John Oliver from "The Daily Show"

Supporting *worthy causes* at the **G20** Pittsburgh Summit:

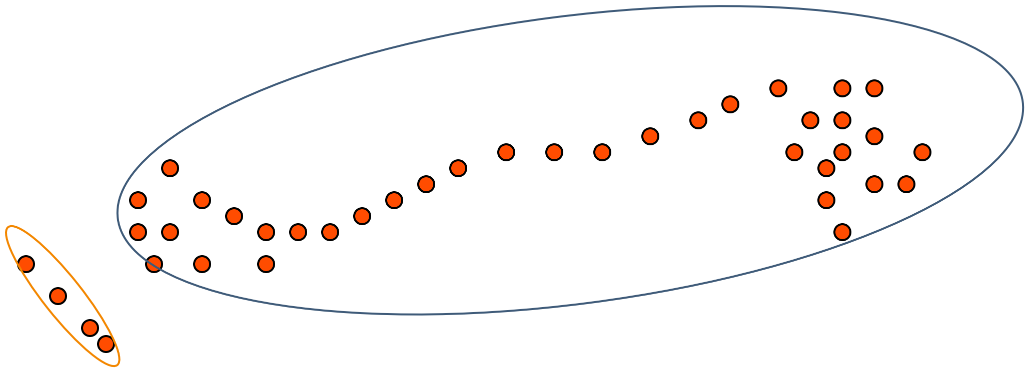
"Bayesians Against Discrimination"

"Ban Genetic Algorithms"

"Support Vector Machines"

Watch out for the protests tonight on The Daily Show!

Picture: Arthur Gretton



TOWARDS A PRINCIPLED THEORY OF CLUSTERING

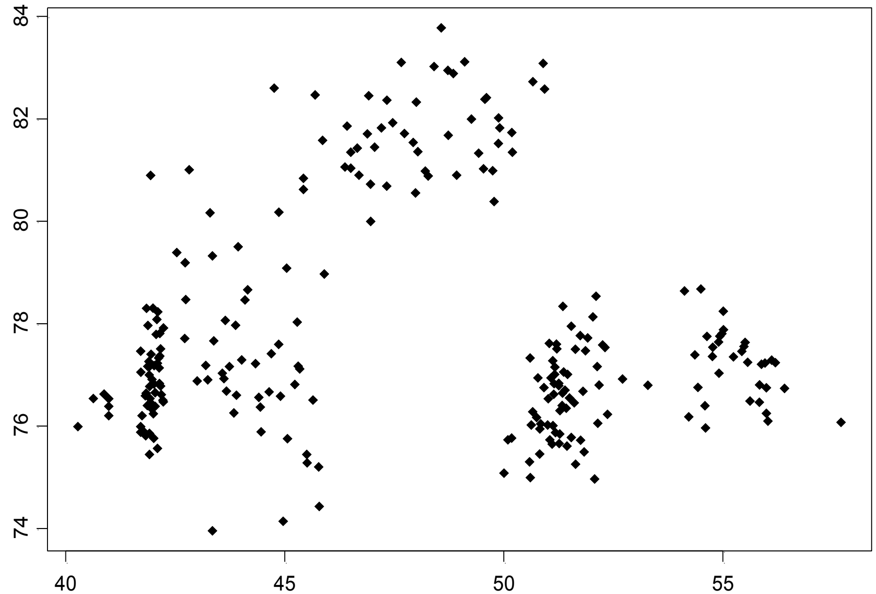
Reza Bosagh Zadeh

(Joint with Shai Ben-David)

CMU Machine Learning Lunch, September 2009

WHAT *IS* CLUSTERING?

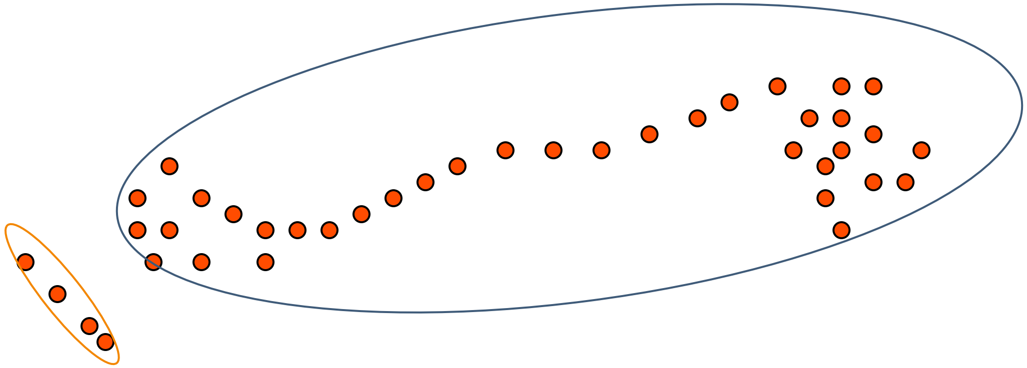
- Given a collection of objects (characterized by feature vectors, or just a matrix of pair-wise similarities), detects the presence of distinct groups, and assign objects to groups.



THERE ARE MANY CLUSTERING TASKS

"Clustering" is an ill-defined problem

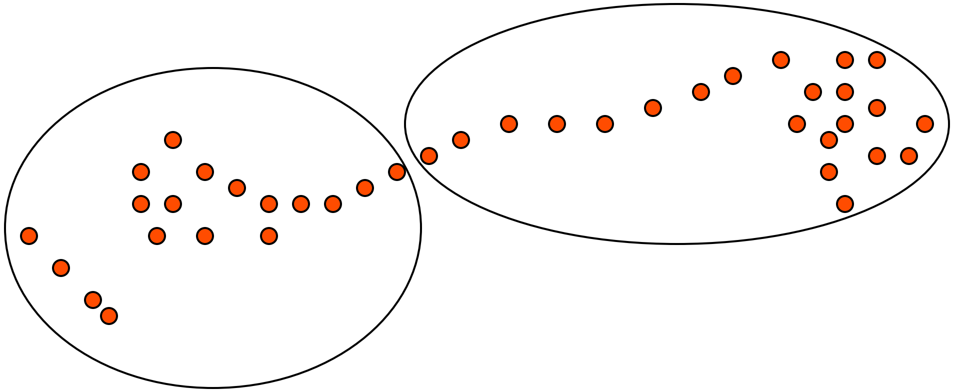
- There are many different clustering tasks, leading to different clustering paradigms:



THERE ARE MANY CLUSTERING TASKS

"Clustering" is an ill-defined problem

- There are many different clustering tasks, leading to different clustering paradigms:



TALK OUTLINE

- Questions being addressed
- Introduce Axioms & Properties
- Characterization for Single-Linkage and Max-Sum
- Taxonomy of Partitioning Functions

SOME BASIC UNANSWERED QUESTIONS

- Are there principles governing ***all*** clustering paradigms?
- Which clustering paradigm should I use for a given task?

WE WOULD LIKE TO DISCUSS THE BROAD NOTION OF CLUSTERING

Independently *of any*

- ✓ particular algorithm,
- ✓ particular objective function, or
- ✓ particular generative data model

WHAT FOR?

- *Choosing a suitable algorithm for a given task.*
  - *Axioms:* to capture intuition about clustering *in general.*
    Expected to be satisfied by all clustering paradigms
  - *Properties:* to capture differences between different clustering paradigms

TIMELINE – AXIOMATIC APPROACH

- Jardine, Sibson 1971
  - Considered only hierarchical functions
- Kleinberg 2003
  - Presented an impossibility result
- Ackerman, Ben-David 2008
  - Clustering Quality measures formalization

These are only axiomatic approaches, there are other ways of building a principled theory for Clustering,
e.g. Balcan, Blum, Vempala STOC 2008

THE BASIC SETTING

- For a finite domain set A, a *similarity function* $s(x,y)$ is a symmetric mapping to a similarity score

  $s(x,y) > 0$, and

  $s(x,y) \rightarrow \infty$ iff $x=y$

- A *partitioning function* takes a similarity function and returns a partition of A.

- *We wish to define axioms that distinguish clustering functions, from other partitioning functions.*

KLEINBERG'S AXIOMS (NIPS 2001)

- *Scale Invariance*

  $F(\lambda s)=F(s)$ for all s and all strictly positive λ.

- *Richness*

  The range of $F(s)$ over all s is the set of all possible partitionings

- *Consistency*

If s' equals s except for increasing similarities within clusters of $F(s)$ or decreasing between-cluster similarities,

then $F(s) = F(s')$.

KLEINBERG' S AXIOMS (NIPS 2001)

- *Scale Invariance*

  $F(\lambda s)=F(s)$ for all s and all strictly positive λ.

- *Richness*

  The range of $F(s)$ over all s is the set of all possible partitionings

- *Consistency*

If s' equals s except for increasing similarities within clusters of $F(s)$ or decreasing between-cluster similarities,

then $F(s) = F(s')$.

Inconsistent! No algorithm can satisfy all 3 of these.

KLEINBERG' S AXIOMS (NIPS 2001)

- *Scale Invariance*
  $F(\lambda s)=F(s)$ for all s and all strictly positive λ.

- *Richness*
  The range of $F(s)$ over all s is the set of all possible partitionings

- *Consistency*

If s' equals s except for increasing similarities within clusters of $F(s)$ or decreasing between-cluster similarities,
then $F(s) = F(s')$.

Proof:

CONSISTENT AXIOMS (UAI 2009)

Fix k

- *Scale Invariance*

  $F(\lambda s, k)=F(s, k)$ for all d and all strictly positive λ.

- *k-Richness*

  The range of $F(s, k)$ over all s is the set of all possible k-partitionings

- *Consistency*

If s' equals s except for increasing similarities within clusters of $F(s, k)$ or decreasing between-cluster similarities,

then $F(s, k)=F(s', k)$.

Consistent! (And satisfied by Single-Linkage, Max-Sum, …)

CLUSTERING FUNCTIONS

- **Definition.** Call any partitioning function which satisfies
  - *Scale Invariance*
  - *k-Richness*
  - *Consistency*

  a ***Clustering Function***

TWO CLUSTERING FUNCTIONS

Single-Linkage

1. Start with with all points in their own cluster
2. While there are more than k clusters

   Merge the two most similar clusters

Similarity between two clusters is the similarity of the most similar two points from differing clusters

Hierarchical

Max-Sum k-Clustering

Find the k-partitioning Γ which maximizes

$$\Lambda_s(\Gamma) = \sum_{C \in \Gamma} \sum_{i,j \in C} s(i,j)$$

Not Hierarchical

(Is NP-Hard to optimize)

Both Functions satisfy:

- *Scale Invariance*
- *k-Richness*
- *Consistency*

Proofs in paper.

CLUSTERING FUNCTIONS

- Single-Linkage and Max-Sum are both Clustering functions.
- How to distinguish between them in an Axiomatic framework? Use *Properties*
- *Not all properties are desired in every clustering situation: pick and choose properties for your task*

PROPERTIES - ORDER-CONSISTENCY

- *Order-Consistency*
  If two datasets ***s*** and ***s'*** have the same ordering of similarity scores, then for all ***k***,
  F(s, k)=F(s', k)

- In other words the clustering function only cares about whether a pair of points are more/less similar than another pair of points.
- i.e. Only relative similarity matters.
- Satisfied by Single-Linkage, Max-Linkage, Average-Linkage…
- NOT satisfied by most objective functions (Max-Sum, k-means, …)

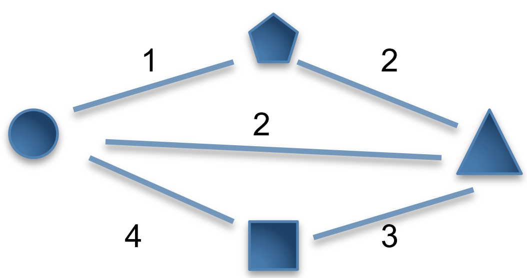
PATH-SIMILARITY

Given a similarity measure, s over some domain set X, we define the *s-induced path similarity*, P_s, by setting, for all $x, y \in X$,

$$P_s(x,y) = \max_{q \in P_{x,y}} \min_{i<|q|} s(q(i), q(i+1))$$

In other words, we find the path from *x* to *y*, which has the largest bottleneck.

e.g.

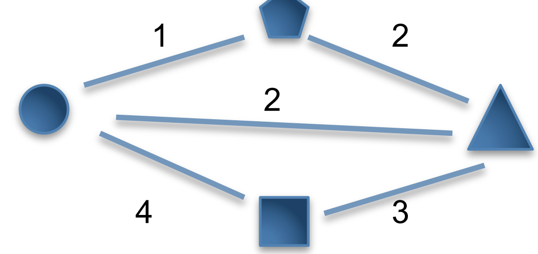


Undrawn edges are small

P_s( , ) = 3

Since the path through the bottom has bottleneck of 3

PATH-SIMILARITY



$P_s($ ● , ▲ $) = 3$

- Imagine each point is an island out in the ocean, with bridges that have some weight restriction, and we would like to go from island ● to island ▲
- Having some mass, we are restricted in which bridges we can take from island to island.
- Path-Similarity would have us find the path with the largest bottleneck, ensuring that we could complete all the crossings successfully, or fail if there is no path with a large enough bottleneck

PROPERTIES – PATH-SIMILARITY COHERENCE

- *Path-Similarity Coherence*

  If two datasets ***s*** and ***s'*** have the same induced-path-similarity edge ordering then for all ***k***, ***F(s, k)=F(s', k)***

UNIQUENESS THEOREM: SINGLE-LINKAGE

- **Theorem** (Bosagh Zadeh 2009)
  - Single-Linkage is the *only clustering function* satisfying Path-Similarity-Coherence

UNIQUENESS THEOREM: SINGLE-LINKAGE

- **Theorem** (Bosagh Zadeh 2009)
  - Single-Linkage is the *only clustering function* satisfying Path-Similarity-Coherence
- Is Path-Similarity-Coherence doing all the work? No.
  - Consistency is necessary for uniqueness
  - k-Richness is necessary for uniqueness
- "X is Necessary": All other axioms/properties satisfied, just X missing, still not enough to get uniqueness

UNIQUENESS THEOREM: MAX-SUM

- Time to characterize another clustering function
- Use a different *property* in lieu of path-similarity
- Turns out generalizing Path-Similarity does the trick.

GENERALIZED PATH SIMILARITY

Given a similarity measure, s over some domain set X, we define the s-*induced generalized path similarity*, P_s, by setting, for all $x, y \in X$,

$$P_s^{\oplus}(x,y) = \max_{a \in E_{x,y}} \bigoplus_{q \in a} \min_{i<|q|} s(q(i), q(i+1))$$

Claims:

- If $\oplus$ is the max operator, then $P_s^{\max}(x,y)$ defines the regular path similarity between x and y.
- If $\oplus$ is the Σ operator, then $P_s^{\Sigma}(x,y)$ defines the maximum flow between x and y.

UNIQUENESS THEOREMS

- **Theorem**
  - Single-Linkage is ***the*** clustering function satisfying $P_s^{\max}$ -Coherence

- **Theorem**
  - Max-Sum is ***the*** clustering function satisfying P_s^{Σ} -Coherence

For two-class Clustering (k=2) only

TWO CLUSTERING FUNCTIONS

Single-Linkage

1. Start with with all points in their own cluster

2. While there are more than k clusters

   Merge the two most similar clusters

Similarity between two clusters is the similarity of the most similar two points from differing clusters

Max-Sum k-Clustering

Find the k-partitioning Γ which maximizes

$$\Lambda_s(\Gamma) = \sum_{C \in \Gamma} \sum_{i,j \in C} s(i,j)$$

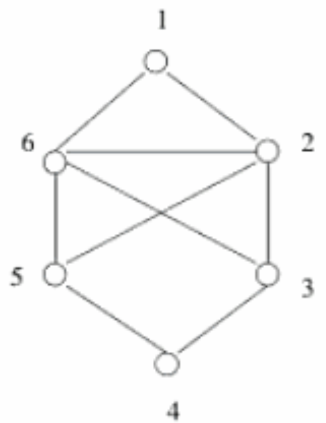
Can use Uniqueness Theorems as alternate definitions to replace these definitions that on the surface seem unrelated.

PRACTICAL CONSIDERATIONS

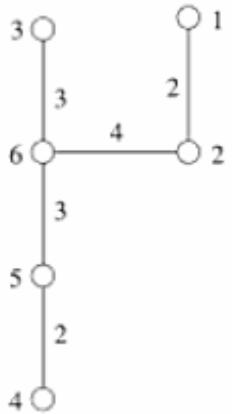
- Single-Linkage, or Max-Sum are *not* always the right functions to use
  - Because Generalized Path-Similarity is not always desirable.
- It's not always immediately obvious when we want a function to focus on the Generalized Path Similarity
  - Introduce a different formulation involving Tree Constructions

ASIDE: MINIMUM CUT TREES

Graph G on 6 nodes



Min-Cut Tree for G



Nodes in Min-Cut tree correspond to nodes in G, but edges do not.

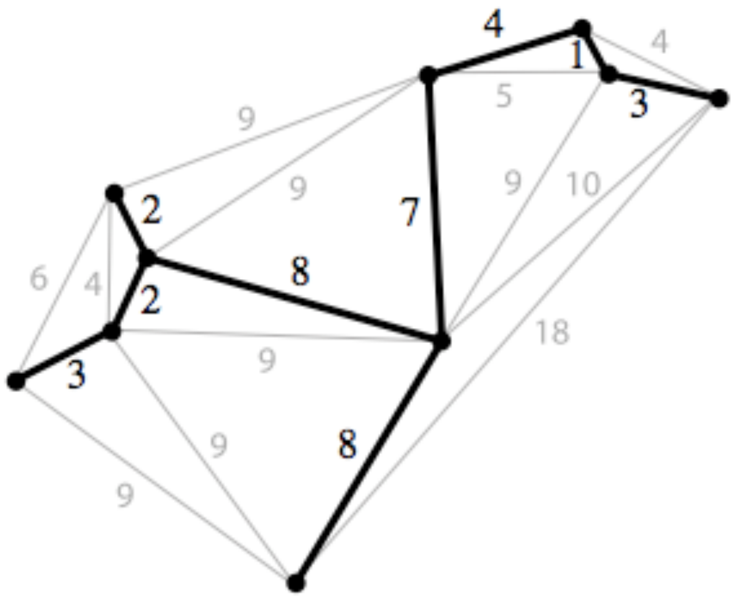
- Min-Cut tree can be computed in at most n-1 Min-Cut Max-Flow computations!

- Weight of Min-Cut between nodes x and y is weight of <u>smallest edge on the unique x-y path</u>

- Cutting that edge will give the two sides of the cut in the original graph

Picture: Encyclopedia of Algorithms

ASIDE: MAXIMUM SPANNING TREES



Bold: Minimum Spanning Tree of the graph

Spanning Tree: Tree Sub-graph of original graph which touches all nodes. Weight of tree is equal to sum of all edge weights.

Spanning Trees ordered by weight, we are interested in the Maximum Spanning Tree

Picture: Wikipedia

PROPERTIES - MST-COHERENCE

- *MST-Coherence*
  If two datasets ***s*** and ***s'*** have the same
  Maximum-Spanning-Tree edge ordering
  then for all ***k***, ***F(s, k)=F(s', k)***

- *MCT-Coherence*
  If two datasets ***s*** and ***s'*** have the same
  Minimum-Cut-Tree edge ordering
  then for all ***k***, ***F(s, k)=F(s', k)***

PROPERTIES - MST-COHERENCE

- *MST-Coherence*
  If two datasets ***s*** and ***s'*** have the same Maximum-Spanning-Tree edge ordering then for all ***k***, ***F(s, k)=F(s', k)***

Characterizes Single-Linkage

- *MCT-Coherence*
  If two datasets ***s*** and ***s'*** have the same Minimum-Cut-Tree edge ordering then for all ***k***, ***F(s, k)=F(s', k)***

Characterizes Max-Sum

The uniqueness theorems apply in the same way to the tree constructions

A TAXONOMY OF CLUSTERING FUNCTIONS

	Scale-Invariance	Consistency	k-Richness	MST-Coherence	Order-Consistency
Single-Linkage	✓	✓	✓	✓	✓
MST cuts family	✓	×	✓	✓	✓
Min-Sum k-clustering	✓	✓	✓	×	×
Constant partitioning	✓	✓	×	✓	✓

- Min-Sum satisfies neither MST-Coherence nor Order-Consistency
- Future work: Characterize other clustering functions

TAKEAWAY LESSONS

- Impossibility result wasn't too bad
- Can go a long way by fixing k
- Uniqueness theorems can help you decide when to use a function
- An axiomatic approach can bring out underlying motivating principles,
  Which in the case of Max-Sum and Single-Linkage are very similar principles

CLUSTERING THEORY WORKSHOP

- Axiomatic Approach is only one approach
- There are other approaches.
- Come hear about them at our workshop

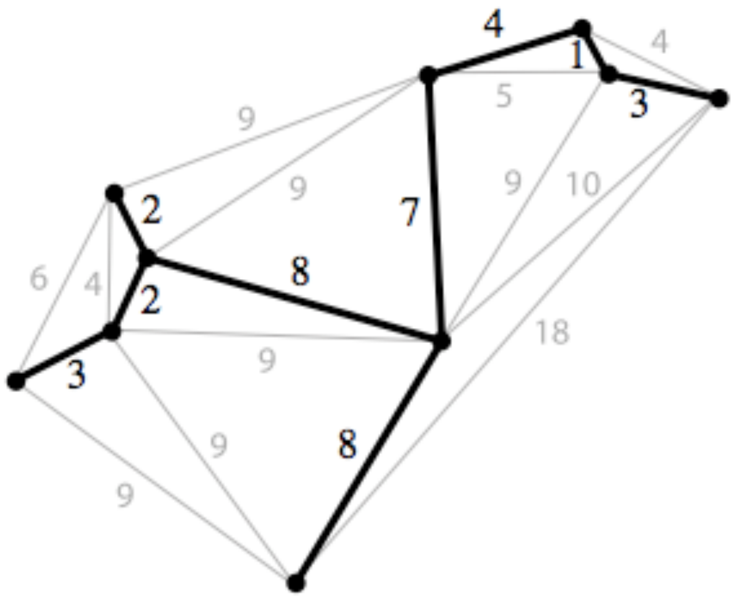
ClusteringTheory.org

NIPS 2009 Workshop

Deadline: 30^{th} October 2009

THANKS FOR YOUR ATTENTION!

ASIDE: MINIMUM SPANNING TREES



Bold: Minimum Spanning Tree of the graph

Spanning Tree: Tree Sub-graph of original graph which touches all nodes. Weight of tree is equal to sum of all edge weights.

Spanning Trees ordered by weight, we are interested in the Minimum Spanning Tree

Picture: Wikipedia

PROOF OUTLINE: CHARACTERIZATION OF SINGLE-LINKAGE

1. Start with arbitrary d, k

2. By k-Richness, there exists a d_1 such that
   $F(d_1, k) = SL(d, k)$

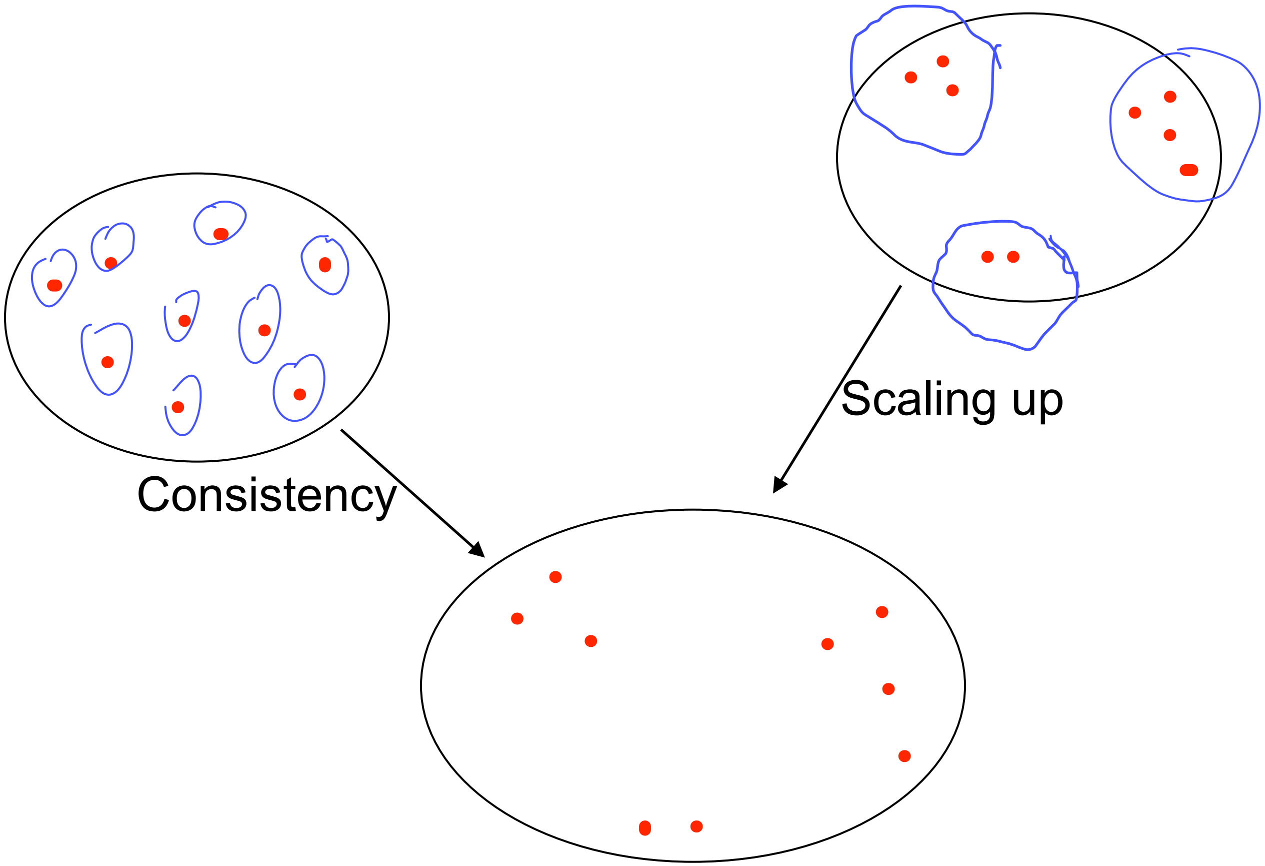
3. Through a series of Consistent transformations, can transform d_1 into d_6, which will have the same MST as d

4. Invoke MST-Coherence to get
   $F(d_1, k) = F(d_6, k) = F(d, k) = SL(d, k)$

KLEINBERG'S IMPOSSIBILITY RESULT

There exist no clustering function all 3 properties

Proof:



AXIOMS AS A TOOL FOR A TAXONOMY OF CLUSTERING PARADIGMS

- The goal is to generate a variety of axioms (or properties) over a fixed framework, so that different clustering approaches could be classified by the different subsets of axioms they satisfy.

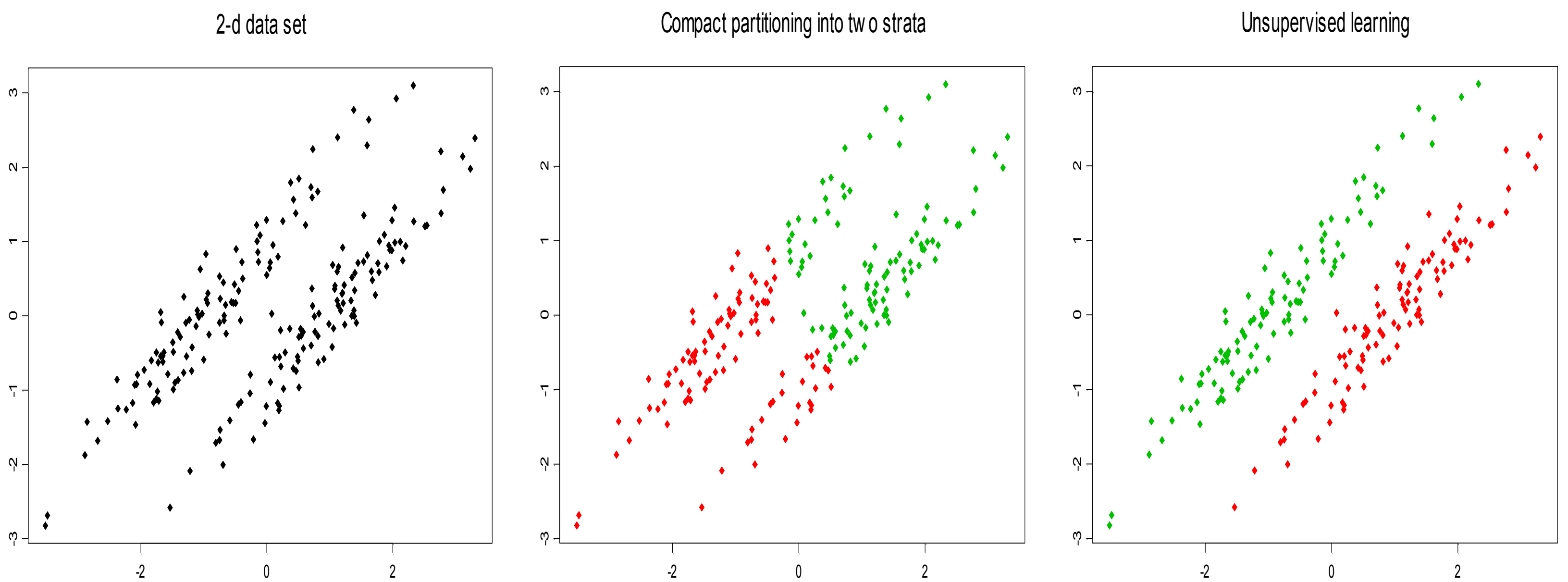
"Axioms" (Scale Invariance, k-Richness, Consistency, Separability) — *"Properties"* (Order Invariance, Hierarchy)

	Scale Invariance	k-Richness	Consistency	Separability	Order Invariance	Hier-archy
Single Linkage	+	+	+	+	+	+
Center Based	+	+	+	+	-	
Spectral	+	+	-	-	-	
MDL	+	+	-			
Rate Distortion	+	+	-			

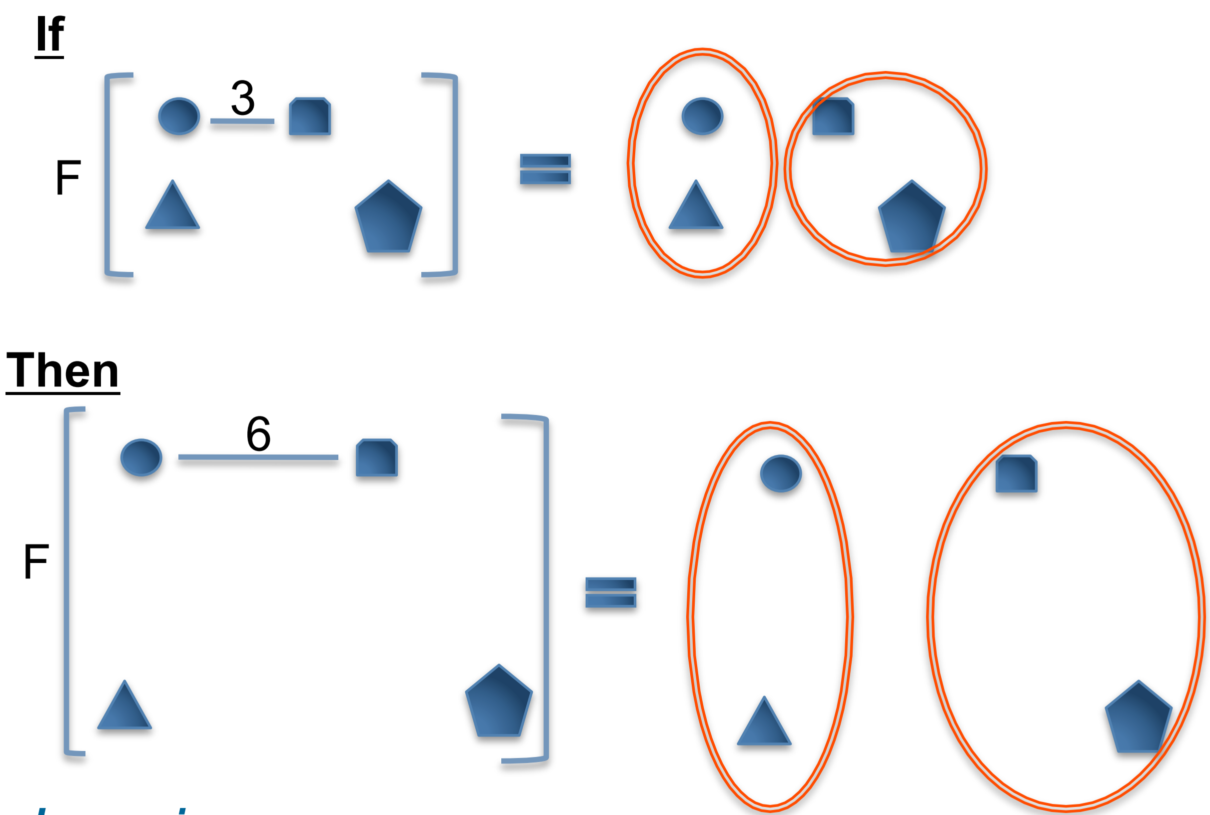
PROPERTIES

- Order-Consistency
  - Function only compares distances together, not using absolute value
- Minimum Spanning Tree Coherence
  - If two datasets d and d' have the same Minimum Spanning Tree, then for all k, F(d, k) = F(d', k)
  - Function makes all its decisions using the Minimum Spanning Tree

SOME MORE EXAMPLES



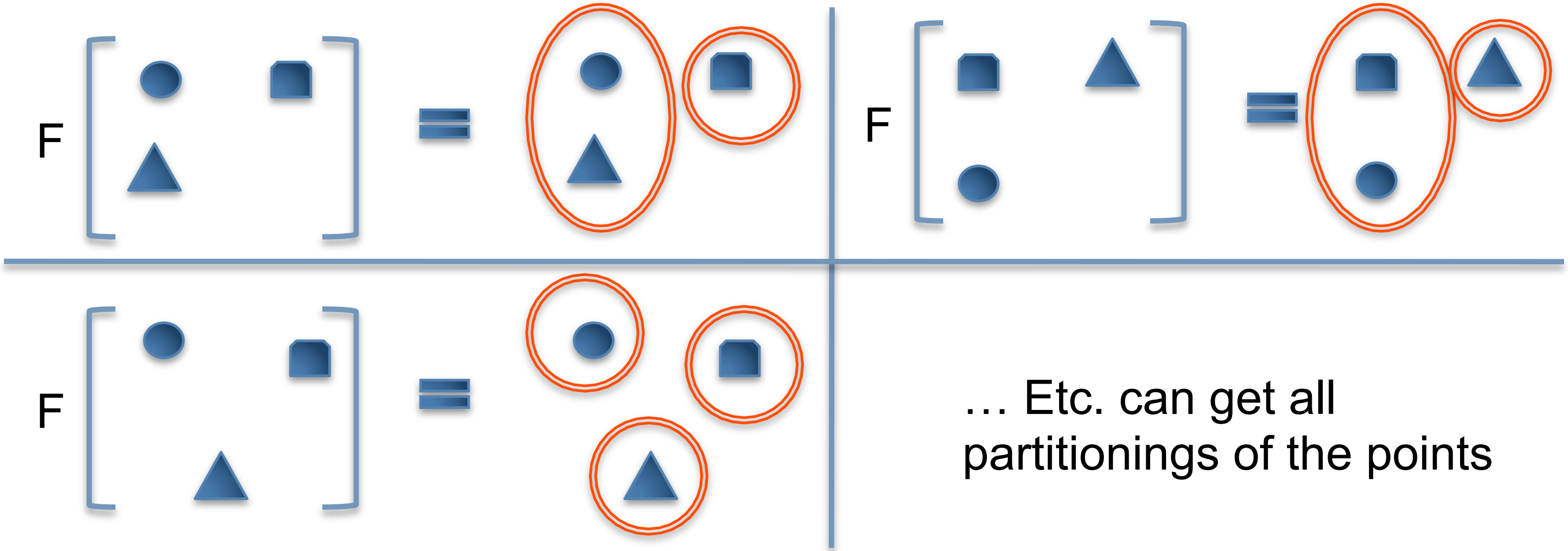
AXIOMS - SCALE INVARIANCE



e.g. double the distances

- *Scale Invariance*
  $F(\lambda d)=F(d)$ for all ***d*** and all strictly positive ***λ***.

AXIOMS - RICHNESS

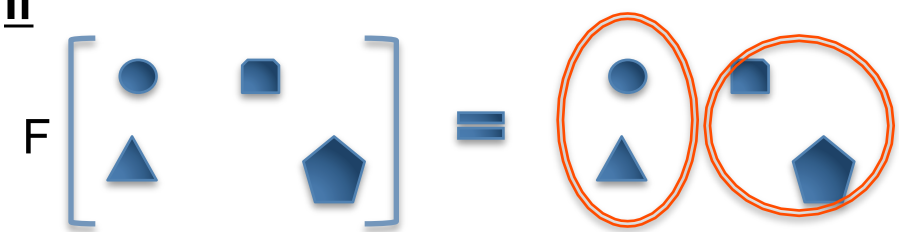


… Etc. can get all partitionings of the points

- *Richness*
  The range of ***F(d)*** over all ***d*** is the set of all possible partitionings

AXIOMS - CONSISTENCY

<u>If</u>

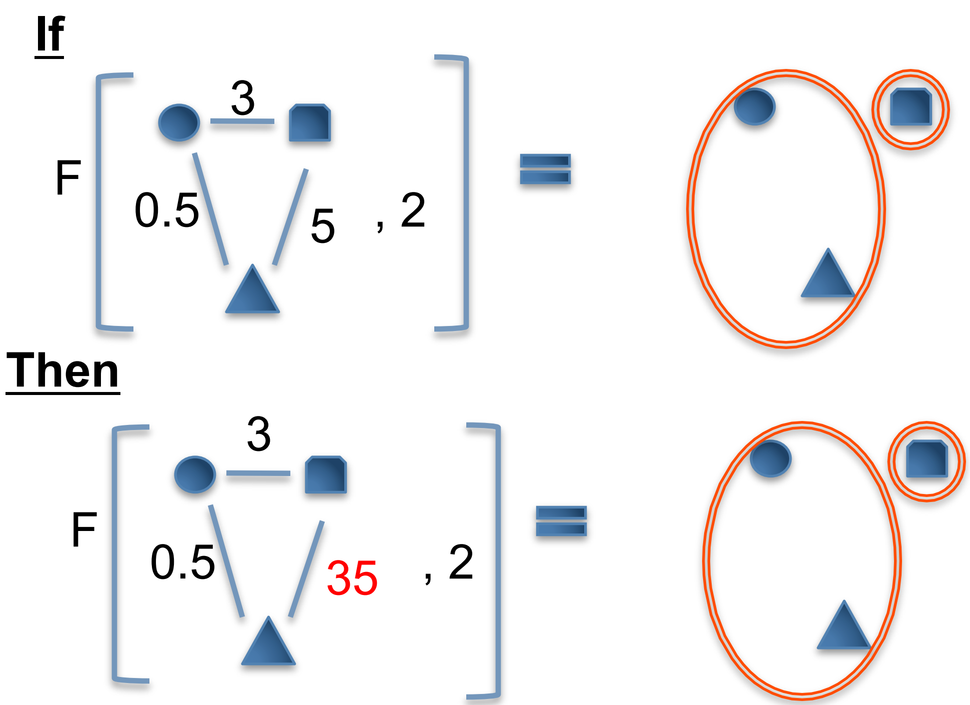


<u>Then</u>

- *Consistency*

  If ***d'*** equals ***d*** except for shrinking distances within clusters of ***F(d)*** or stretching between-cluster distances, then ***F(d)=F(d')***.

PROPERTIES - ORDER-CONSISTENCY



Maintain edge ordering

- *Order-Consistency*

  If two datasets ***d*** and ***d'*** have the same ordering of the distances, then for all ***k***, ***F(d, k)=F(d', k)***