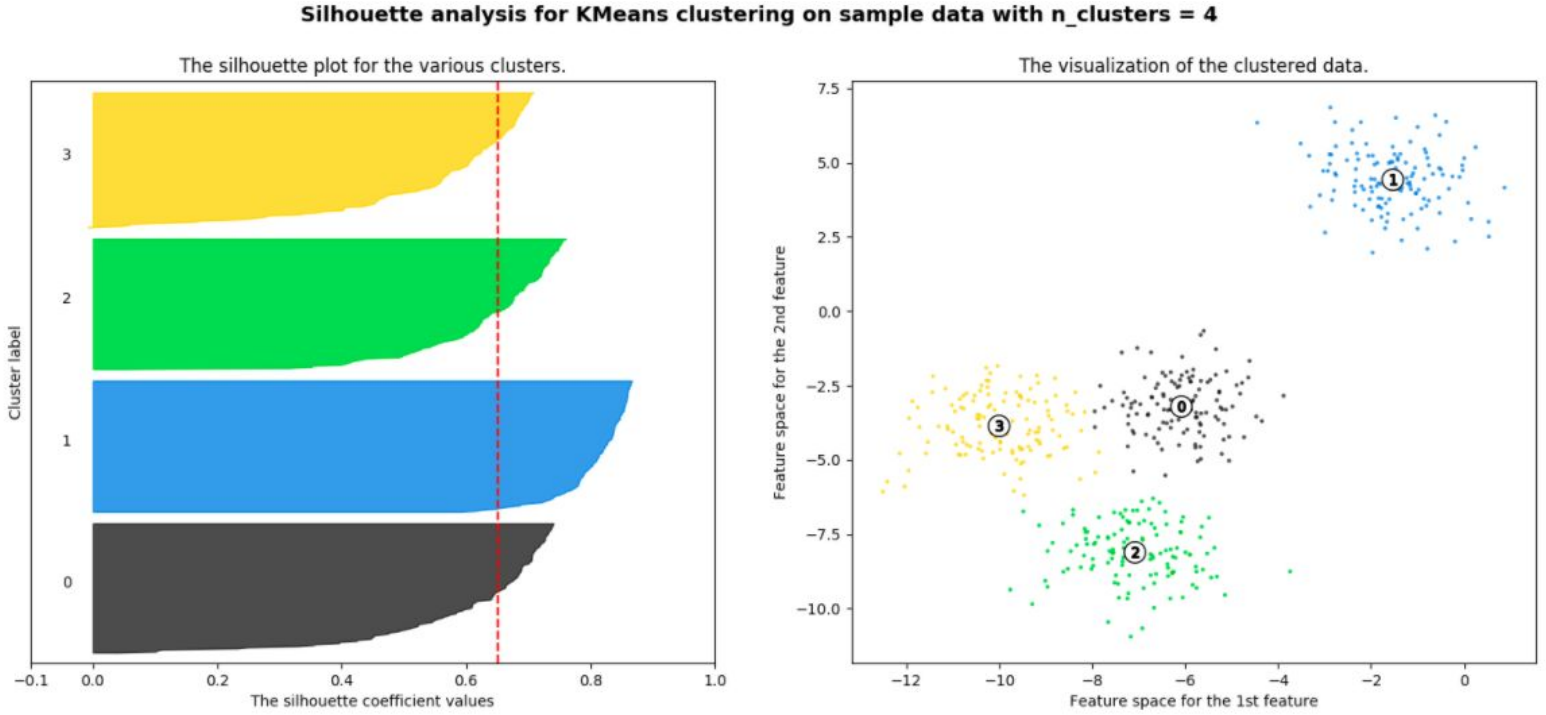
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots (**average sil = 0.59**)



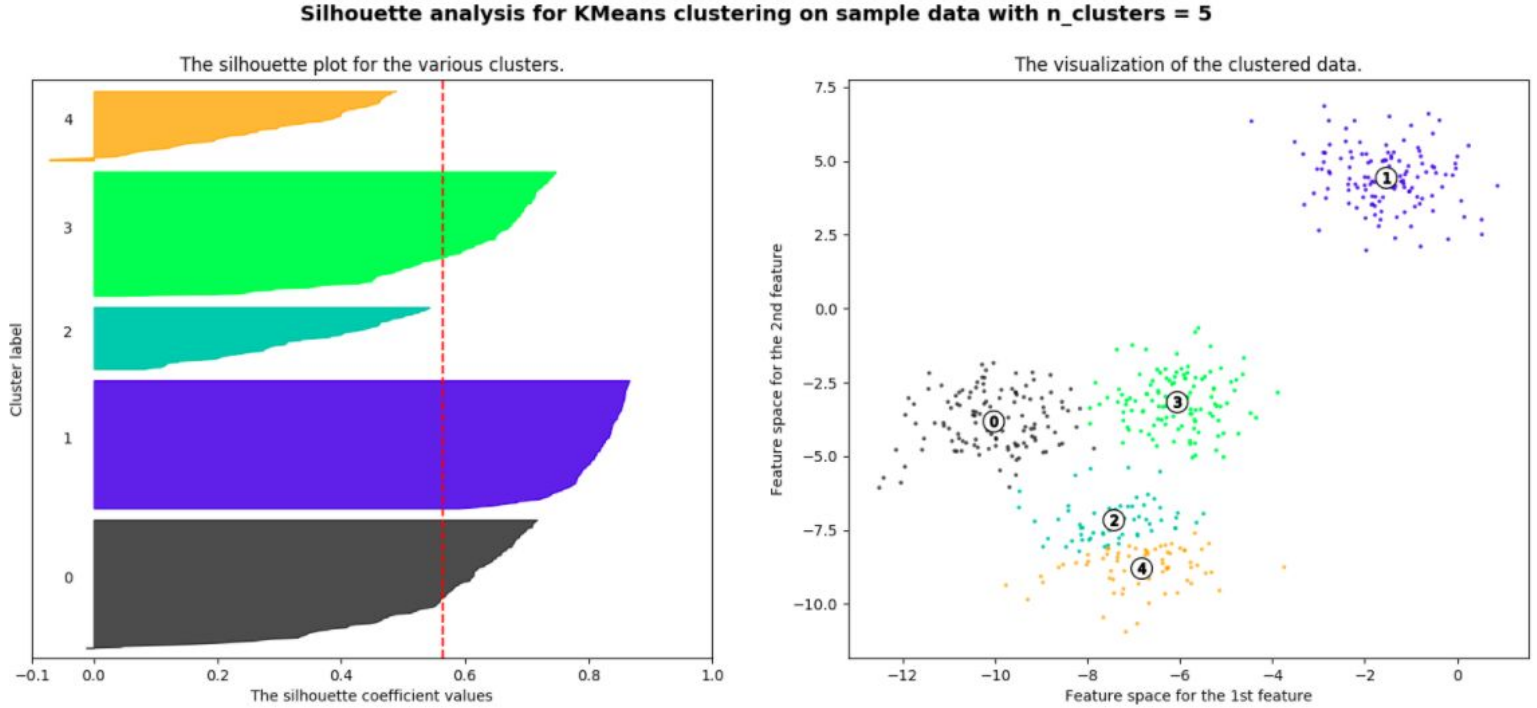
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots (**average sil = 0.65**)



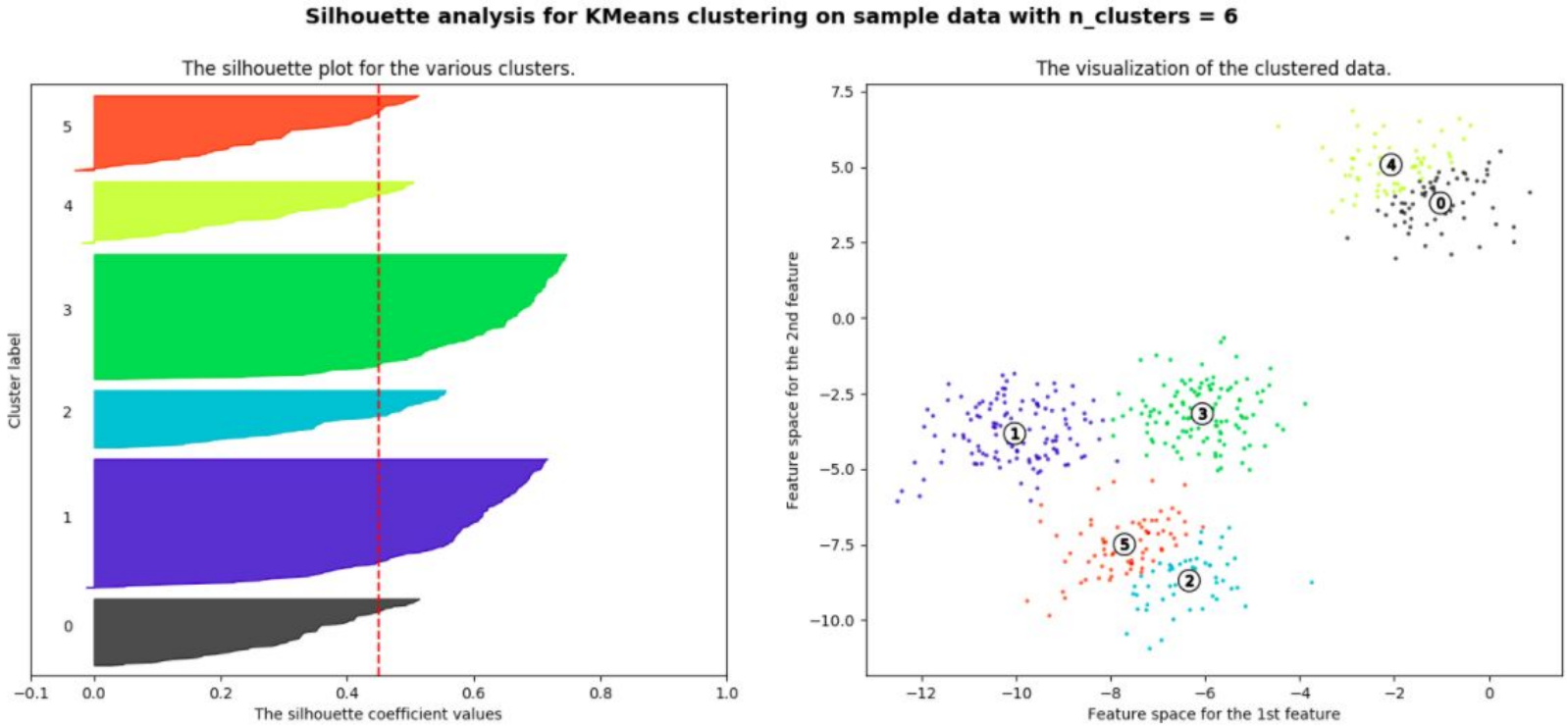
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots (**average sil = 0.56**)



http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots (**average sil = 0.45**)



http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette plots (**k = 2 is the best!**)

```
For n_clusters = 2 The average silhouette_score is : 0.704978749608
For n_clusters = 3 The average silhouette_score is : 0.588200401213
For n_clusters = 4 The average silhouette_score is : 0.650518663273
For n_clusters = 5 The average silhouette_score is : 0.563764690262
For n_clusters = 6 The average silhouette_score is : 0.450466629437
```

http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Clustering exercises

1. Load wine.csv (14 characteristics of 178 wines from 3 different cultivars)
2. Plot the wines in the space defined by the first two principal components. Color each wine by its cultivar (type).

Clustering exercises

1. Load wine.csv (14 characteristics of 178 wines from 3 different cultivars)
2. Plot the wines in the space defined by the first two principal components. Color each wine by its cultivar.
3. **Run k-means with 3 cluster centers using all variables (except cultivar). Color each point in your previous plot by cluster.**

Clustering exercises

1. Load wine.csv (14 characteristics of 178 wines from 3 different cultivars)
2. Plot the wines in the space defined by the first two principal components. Color each wine by its cultivar.
3. Run k-means with 3 cluster centers using all variables (except cultivar). Color each point in your previous plot by cluster.
4. **Run k-means using the first two principal components only. Color each point in your plot by cluster. Compare the spectral clustering to the standard k-means clustering.**

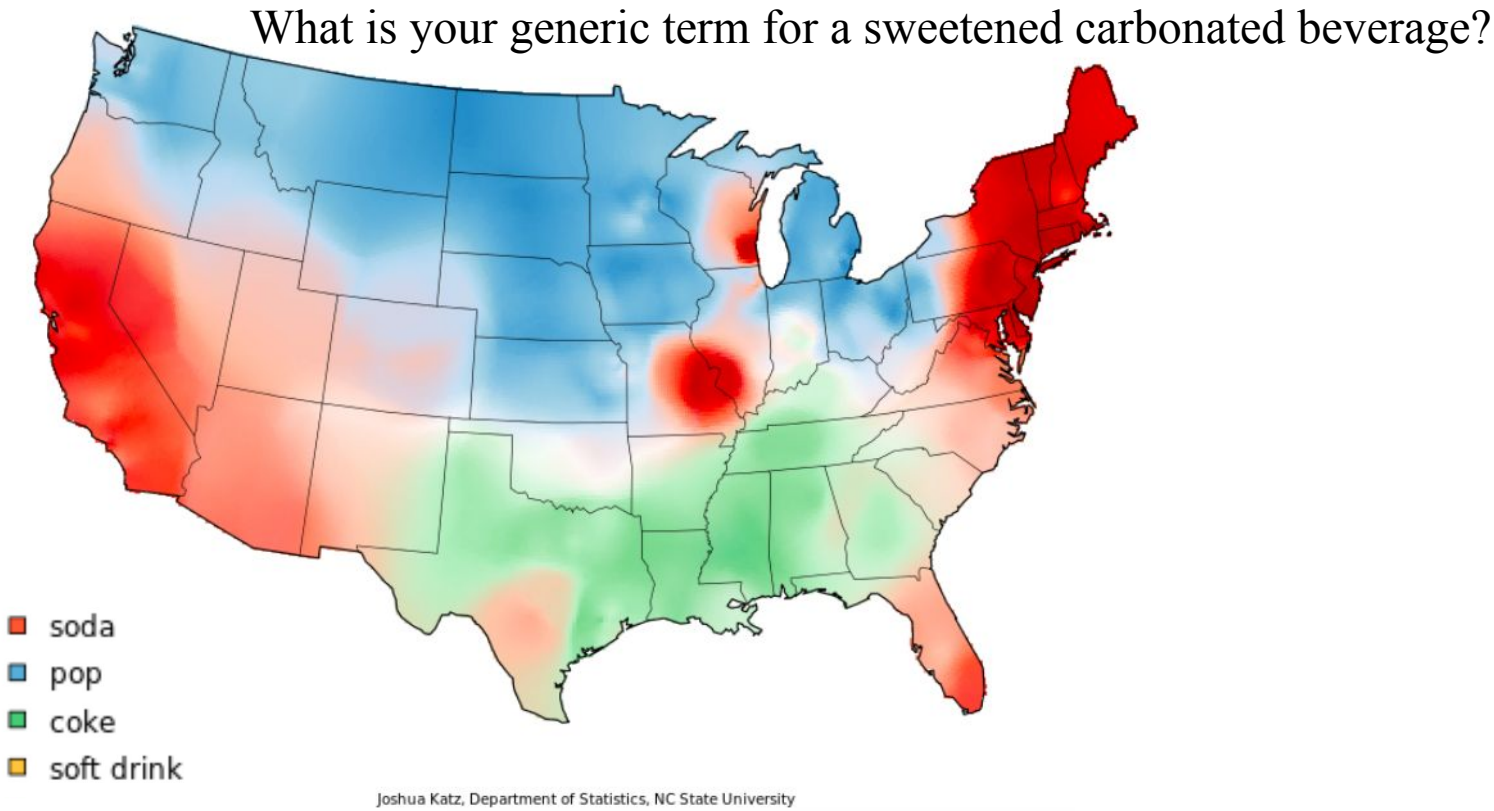
Clustering exercises

1. Load wine.csv (14 characteristics of 178 wines from 3 different cultivars)
2. Plot the wines in the space defined by the first two principal components. Color each wine by its cultivar.
3. Run k-means with 3 cluster centers using all variables (except cultivar). Color each point in your previous plot by cluster.
4. Run k-means using the first two principal components only. Color each point in your plot by cluster. Compare the spectral clustering to the standard k-means clustering.
5. **Re-run steps 3 and 4 each four times. Do the results change?**

Clustering exercises

1. Load wine.csv (14 characteristics of 178 wines from 3 different cultivars)
2. Plot the wines in the space defined by the first two principal components. Color each wine by its cultivar.
3. Run k-means with 3 cluster centers using all variables (except cultivar). Color each point in your previous plot by cluster.
4. Run k-means using the first two principal components only. Color each point in your plot by cluster. Compare the spectral clustering to the standard k-means clustering.
5. Re-run steps 3 and 4 each four times. Do the results change?
6. **Re-run steps 3-5 with 10 cluster centers. Compare silhouette plots.**

Introducing Lab 2



http://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6