Unsupervised Model Selection and Evaluation

February 17, 2026

Today's Plan

0 Review of Last Time

1 Supervised vs **Unsupervised** Learning

2 Overview of Popular **Dimension Reduction** Methods

3 Overview of Popular **Clustering** Methods

4 **Model Selection** and **Evaluation**

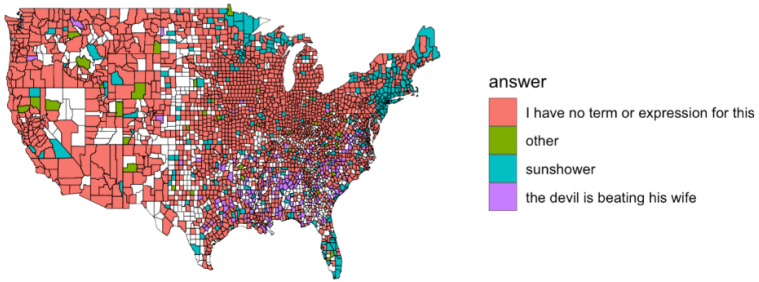
5 **In-Class Lab:** Linguistics Data

Review of Last Time

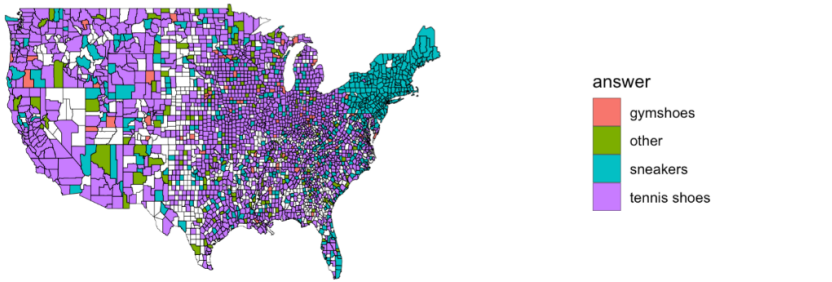
Introduction to Linguistics Case Study

Domain question: Are there geographical regions in the US with distinctive dialects? Can we find clusters of people who speak similarly in the US?

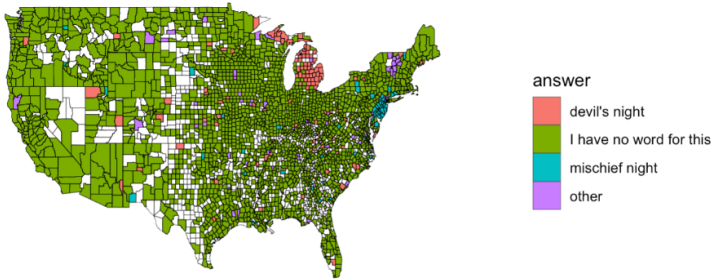
Q080: What do you call it when rain falls while the sun is shining?



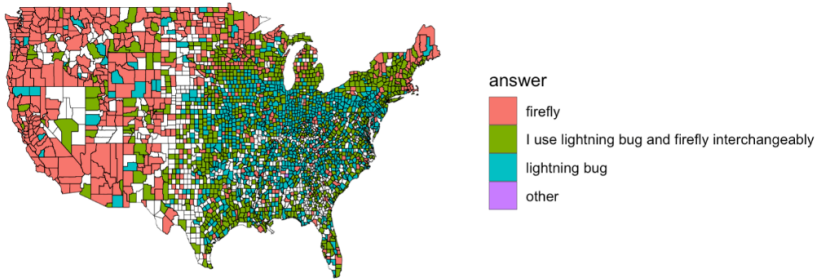
Q073: What is your *general* term for the rubber-soled shoes worn in gym class, for athle



Q110: What do you call the night before Halloween?

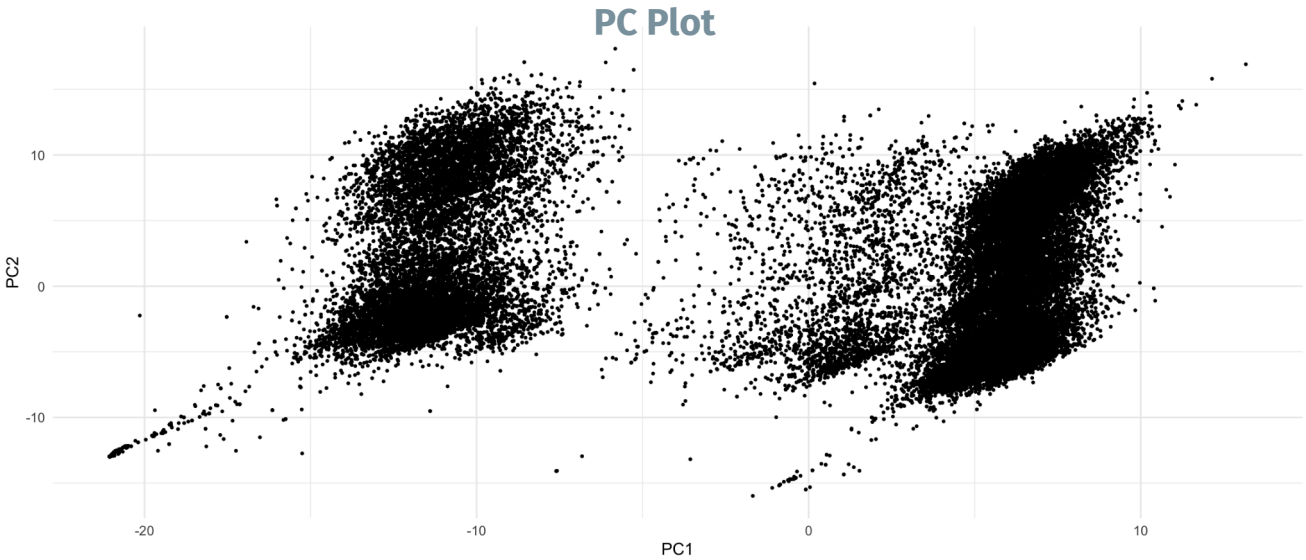


Q065: What do you call the insect that flies around in the summer and has a rear section



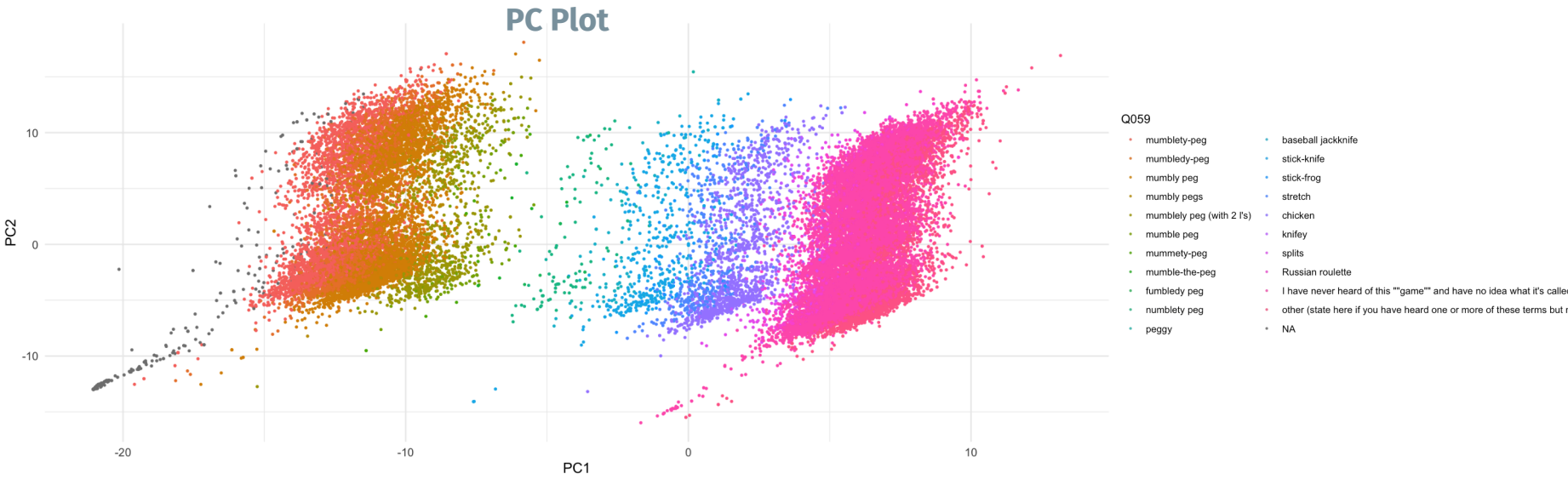
Our journey so far

PCA Attempt 1: PCA on the raw survey response data



Our journey so far

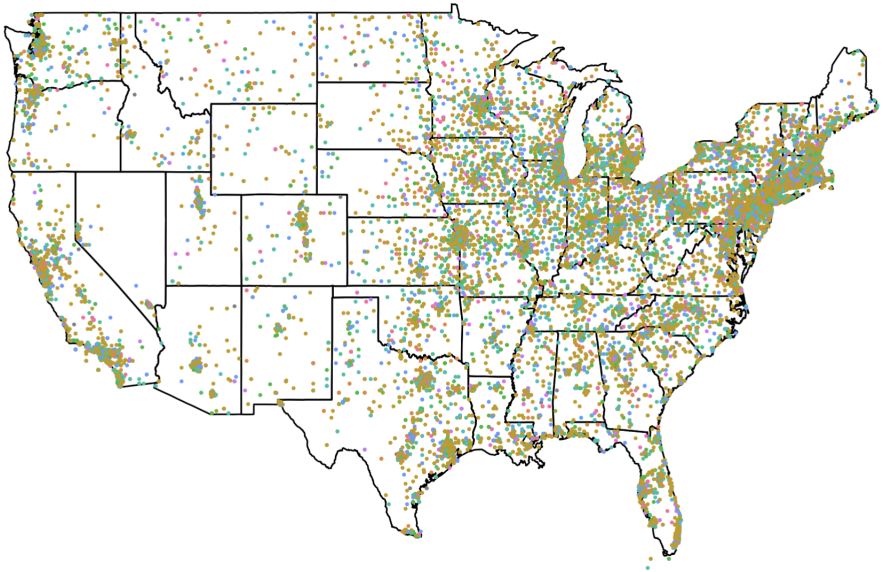
PCA Attempt 1: PCA on the raw survey response data

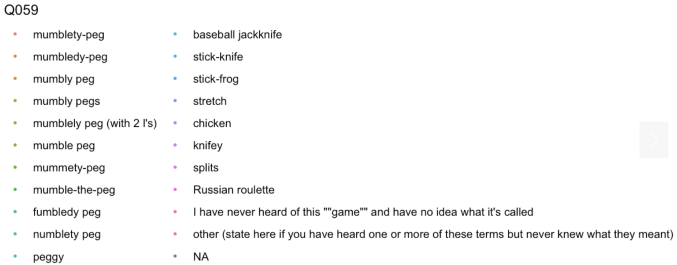


PC1 predominantly captures the variation in Q059 which has the most options

Our journey so far

What if we go back and look at the Q59 map?





Q59 is one of the few questions that doesn't show any geographical patterns.

Our journey so far

At this point, PCA has told us that:

- The "pattern" that explains the most variation in our data is not related to geographical differences

This seems contradictory to our EDA, where most questions show some type of geographical variation. (Think of this EDA as our sanity check)

What might be going on?

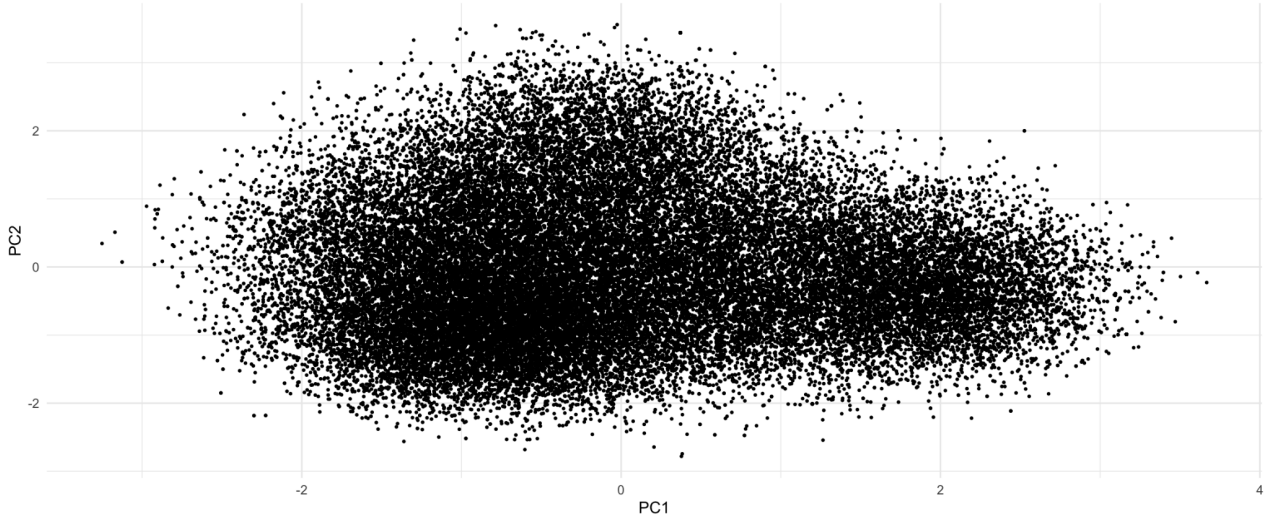
- Currently, PCA is being influenced by our **arbitrary** encoding of the data

How did we proceed?

- One-hot encoded categorical data and removed non-responders

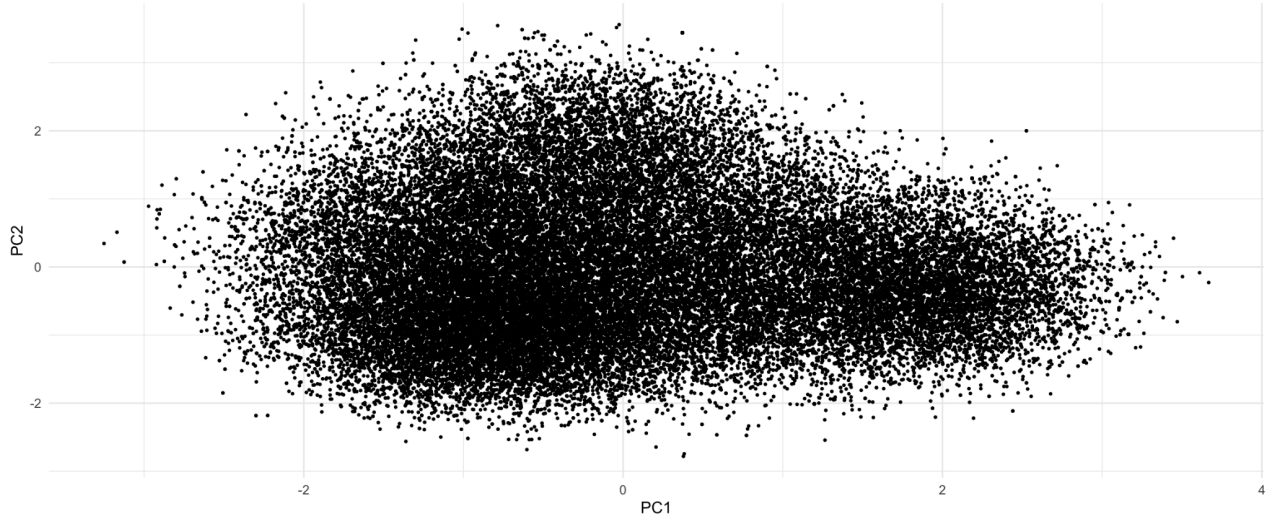
Continuing onwards...

PCA on the *cleaned one-hot-encoded* survey response data



Continuing onwards...

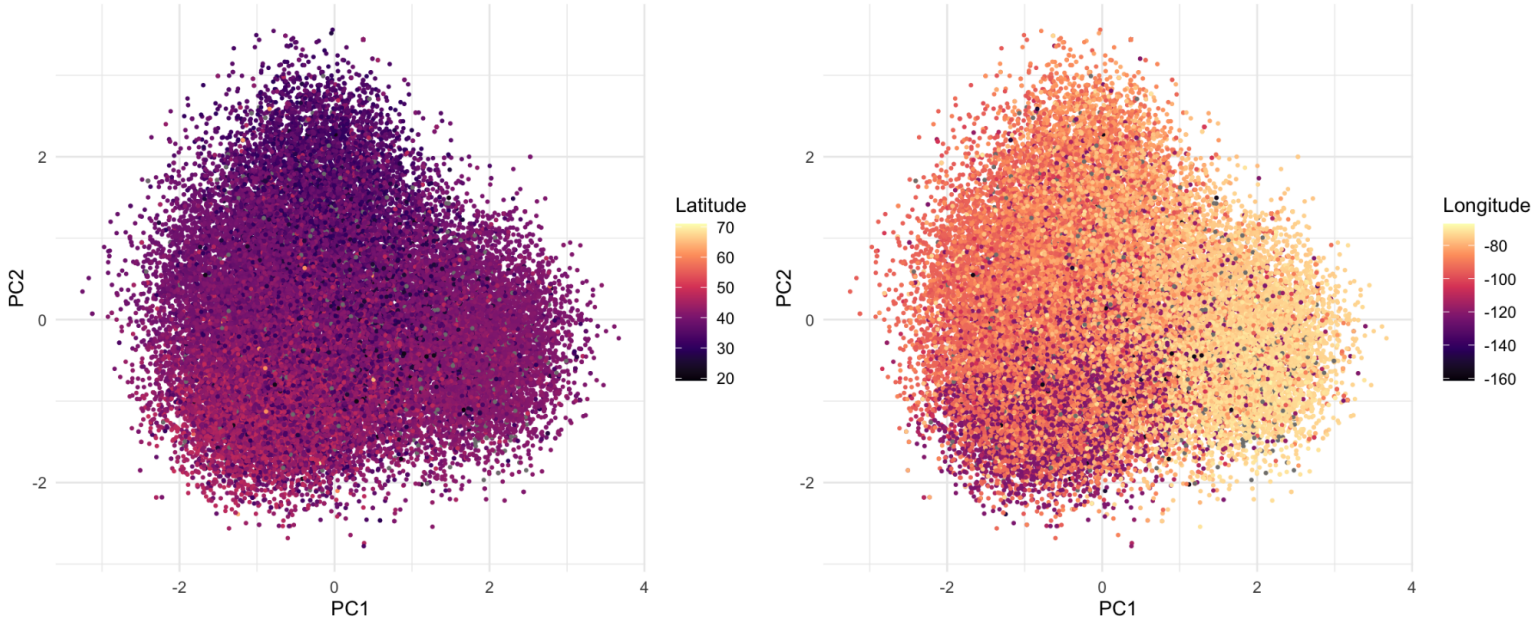
PCA on the *cleaned one-hot-encoded* survey response data



Is this good? We can't tell based upon this information

Continuing onwards...

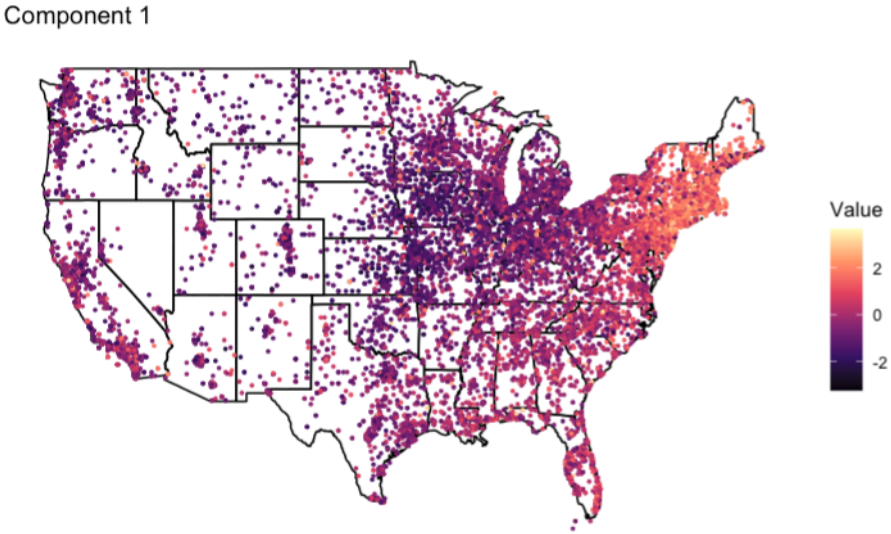
PCA on the *cleaned one-hot-encoded* survey response data

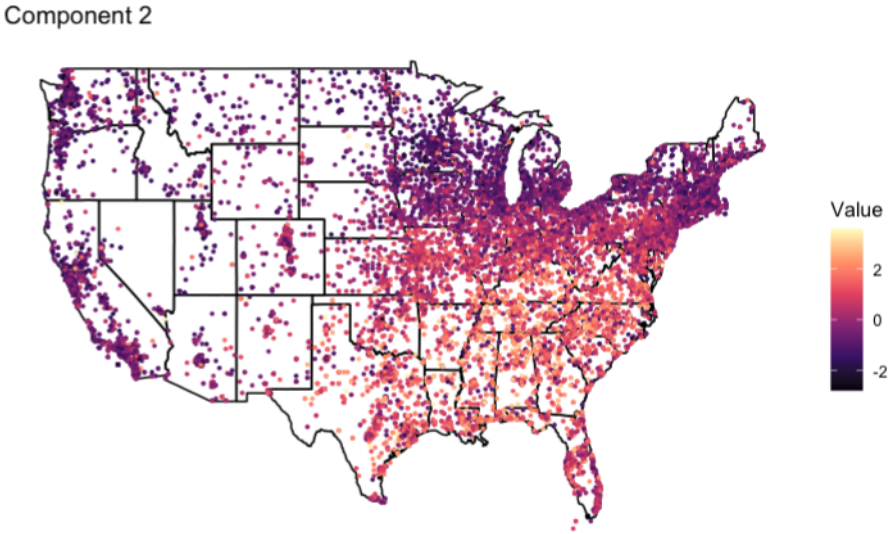


Need to leverage some form of domain information to assess whether or not this is a "good" dimension reduction

Continuing onwards...

PCA on the *cleaned one-hot-encoded* survey response data

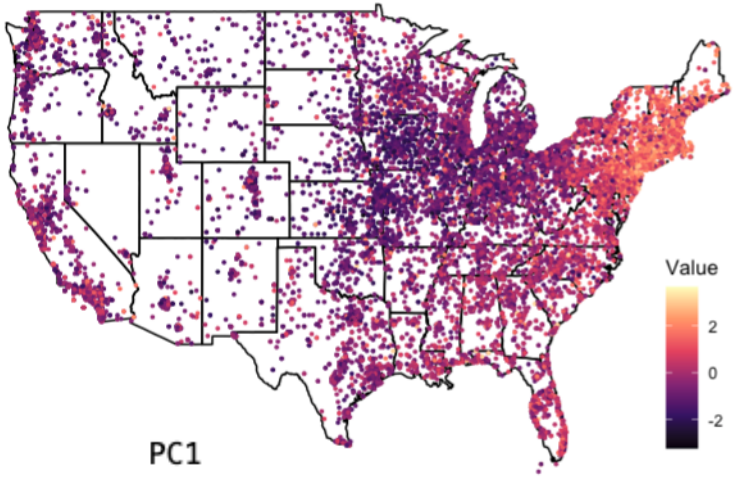




Need to leverage some form of domain information to assess whether or not this is a "good" dimension reduction

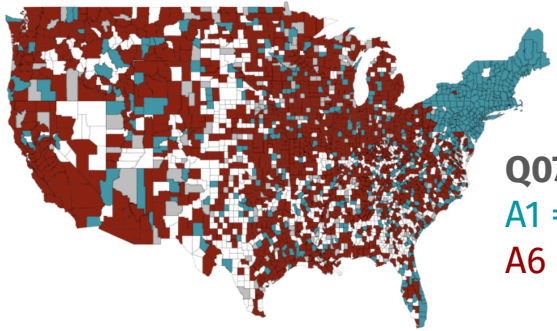
Interpreting PCA

Component 1

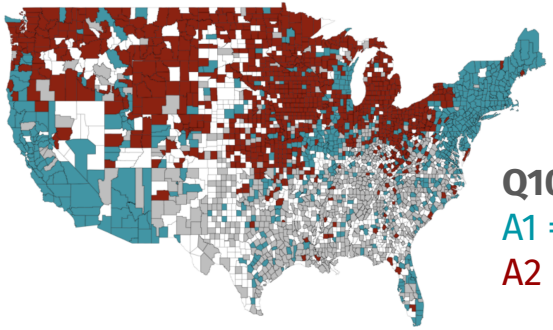


PC1
Q073.1 (0.27)
Q073.6 (−0.24)
Q105.1 (0.20)
Q080.1 (0.18)
Q080.8 (−0.18)
Q105.2 (−0.17)
⋮

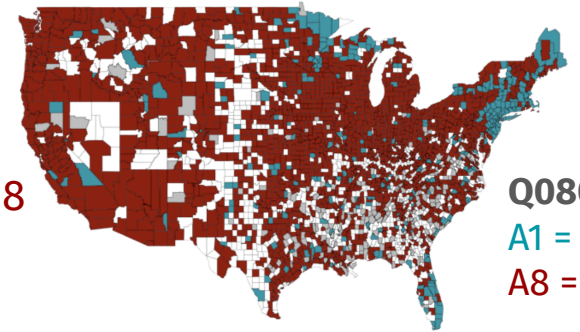
PC1 = 0.27 × Q073.1 - 0.24 × Q073.6
+ 0.2 × Q105.1 - 0.17 × Q105.2
+ 0.18 × Q080.1 - 0.18 × Q080.8
+ …



Q073: "Shoes" question
A1 = sneakers
A6 = tennis shoes

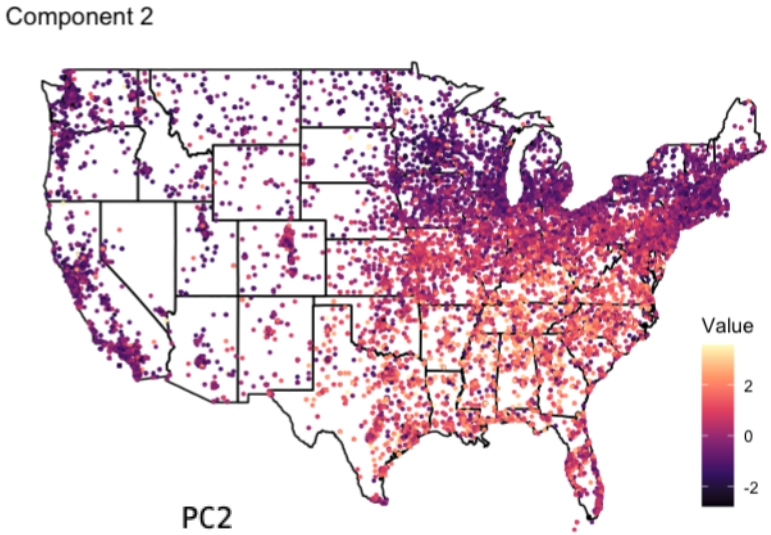


Q105: "Soda" question
A1 = soda
A2 = pop



Q080: "Sunshower" quest
A1 = sunshower
A8 = no term for this

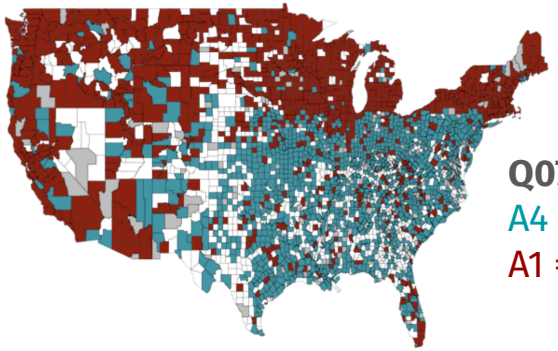
Interpreting PCA



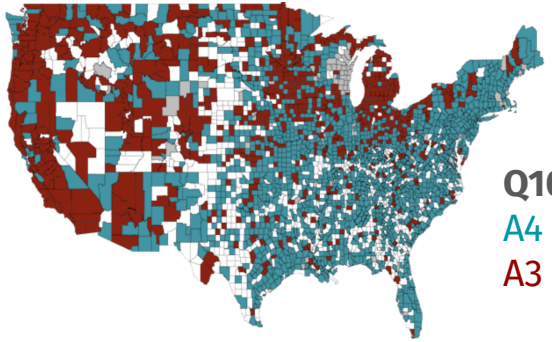
PC2

Q076.1 (−0.25)
Q076.4 (0.24)
Q103.4 (0.22)
Q050.9 (0.19)
Q103.3 (−0.19)
Q071.5 (0.16)
⋮

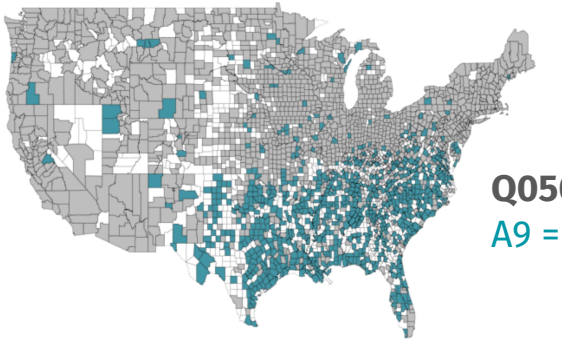
PC2 = 0.24 × Q076.4 - 0.25 × Q076.1
+ 0.22 × Q103.4 - 0.19 × Q103.3
+ 0.19 × Q050.9
+ 0.16 × Q071.5
+ …



Q076:
A4 = catty-corner
A1 = kitty-corner



Q103:
A4 = water fountain
A3 = drinking fountain

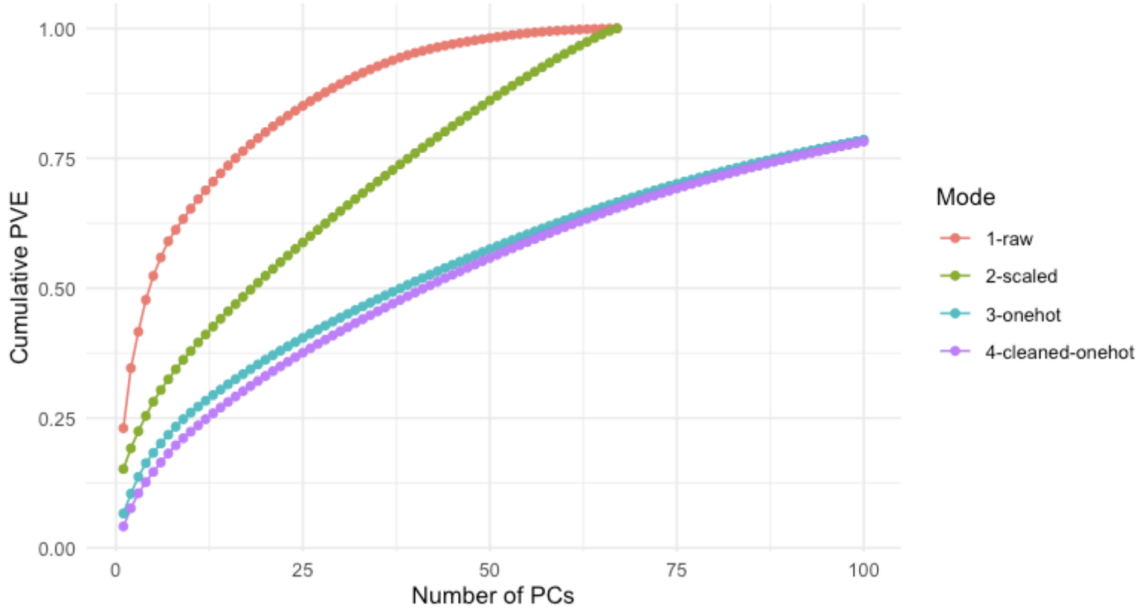


Q050:
A9 = y'all

Interpreting PCA

Does the proportion of variance explained (PVE) tell us anything about the "quality" of the PCA results? No!

- PC1 from attempt #1 explained the most variance, but this was variance in the data that we did not care about
- PVE is relative to the *total amount of variation* in the data, which differs across datasets (or even the same dataset, scaled differently)



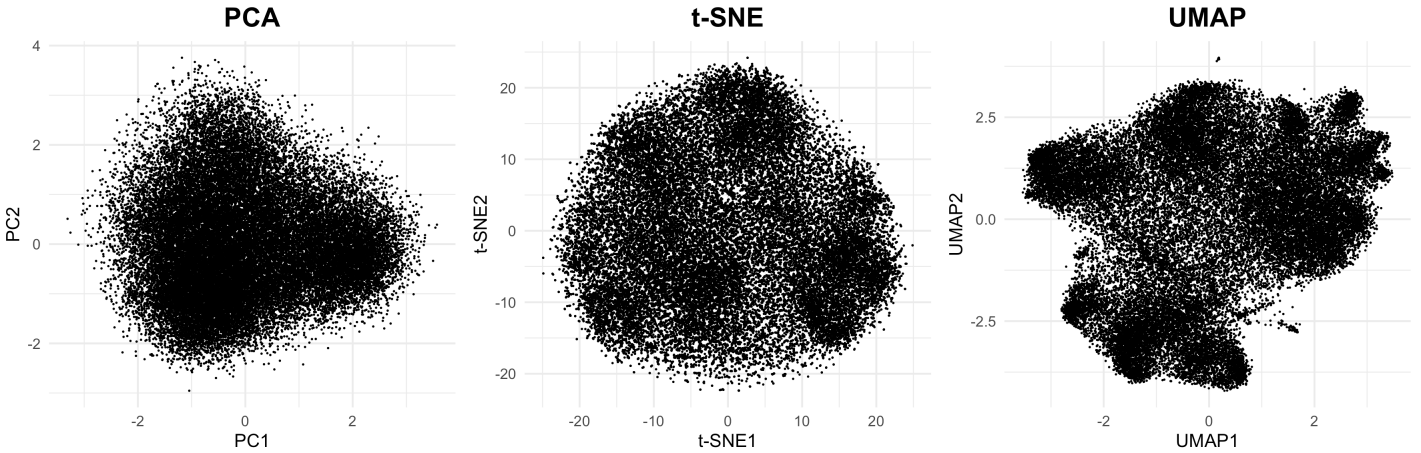
Today's Plan

1. Supervised vs **Unsupervised** Learning
2. Overview of Popular **Dimension Reduction** Methods
3. Overview of Popular **Clustering** Methods
4. **Model Selection** and **Evaluation**
5. **In-Class Lab:** Linguistics Data

Dimension Reduction: Model Selection and Evaluation

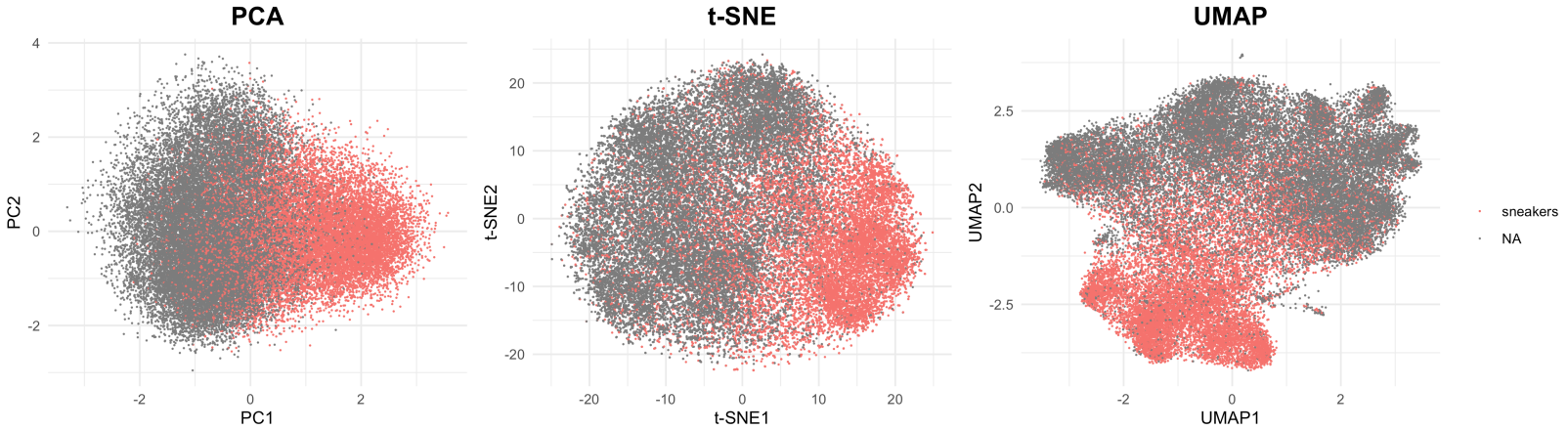
Comparing/Evaluating Dimension Reduction Methods

If we don’t have “labels”, how do we know whether a dimension reduction method is good?



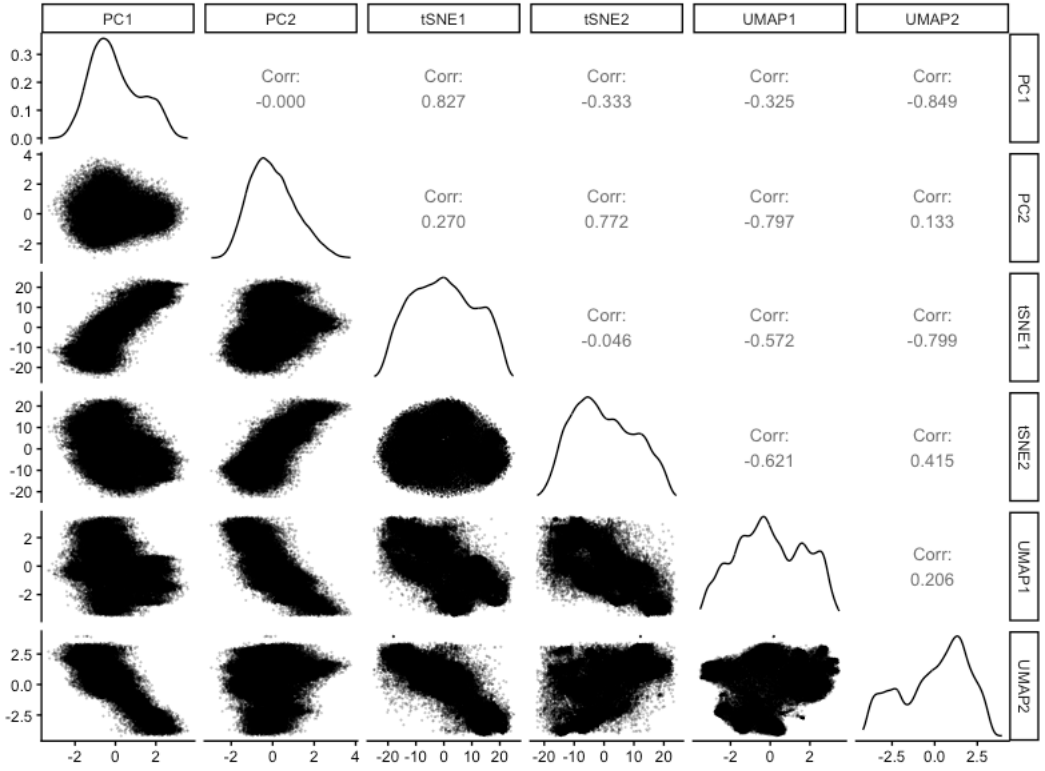
Comparing/Evaluating Dimension Reduction Methods

Heuristic 1: color points based on important features



Comparing/Evaluating Dimension Reduction Methods

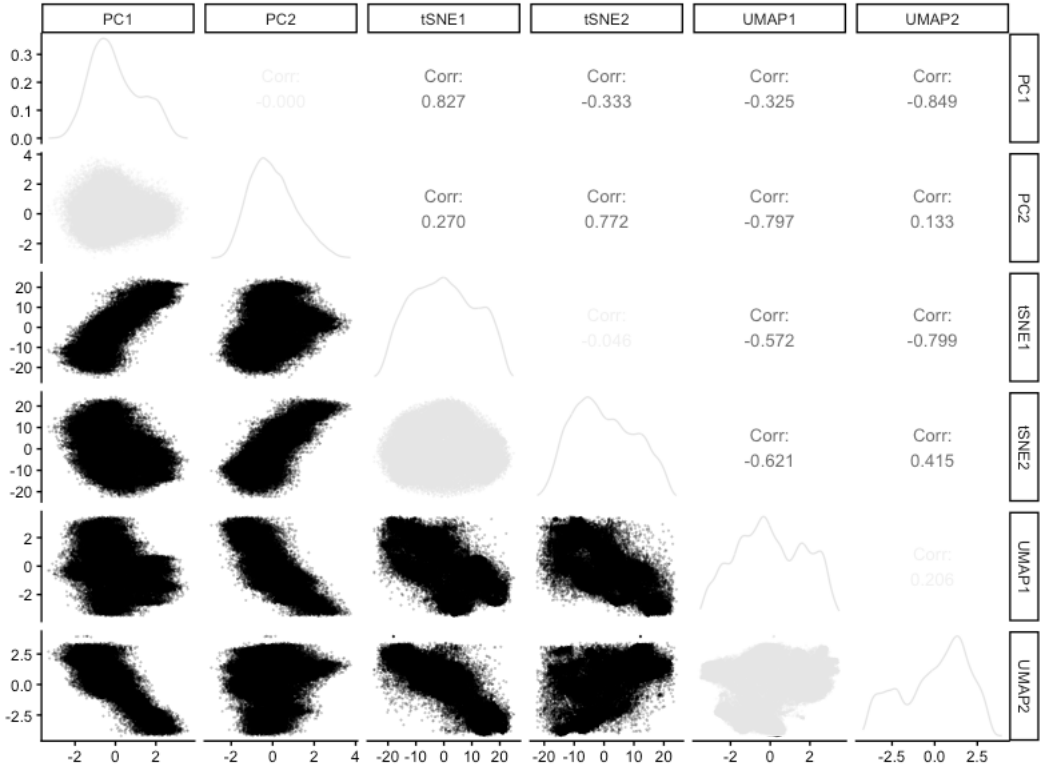
Heuristic 2: compare/contrast dimension reduction results



Note: can also use **procrustes analysis** to "align" dimension reduction methods first

Comparing/Evaluating Dimension Reduction Methods

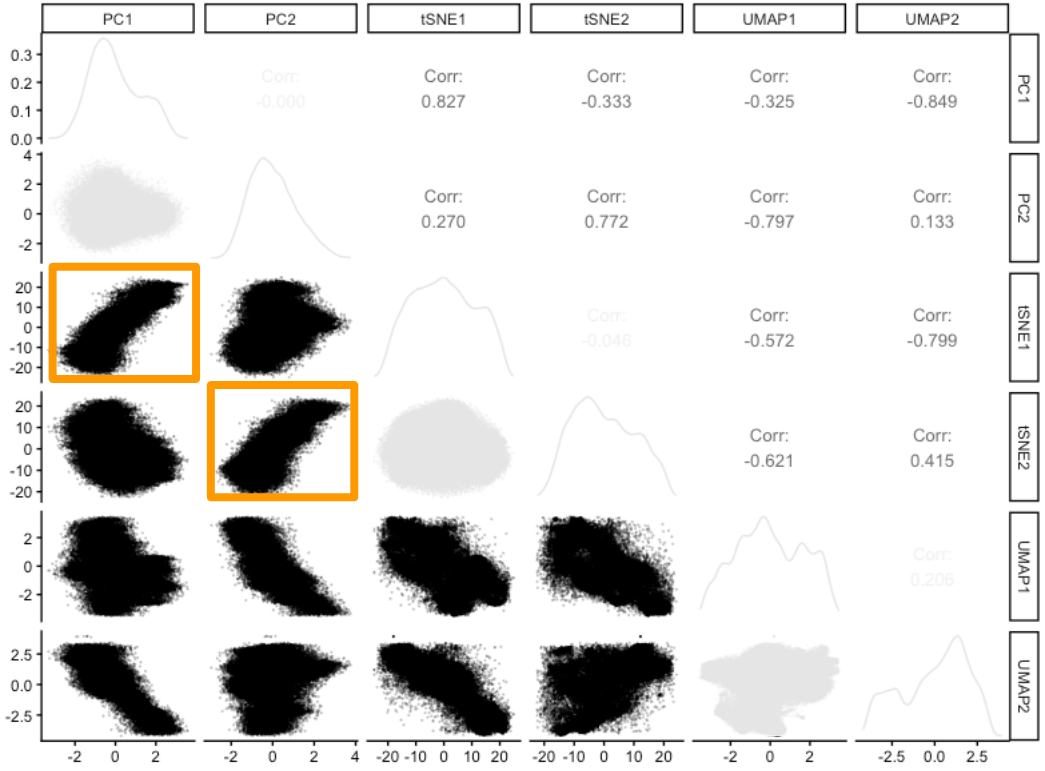
Heuristic 2: compare/contrast dimension reduction results



Note: can also use **procrustes analysis** to "align" dimension reduction methods first

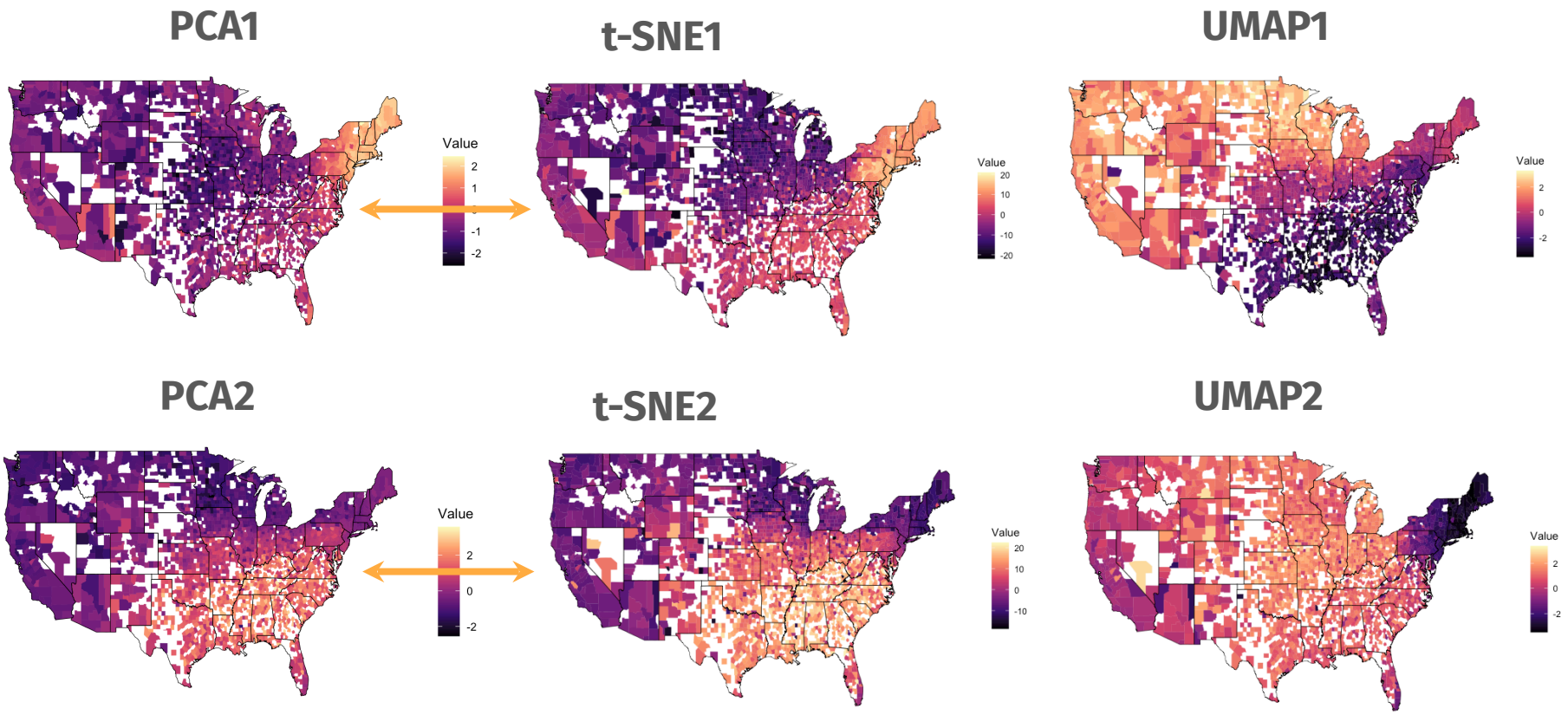
Comparing/Evaluating Dimension Reduction Methods

Heuristic 2: compare/contrast dimension reduction results



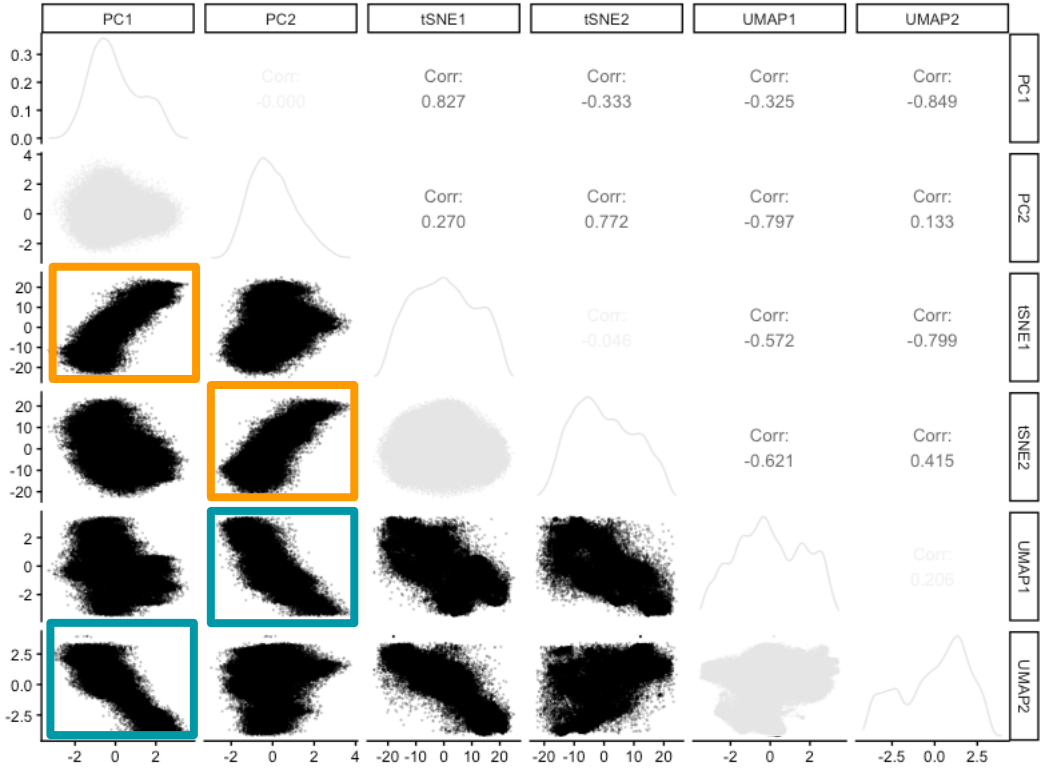
Note: can also use **procrustes analysis** to "align" dimension reduction methods first

Comparing/Evaluating Dimension Reduction Methods



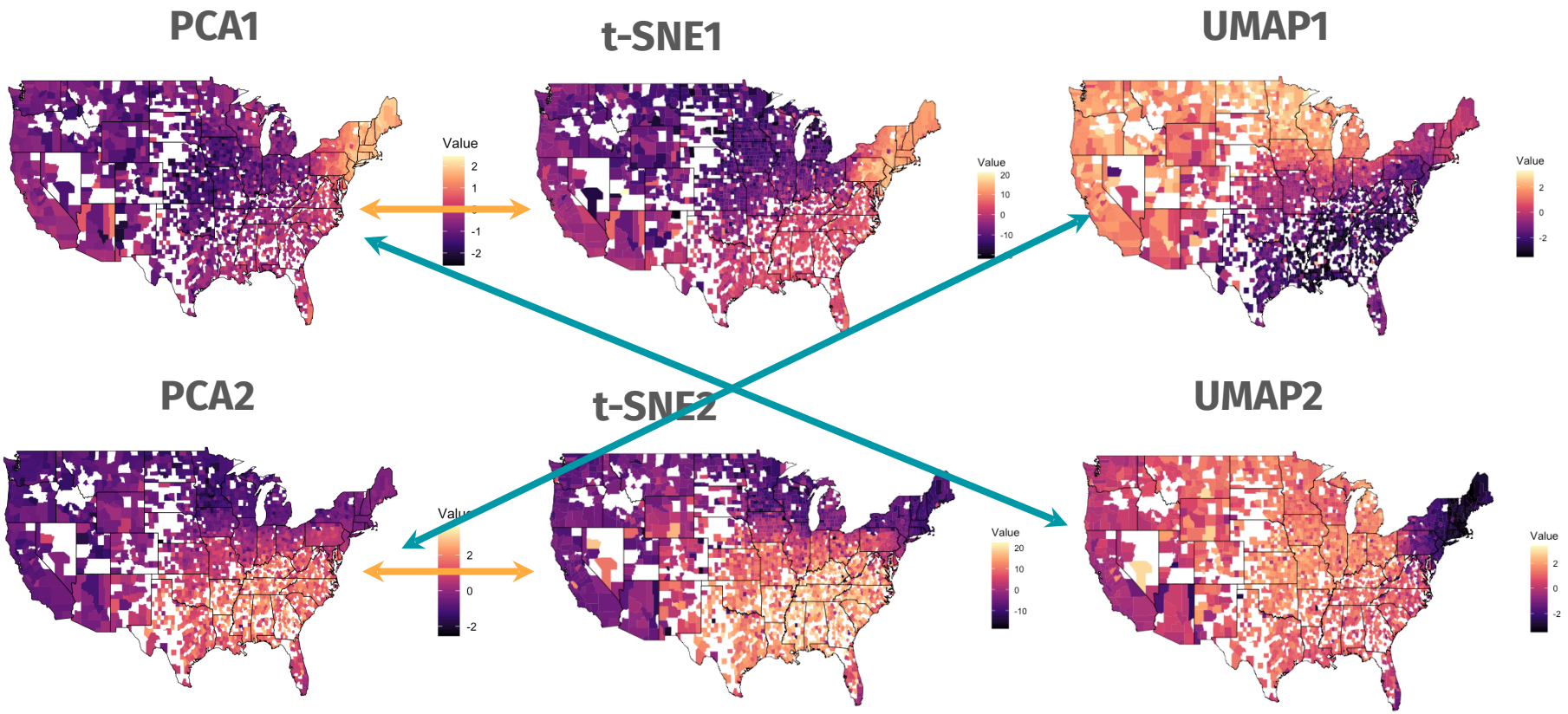
Comparing/Evaluating Dimension Reduction Methods

Heuristic 2: compare/contrast dimension reduction results



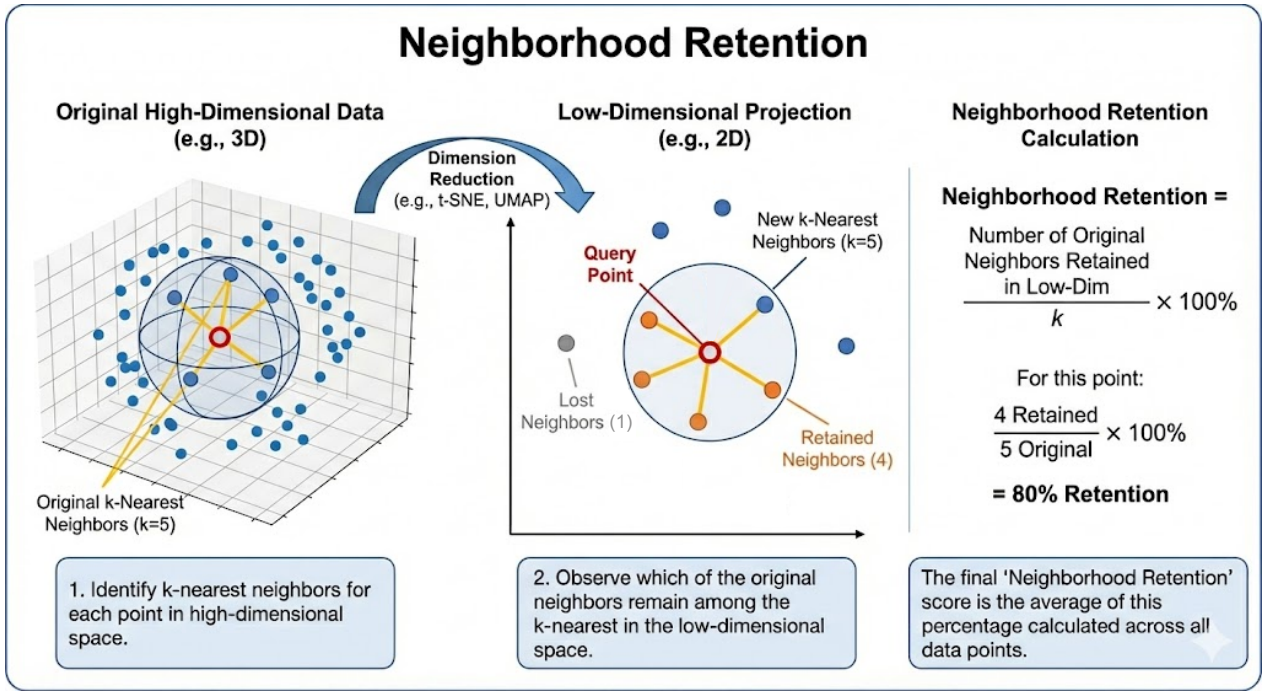
Note: can also use **procrustes analysis** to "align" dimension reduction methods first

Comparing/Evaluating Dimension Reduction Methods



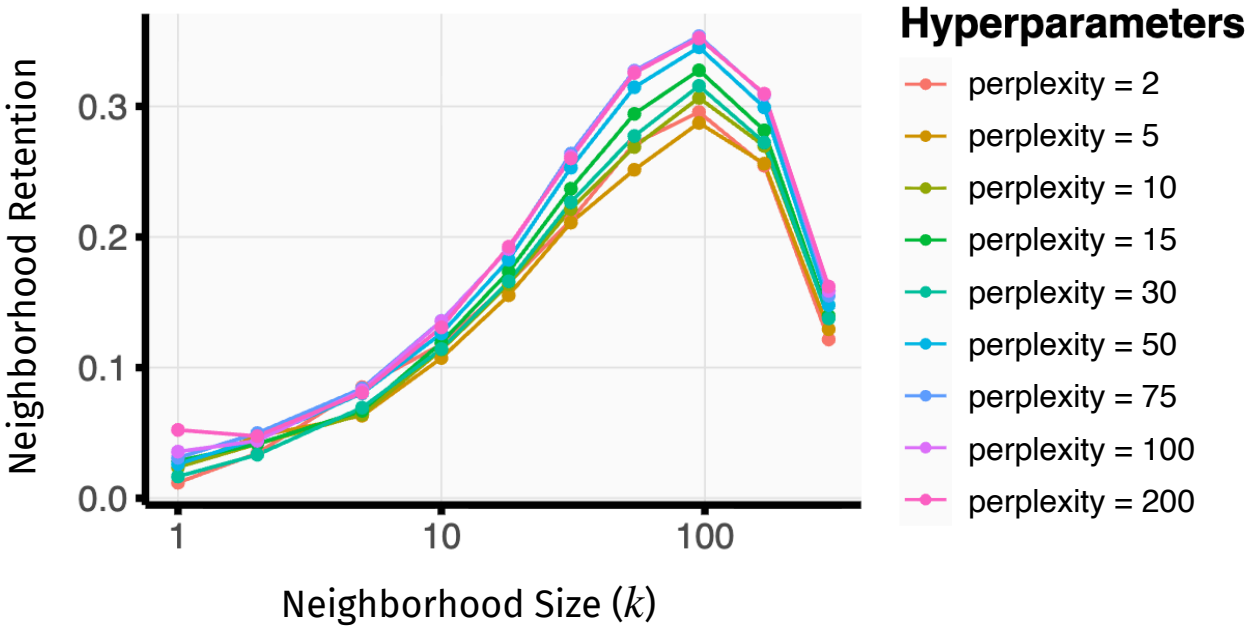
Comparing/Evaluating Dimension Reduction Methods

Heuristic 3: neighborhood retention = average proportion of k-nearest neighbors in original data that are retained in low-dimensional data



Comparing/Evaluating Dimension Reduction Methods

Heuristic 3: neighborhood retention = average proportion of k-nearest neighbors in original data that are retained in low-dimensional data



Clustering: Model Selection and Evaluation

How to choose the number of clusters *K*

- **Method 1: Silhouette Index**
- **Method 2: Stability****

Both work for any clustering algorithm.

Can also be used to select appropriate hyperparameters for your model.

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters
- **Formally:** Given a data point *i*,
  the Silhouette index is defined as

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

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters
- **Formally:** Given a data point *i*,
  the Silhouette index is defined as

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

where $$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j)$$

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters
- **Formally:** Given a data point *i*,
  the Silhouette index is defined as

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

where $a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j)$ (i.e., Mean distance to other points in its own cluster C_I)

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters
- **Formally:** Given a data point *i*,
  the Silhouette index is defined as

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

where $$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j)$$ (i.e., Mean distance to other points in its own cluster C_I)

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)$$

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters
- **Formally:** Given a data point *i*, the Silhouette index is defined as

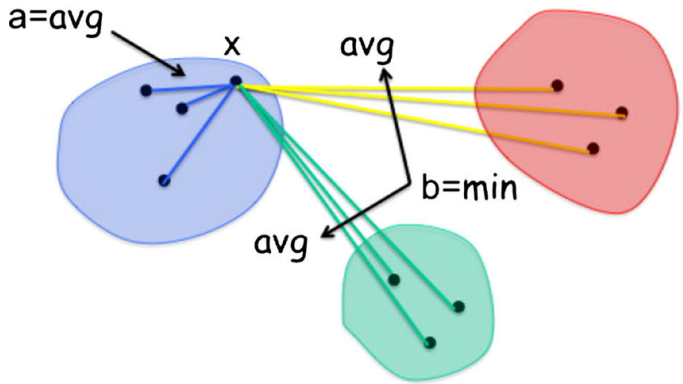
$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where $a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j)$ (i.e., Mean distance to other points in its own cluster C_I)

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)$$ (i.e., Mean distance to points in the nearest other cluster C_J)

Silhouette Index

- **Idea:** measures how similar a point is to other points in its cluster, as opposed to other points in different clusters
- **Formally:** Given a data point *i*, the Silhouette index is defined as



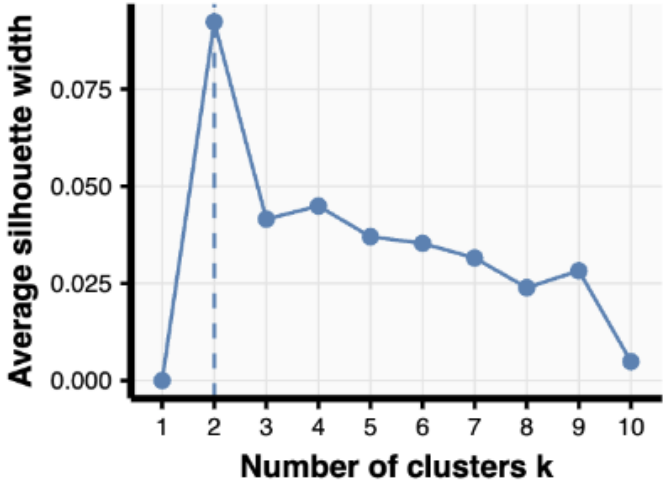
$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where $a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$ (i.e., Mean distance to other points in its own cluster C_I)

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$ (i.e., Mean distance to points in the nearest other cluster C_J)

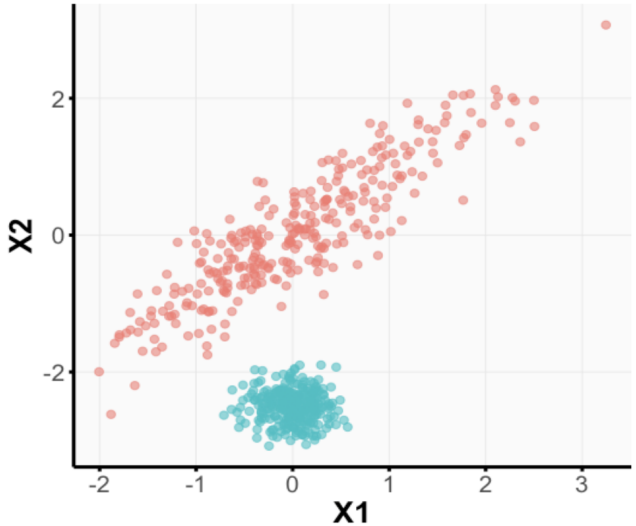
Silhouette Index

- Silhouette index is always between -1 and 1, with higher = better
- Can look at
  - Distribution of silhouette indices across all samples (e.g., boxplots)
  - Mean of silhouette indices across all samples
- Generally choose number of clusters *k* that gives the highest mean silhouette index



Silhouette Index

- Favors spherical, equally-sized clusters
- Does not work very well when “true” clusters are not spherical, e.g.,



Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

A stability based method for discovering structure in clustered data

Asa Ben-Hur*, Andre Elisseeff† and Isabelle Guyon*
BioWulf Technologies LLC
*2030 Addison st. Suite 102 †305 Broadway (9th Floor)
Berkeley, CA 94704 New-York, NY 10007

Abstract

We present a method for visually and quantitatively assessing the presence of structure in clustered data. The method exploits measurements of the stability of clustering solutions obtained by perturbing the data set. Stability is characterized by the distribution of pairwise similarities between clusterings obtained from sub samples of the data. High pairwise similarities indicate a stable clustering pattern. The method can be used with any clustering algorithm; it provides a means of rationally defining an optimum number of clusters, and can also detect the lack of structure in data. We show results on artificial and microarray data using a hierarchical clustering algorithm.

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

for $k = 2, \ldots, k_{\max}$ **do**
  for $b = 1, \ldots, B$ **do**
    $sub_1 = \text{subsample}(X, \pi)$
    $sub_2 = \text{subsample}(X, \pi)$
    $L_1 = \text{cluster}(sub_1, k)$
    $L_2 = \text{cluster}(sub_2, k)$
    $S(k, b) = similarity(L_1[sub_1 \cap sub_2], L_2[sub_1 \cap sub_2])$
  end
  $S(k) = \frac{1}{B} \sum_{b=1}^{B} S(k, b)$
end
Select k with the highest $S(k)$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

```
for k = 2, ..., k_max do
    for b = 1, ..., B do
        sub_1 = subsample(X, π)
        sub_2 = subsample(X, π)
        L_1 = cluster(sub_1, k)
        L_2 = cluster(sub_2, k)
        S(k, b) = similarity(L_1[sub_1 ∩ sub_2], L_2[sub_1 ∩ sub_2])
    end
    S(k) = (1/B) Σ_{b=1}^{B} S(k, b)
end
Select k with the highest S(k)
```

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

```
for k = 2, ..., k_max do
    for b = 1, ..., B do
        sub_1 = subsample(X, π)
        sub_2 = subsample(X, π)
        L_1 = cluster(sub_1, k)
        L_2 = cluster(sub_2, k)
        S(k, b) = similarity(L_1[sub_1 ∩ sub_2], L_2[sub_1 ∩ sub_2])
    end
    S(k) = (1/B) Σ_{b=1}^{B} S(k, b)
end
Select k with the highest S(k)
```

Take two subsamples (with subsampling fraction π)

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

```
for k = 2, ..., k_max do
    for b = 1, ..., B do
        sub_1 = subsample(X, π)
        sub_2 = subsample(X, π)
        L_1 = cluster(sub_1, k)
        L_2 = cluster(sub_2, k)
        S(k, b) = similarity(L_1[sub_1 ∩ sub_2], L_2[sub_1 ∩ sub_2])
    end
    S(k) = (1/B) Σ_{b=1}^{B} S(k, b)
end
Select k with the highest S(k)
```

Take two subsamples (with subsampling fraction π)

Run clustering on each subsample to get cluster labels L_1 & L_2

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

```
for k = 2, ..., k_max do
    for b = 1, ..., B do
        sub_1 = subsample(X, π)
        sub_2 = subsample(X, π)
        L_1 = cluster(sub_1, k)
        L_2 = cluster(sub_2, k)
        S(k, b) = similarity(L_1[sub_1 ∩ sub_2], L_2[sub_1 ∩ sub_2])
    end
    S(k) = (1/B) Σ_{b=1}^{B} S(k, b)
end
Select k with the highest S(k)
```

Take two subsamples (with subsampling fraction π)

Run clustering on each subsample to get cluster labels L_1 & L_2

Compute similarity between clusters using points common to both subsamples

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

1 for $k = 2, \ldots, k_{\max}$ do
2 **for** $b = 1, \ldots, B$ **do**
3 $sub_1 = \text{subsample}(X, \pi)$
4 $sub_2 = \text{subsample}(X, \pi)$
5 $L_1 = \text{cluster}(sub_1, k)$
6 $L_2 = \text{cluster}(sub_2, k)$
7 $S(k, b) = similarity(L_1[sub_1 \cap sub_2], L_2[sub_1 \cap sub_2])$
8 **end**
9 $S(k) = \frac{1}{B} \sum_{b=1}^{B} S(k, b)$
10 end
11 Select k with the highest $S(k)$

Repeat across many many different subsamples

Take two subsamples (with subsampling fraction π)

Run clustering on each subsample to get cluster labels L_1 & L_2

Compute similarity between clusters using points common to both subsamples

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

1 **for** $k = 2, \ldots, k_{\max}$ **do**
2 **for** $b = 1, \ldots, B$ **do**
3 $sub_1 = \text{subsample}(X, \pi)$
4 $sub_2 = \text{subsample}(X, \pi)$
5 $L_1 = \text{cluster}(sub_1, k)$
6 $L_2 = \text{cluster}(sub_2, k)$
7 $S(k, b) = similarity(L_1[sub_1 \cap sub_2], L_2[sub_1 \cap sub_2])$
8 **end**
9 $S(k) = \frac{1}{B} \sum_{b=1}^{B} S(k, b)$
10 **end**
11 Select k with the highest $S(k)$

Repeat across many many different subsamples

Take two subsamples (with subsampling fraction π)

Run clustering on each subsample to get cluster labels L_1 & L_2

Compute similarity between clusters using points common to both subsamples

Choose k that is most stable

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

1 **for** $k = 2, \ldots, k_{\max}$ **do**
2 **for** $b = 1, \ldots, B$ **do**
3 $sub_1 = \text{subsample}(X, \pi)$
4 $sub_2 = \text{subsample}(X, \pi)$
5 $L_1 = \text{cluster}(sub_1, k)$
6 $L_2 = \text{cluster}(sub_2, k)$
7 $S(k, b) = similarity(L_1[sub_1 \cap sub_2], L_2[sub_1 \cap sub_2])$
8 **end**
9 $S(k) = \frac{1}{B} \sum_{b=1}^{B} S(k, b)$
10 **end**
11 Select k with the highest $S(k)$

Repeat across many many different subsamples

Take two subsamples (with subsampling fraction π)

Run clustering on each subsample to get cluster labels L_1 & L_2

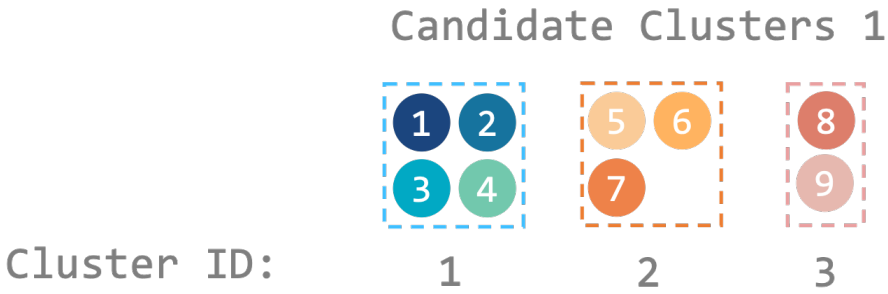
??

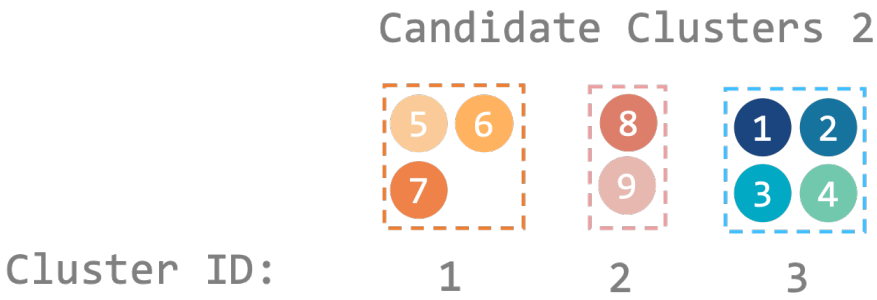
Compute similarity between clusters using points common to both subsamples

Choose k that is most stable

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

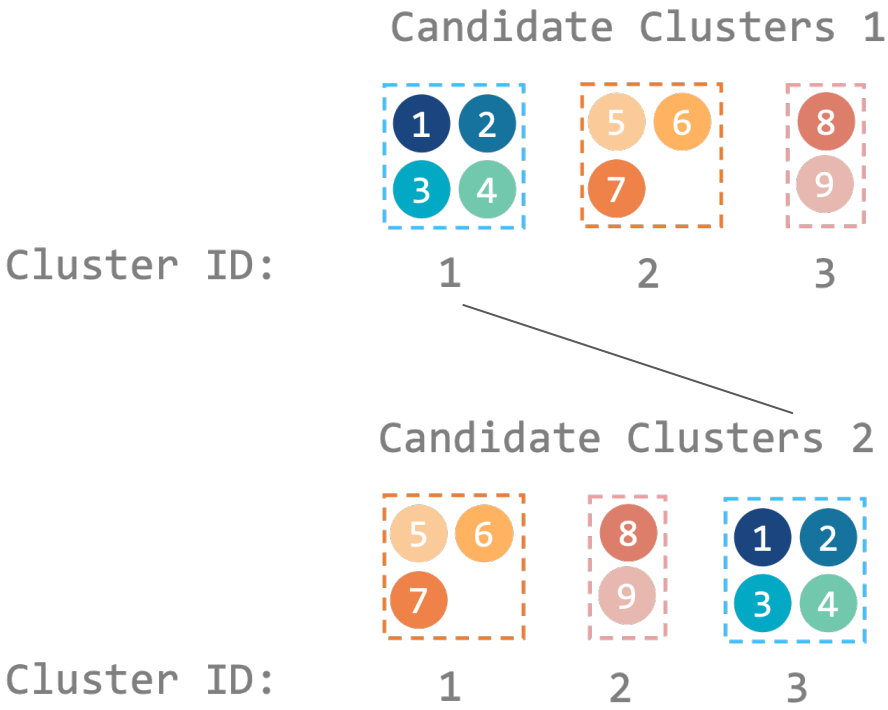
- **How do we measure similarity between two clusters?**





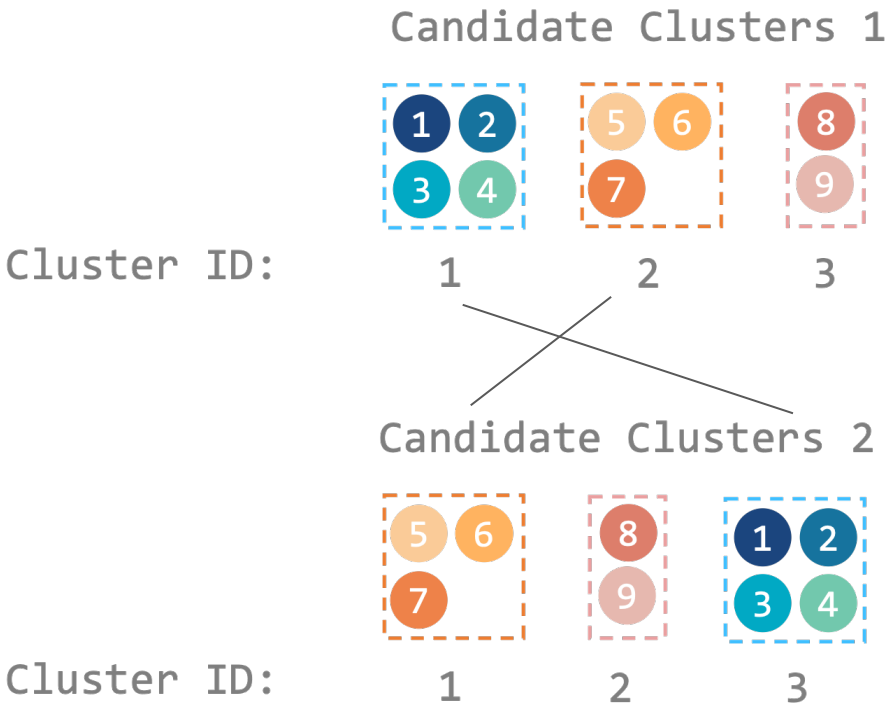
Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**



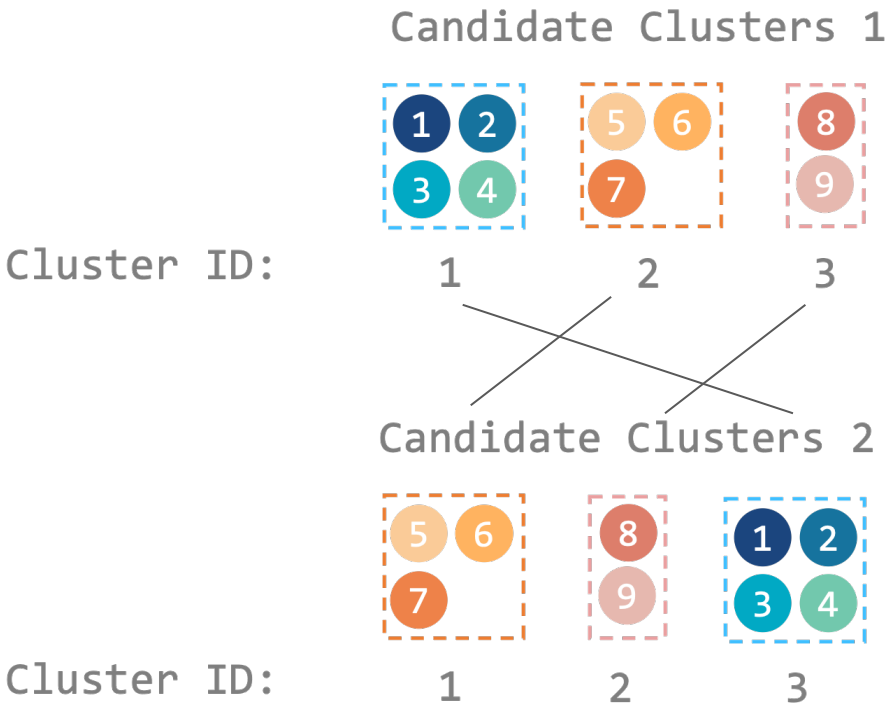
Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**



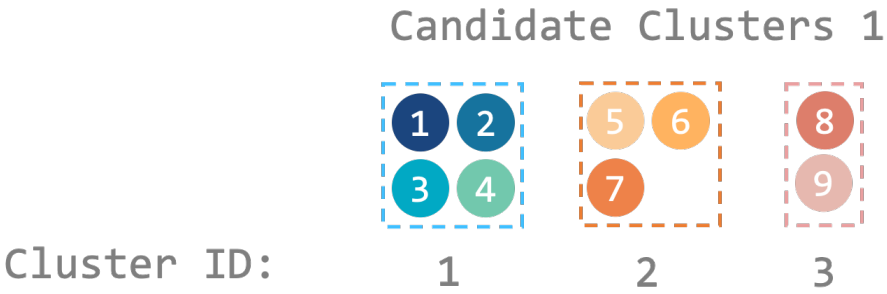
Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

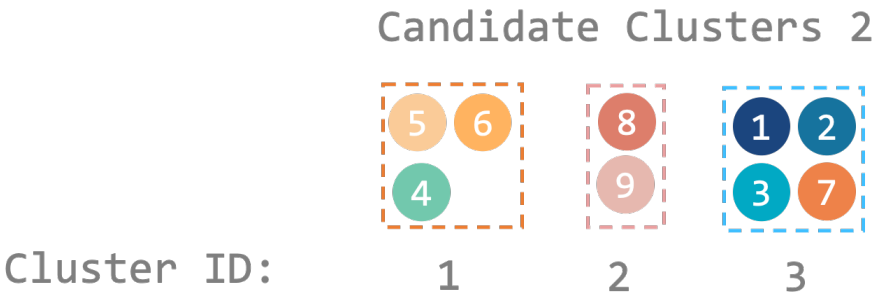
- **How do we measure similarity between two clusters?**



Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

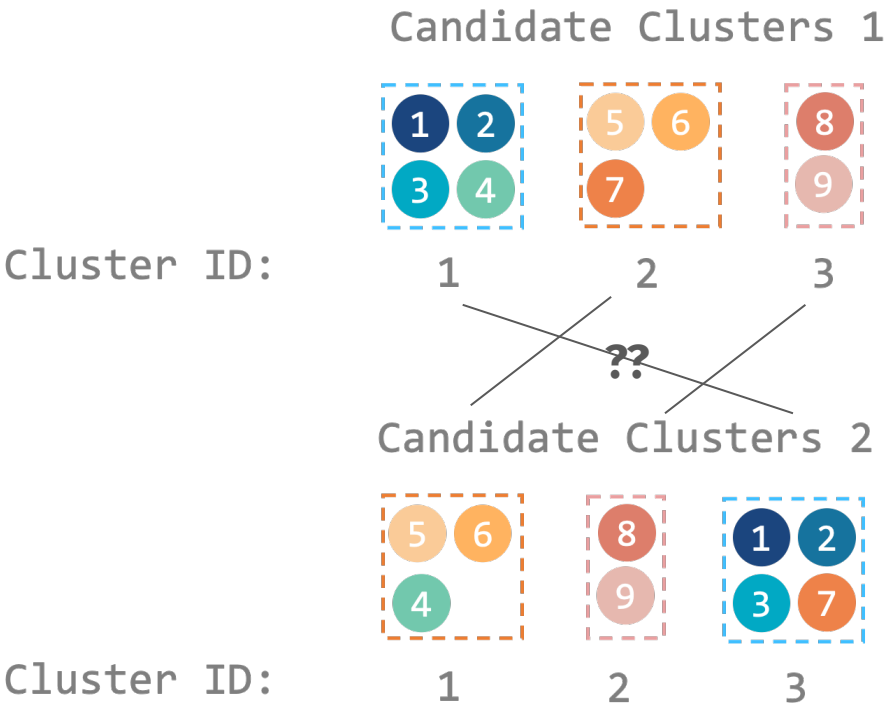
- **How do we measure similarity between two clusters?**





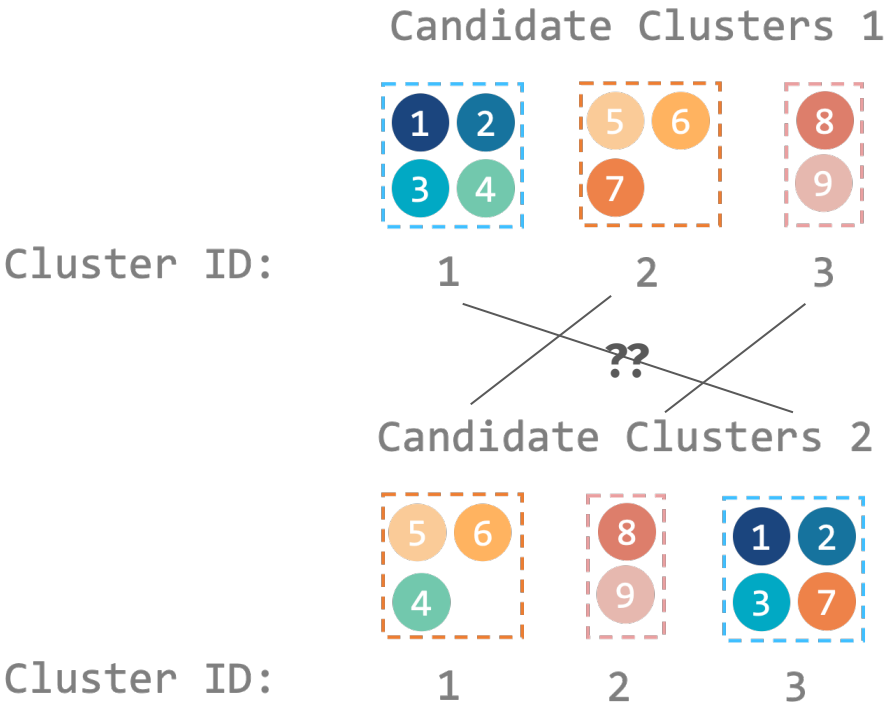
Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**



Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**



Rather than comparing cluster IDs directly, let's compare each point's

neighbors

- *Main Idea:* If my neighbors are similar in the two clusters, then the two clusters are stable.
- *Two metrics:*
  (1) Adjusted Rand Index
  (2) Jaccard Index

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

$$C^{(1)} =$$

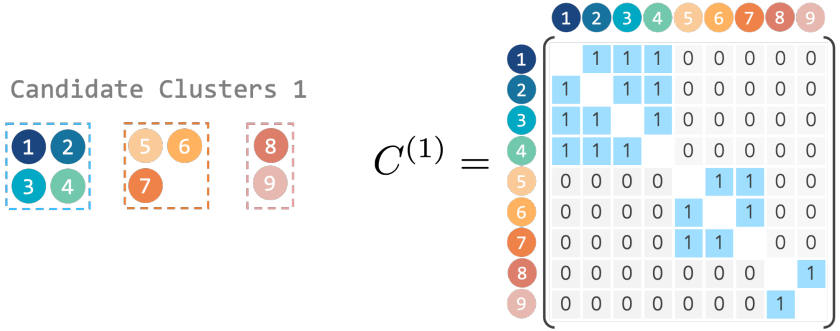
	1	2	3	4	5	6	7	8	9
1		1	1	1	0	0	0	0	0
2	1		1	1	0	0	0	0	0
3	1	1		1	0	0	0	0	0
4	1	1	1		0	0	0	0	0
5	0	0	0	0		1	1	0	0
6	0	0	0	0	1		1	0	0
7	0	0	0	0	1	1		0	0
8	0	0	0	0	0	0	0		1
9	0	0	0	0	0	0	0	1	

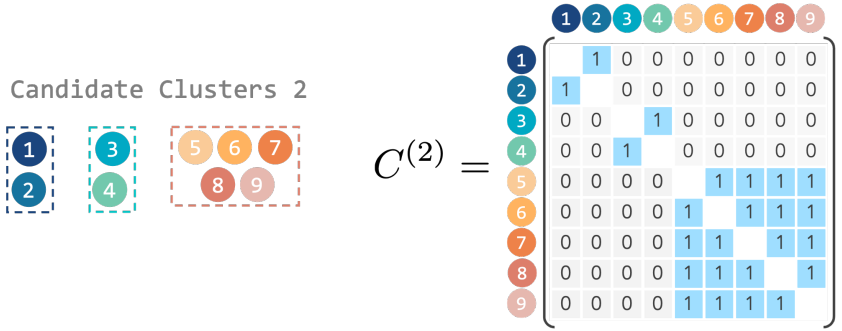
Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

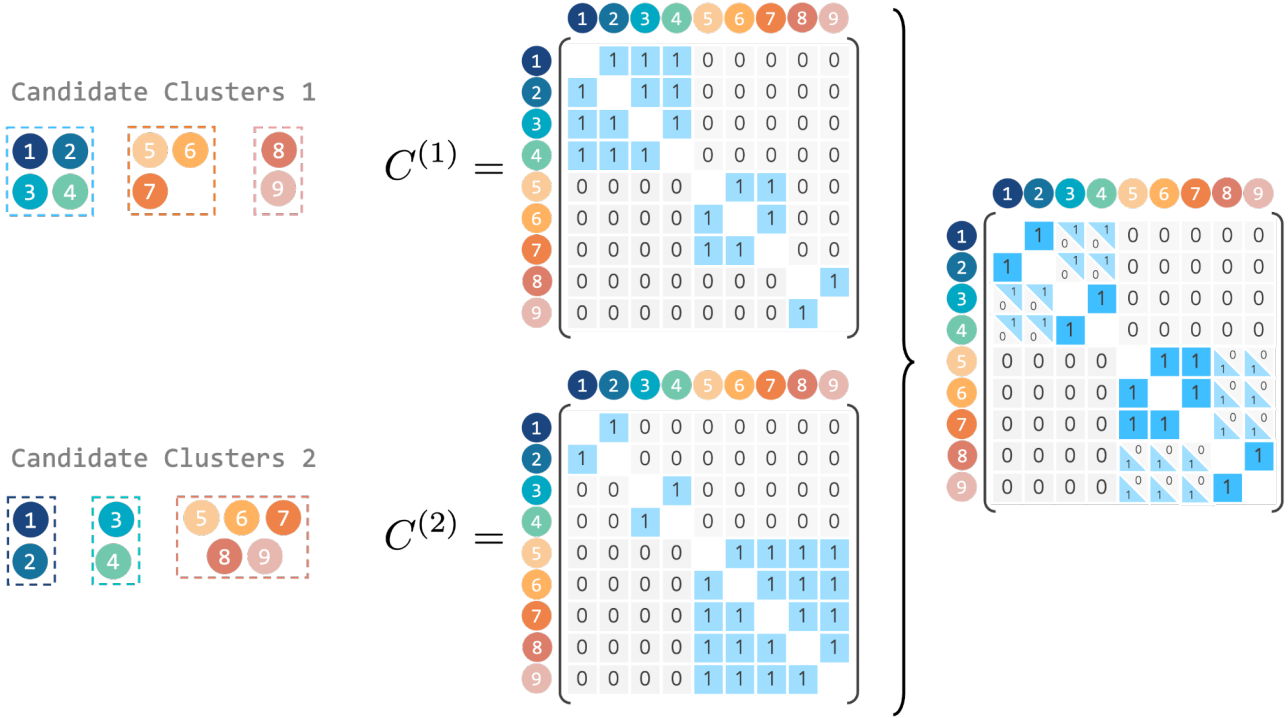
Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.





Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

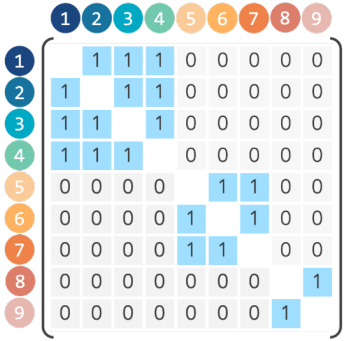


Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

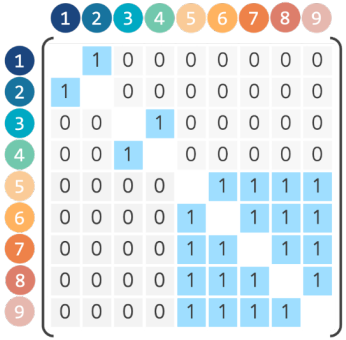
Candidate Clusters 1

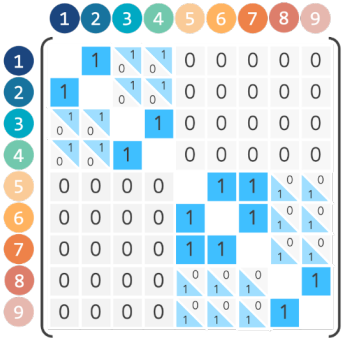
$$C^{(1)} =$$



Candidate Clusters 2

$$C^{(2)} =$$





$N_{11} = 12$ → # of sample pairs that fall in the **same** cluster in **both** $C^{(1)}$ and $C^{(2)}$

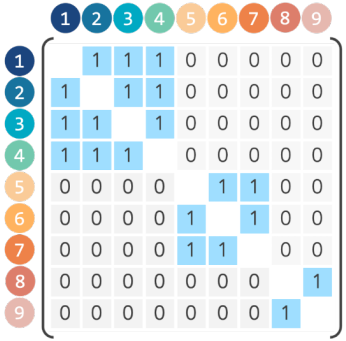
$N_{00} = 40$ → # of sample pairs that fall into **different** clusters in **both** $C^{(1)}$ and $C^{(2)}$

$N_{01} + N_{10} = 20$

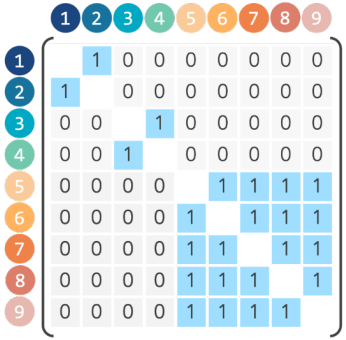
Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

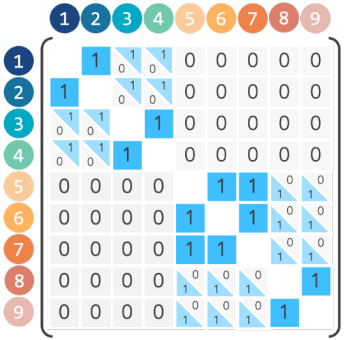
- **How do we measure similarity between two clusters?**
  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

$C^{(1)} =$



$C^{(2)} =$





$N_{11} = 12$

$N_{00} = 40$

$N_{01} + N_{10} = 20$ → # of sample pairs that fall in the same cluster in one $C^{(1)}$ and $C^{(2)}$ but different clusters in the other $C^{(\cdot)}$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

     Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

     Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

     Rand Index: $$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

     Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Rand Index:

$$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

  Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Rand Index:

$$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Number of sample pairs that fall in the same cluster in both $C^{(1)}$ and $C^{(2)}$

where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q,\, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

  Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Rand Index:

$$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Number of sample pairs that fall in the same cluster in both $C^{(1)}$ and $C^{(2)}$

Number of sample pairs that fall into different clusters in both $C^{(1)}$ and $C^{(2)}$

where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q,\, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

  Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Rand Index:

$$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Number of sample pairs that fall in the same cluster in one $C^{(1)}$ and $C^{(2)}$ but different clusters in the other $C^{(.)}$

where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q,\, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

  Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Rand Index: $$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Problem: rand index can be easily dominated by N_{00}

where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

  Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

Rand Index:

$$RI(C^{(1)}, C^{(2)}) = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Jaccard Coefficient:

$$J(C^{(1)}, C^{(2)}) = \frac{N_{11}}{N_{01} + N_{10} + N_{11}}$$

where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q,\, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **How do we measure similarity between two clusters?**
  1. Take each cluster label vector and represent it as an adjacency matrix

$$C_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster and } i \neq j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

  Let $C^{(1)}$ and $C^{(2)}$ be the "C" representations for the two different clusters.

  2. Compute "similarity" between $C^{(1)}$ and $C^{(2)}$

  Adjusted Rand Index: rand index, adjusted for random chance

  Jaccard Coefficient: $J(C^{(1)}, C^{(2)}) = \dfrac{N_{11}}{N_{01} + N_{10} + N_{11}}$

  where $N_{qr} = |\{(i,j) \in [n] \times [n] : C_{ij}^{(1)} = q,\, C_{ij}^{(2)} = r\}|$

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- **Idea:**
  - Measure how clusters change if we had used different subsamples of the data to do the clustering
  - More stable clusters = better

Algorithm 1: Stability Selection for Choosing Number of Clusters

Input: data X, maximum number of clusters under consideration $k_{\max}$, subsampling fraction π, number of repeated subsamples B

1 **for** $k = 2, \ldots, k_{\max}$ **do**
2 **for** $b = 1, \ldots, B$ **do**
3 $sub_1 = \text{subsample}(X, \pi)$
4 $sub_2 = \text{subsample}(X, \pi)$
5 $L_1 = \text{cluster}(sub_1, k)$
6 $L_2 = \text{cluster}(sub_2, k)$
7 $S(k, b) = similarity(L_1[sub_1 \cap sub_2], L_2[sub_1 \cap sub_2])$
8 **end**
9 $S(k) = \frac{1}{B} \sum_{b=1}^{B} S(k, b)$
10 **end**
11 Select k with the highest $S(k)$

Repeat across many many different subsamples

Take two subsamples (with subsampling fraction π)

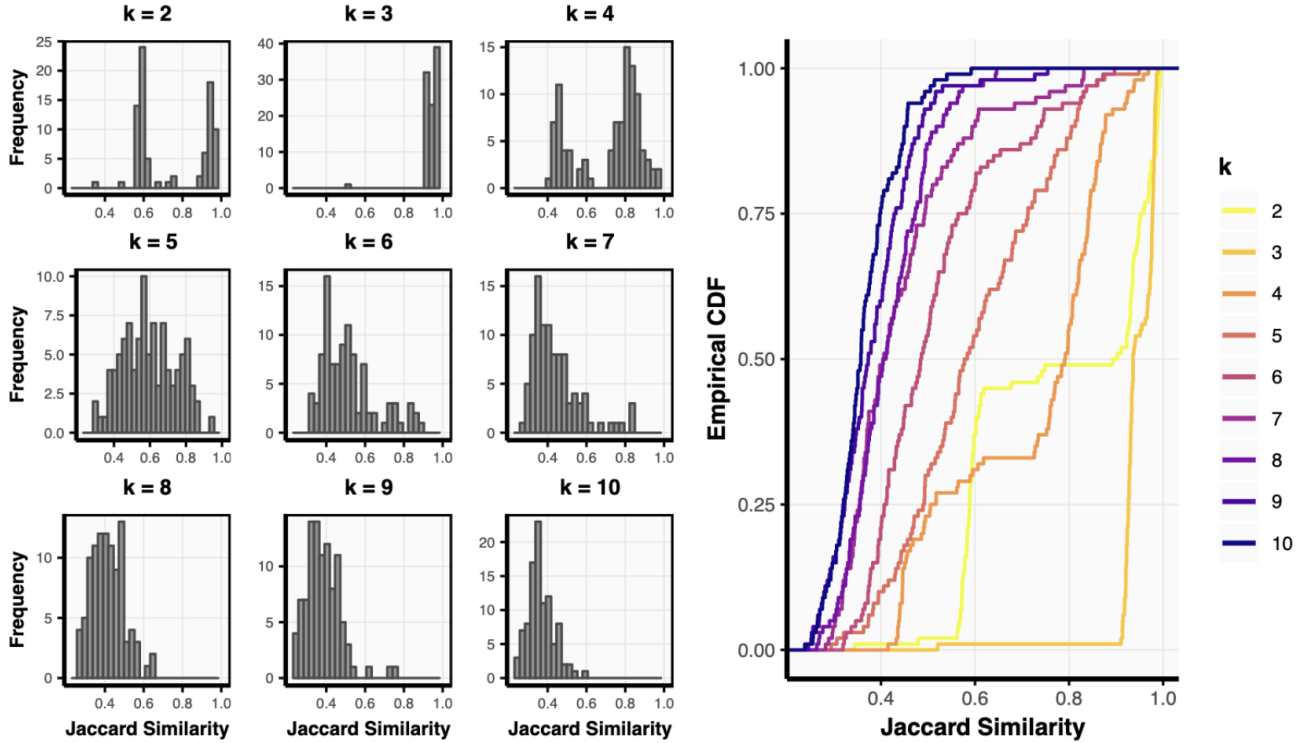
Run clustering on each subsample to get cluster labels L_1 & L_2

Compute similarity between clusters using points common to both subsamples

Choose k that is most stable

Stability Selection for Choosing Number of Clusters [Ben-Hur 2002]

- Higher similarity indicates more stable clusters



There were many human judgment calls throughout this process

- Which method?
- Which hyperparameters?
- How many components?
- How to preprocess our data?
- And more…

It is important to investigate how these decisions impact the dimension reduction and clustering results (i.e., repeat our analysis for different choices)

Linguistics Data Lab

Clustering Activity

Go to https://tiffanymtang.shinyapps.io/dsip_linguistics/

Goals:

1. Determine the number of dialects in the US
2. Support your choice with evidence.

Supervised Learning Questionnaire