Agglomeration

Agenda

- Review
- Correlations
- Normalizing Categorical Data
- Breaking Ties
- Describing Clusters
- Types of clusters
  - Hierarchical
  - Partitioned
- Silhouette Coefficient
- Linkages between Clusters
- Example

Looking Back – k-NN

- What issues are there in using k-NN?
  - value of k
  - distance metric to use
- Which value of k is best?
  - You must build and test
  - use n-fold cross validation to find this
- What is a Voronoi polygon?
  - the graphical solution for 1-NN
  - given a set of points, you should be able to sketch

Review – Data Visualization

- What attributes for data can you use to make patterns pop-out?
  - See notes. There are about 18 of them listed.
- What is the fusiform gyrus?
  - Part of brain used for object recognition and handling
  - You only have one fusiform gyrus
  - One piece of the brain is responsible for handling objects
  - This is why you cannot talk on the phone and drive simultaneously
  - This is why multi-tasking (doing two things at the same time) does not work.
  - There was a talk on this at RIT last Friday by Christopher Chabris "The Illusion of Attention", author of "The Invisible Gorilla"

Interesting Correlations Found

GRADE	TIME	ORD	PEN	GLS	LH	SICK	DVUSR
1	0.61	0.15	0.16	-0.02	0.05	-0.11	-0.21
	1	0.67	-0.04	0.02	0.08	0.18	0.13
		1	-0.19	0.11	0.14	0.27	0.01
			1	-0.17	-0.32	0.04	0.08
				1	0.02	0.02	0.28
					1	-0.12	-0.06
						1	-0.06
							1

GRADE	GRADE on EXAM
TIME	TIME to TAKE EXAM
ORD	ORDER RETURNED
PEN	PEN USER vs PENCIL
GLS	WEARS GLASSES
LH	LEFT HANDED
SICK	SICK DURING EXAM
DVUSR	ELECTRONIC DEVICE USE IN CLASS

So, if you are going to text during class,
you would be better off taking the exam while ill.

Looking Back – Data Normalization

- What are three methods for normalizing numerical Data?
  - z-score (subtract mean, divide by standard deviation
  - dynamic ranging to range of [0,1]
  - center ranging to range of [-1,1]
- How do you normalize categorical data?
  - see next page

More on Normalizing Data

Last lecture we discussed normalizing numerical data.

How do you handle Categorical Data?

Normalizing Categorical Data

1. Nominal
   - Vanilla, chocolate, strawberry
   - In theory, doesn't matter, just give each nominal category a number
   - In reality, you might want to do some pre-processing of your own
   - Put the most common category in the middle
     - ★This way you have differences from the normal

Normalizing Categorical Data

2. Ordinal Data:
   - Assign a number
   - Again, might want to impose a distance between the numbers

Breaking Ties

Kinsman's General Theory of Algorithm Improvement

Any algorithm that relies on
a <= or >= condition,
can be improved by
considering the == case.

Breaking Ties – Similar Concepts

Breaking Ties - Twins

10/21/15 Copyright 2015, T. Kinsman 13

How to break ties, 1

1. Flip a coin (use a random choice)

   No joke.

   Sometimes you break a tie using a random choice.

10/21/15 Copyright 2015, T. Kinsman 14

How to break ties, 2

2. Defer to a secondary (backup) distance metric
   A. A completely different proximity metric
      - King's Move distance
        – Modification of the Manhattan distance
        – One step in any direction (horizontal, vertical, diagonal)
      - cos similarity
      - what else might apply for spell checking?
      - what else would apply for facebook?
      - clustering videos on youtube?
      - clustering songs on ...
   B. OR the distance of a sub-set of *critical* attributes
      - For cars with same mileage and cost, favor cost
   C. OR involve a previously non-critical attribute
      - For cars with same mileage and cost, consider color

How to break ties, 3

3. Methods for breaking ties may differs depending on application:
   - **clustering** – kMeans or kMedoids
     Use another algorithm
   - **classifying** – kNN
     Use another heuristic, such as a larger value of k.

How to break ties, 4

4. Defer to a secondary (backup) attribute:
   - For cars with same mileage and cost, favor color
   - For cameras with the same features, use weight of camera

–The Kids differed to a different attribute ... jar diameter.

DESCRIBING CLUSTERS

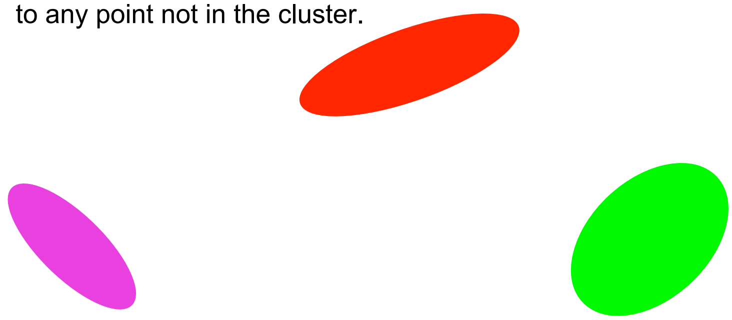
Cluster Vocabulary

Describing Clusters

1. Are they well-separated?
2. Are they center-based clusters?
3. Are they contiguous clusters?
4. Are they density-based clusters?
5. Are they described by an Objective Function?
6. Property or Concept based?

Types of Clusters: Well-Separated

- **Well-Separated Clusters:**
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



–3 well-separated clusters

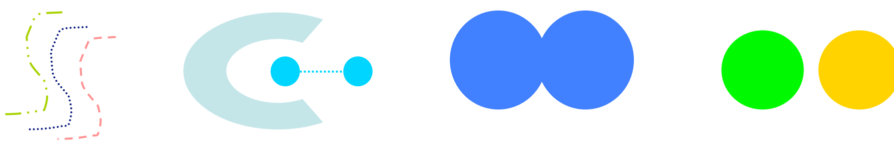
Types of Clusters: Center-Based

- **Center-based**
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - Clusters are circular or elliptical
  - The prototype center of a cluster is often either:
    - a centroid, the average of all the points in the cluster
    - or a medoid, the most central data point

–4 center-based clusters

Types of Clusters: Contiguity-Based

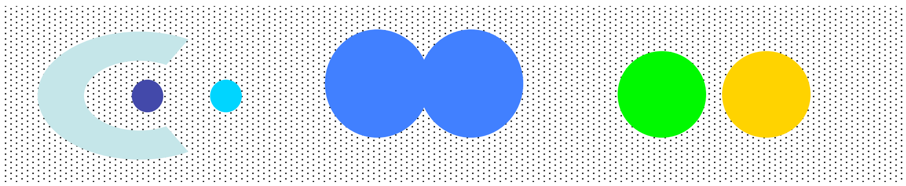
- **Contiguous Cluster**
- A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



–8 contiguous clusters

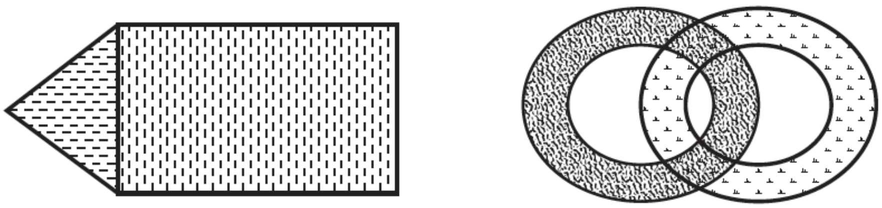
Types of Clusters: Density-Based

- **Density-based**
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.
  - Think "contour map"



–6 density-based clusters

Conceptual Clusters

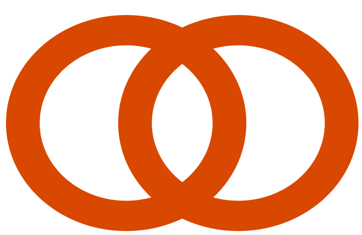


(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

Types of Clusters: Conceptual Clusters

Shared Property or Conceptual Clusters

- Finds clusters that share some common property or represent a particular concept.



–2 Overlapping Circles

- **Examples:**
  - **taxonomies**
  - **meteorological models**
  - **genetic clustering**
  - **Business modeling**

Types of Clusters: Objective Function Based

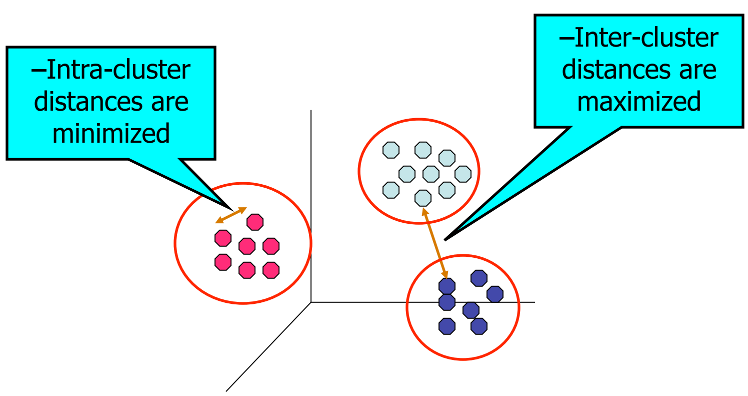
Shared attempt to minimize, or maximize, some mathematical measure of a function.

- Maximize inter-cluster distance
- Minimize intra-cluster distance
- Etc...
- Minimize the Sum of the Squared Errors

- **Ward's method** is one method that uses this, we mention it here briefly, and hit it later.

As before – Ideal Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



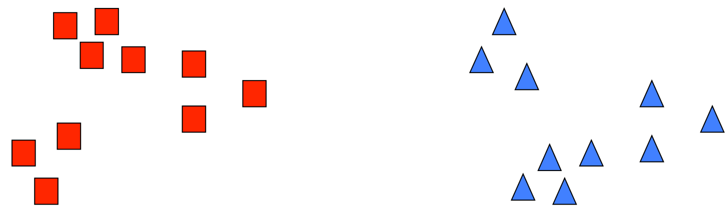
What is not Clustering?

- Supervised classification
  - Uses a class label
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
  - Tells us little about the structure of the data
  - Does not especially use a *measurement* of the students
- Results of a query
  - Groupings are a result of an external specification

The Ambiguity of Clusters

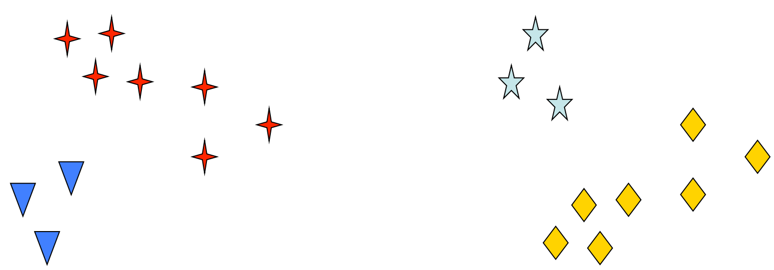
–How many clusters?

The Ambiguity of Clusters



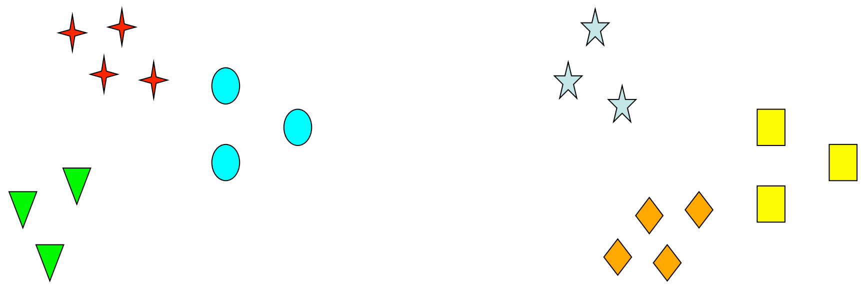
–Two Clusters

The Ambiguity of Clusters



–Four Clusters

The Ambiguity of Clusters



–Six Clusters

The Ambiguity of Clusters

–How many clusters?

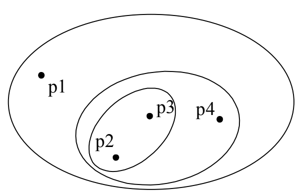
– Obi-Wan: Luke, you're going to find that many of the truths we cling to depend greatly on our own point of view. Anakin was a good friend. When I first met him, your father was already a great pilot. But I was amazed how strongly the Force was with him. I took it upon myself to train him as a Jedi. I thought that I could instruct him just as well as Yoda. I was wrong.

–The number of clusters depends greatly on what you want to do with them.

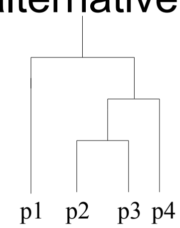
Clustering Vocabulary

- A clustering is a set of clusters
- Distinction between hierarchical and partitional clusters:
  - Hierarchical clustering:
    - A set of nested clusters organized as a hierarchical tree
    - Agglomerative clustering is an example
    - Emphasizes the inclusion in the clusters
  - Partitional Clustering:
    - Division of data points into non-overlapping clusters such that each data point is in exactly one cluster
    - Emphasizes the boundaries between clusters

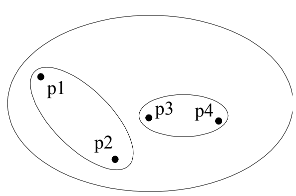
Two Hierarchical Clusterings: traditional versus alternative



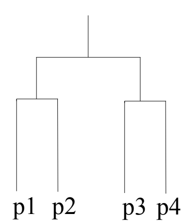
Traditional Hierarchical Clustering



Traditional Dendrogram

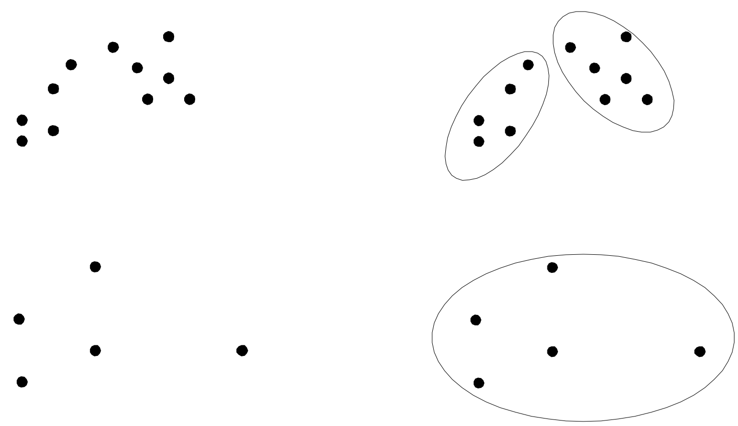


Alternative Hierarchical Clustering



Alternative Dendrogram

Partitional Clustering – unique clusters



–Original Points

–A Partitional Clustering

Other Distinctions Between Sets of Clusters

- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or ‘border’ points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Cluster of widely different sizes, shapes, and densities

Internal Measures: Cohesion vs. Separation

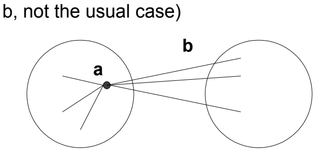
- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters

Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points
- For an individual point, *i*
  - Calculate *a* = **average** distance of *i* to the points in the same cluster
  - Calculate *b* = **min**(average distance of *i* to points in any other cluster)
  - The silhouette coefficient for a point is then given by

    s = 1 – a/b if a < b, (or s = b/a - 1 if a ≥ b, not the usual case)

  - Typically between 0 and 1.
  - The closer to 1 the better.



Review of Measures of Central Tendency

1. **Mean** – the "average"
   - The mean can result in a representative value that is not in your data set.
   - Quick example: avg(1,2) = 1.5 → a non-integer number.
   - And you will want a std to go with that.
2. **Mode** – the most common value.
3. **Median** – the central value.
4. **Centroid** – let's talk about the centroid.
5. **Medoid** – what's that?

The two Hierarchical Clustering Approaches

Goal is to "natural" groupings

- **Two approaches:**

  1. **Top down – Divisive**
     Assign all data to one cluster, and divide into smaller and smaller pieces.

     Example: divisive k-Means.

  2. **Bottom up – Agglomerative - Today**
     Each record or data point starts with its own data cluster. Clusters are then combined into bigger clusters.

How do most people solve jig-saw puzzles?

- If you ask them – they will say that they separate out the edge pieces first.
- BUT FIRST THEY DO SOMETHING MYSTERIOUS –
  1. They turn all of the pieces picture-side up.
  2. They also twist each the piece so that each piece is as “right side” up as possible. In other words, they set each piece down so that the “top” is towards the “top” of the picture they are building.
- They actually do some feature selection that they are not aware of!
- These are decisions based on knowledge they are not aware they have – a hunch.

How do most people solve jig-saw puzzles –

1. They turn all of the pieces picture-side up.
2. They also twist each the piece so that each piece is as “right side” up as possible. In other words, they set each piece down so that the “top” is towards the “top” of the picture they are building.
3. They separate out the “straight edges”.
4. They group the pieces by colors.
5. They group the pieces by texture patterns.
6. This is all divisive clustering – breaking down the entire set of pieces into smaller manageable clusters.
7. THEN –
   they start assembling the pieces into one picture.
   This is agglomerative.
8. One of the features used here is the “shape context”
   of each piece.

How do most people solve jig-saw puzzles –

1. They turn all of the pieces picture-side up.
2. They also twist each the piece so th… is as "right side" up as possible. In other… own so that the "top" i… uilding.
3.
4.
5. T
6. T… ire set of pieces in
7. Th… they start assembling the pieces into one picture. This is agglomerative.
8. One of the features used here is the "shape context" of each piece.

- **There are actual algorithms that switch from divisive to agglomerative clustering.**
- **They stop when things "settle down".**

The Generalized Theory of Clustering

- **IN THEORY** –
  In an ideal world…
  If the data was perfect,
  and the clustering methods were perfect,
  then it would not matter which clustering technique was used.
- This is important, because it allows us to compare clustering methods, even when nobody has any idea what the correct classification is.
- So, don't worry if you choose the "wrong" clustering method, choose a several of them and compare results.

Generalized Clustering

- Select best features to measure
- Start the clustering process
- Refine the clustering:
  - Decide how to form a new cluster
  - Decide how to measure things
- Decide if you want to refine your clustering – has it become lop-sided?.
- Decide if you are ready to stop
- Repeat if necessary

Agglomerative Clustering

1. Select most relevant attributes or features
2. Assign each item to its own cluster of one.
3. Find the <u>most similar</u> pair (usually the closest pair, but not always).
   a. Merge them into a single cluster
   b. Update the cluster prototype
4. Compute the new distances between the new cluster and every other cluster. (Update all distances)
5. Repeat steps 3 and 4 until all items are clustered.
   - Forms a <u>dendrogram</u>.
   - How you do steps 3 and 4 here is the difference between single-link, complete-link, and average link.

Agglomerative Design Decisions

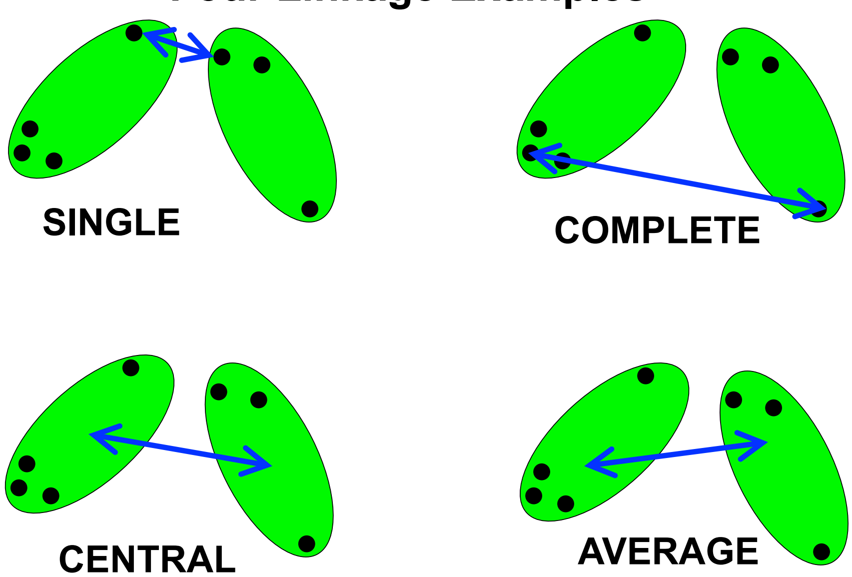
1. What features will you use?
2. What distance metric will you use?
3. How do you describe clusters?
   What cluster prototype will you use?
4. How do you compute the most similar pair?
   What linkage method will you use?
5. When do you stop?

Linkages

Agglomerative (Bottom Up) Hierarchical Clustering

- **Inter-cluster distances – the the most common ones:**
  1. Single link -- shortest distance
  2. Complete link -- longest distance
  3. Average link -- average distance
  4. Central linkage -- between centers
- Called the “linkage” between clusters.

Four Linkage Examples



Clustering Cities –

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS		206	429	1504	963	2976	3095	2979	1949
NY	206		233	1308	802	2815	2934	2786	1771
DC	429	233		1075	671	2684	2799	2631	1616
MIA	1504	1308	1075		1329	3273	3053	2687	2037
CHI	963	802	671	1329		2013	2142	2054	996
SEA	2976	2815	2684	3273	2013		808	1131	1307
SF	3095	2934	2799	3053	2142	808		379	1235
LA	2979	2786	2631	2687	2054	1131	379		1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	

Single Link Clustering Example

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS		206	429	1504	963	2976	3095	2979	1949
NY	206		233	1308	802	2815	2934	2786	1771
DC	429	233		1075	671	2684	2799	2631	1616
MIA	1504	1308	1075		1329	3273	3053	2687	2037
CHI	963	802	671	1329		2013	2142	2054	996
SEA	2976	2815	2684	3273	2013		808	1131	1307
SF	3095	2934	2799	3053	2142	808		379	1235
LA	2979	2786	2631	2687	2054	1131	379		1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	

Single Link Clustering Example

	BOS /NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY		233	1308	802	2815	2934	2786	1771
DC	233		1075	671	2684	2799	2631	1616
MIA	1308	1075		1329	3273	3053	2687	2037
CHI	802	671	1329		2013	2142	2054	996
SEA	2815	2684	3273	2013		808	1131	1307
SF	2934	2799	3053	2142	808		379	1235
LA	2786	2631	2687	2054	1131	379		1059
DEN	1771	1616	2037	996	1307	1235	1059	

Single Link Clustering Example

	3OS /NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/D|)		1075	671	2684	2799	2631	1616
MIA	1075		1329	3273	3053	2687	2037
CHI	671	1329		2013	2142	2054	996
SEA	2684	3273	2013		808	1131	1307
SF	2799	3053	2142	808		379	1235
LA	2631	2687	2054	1131	379		1059
DEN	1616	2037	996	1307	1235	1059	

Single Link Clustering Example

	3OS /NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/D	)	1075	671	2684	2631	1616
MIA	1075		1329	3273	2687	2037
CHI	671	1329		2013	2054	996
SEA	2684	3273	2013		808	1307
SF/LA	2631	2687	2054	808		1059
DEN	1616	2037	996	1307	1059	

Single Link Clustering Example

	3OS /NY/DC	MIA	SEA	SF/LA	DEN
BOS/NY/D	)	1075	2013	2054	996
MIA	1075		3273	2687	2037
SEA	2013	3273		808	1307
SF/LA	2054	2687	808		1059
DEN	996	2037	1307	1059	

Single Link Clustering Example

	BOS /NY/D	MIA	SF/LA	DEN
BOS/NY/D	)	1075	2013	996
MIA	1075		2687	2037
SF/LA	2054	2687		1059
DEN	996	2037	1059	

Single Link Clustering Example

Merge BOS/NY/DC/CHI/DEN

	BOS /NY/D	MIA	SF/LA
BOS/NY/D	)	1075	1059
MIA	1075		2687
SF/LA	1059	2687	

Single Link Clustering Example

	BOS /NY/D	MIA
BOS/NY/D	)	1075
MIA	1075	

Real Distances aren't always nice

DEST. CITY OF ORIGIN

	BOS	NYC	DC	CHI	MIA	SEA	SFO	LAX	DEN
BOS	-	70	103	132	185	[illegible]	325	325	226
NYC	70	-	85	[illegible]	[illegible]	[illegible]	322	312	215
DC	105	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	294	195
CHI	167	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	239	140
MIA	190	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	310	285	221
SEA	378	372	[illegible]	[illegible]	405	-	130	160	173
SFO	380	370	[illegible]	270	360	130	-	81	163
LAX	370	372	348	265	325	156	80	-	150
DEN	277	286	241	150	250	155	160	142	-

What is this distance metric?

It is supposed to be symmetric!!

Real Distances aren't always nice –

DEST.	CITY OF ORIGIN								
	BOS	NYC	DC	CHI	MIA	SEA	SFO	LAX	DEN
BOS	-	70	103	132	185	318	325	325	226
NYC	70	-	85	126	170	299	322	312	215
DC	105	89	-	109	145	289	[illegible]	[illegible]	[illegible]
CHI	167	151	119	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
MIA	190	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
SEA	378	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
SFO	380	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
LAX	370	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DEN	277	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	160	142	-

- These are the domestic air flight times, in minutes!
- The table is not symmetric due to prevailing westerly winds.

74

Review

- Clustering is <u>un</u>supervised learning
- Two hierarchical clustering algorithms:
  - Top down – divisive
  - Bottom up – agglomerative, discussed today

Review

- Five decisions to make:
  1. Attributes / features to use?
  2. Distance metric to use?
  3. How to assign new cluster center?
  4. How to determine distance between clusters? Linkage
  5. When do you stop?

Review

- Agglomerative Clustering:
  1. Start with each point in its own cluster
  2. Merge the two closest points into a cluster based on your pre-defined concept of proximity
  3. Update the prototype or model for that new cluster
  4. Update the distances from this new cluster to all other clusters

END