STAT 215A Fall 2019
Week 5

Tiffany Tang

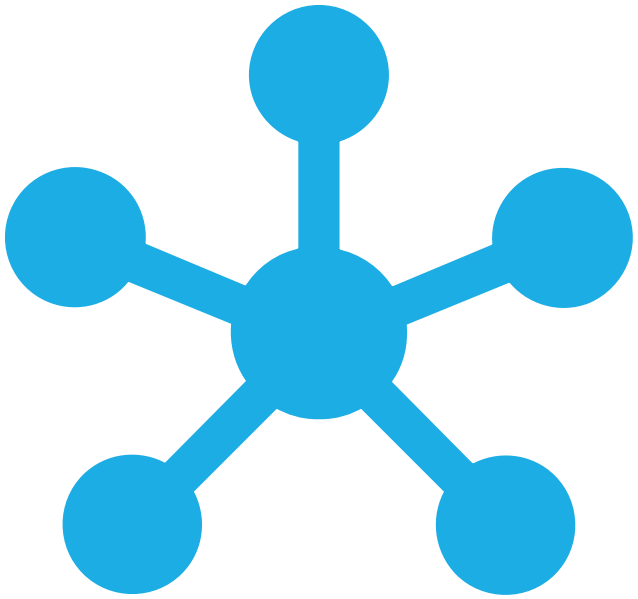
9/27/19

Announcements

- Peer reviews **due today**
- Lab 2 + Homework 2 will be released today
  - Due in two weeks: **October 10 11:59pm**

Plan for Today:

- K-means
- Hierarchical Clustering
- NMF
- Centering and Scaling?
- Discuss Lab 1 and introduce Lab 2



Clustering

k-Means (with l_2 distance)

Idea: find clusters C which minimize the within-cluster sum of squares

$$\underset{C}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2, \quad \text{where} \quad \boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

k-Means (with l_2 distance)

Advantages

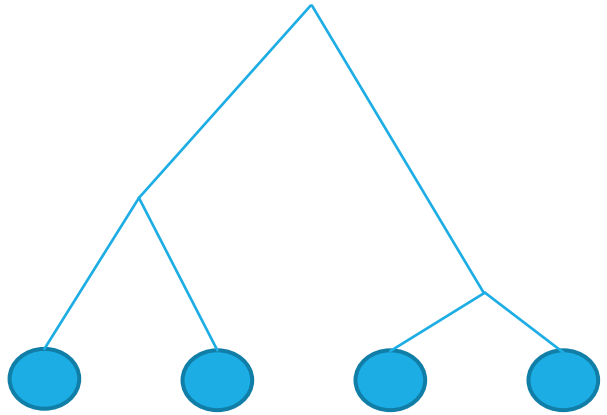
- Fast
- Good when clusters are spherical balls and linearly separable

Disadvantages

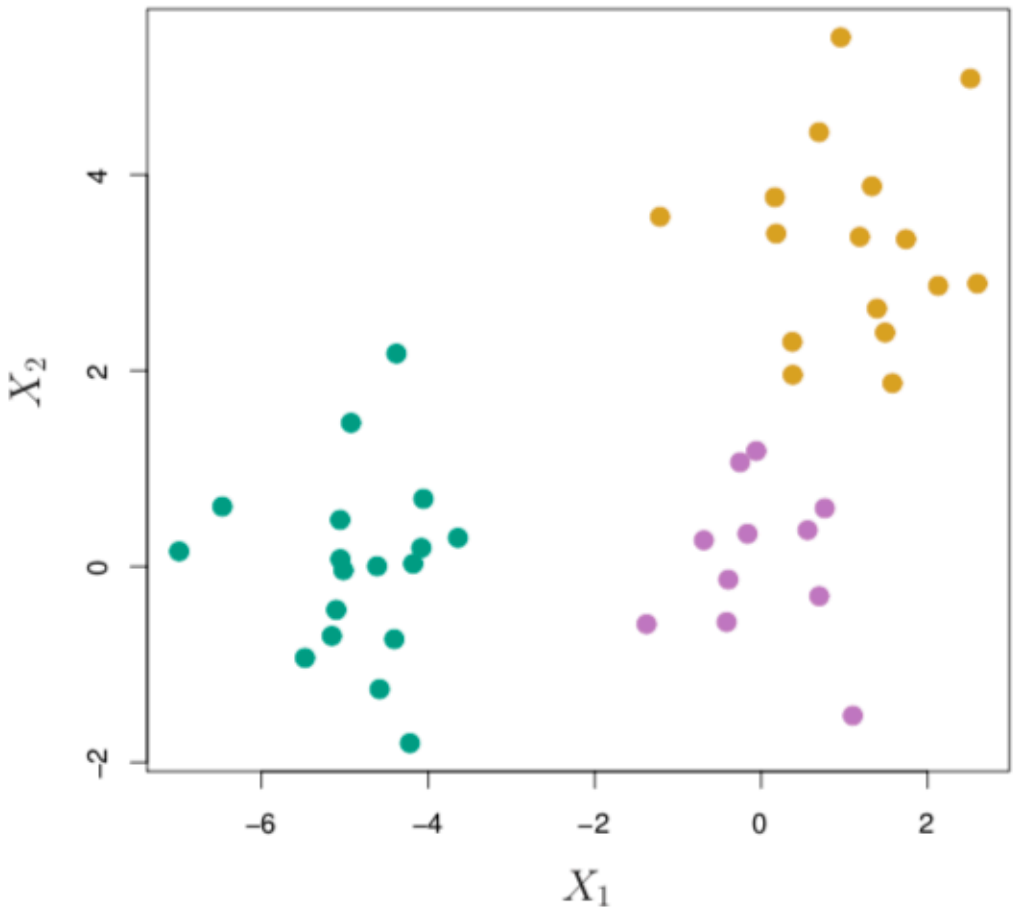
- Bad when clusters not spherical
- Bad when clusters have different variances
- p >> n (“Curse of Dimensionality”)
- Irrelevant variables are treated as equals with relevant ones
- Algorithmic solution depends on initialization
- How to choose k?
  - Stability
  - Silhouette statistic: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

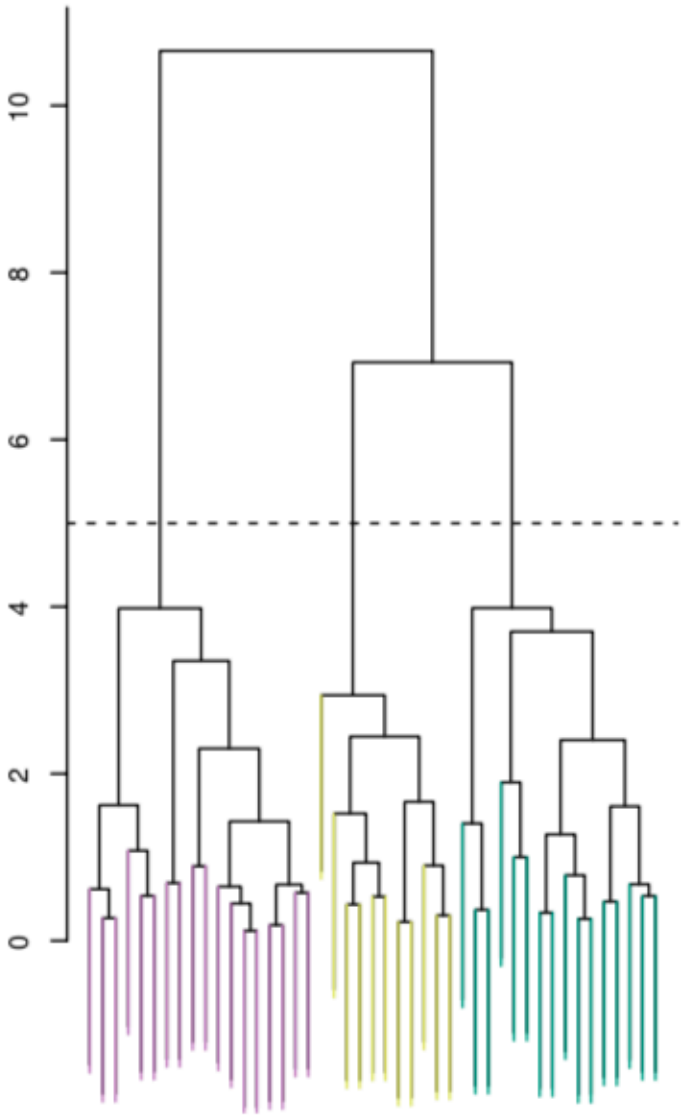
Hierarchical Clustering

- Gives family of nested clusterings, presented as a tree
- A greedy, agglomerative algorithm, not based upon an optimization problem
- At the lowest level, each cluster contains a single observation
- As we move up the tree, some leaves begin to fuse into branches – these are observations that are similar to each other.
- The lower in the tree fusions occur, the more similar the groups of observations are to each other
- At the highest level, there is only one cluster containing all observations



Interpreting Dendrograms

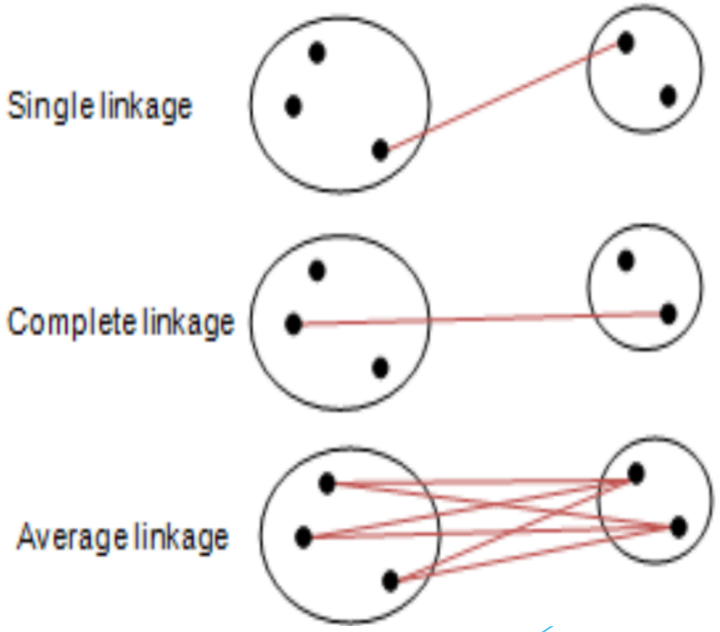




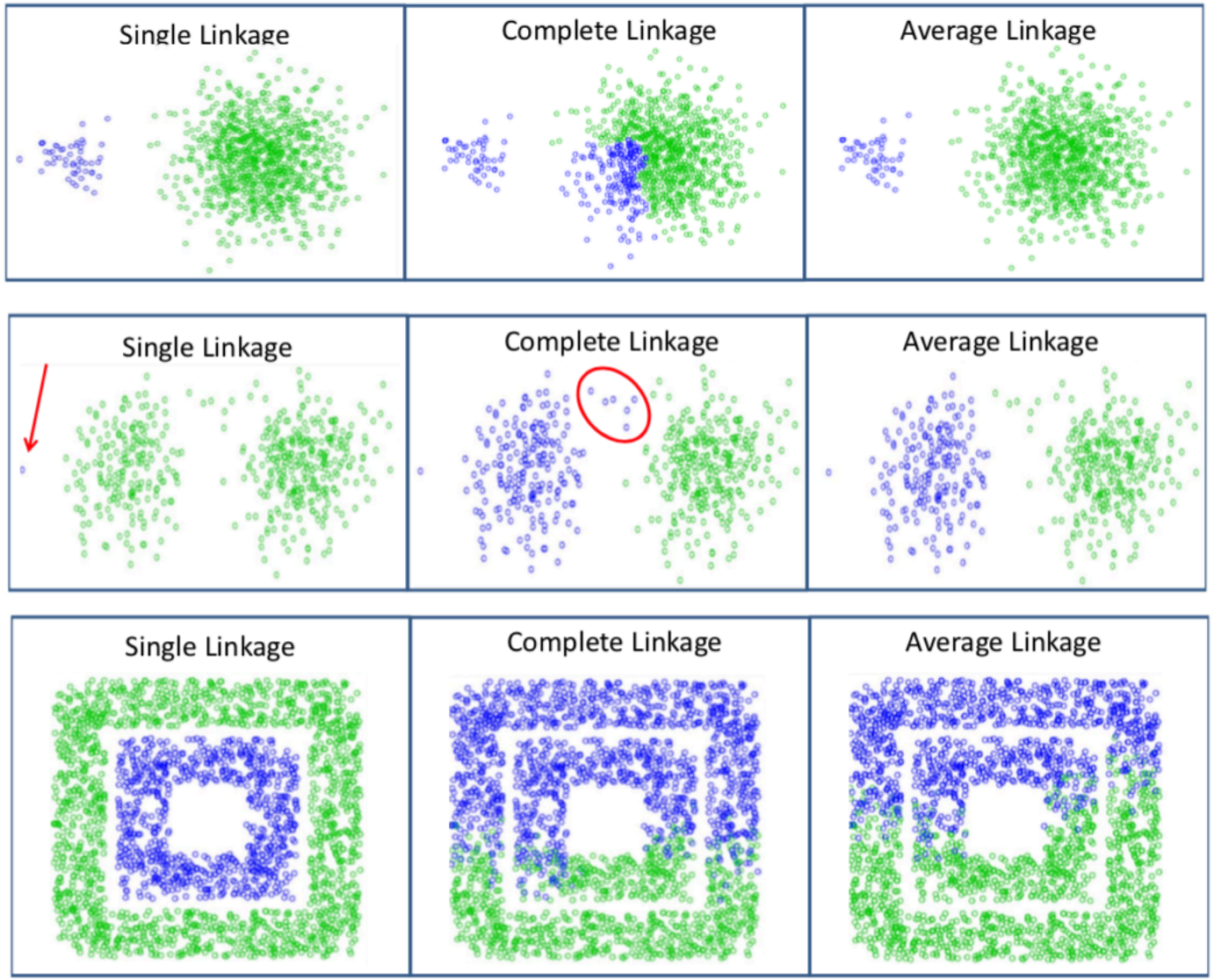
How to join clusters/observations

1. **Distance metric:** a measure of dissimilarity between two observations $d(x, y)$
   - *Examples:* l_2, l_1, any of your favorite norms, $1 - cor(x, y)$

2. **Linkage metric:** rule for joining two clusters
   - Single Linkage
   - Complete Linkage
   - Average Linkage
   - Ward's Linkage



Linkage Examples



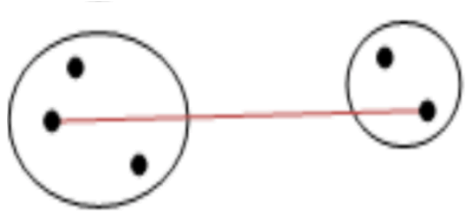
Linkages

Single Linkage (min)

- Can handle diverse shapes
- Very sensitive to outliers or noise
- Often results in unbalanced clusters
- Extended, trailing clusters in which observations fused one at a time – chaining

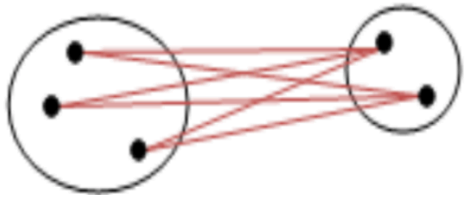
Complete Linkage (max)

- Often gives cluster with similar sizes
- Less sensitive to outliers
- Works better with spherical distributions



Average Linkage

- A compromise between single and complete linkage
- Less sensitive to outliers than complete linkage, but not as robust as single linkage
- Works better with spherical distributions



- **Ward's Linkage:** join sets that minimize the Euclidean distance between all pairs of points
- Average and Ward's linkages are most widely used

Hierarchical Clustering

Advantages

- Gives nested family of clusterings
- Convenient visualizations with dendrograms

Disadvantages

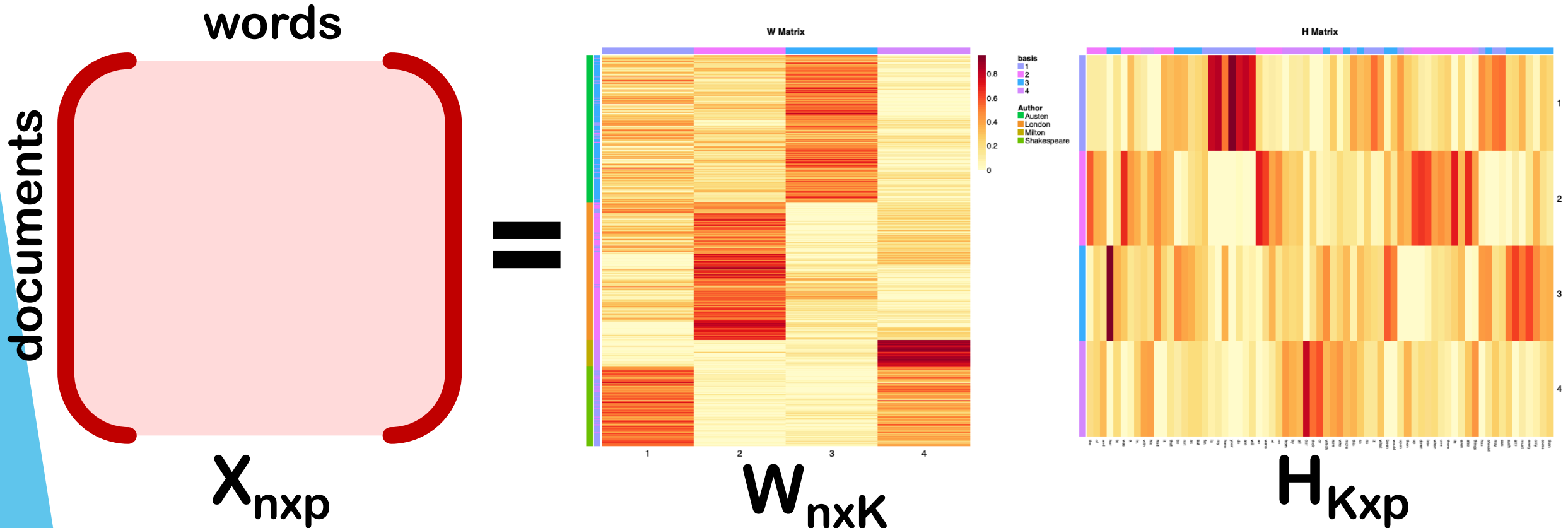
- Depends heavily on linkage
- p >> n (“Curse of Dimensionality”)
- Local solution

Nonnegative Matrix Factorizations (NMF)

- Given a non-negative matrix **X**, NMF solves

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} \|\mathbf{X} - \mathbf{W}\,\mathbf{H}\|_F^2$$

- Tool for dimension reduction, pattern recognition, and soft clustering with positive data



Nonnegative Matrix Factorizations (NMF)

Advantages

- Soft clustering
- Inherently gives sparse feature matrix H (unlike PCA)
- Great for pattern recognition with positive data

Disadvantages

- Not a convex problem → depends on initialization of algorithm
- Components are unordered and not nested
  - Change number of components K can give vastly different results
- Can't find *strength* of patterns as in PCA

To center/scale or not to center/scale…

- By **centering**, I mean subtracting the sample mean from each column in your data matrix so that the mean of each column/feature is 0
- By **scaling**, I mean dividing each column by some constant so that the 2-norm of each column/feature is 1

- Very subjective…
- If it is meaningful to compare the variance of different features in your data matrix, don't *need* to scale
  - Gene expression data
- If features are measured on different scales, definitely need to scale (and maybe center?)
  - Income and number of kids
- Centering may result in a loss in interpretability (e.g., with positive data)
- What most people do in practice: try both

Go to lab_week5/ folder and work in groups

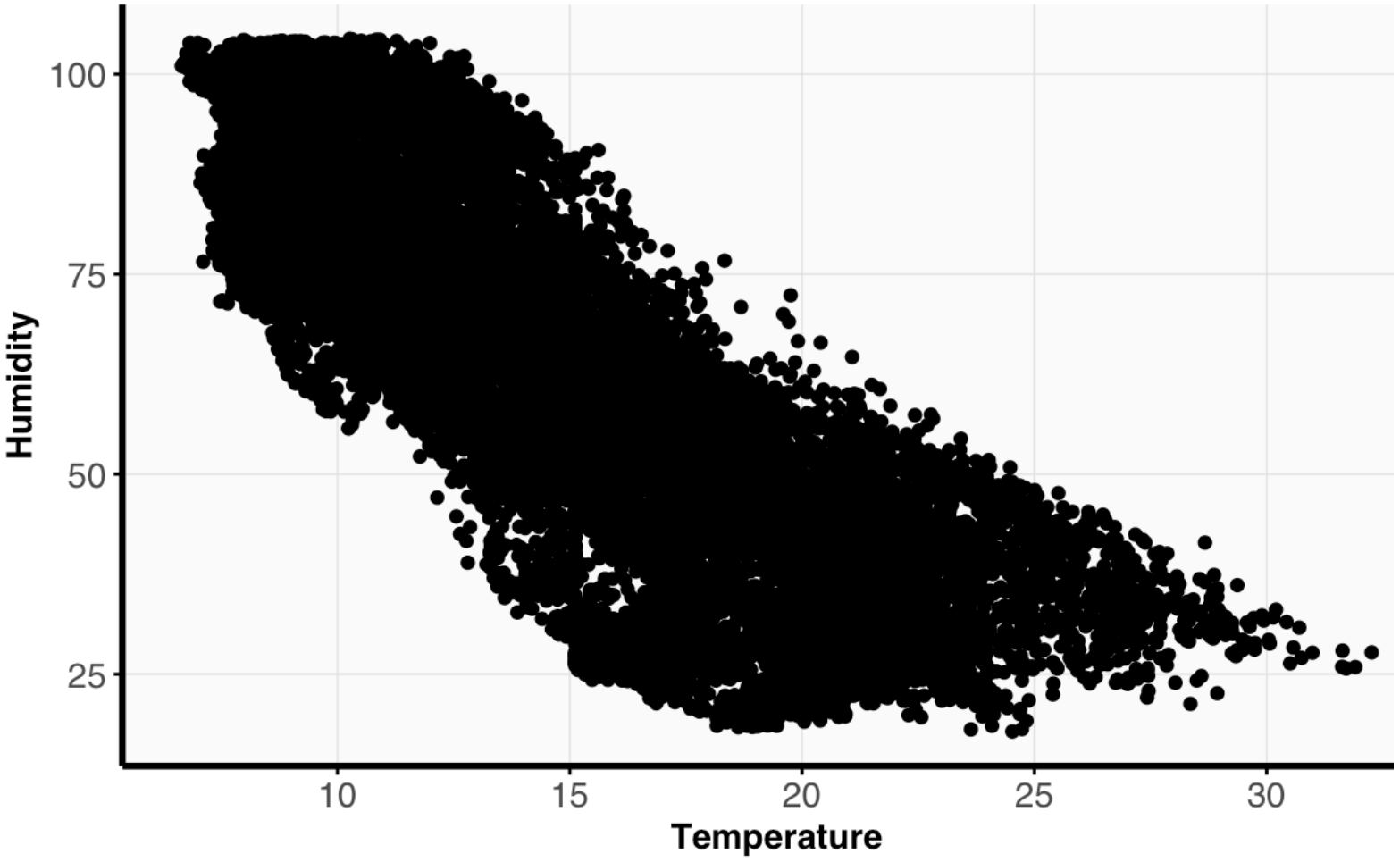
Lab 1
Redwood Trees

Any problems with the data?

Lab 1: Some tips and tricks

- Adding figure captions and labels using Rmd to enable cross-referencing
  - For example, see **week2/** folder: lab_gapminder_solutions.Rmd
- Change axis labels (e.g., capitalize titles, don't use "_")
- Spell check for rmd in Rstudio!
- When writing, refer to the English variable name instead of the R variable name
  - E.g., say temperature instead of humid_temp
- Code: use consistent spacing; see R Tidyverse style guide
  - Put spaces around = and other binary operators, e.g., +, -, &, >=
  - Put space after comma
- Don't push data/ folder and other files not needed to compile your report
  - .gitignore file

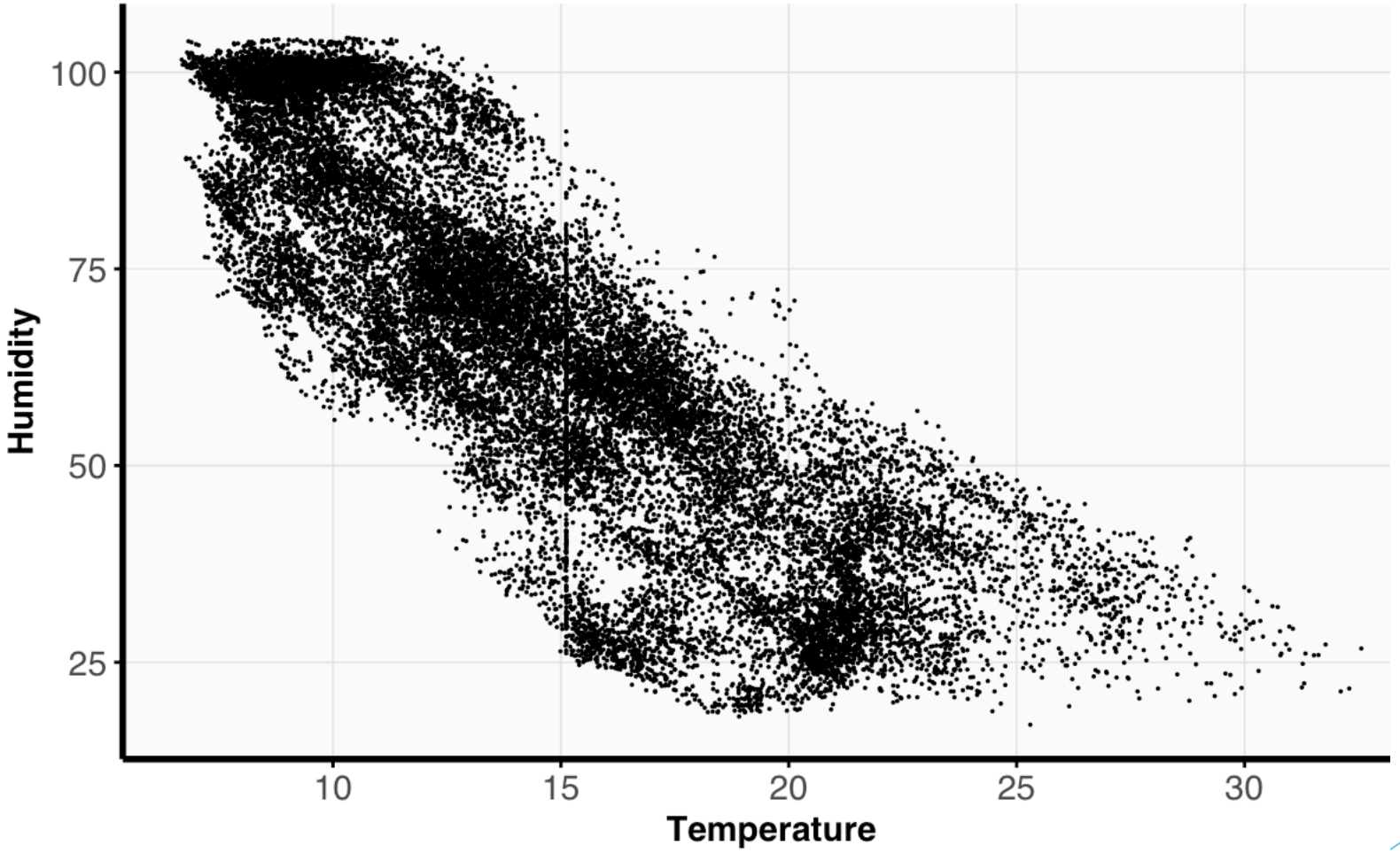
Overplotting



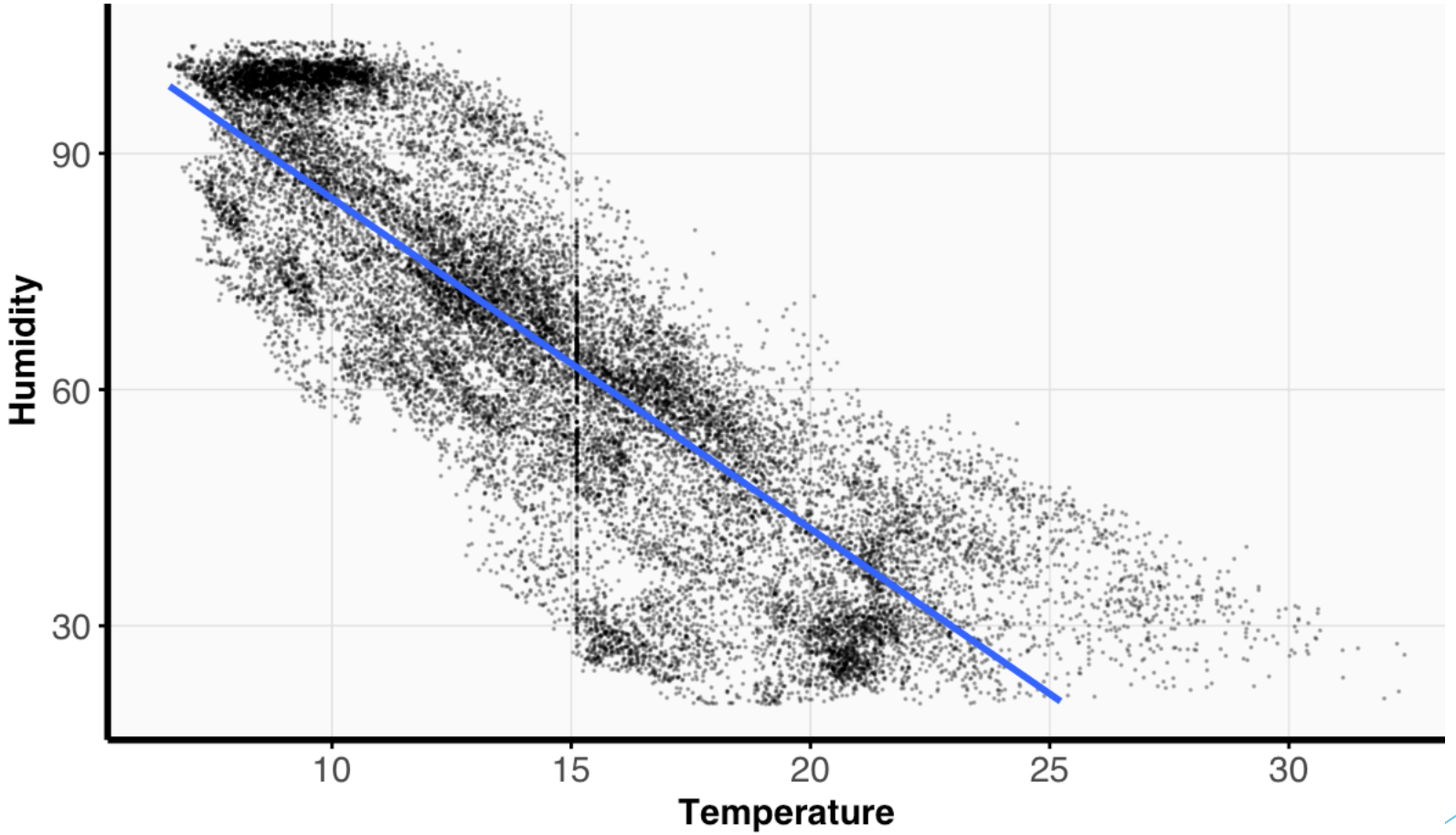
Potential Solutions

- Subsampling
- Use transparency (alpha)
- Smaller point sizes

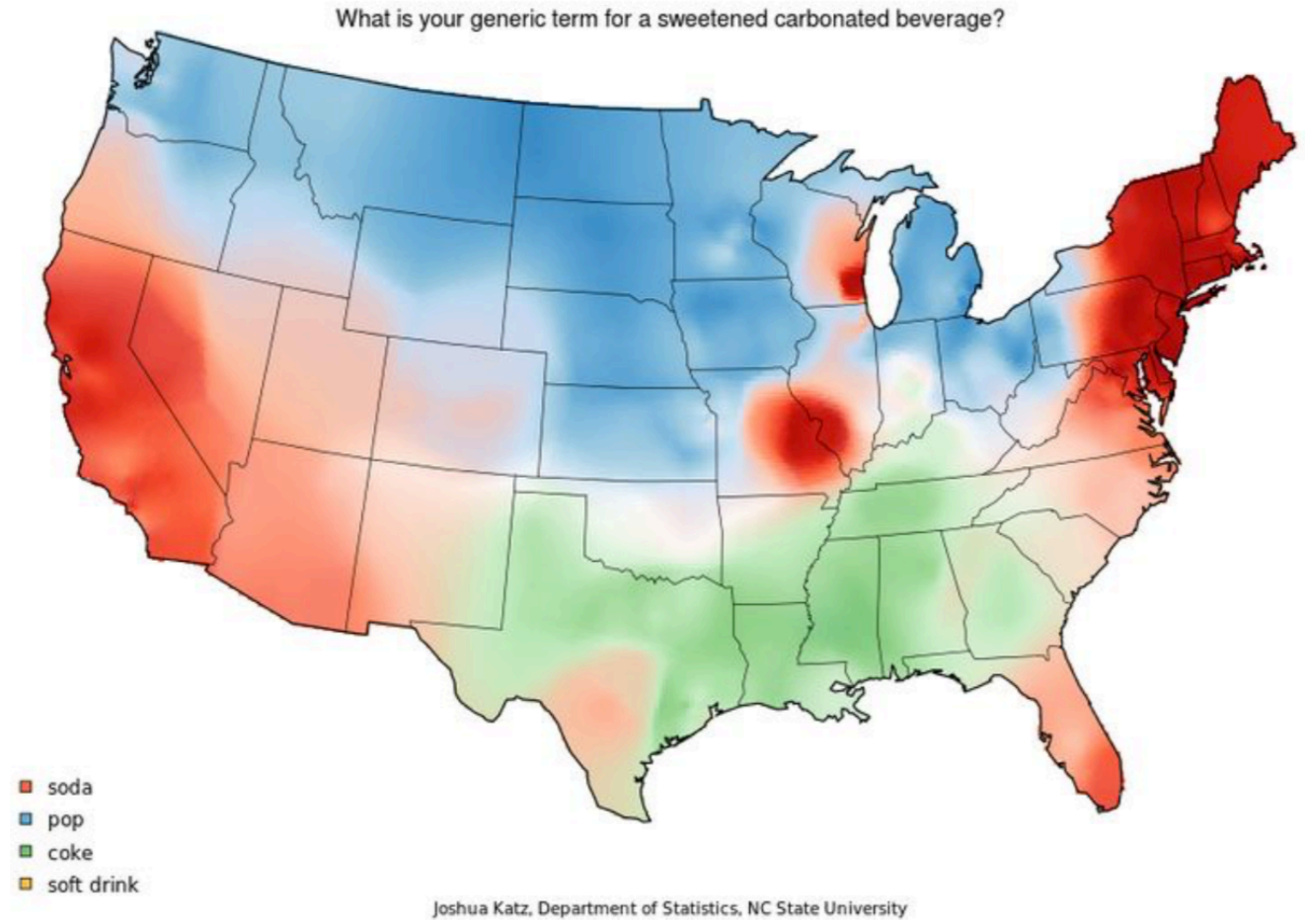
Overplotting



Overplotting



Lab 2 – Linguistics Survey (Due October 10 11:59pm)



https://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6#ok-this-one-is-crazy-everyone-pronounces-pecan-pie-differently-10

Lab 2 – Linguistics Survey

- Your primary goal is to:
  - Perform EDA/dimension reduction and clustering
  - Evaluate stability of clustering
- Other things to keep in mind
  - Readability of narrative and code
  - Clear and effective visualizations
  - Clarity of folder structure
    - Only push files required to reproduce the report
- Don’t forget about HW

Start early!!!!!!